ISSN 2079-3316 PROGRAM SYSTEMS: THEORY AND APPLICATIONS no.1(28), 2016, pp. 201–208

S. V. Znamenskij

A picture of common subsequence length for two random strings over an alphabet of 4 symbols

Abstract. The maximal length of longest common subsequence (LCS) for a couple of random finite sequences over an alphabet of 4 characters was considered as a random function of the sequences lengths m and n. Exact probability distributions tables are presented for all couples of length in a range $2 < m + n < 19$.

The graphs of expected value and standard deviation as a functions of length are shown in linear perspective which presents the behaviour of large lengths at the horizon. In order to illustrate behaviour on large lengths, the results of numeric simulation for $m + n = 32$, 512, 8192 and 131072 are also shown on the same graphs. The presented graph of expected value dependency of m and n looks to have asymptotic right circular cone. The variance looks alike growing as $(n + m)^{\frac{3}{4}}$.

Key words and phrases: similarity of strings, sequence alignment, edit distance, LCS, Levenshtein metric.

2010 *Mathematics Subject Classification*: 68T37; 68P10, 68W32.

The need for a theoretical estimation of occasional wrong match probability in fussy search with longest common subsequence (LCS) related algorithm raises the difficult problem of estimation the LCS length of two random symbol strings.

For practically important for applications in bioinfiormatics case of alphabet $\Sigma_4 = A, C, G, T$ [1] it was found in [2] that natural DNA-sequences from a database and true random sequences show the same statistical behavior in terms of similarity scores. The aim of this paper is to get a brief overall picture of functional dependency from both length of strings instead in spite of know for binary alphabet results [3].

Let's random variable $L_{m,n}$ notes the length of the longest common subsequence of random strings over the alphabet Σ with lengths m and n. The behaviour of its expected value $E(L_{m,n})$ and variance $V(L_{m,n})$ as a function of m and n remains dark and unclear for many decades. After [4] investigations for this alphabet are mostly concerned on bounding of

Chvátal-Sankoff constant $\gamma_4 = \lim\limits_{n\to\infty} \frac{E(L_{n,n})}{n}$ and detection of convergence rate for the limit value [5].

1. Exact probability mass function

We start from calculation of the exact probability mass function

$$f_{L_{m,n}}(k) = P(L_{m,n} = k),$$

where P is probability for random variable $L_{m,n}$ to get a value $k = 0, 1, \ldots, \min(m, n)$.

THEOREM 1. *Let $m < n$ are natural numbers. In the following simple cases the probability mass function $f_{L_{m,n}}(k)$ can be calculated by formulae:*

$p_{1,n}(0) = \left(\frac{3}{4}\right)^n$;
$p_{1,n}(1) = 1 - \left(\frac{3}{4}\right)^n$;
$p_{2,n}(0) = \frac{1}{4}\left(\frac{3}{4}\right)^n + \frac{3}{2^{n+2}}$ *for* $n > 1$;
$p_{2,n}(1) = \left(\frac{3}{4}\right)^{n+1} + \frac{n}{4}\left(\frac{3}{4}\right)^{n-1} - \frac{3}{2^{n+2}}$ *for* $n > 1$;
$p_{2,n}(2) = 1 - \left(\frac{3}{4}\right)^{n+1} - \frac{1}{4}\left(\frac{3}{4}\right)^n - \frac{n}{4}\left(\frac{3}{4}\right)^{n-1}$ *for* $n > 1$;
$p_{3,n}(0) = \frac{1}{16}\left(\frac{3}{4}\right)^n + \frac{9}{2^{n+4}} + \frac{6}{4^{n+2}}$ *for* $n > 2$;

The proof can be obtained in a standard way by the law of total probability application to following conditions

- shorter string consists of identical symbols,
- shorter string contains exactly two different symbols,
- shorter string contains three different symbols

and counting all the cases with the use of geometric series formula. Unfortunately for larger n values calculations became too complicated and huge to get through.

The Monte Carlo simulation never gives exact information. Nevertheless, if lengths are small then computer can just find LCS for each couple of strings with given lengths and *exactly* calculate the probabilities by the formula $p_{n,m}(k) = \frac{X_{n,m}(k)}{4^{n+m}}$ where $X_{n,m}(k)$ is the total number of string couples with lengths n and m and LCS length k. Any modern PC allows to calculate probability mass function for any length with $m + n < 16$ in a time less then one hour. But the performance time grows very fast, always more then 4 times with each added to total length symbol. Even if the Moore's Law continue its action, we shall never get a computational resource enough to proceed with $m + n = 100$. So as the table will not grow much we publish it except lines of the values that

TABLE 1. Exact values $x_{m.n}(k)$ for $m = 3, \ldots, 5$

m	n	$x_{m,n}(0)$	$x_{m,n}(1)$	$x_{m,n}(2)$	$x_{m,n}(3)$	$x_{m,n}(4)$	$x_{m,n}(5)$
3	3	105	570	333	16		
3	4	231	1797	1860	208		
3	5	537	5436	8715	1696		
3	6	1311	16131	36990	11104		
3	7	3345	47502	147441	63856		
3	8	8871	139713	562920	337072		
3	9	24297	411888	2083239	1674880		
3	10	68271	1219263	7530450	7959232		
3	11	195585	3626226	26726205	36560848		
3	12	568311	10834653	93468060	163564432		
3	13	1668057	32507028	322953267	716613472		
3	14	4930431	97874331	1104629190	3087533344		
4	4	453	4800	8742	2325	64	
4	5	951	12537	34737	16287	1024	
4	6	2109	32688	125919	91572	9856	
4	7	4911	85983	431556	452142	73984	
4	8	11973	229512	1426380	2049063	477376	
4	9	30471	623565	4602219	8740545	2780416	
4	10	80589	1726080	14611437	35649990	15040768	
4	11	220191	4864899	45895014	140496120	76959232	
4	12	617493	13937016	143156802	538909425	377121088	
4	13	1766391	40488273	444586053	2022505587	1785620992	
5	5	1833	28890	118404	98340	14421	256
5	6	3759	67587	372240	478146	121980	4864
5	7	8097	161496	1118523	2048547	803625	54016
5	8	18231	395277	3276435	8083134	4546155	457984
5	9	42873	992934	9470214	30124806	23194581	3283456
5	10	105279	2562903	27226098	107738208	109830936	20972032
5	11	269457	6797844	78276255	373719987	491599113	123079168
5	12	715911	18504897	225874791	1266831732	2105965533	677074432

consists of calculated by the theorem values. Table 1 and Table 2 contains the numbers $\frac{x_{n,m}(k)}{=}\frac{X_{n,m}(k)}{4}$ to avoid rounding errors completely.

TABLE 2. Exact values $x_{m.n}(k)$ for $m = 6, \ldots, 9$

m	n	$x_{m,n}(0)$	$x_{m,n}(1)$	$x_{m,n}(2)$	$x_{m,n}(3)$	$x_{m,n}(4)$	$x_{m,n}(5)$	$x_{m,n}(6)$	$x_{m,n}(7)$	$x_{m,n}(8)$	$x_{m,n}(9)$
6	6	7221	144216	1028931	2022960	907179	82773	1024			
6	7	14631	317769	2753415	7587861	5259105	821907	22528			
6	8	30909	721704	7267770	26375340	26208204	6221289	283648			
6	9	67839	1687911	19132338	87148998	117952467	39763023	2682880			
6	10	154821	4067112	50587443	278159046	493461417	226129521	21182464			
6	11	368151	10107261	134995695	867001353	1954801602	1180257714	147435520			
7	7	28185	653256	6653442	25126110	26901270	7292628	449877	4096		
7	8	56751	1393707	16048740	77752320	118644660	49303485	5133393	102400		
7	9	118257	3069144	38981253	230646870	474140310	281409102	43935096	1441792		
7	10	254391	6959589	95813703	666472293	1767695262	1429288875	313426287	15056896		
7	11	566025	16255368	239097306	1896158694	6264656730	6664862241	1968372276	129900544		
8	8	109893	2818902	35794041	216361560	465327408	297936519	53028744	2348373	16384	
8	9	220551	5910921	81120723	582733095	1662146217	1521603615	410485389	30288033	458752	
8	10	454989	12762708	187198638	1542583476	5569004667	6924999531	2646395274	289326477	7143424	
9	9	429513	11892570	173097576	1439627910	5342669283	6969837924	2872947741	357417270	11883861	65536
9	10	860559	24709083	378466404	3527391882	16219427187	28529534742	16745716626	3120355458	170983179	2031616

TABLE 3. Parameters of numeric simulation

$m+n$	step of n	step execution time	repetitions for step
32	1	20 sec	1111275–1173205
512	8	20 sec	49334–161188
8192	128	20 min	15240–185462
131072	2048	200 min	535–4391
2097152	262144	10000 sec	2–5

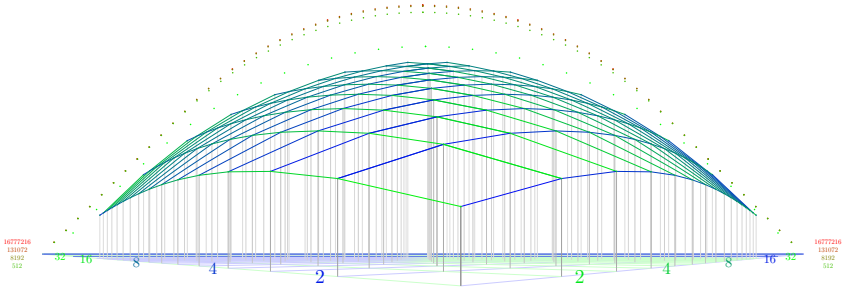

FIGURE 1. Expected value of LCS as a function of the sequence lengths

2. Design and interpretation of expected value and variance graphs

We believe that the function $E(L_{m,n})$ of two string length is concave and grows nearly linear with lengths grows. This is a reason to use projective geometry to draw graphs in linear perspective and see the limit values in infinity as the horizon.

The calculated exact values does not represent behaviour on large lengths. So numeric simulation was used to get a full picture. It used fixed total length $n+m$ and a fixed step on n with fixed stepwise execution time. Table 3 shows the given parameters of numeric simulation and number of repetitions for each step. The graph of expected value on the Picture 1 use the projective transform $\left(\frac{2x-8y}{2+5x+5y}, \frac{10z-1}{2+5x+5y}\right)$ mapping numbers to centimeters. The point of view was selected close to a vertex $(-0.25, -0.25, 0.1)$ of approximate asymptotic cone by a series of iterations aimed to get images of far points proportionally with respect to logarithms of numbers closer to some limit line. Our experiment detects $\gamma_4 = 0.6542$ to be compatible with the published in [5] more precise value $\gamma_4 = 0.654361$ obtained in very smart numeric simulation.

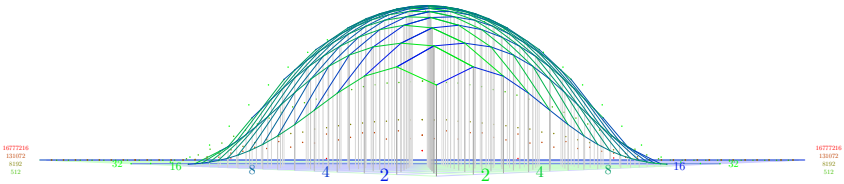

FIGURE 2. Standard deviation of LCS as a function of the sequence lengths

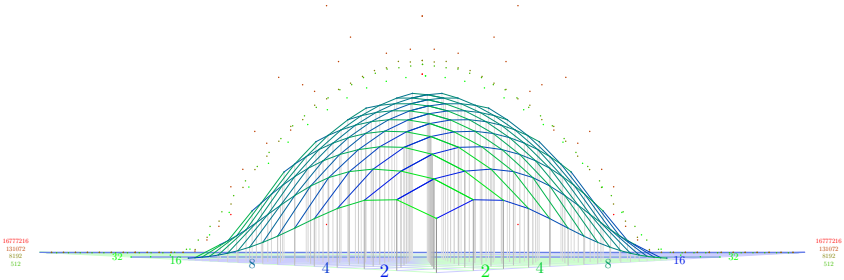

FIGURE 3. Standard deviation of LCS scaled by $\sqrt[4]{m+n}$

The graph show unexpected structure of dotted consisting from central circle arc extended by two line segments. So the graph visualy looks smoothly combined from two plains $z = x$, $z = y$ and surface $z = s - \sqrt{2}r + \sqrt{r^2 - s^2}$ where $s = \frac{x+y}{2}$, $r = (\sqrt{2}+1)(s - e(s)))$ and the diagonal function $e(n) = E(L_{n,n})$ is monotonic, concave and unknown, but for large m and n it is close to linear and therefore surface is close to asymptotic right circular cone.

The graph of variance on the Picture 2 use another projective transform $\left(\frac{x-1.1y}{7+x+y}, \frac{12z-0.4}{7+x+y}\right)$ mapping numbers to centimeters to reflect differences from conical form. We see that variance grows more slow then linear. The asymptotic (non-circle) similar to cone surface appears visible on the Picture 3 for magnified variance $\sqrt[4]{m+n} \cdot V(L_{m,n})$ shown under $\left(\frac{x-1.1y}{7+x+y}, \frac{6z-0.6}{7+x+y}\right)$ transform.

References

[1] R. Durbin, S. Eddy, A. Krogh, G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998, 370 c. ↑ 201

[2] J. G. Reich, H. Drabsch, A. Däumler. «On the statistical assessment of similarities in DNA sequences», *Nucleic acids research*, **12**:13 (1984), c. 5529–5543. ↑ 201

[3] K. Ning, K. P. Choi. *Systematic assessment of the expected length, variance and distribution of Longest Common Subsequences*, 2013, arXiv: 1306.4253. ↑ 201

[4] V. Chvátal, D. Sankoff. «Longest Common subsequences of two random sequences», *J. Appl. Probability*, **12**:2 (1975), c. 306–315. ↑ 201

[5] R. Bundschuh. «High precision simulations of the longest common subsequence problem», *The European Physical Journal B-Condensed Matter and Complex Systems*, **22**:4 (2001), c. 533–541. ↑ 202,205

Submitted by *dr E. P. Kurshev*

About the author:

Foto by A. Yu. Fomenko, CC-BY-SA

Sergej Vital'evich Znamenskij

Chair of Mathematics in the Ailamazyan Pereslavl University, head of laboratory in Ailamazyan Program Systems Institute of RAS. Research interests migrate from research in Functional Analysis, Complex Analysis and finite-dimensional Projective Geometry (analogues of Convexity) to the foundations of Collaborative Software Development.

e-mail: svz@latex.pereslavl.ru

Sample citation of this publication:

S. V. Znamenskij. "A picture of common subsequence length for two random strings over an alphabet of 4 symbols", *Program systems: theory and applications*, 2016, **7**:1(28), pp. 201–208.

URL: http://psta.psiras.ru/read/psta2016_1_201-208.pdf

УДК 004.416

С. В. Знаменский. *Картина наибольшей длины общих подпоследовательностей пары случайных строк 4 буквенного алфавита.*

Аннотация. Наибольшая длина (LCS) общей подпоследовательности пары случайных конечных последовательностей из 4 букв рассмотрена как случайная функция от длин m и n этих двух последовательностей. Представлены таблицы точных значений вероятностей для всех пар конкретных длин в диапазоне $2 < m+n < 19$.

Графики зависимости математического ожидания и дисперсии показаны в линейной перспективе, позволяющей просматривать на горизонте поведение при растущих длинах. Для иллюстрации поведения при больших значениях длин на этих же графиках показаны результаты численного эксперимента для больших значений $m + n = 32$, 512, 8192 и 131072. Представленный график зависимости математического ожидания от m и n выглядит имеющим асимптотический прямой круговой конус. Дисперсия выглядит растущей как $(n + m)^{\frac{3}{4}}$.

Ключевые слова и фразы: сходство строк, выравнивание последовательностей, случайные общие подпоследовательности, LCS, метрика Левенштейна.

Пример ссылки на эту публикацию:

С. В. Знаменский. «Картина наибольшей длины общих подпоследовательностей пары случайных строк 4 буквенного алфавита», *Программные системы: теория и приложения*, 2016, **7**:1(28), с. 201–208. (*Англ*).

URL: http://psta.psiras.ru/read/psta2016_1_201-208.pdf