Expected length of the longest common subsequence for large alphabets*

Marcos Kiwi[†]
Depto. Ing. Matemática and
Ctr. Modelamiento Matemático UMR 2071,
University of Chile
Correo 3, Santiago 170–3, Chile
e-mail: mkiwi@dim.uchile.cl

Martin Loebl[‡]
Dept. of Applied Mathematics and
Institute of Theoretical Computer Science (ITI)
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic
e-mail: loebl@kam.mff.cuni.cz

Jiří Matoušek[§]
Dept. of Applied Mathematics and
Institute of Theoretical Computer Science (ITI)
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic
e-mail: matousek@kam.mff.cuni.cz

Presented by J. Matoušek

We investigate the distribution of the length L of the longest common subsequence of two randomly uniformly and independently chosen n character words $u = u_1u_2 \ldots u_n$ and $v = v_1v_2 \ldots v_n$ over a k-ary alphabet. That is, L is the maximum integer such that there exist indices $i_1 < i_2 < \cdots < i_L$ and $j_1 < j_2 < \cdots < j_L$ with $u_{i_q} = v_{j_q}$, $q = 1, 2, \ldots, L$. This problem has emerged more or less independently in several remarkably disparate areas, including the comparison of versions of computer programs, cryptographic snooping, and molecular biology.

By a well-known subadditivity argument, $\mathbf{E}[L]/n$ converges to a constant γ_k. The value of γ_k is not known for any particular value of k, although much effort has been spent in finding good upper an lower bounds for it (see, for example, [2] and references therein).

We analyze the behavior of γ_k for $k \to \infty$, and more generally, we consider the expectation of L when k is an (arbitrarily slowly growing) function of n and $n \to \infty$. In particular, we prove a conjecture of Sankoff and Mainville from the early 80's [7] stating that

$$\lim_{k\to\infty} \gamma_k \sqrt{k} = 2. \tag{1}$$

(See [6, § 6.8] for a discussion of work on lower and upper bounds on γ_k as well as a stronger version, due to Arratia and Steele, of the above stated conjecture.)

*An extended abstract of this work appears in *Proc. 6th Latin American Theoretical Informatics Symposium* (LATIN 2004), LNCS series, Springer, Berlin. A full version is available at the web page of the third author.

†Gratefully acknowledges the support of ICM P01–05 and Fondecyt 1010689.

‡Gratefully acknowledges the support of ICM-P01-05. This work was done while visiting the Dept. Ing. Matemática, U. Chile.

§This research was done while visiting the Ctr. de Modelamiento Matemático, UMR–UChile 2071, U. Chile, Santiago, supported by Fondap in Applied Mathematics 2000–05.

The constant 2 in (1) arises from a connection with another celebrated problem, the distribution of LIS_N, the length of the longest increasing subsequence in a (uniform) random permutation of $\{1, 2, \ldots, N\}$. Hammersley [4] proved the existence of $\beta = \lim_{N\to\infty}(\mathbf{E}[\mathrm{LIS}_N]/\sqrt{N})$ and conjectured that $\beta = 2$. Later, Logan and Shepp [5], based on a result by Schensted [8], proved $\beta \leq 2$, and finally, Vershik and Kerov [10] showed $\beta = 2$. In a major recent breakthrough Baik, Deift, Johansson [3] described explicitly the asymptotic distribution of LIS_N (for $N \to \infty$). For a detailed account of these results, history, and related work we can recommend the surveys of Aldous and Diaconis [1] and Stanley [9]; the methods used in attacking this problem are of remarkable beauty and diversity.

Our main result about the longest common subsequence can be stated as follows.

Theorem 1 *For every $\varepsilon > 0$ there exist k_0 and C such that for all $k > k_0$ and all n with $n/\sqrt{k} > C$ we have*

$$(1-\varepsilon)\cdot\frac{2n}{\sqrt{k}} \;\leq\; \mathbf{E}[L] \;\leq\; (1+\varepsilon)\cdot\frac{2n}{\sqrt{k}}$$

where, as above, L is the length of the longest common subsequence of two random words of length n over an alphabet of size k. Moreover, there is an exponentially small tail bound; namely, for every $\varepsilon > 0$ there exists $c > 0$ such that for k and n as above,

$$\mathbf{P}\left[\left|L-\frac{2n}{\sqrt{k}}\right| \geq \varepsilon\frac{2n}{\sqrt{k}}\right] \leq e^{-cn/\sqrt{k}}.$$

In the rest of this extended abstract, we outline the main tools and ideas of the proof, referring to the full version for precise formulations and further details.

First we reformulate the problem a little. Given the two random words $u = u_1u_2\ldots u_n$ and $v = v_1v_2\ldots v_n$, let us draw two horizontal lines in the plane and place n points $a_1, a_2, \ldots, a_n$ in this order on the top line and n points $b_1, b_2, \ldots, b_n$ in this order on the bottom line. Then we connect a_i to b_j by an edge (straight segment) iff $u_i = v_j$, obtaining a drawing of a bipartite graph G (which is a disjoint union of complete bipartite graphs). A common subsequence of the words u and v corresponds to a *planar matching* in G (a matching in which no two edges cross).

Although we want to deal mainly with the case of n arbitrarily large compared to k, which is the situation in the Sankoff–Mainville conjecture, we first consider a seemingly opposite setting: when k is large and n is also large but considerably smaller than n. For definiteness, we set $n = k^{0.7}$. Then we expect G to have about $n^2/k = k^{0.4}$ edges, and most of these edges connect vertices of degree 1. If we let G' be the subgraph of G obtained by deleting all edges incident to vertices of degree greater than 1, then G' is a matching (plus some isolated vertices). The number N of edges of G' is typically very close to $k^{0.4}$. The matching determines a permutation of $\{1, 2, \ldots, N\}$, and by a symmetry argument, it can be seen that, for a given N, all permutations of $\{1, 2, \ldots, N\}$ have the same probability of being obtained in this way. Moreover, the longest increasing subsequence of the permutation corresponds exactly to the largest planar matching in G'. Therefore, up to a small error, the longest common subsequence of u and v is distributed as LIS_N. Then one can derive from the known results about LIS_N that $\mathbf{E}[L] = (2 + o(1))n/\sqrt{k}$ holds in this situation. For the rest of the proof, we also need tail estimates for large deviations of L, and these are conveniently obtained from Talagrand's inequality applied to L (we cannot directly use known tail estimates for LIS_N, for example because of the vertices of degree larger than 1 in G).

Now we consider n very large compared to k (and k still large). A lower bound for $\mathbf{E}[L]$ is straightforward: We partition the words u and v into segments of length $k^{0.7}$ each, and we use the previously derived result separately for each block (the ith block consists of the ith segment of the word u plus the ith segment of the word v). Thus, the lower bound is provided by a common subsequence, or planar matching in the graph language, that never crosses a block boundary.

An upper bound for $\mathbf{E}[L]$ is more demanding, since the largest planar matching need not respect any partition into blocks fixed in advance; there could be "very skew" edges. Our strategy is to simultaneously consider many different partitions into blocks. The blocks have upper and lower segments of size about $k^{0.7}$, but they can be very skew; the segment of u starting at a position i can form a block with a segment of v starting at position j, with i and j differing by a large amount. Supposing that there is a planar matching with at least $m = (1+\varepsilon)2n/\sqrt{k}$ edges, it "fits" at least one of the block partitions, meaning that it respects its block boundaries. For each fixed block partition and each fixed distribution of the numbers of edges of the planar matching among the blocks, we bound above the probability that there is a planar matching with m edges that fits that block partition; this relies on independence among the blocks. Then we sum up over all possible block partitions and show that with high probability, there is no planar matching with m edges at all.

References

[1] D. Aldous and P. Diaconis. Longest increasing subsequences: From patience sorting to the Baik–Deift–Johansson theorem. *Bulletin of the AMS*, 36(4):413–432, 1999.

[2] R. Baeza-Yates, G. Navarro, R. Gavaldá, and R. Schehing. Bounding the expected length of the longest common subsequences and forests. *Theory of Computing Systems*, 32(4):435–452, 1999.

[3] J. Baik, P. A. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutaions. *J. Amer. Math. Soc.*, 12:1119–1178, 1999.

[4] J. M. Hammersley. A few seedlings of research. In *Proc. Sixth Berkeley Sympos. Math. Stat. Prob.*, pages 345–394, Berkeley, Calif., 1972. Univ. of California Press.

[5] B. F. Logan and L. A. Shepp. A variational problem or random Young tableaux. *Adv. in Math.*, 26:206–222, 1977.

[6] P.A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach.* MIT Press, 2000.

[7] D. Sankoff and J. Kruskal, editors. *Common subsequences and monotone subsequences*, chapter 17, pages 363–365. Addison–Wesley, Reading, Mass., 1983.

[8] C. Schensted. Longest increasing and decreasing subsequences. *Canad. J. Math.*, 13:179–191, 1961.

[9] R. P. Stanley. Recent progress in algebraic combinatorics. *Bulletin of the AMS*, 40(1):55–68, 2002.

[10] A. M. Vershik and S. V. Kerov. Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tableaux. *Dokl. Akad. Nauk SSSR*, 233:1024–1028, 1977.