Discrete Applied Mathematics 210 (2016) 75–87

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Repetition-free longest common subsequence of random sequences

Cristina G. Fernandes [a], Marcos Kiwi [b,*]

[a] Computer Science Department, Universidade de São Paulo, Brazil

[b] Depto. Ingeniería Matemática & Ctr. Modelamiento Matemático UMI 2807, U. Chile, Chile

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form 2 July 2015
Accepted 6 July 2015
Available online 6 August 2015

Keywords:
Repetition-free subsequence
Common subsequence
Random sequences

ABSTRACT

A repetition-free Longest Common Subsequence (LCS) of two sequences x and y is an LCS of x and y where each symbol may appear at most once. Let R denote the length of a repetition-free LCS of two sequences of n symbols each one chosen randomly, uniformly, and independently over a k-ary alphabet. We study the asymptotic, in n and k, behavior of R and establish that there are three distinct regimes, depending on the relative speed of growth of n and k. For each regime we establish the limiting behavior of R. In fact, we do more, since we actually establish tail bounds for large deviations of R from its limiting behavior.

Our study is motivated by the so called exemplar model proposed by Sankoff (1999) and the related similarity measure introduced by Adi et al. (2010). A natural question that arises in this context, which as we show is related to long standing open problems in the area of probabilistic combinatorics, is to understand the asymptotic, in n and k, behavior of parameter R.

1. Introduction

Several of the genome similarity measures considered in the literature either assume that the genomes do not contain gene duplicates, or work efficiently only under this assumption. However, several known genomes do contain a significant amount of duplicates. (See the review on gene and genome duplication by Sankoff [19] for specific information and references.) One can find in the literature proposals to address this issue. Some of these proposals suggest to filter the genomes, throwing away part or all of the duplicates, and then applying the desired similarity measure to the filtered genomes. (See [2] for a description of different similarity measures and filtering models for addressing duplicates.)

Sankoff [18], trying to take into account gene duplication in genome rearrangement, proposed the so called exemplar model, which is one of the filtering schemes mentioned above. In this model, one searches, for each family of duplicated genes, an exemplar representative in each genome. Once the representative genes are selected, the other genes are disregarded, and the part of the genomes with only the representative genes is submitted to the similarity measure. In this case, the filtered genomes do not contain duplicates, therefore several of the similarity measures (efficiently) apply. Of course, the selection of the exemplar representative of each gene family might affect the result of the similarity measure. Following the parsimony principle, one wishes to select the representatives in such a way that the resulting similarity is as good as possible. Therefore, each similarity measure induces an optimization problem: how to select exemplar representatives of each gene family that result in the best similarity according to that specific measure.

* Corresponding author.
E-mail addresses: cris@ime.usp.br (C.G. Fernandes), mkiwi@dim.uchile.cl (M. Kiwi).

http://dx.doi.org/10.1016/j.dam.2015.07.005
0166-218X/

The length of a Longest Common Subsequence (LCS) is a well-known measure of similarity between sequences. In particular, in genomics, the length of an LCS is directly related to the so called edit distance between two sequences when only insertions and deletions are allowed, but no substitution. This similarity measure can be computed efficiently in the presence of duplicates (the classical dynamic programming solution to the LCS problem takes quadratic time, however, improved algorithms are known, specially when additional complexity parameters are taken into account — for a comprehensive comparison of well-known algorithms for the LCS problem, see [4]). Inspired by the exemplar model above, some variants of the LCS similarity measure have been proposed in the literature. One of them, the so called *exemplar* LCS [6], uses the concept of mandatory and optional symbols, and searches for an LCS containing all mandatory symbols. A second one is the so called *repetition-free* LCS [1], that requires each symbol to appear at most once in the subsequence. Some other extensions of these two measures were considered under the name of *constrained* LCS and *doubly-constrained* LCS [7]. All of these variants were shown to be hard to compute [1,5–7], so some heuristics and approximation algorithms for them were proposed and experimentally tested [1,6,14,10].

Specifically, the notion of repetition-free LCS was formalized by Adi et al. [1] as follows. They consider finite sets, called *alphabets*, whose elements are referred to as *symbols*, and then they define the RFLCS problem as: Given two sequences x and y, find a repetition-free LCS of x and y. We write RFLCS (x, y) to refer to the RFLCS problem for a generic instance consisting of a pair (x, y), and we denote by Opt(RFLCS(x, y)) the length of an optimal solution of RFLCS (x, y). In their paper, Adi et al. showed that RFLCS is MAX SNP-hard, proposed three approximation algorithms for RFLCS, and presented an experimental evaluation of their proposed algorithms, using for the sake of comparison an exact (computationally expensive) algorithm for RFLCS based on an integer linear programming formulation of the problem.

Whenever a problem such as the RFLCS is considered, a very natural question arises: What is the expected value of Opt(RFLCS(x, y))? (where expectation is taken over the appropriate distribution over the instances (x, y) one is interested in). It is often the case that one has little knowledge of the distribution of problem instances, except maybe for the size of the instances. Thus, an even more basic and often relevant issue is to determine the expected value taken by Opt(RFLCS(x, y)) for uniformly distributed choices of x and y over all strings of a given length over some fixed size alphabet (say each sequence has n symbols randomly, uniformly, and independently chosen over a k-ary alphabet Σ). Knowledge of such an average case behavior is a first step in the understanding of whether a specific value of Opt(RFLCS(x, y)) is of relevance or could be simply explained by random noise. The determination of this latter average case behavior in the asymptotic regime (when the length n of the sequences x and y go to infinity) is the main problem we undertake in this article. Specifically, let $R_n = R_n(x, y)$ denote the length of a repetition-free LCS of two sequences x and y of n symbols randomly, uniformly, and independently chosen over a k-ary alphabet. Note that the random variable R_n is simply the value of Opt(RFLCS(x, y)). We are interested in determining (approximately) the value of $\mathbb{E}\left(R_n\right)$ as a function of n and k, for very large values of n.

One of the results established in this article is that the behavior of $\mathbb{E}\left(R_n\right)$ depends on the way in which n and k are related. In fact, if k is fixed, it is easy to see that $\mathbb{E}\left(R_n\right)$ tends to k when n goes to infinity (simply because any fix permutation of a k-ary alphabet will appear in a sufficiently large sequence of uniformly and independently chosen symbols from the alphabet). Thus, the interesting cases arise when $k = k(n)$ tends to infinity with n. However, the speed at which $k(n)$ goes to infinity is of crucial relevance in the study of the behavior of $\mathbb{E}\left(R_n\right)$. We identify three distinct growth regimes depending on the asymptotic dependency between n and $k\sqrt{k}$. Specifically, we establish the next result[1]:

Theorem 1. *The following holds:*

- *If* $n = \omega(\sqrt{k})$ *and* $n = o(k\sqrt{k})$, *then* $\lim_{n\to\infty} \frac{\mathbb{E}\left(R_n\right)}{n/\sqrt{k(n)}} = 2$.
- *If* $n = \frac{1}{2}\rho k\sqrt{k}$ *for* $\rho > 0$, *then* $\liminf_{n\to\infty} \frac{\mathbb{E}\left(R_n\right)}{k(n)} \geq 1 - e^{-\rho}$. *(By definition* $R_n \leq k(n)$.*)*

  Moreover, if $n = \omega(k\sqrt{k})$, *then* $\lim_{n\to\infty} \frac{\mathbb{E}\left(R_n\right)}{k(n)} = 1$.

The main results of this article are obtained by relating the asymptotic average case behavior of $\mathbb{E}\left(R_n\right)$ with that of the length $L_n = L_n(x, y)$ of a Longest Common Subsequence (LCS) of two sequences x and y of n symbols chosen randomly, uniformly, and independently over a k-ary alphabet. A simple (well-known) fact concerning L_n is that $\mathbb{E}\left(L_n\right)/n$ tends to a constant, say γ_k, when n goes to infinity. The constant γ_k is known as the Chvátal–Sankoff constant. A long standing open problem is to determine the exact value of γ_k for any fixed $k \geq 2$. However, Kiwi, Loebl, and Matoušek [17] proved that $\gamma_k\sqrt{k} \to 2$ as $k \to \infty$ (which positively settled a conjecture due to Sankoff and Mainville [20]).

We now give an informal and intuitive justification for each of the claims stated in Theorem 1. As pointed out above, in [17], it was shown that, under some conditions on the speed of growth of $k = k(n)$, the expected length of an LCS of two length n sequences randomly, uniformly, and independently chosen over a k-ary alphabet, is roughly $2n/\sqrt{k}$. When $n = \omega(\sqrt{k}) \cap o(k\sqrt{k})$, we see that $2n/\sqrt{k} = \omega(1) \cap o(k)$. If the k-ary symbols that belong to an LCS show up more or

[1] Adhering to standard notation, for functions f and g defined over the non-negative integers, g always non-zero, we say that $f(n) = \omega(g(n))$ if $|f(n)|/|g(n)|$ tends to infinity when $n \to \infty$.