Information and Computation 241 (2015) 215–226

Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Approximate periodicity ☆

Amihood Amir [a,d,1], Estrella Eisenberg [a], A. Levy [b,c,*]

[a] Department of Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel
[b] Department of Software Engineering, Shenkar College, 12 Anna Frank, Ramat-Gan, Israel
[c] CRI, University of Haifa, Mount Carmel, Haifa 31905, Israel
[d] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, United States

ARTICLE INFO

Article history:
Received 10 July 2012
Received in revised form 6 November 2013
Available online 20 February 2015

Keywords:
String algorithms
Periodicity
Approximate periodicity
Approximate matching
Hamming distance
Swap distance

ABSTRACT

Finding an approximate period in a given string S of length n is defined as follows. Let S' be a periodic string closest to S under some distance metric, find the smallest period of S'. This period is called *an approximate period* of S under the given metric. Let the distance between the input string S and a closest periodic string under the Hamming distance S' be k. We develop algorithms that construct an approximate period of S under the Hamming distance in time $O(nk \log\log n)$ and under the swap distance in time $O(n^2)$. Finally, we show an $O(n \log n)$ algorithm for finite alphabets, and an $O(n \log^3 n)$ algorithm for infinite alphabets, that approximate the minimum number of mismatches between the input string and a closest periodic string under the Hamming distance.

1. Introduction

String Periodicity is a classic topic in Computer Science. It has been extensively studied over the years [31] and linear time algorithms for exploring the periodic nature of a string were suggested (e.g. [17]). Multidimensional periodicity [3,21,34] and periodicity in parameterized strings [11] were also explored. In addition, periodicity has played a role in efficient parallel string algorithms [19,4,5,15].

Furthermore, periodicity models cyclic natural phenomena. These phenomena abound in such diverse areas as Astronomy, Geology, Earth Science, Oceanography, Meteorology, Biological Systems, the Genome, Economics, and more. Realistic data may contain errors. Such errors may be caused by the process of gathering the data which might be prone to transient errors. Errors can also be an inherent part of the data because the periodic nature of the data represented by the string may be inexact. However, it is still valuable to detect and utilize the underlying periodicity. This calls for the notion of *approximate periodicity*. To our knowledge, although there has been quite an extensive research on the related notion of approximate multiple tandem repeats as discussed below in the related work subsection, all previous work on (full) periodicity dealt with exact periodicity. Approximate tandem repeats differ from the problem of approximate periodicity as defined in this paper (see Subsection 1.1).

☆ A preliminary version appeared in the proceedings of ISAAC 2010.

* Corresponding author at: Department of Software Engineering, Shenkar College, 12 Anna Frank, Ramat-Gan, Israel.
E-mail addresses: amir@cs.biu.ac.il (A. Amir), estchoc@gmail.com (E. Eisenberg), avivitlevy@shenkar.ac.il (A. Levy).

[1] Partly supported by NSF grant CCR-09-04581 and ISF grant 571/14.

http://dx.doi.org/10.1016/j.ic.2015.02.004
0890-5401/

A natural way of handling errors is the following. Since we may not be confident of our measurement or suspect the periodic process to be inexact, then we may be interested in finding a current *approximate periodic* nature of the string, i.e., what is a smallest length period of a periodic string S' that is closest to the input string S. We define such a period of S' as *an approximate period* of S. Note that, the approximate period in the meaning of "length" is uniquely defined, but for the same optimal length, several words may be called a period each, since there can be several strings S' that have the minimal distance to the input string S. It is natural to ask if such an approximate period can be found efficiently.

Different phenomena may be plagued by different types of errors. Thus the notion of "closeness" should be defined according to the different error causes. This is formalized by considering different distance metrics. In this paper we study approximate periodicity under two classical metrics: the *Hamming Distance*, where strings are possibly corrupted with *substitution* errors, i.e., a character may be substituted by a different character, and the *swap distance*, where the errors are the exchange of two adjacent symbols (with no symbol participating in more than one exchange). This definition of the swap distance is inspired by common text editing errors and was extensively used in the literature, e.g., in [2,16,10,14,23,1,9].

It may also be interesting to get an approximation of the number of errors, assuming approximate periodicity, even if the period or the exact number of errors are unknown. Such a number can either indicate that any cyclic phenomena is not interesting because it has too many errors, or may identify a small number of errors that may indeed define periodicity in the data. Can such an approximation be achieved quickly?

Results. Let S be a string of length n over alphabet Σ. We prove the following:

Finding an Approximate Period Under the Hamming Distance: An approximate period of S under Hamming distance can be found in time $O(nk \log\log n)$, where k is the distance between S and S' (Theorem 2).

Note that k is not known a priori.

This is done in Section 3.

Finding an Approximate Period Under the Swap Distance: An approximate period of S under the swap distance can be found in time $O(n^2)$ (Theorem 3).

This is done in Section 4.

Fast Approximation of the Error Bound For Hamming Distance: The number of mismatches between S and a closest periodic string under Hamming distance can be approximated to within a factor of 2 in time $O(|\Sigma| n \log n)$ (Theorem 4).

A surprising and important feature of our algorithm is the independence of the actual bound, i.e., its complexity is (almost) linear even if many errors are needed in order to assume periodicity in the input string.

For infinite alphabets, for every $\epsilon > 0$, the number of mismatches between S and a closest periodic string under Hamming distance can be approximated to a factor of $2(1+\epsilon)$ in time $O(\frac{1}{\epsilon} \cdot n \log^3 n)$ (Theorem 5).

This is done in Section 5.

This paper is a full version of [6]. In particular, it includes the proofs of all lemmas, which were all omitted from [6] except for the proof of Lemma 2. The authors of this paper also studied the related problem of finding closest periodic vectors under L_p distances [7], and the related but different question of period recovery, i.e., can we tell something about the period of the original string given a possibly corrupted string [8].

1.1. Related work

The notion of approximate periodicity as defined in this paper is *related but different* from the known and studied notion of *approximate tandem repeats* (alternatively, *squares*, *powers* or *runs*). We do not attempt to give a comprehensive presentation of the work on tandem repeats, but rather to emphasize the difference of this paper. A *perfect single tandem repeat* is defined as a nonempty string that can be divided into two identical sub-strings, e.g., *abcabc*. It is a well-studied problem motivated first from research in formal languages [33]. Main and Lorentz [32] present an $O(n \log n)$ algorithm which reports, for a given input string, all substrings that are perfect tandem repeats. Repeats also occur frequently in biological sequences, yet they are seldom exact. This motivated the study of approximate tandem repeats. An *approximate single repeat* is a nonempty string that can be divided into two similar substrings. The distance between the two substrings must be less than a given threshold k, in order for the two parts to be considered similar. Common studied distances are the Hamming distance and the (possibly weighted) edit distance (e.g. [24,12,35,13,30,28]).

The problem of finding all approximate single repeats is a sub-problem of finding all approximate multiple repeats in a string. A *perfect multiple repeat* is a nonempty string that can be divided into a number of identical adjacent substrings, or periods. The last period of the repeat can be partial. Note that the related problem of finding the (exact) period of a given string is a simpler problem, because this period should repeat from beginning to end. It is, therefore, expected that better algorithms can be found for this problem. An *approximate multiple repeat* is a multiple repeat in which the periods of the repeat are approximate.

A number of different definitions of the approximate multiple repeat were studied, even for variations that are all based on the Hamming distance (e.g. [30,36,28]). All these definitions attempt to capture the biologically relevant relationships between biosequences, and have, in some sense, a local approach to defining the "approximate repetitions". For example, [30]