Discrete Applied Mathematics 161 (2013) 1556–1575

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

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

# Fast algorithms for finding a minimum repetition representation of strings and trees☆

Atsuyoshi Nakamura [a,*], Tomoya Saito [a], Ichigaku Takigawa [a], Mineichi Kudo [a], Hiroshi Mamitsuka [b]

[a] Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan

[b] Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

ARTICLE INFO

*Article history:*
Received 4 June 2011
Received in revised form 13 September 2012
Accepted 18 December 2012
Available online 21 January 2013

*Keywords:*
Tandem repeat
String algorithm
Labeled ordered trees

ABSTRACT

A string with many repetitions can be represented compactly by replacing $h$-fold contiguous repetitions of a string $r$ with $(r)^h$. We present a compact representation, which we call a *repetition representation (of a string)* or RRS, by which a set of disjoint or nested tandem arrays can be compacted. In this paper, we study the problem of finding a *minimum RRS* or MRRS, where the size of an RRS is defined by the sum of the length of component letters and the description length of the component repetitions $(\cdot)^h$ which is defined by $w_{\mathcal{R}}(h)$ using a repetition weight function $w_{\mathcal{R}}$. We develop two dynamic programming-based algorithms to solve this problem: CMR, which works for any type of $w_{\mathcal{R}}$, and CMR-C, which is faster but can be applied to a constant $w_{\mathcal{R}}$ only. CMR-C is an $O(n^2 \log n)$-time $O(n \log n)$-space algorithm, which is more efficient in both time and space than CMR by a $((\log n)/n)$-factor, where $n$ is the length of the given string. The problem of finding an MRRS for a string can be extended to that of finding a *minimum repetition representation (of a tree)* or MRRT for a given labeled ordered tree. For this problem, we present two algorithms, CMRT and CMRT-C, by using CMR and CMR-C, respectively, as a subroutine. As well as the theoretical analysis, we confirmed the efficiency of the proposed algorithms by experiments, which consist of the following three parts: First we demonstrated that CMR-C and CMRT-C are fast enough for large-scale data by using synthetic strings and trees, respectively. The size of an MRRS for a given string can be a measure of how compactly the string can be represented, meaning how well the string is structurally organized. This is also true of trees. To check such ability of MRRS-size, second we measured the size of an MRRS for chromosomes of nine different species. We found that all the chromosomes of the same species have a similar compression rate when realized by an MRRS. Run length encoding (RLE) was also shown to have species-specific compression rate, but species were separated more clearly by MRRS than by RLE. Third we examined the size of an MRRT for web pages of world-leading companies by using the tag trees, showing a consistency between the compression rate by an MRRT and visual web page structures.



## 1. Introduction

Repetitions have special meanings in a lot of real-world applications. For example, it is already known that the contiguous repetitions in DNA sequences are related to human diseases [1], and we can recognize a number of data records embedded in web pages by detecting contiguously repeated HTML tag structures [10].

☆ The preliminary version of this paper has appeared in [12].

* Corresponding author. Tel.: +81 11 706 6806.
*E-mail address:* atsu@main.ist.hokudai.ac.jp (A. Nakamura).

0166-218X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2012.12.013

We first recall some notions specific to repetitions in a string as follows: A contiguous repetition is called a *tandem array*, and it is also called a *tandem repeat* if the number of repetitions is only two. The *repetition string* of a tandem array is the unit of the repetition. A tandem array (repeat) is *primitive* if the repetition string itself is not a tandem array. Finding repetitions in a given string has been studied for more than two decades in many fields including computer science, mathematics and biology [15]. Thus a lot of efficient algorithms for this problem have been already developed [15,6,11].

Tandem arrays in a given string allow us to represent the string with repetitions. However we note that the string might have a variety of repetitions, which can overlap with each other. For example, the string *abaababaab* can be regarded as a repetition of *abaab*, and another interpretation of *abaababaab* is a concatenation of *abaaba* and *baab*, where *abaaba* is a repetition of *aba*. The best representation of the string might depend upon the situation, but some consistent criterion to evaluate the representation of a given string would be useful and imperative.

For this issue, we present one notion, which we call *repetition representation (of a string)* or RRS, in which we use $(r)^h$ instead of $rr\cdots r$, i.e., that $r$ is repeated $h$ times. An RRS is a representation using a set of *disjoint* or *nested* tandem arrays in a given string $s[1..n]$, where two substrings $s[i_1..j_1]$ and $s[i_2..j_2]$ are disjoint if they have no intersection (i.e., $j_1 < i_2$ or $j_2 < i_1$) while two substrings $s[i_1..j_1] = (r)^h$ and $s[i_2..j_2] = (r')^{h'}$ are nested if $r$ contains $(r')^{h'}$ (i.e., $i_1 \le i_2 < j_2 < i_1 + (j_1 - i_1 + 1)/h$) or vice versa. An RRS cannot always represent all tandem arrays, which appear in one string, because tandem arrays in one string may overlap and we have to select some of them, resulting in different RRSs for one string depending on the selected tandem arrays. We compare different RRSs with each other by using the size of the RRS, and select a *minimum RRS* (or MRRS) as the best RRS in terms of the size. That is, we consider the problem of finding a minimum RRS of a given string. We define the size of an RRS by the sum of the length of component letters and the description length of the component repetitions $(\cdot)^h$, which is defined to be $w_{\mathcal{R}}(h)$ using the function $w_{\mathcal{R}}$, which we call a *repetition weight function*. Thus the idea of selecting an MRRS (which must represent the most important RRS) is consistent with the well-known MDL (Minimum Description Length) principle [14], which says that "the success in finding regularities can be measured by the length with which the data can be described". Note that the same problem was already studied for the case with $w_{\mathcal{R}}(h) = \log_{10}(h+1)$ using the logarithmic cost model in [8] while, in this paper, we deal with another interesting case of $w_{\mathcal{R}}(h)$, i.e., the case with constant repetition weight functions as well as a general case using the uniform cost model.

For the problem of finding an MRRS for a given string, we propose two algorithms: CMR and CMR-C. CMR works under any repetition weight function $w_{\mathcal{R}}$, though it is slow; it runs in $O(n^3)$ time and $O(n^2)$ space, where $n$ is the length of a given string. On the other hand, CMR-C works under a constant $w_{\mathcal{R}}$ only, but it is faster; it is an $O(u(n+z))$ time and $O(n+z)$ space algorithm, where $z$ is the number of (occurrences of) primitive tandem repeats in a string and $u$ is the number of unique primitive tandem repeats. The worst case setting of $u$ and $z$ is $u = O(n)$ and $z = O(n \log n)$ [6,4], resulting in $O(n^2 \log n)$ time and $O(n \log n)$ space, which are smaller than those of CMR by a $(\log n)/n$ factor.

We can extend the RRS to the *repetition representation (of a tree)* or RRT, in which a labeled ordered tree can be compactly represented by incorporating special nodes, which we call *repetition information nodes*, where a repetition information node indicates the number of repetitions for contiguous repetitions of subtree sequences. More concretely, the repetition of subtree sequences can be represented by a repetition information node and the subtree sequences, which are rooted in the child nodes of the repetition information node. The algorithm for finding an MRRS can be used as a subroutine of an algorithm for finding a *minimum RRT* or MRRT, which yields two algorithms: CMRT and CMRT-C, for finding an MRRT. CMRT uses CMR under any repetition weight function $w_{\mathcal{R}}$ which runs in $O(n^3)$ time and $O(n^2)$ space, and CMRT-C uses CMR-C under a constant $w_{\mathcal{R}}$ which runs in $O(n^2 \log n)$ time and $O(n \log n)$ space, where $n$ is the number of nodes in a given labeled ordered tree.

As well as the theoretical analysis, we empirically checked the computational efficiency of the proposed algorithms by using synthesized data. We generated binary strings randomly, keeping the amount of primitive tandem repeats at around $0.82 \times n$ for the entire string length of $n$, this value being larger than that of real biological data, which will be examined later. Experimental results showed that CMR-C ran in almost linear time with respect to the string length, while CMR needed approximately cubic time, being consistent with the theoretical upper bound. For example, CMR-C used only 0.027 s for a random binary string of 12,800 letters, while CMR spent one hour and 12 min for the same string. In addition, CMR-C needed only 8.2 s even for a string of 1,638,400 letters. This result indicates that CMR-C can be applied to very long, large-scale and real datasets. In fact, we could apply CMR-C to DNA sequences with around $8 \times 10^7$ letters within a practical amount of computation time. Regarding labeled ordered trees, we generated synthetic tree data randomly and demonstrated that CMRT-C was faster than CMRT, particularly in the case that the average number of children of internal nodes is relatively large.

The size of an MRRS can be a measure of how compactly a string can be represented by repetition structures, and this is also true for labeled ordered trees. We used DNA sequences, i.e., chromosomes, of nine different species as real strings, and measured the size of an MRRS of each chromosome. Experimental results indicate that all the chromosomes of the same species have a similar compression rate when realized by an MRRS. Run length encoding (RLE) was also shown to have species-specific compression rate, but species were separated more clearly by MRRS than by RLE. Interestingly, the compression rate by an MRRS over the chromosome of yeast was totally different from that of chicken, despite that the chromosomes of these two species have almost the same density of primitive tandem repeats. This indicates an essential difference of repetition structures between these two species. We further used the web pages of world-leading companies and examined the size of an MRRT of the HTML tag trees in these web pages. The results indicate that web pages with relatively larger or more uniformly structured areas are compressed more by an MRRT, implying that the compression rate by an MRRT can be a measure of the complexity of visual web page structures.