International Workshop On Combinatorial Algorithms
IWOCA
`http://www.iwoca.org` → Problems Section

From String Attractors to Strings

Dominik Köppl

Tokyo Medical and Dental University, Japan,
`koeppl.dsc@tmd.ac.jp`

presented at IWOCA 2022 on June 10
last update: November 2, 2022

Given a string $T[1..n]$, a *string attractor* Γ is a set of positions $\Gamma \subset [1..n]$ such that every substring S of T has an occurrence $T[i..i+|S|-1]$ in T such that $[i..i+|S|-1] \cap \Gamma \neq \emptyset$, see also [2].

Example 1. For $T = banana$, a minimal string attractor is $\Gamma = \{1, 2, 3\}$ since all substrings of T have an occurrence that intersects with $T[1..3]$. For instance, the suffix na has another occurrence starting at position 3, and therefore is "hit" by Γ.

Problem 1. For a given set $\Gamma \subset [1..n]$, find all strings whose smallest string attractor is Γ.

Comment 1. Already for $\Gamma = \{1\}$, there can be infinitely many strings such as `a`, `aa`, `aaa`... having Γ as smallest string attractor. However, if such a string becomes too long, then it becomes *ultimately periodic* [4], meaning that it is a prefix of SP^∞, where S and P are finite strings. So these strings can be classified by S and P. Hence, we can classify a string derived from $\Gamma = \{1\}$ just by the first letter a and its length.

Definition 1. We represent a string T by the triplet (S, P, ℓ) such that $S \cdot P^\ell = T$, S and P are strings, and ℓ a rational number. Further, no rotation of P is a suffix of S (otherwise we could increase ℓ), and P is the shortest possible such string. We say that the triplet is the *ultimately periodic representation* of T.

Example 2.

string	ultimately periodic representation
abbb	(a,b,3)
abcbcbc	(a,bc,3)
abcabab	(abc,ab,2)

We reformulate Problem 1 as follows:

Problem 2. For a given set $\Gamma \subset [1..n]$, what is the number of different ultimately periodic representation when neglecting the length ℓ? (meaning that we count (S, P, ℓ) and (S, P, ℓ') for $\ell \neq \ell'$ only once)

Comment 2. It is still unknown whether we can represent every string T in space $\mathcal{O}(\gamma_T)$, where γ_T is the size of a smallest string attractor of T [3]. However, [1] showed that we can compress every string T of length n into $\mathcal{O}(\gamma_T \log \frac{n}{\gamma_T})$ space.

Assume that we have a representation of a string of length n within $c\gamma \log n$ bits, for $\gamma := \gamma_T$. Then this number of bits is enough to enumerate solutions from 1 to $2^{c\gamma \log n} = n^{c\gamma}$. This leads to another problem:

Problem 3. Prove or disprove: There is a constant c (depending on the alphabet size) such that for any length n, the number of strings of length n having a string attractor of size γ is at most $n^{c\gamma}$.

References

[1] Anders Roy Christiansen, Mikko Berggren Ettienne, Tomasz Kociumaka, Gonzalo Navarro, and Nicola Prezza. Optimal-time dictionary-compressed indexes. *ACM Trans. Algorithms*, 17(1):8:1–8:39, 2021. `doi:10.1145/3426473`.

[2] Dominik Kempa and Nicola Prezza. At the roots of dictionary compression: string attractors. In Ilias Diakonikolas, David Kempe, and Monika Henzinger, editors, *Proc. STOC*, pages 827–840. ACM, 2018. URL: `http://doi.acm.org/10.1145/3188745.3188814`, `doi:10.1145/3188745.3188814`.

[3] Gonzalo Navarro. Indexing highly repetitive string collections, part I: repetitiveness measures. *ACM Comput. Surv.*, 54(2):29:1–29:31, 2021. `doi:10.1145/3434399`.

[4] Antonio Restivo, Giuseppe Romana, and Marinella Sciortino. String attractors and infinite words. In *Proc. LATIN*, pages 426–442, 2022. `doi:10.1007/978-3-031-20624-5_26`.