Using Automata and a Decision Procedure to Prove Results in Pattern Matching

Jeffrey Shallit
School of Computer Science, University of Waterloo, Canada

Abstract

The first-order theory of automatic sequences with addition is decidable, and this means that one can often prove combinatorial properties of these sequences "automatically", using the free software `Walnut` written by Hamoon Mousavi. In this talk I will explain how this is done, using as an example the measure of minimize size string attractor, introduced by Kempa and Prezza in 2018.

Using the logic-based approach, we can also prove more general properties of string attractors for automatic sequences. This is joint work with Luke Schaeffer.

2012 ACM Subject Classification Mathematics of computing → Combinatorics on words; Theory of computation → Regular languages; Theory of computation → Logic and verification

Keywords and phrases finite automata, decision procedure, automatic sequence, Thue-Morse sequence, Fibonacci word, string attractor

Digital Object Identifier 10.4230/LIPIcs.CPM.2022.2

Category Invited Talk

Related Version *Full Version*: `https://arxiv.org/abs/2012.06840`

Funding Research supported by NSERC 2018-04118.

1 Introduction

Many famous sequences, such as the Thue-Morse sequence $\mathbf{t} = \texttt{01101001}\cdots$ and the Fibonacci infinite word $\mathbf{f} = \texttt{01001010}\cdots$ appear as fundamental examples in combinatorial pattern matching.

As just a few examples, I point to [5, 1, 12], where the Thue-Morse sequence makes an appearance, and [13], where the Fibonacci infinite word is studied.

A fundamental result, essentially due to Büchi [4] and Bruyère et al. [3], tells us that the first-order theory of such sequences, with addition, is decidable, and there is a relatively simple decision procedure based on automata. This decision procedure has been implemented in free software called `Walnut`, originally created by Hamoon Mousavi [11]. Therefore, in many cases, we can prove properties of such sequences of interest to the CPM community "automatically", merely by stating the desired property in first-order logic, and invoking `Walnut`.

Recently there has been interest in a certain measure of repetitivity, based on string attractors, originally introduced by Kempa and Prezza [6], and studied further in [9, 7, 8, 10, 2]. A *string attractor* of a finite word $w = w[0..n-1]$ is a subset $S \subseteq \{0, 1, \ldots, n-1\}$ such that every nonempty factor f of w has an occurrence that touches at least one of the indices of S. For example, $\{2, 3, 4\}$ is a string attractor of minimum size for the French word `entente`.

In this talk I will introduce `Walnut`, and explain how to obtain results on string attractors using it and the theory behind it. This is joint work with Luke Schaeffer [14].

2 Results

As an example of the kind of thing we can prove with `Walnut`, here is one theorem:

▶ **Theorem 1.** *Let a_n denote the size of the smallest string attractor for the length-n prefix of the Thue-Morse word* $\mathbf{t}$*. Then*

$$a_n = \begin{cases} 1, & \text{if } n = 1; \\ 2, & \text{if } 2 \leq n \leq 6; \\ 3, & \text{if } 7 \leq n \leq 14 \text{ or } 17 \leq n \leq 24; \\ 4, & \text{if } n = 15, 16 \text{ or } n \geq 25. \end{cases}$$

More generally, we can prove

▶ **Theorem 2.** *Let* $\mathbf{w}$ *be a k-automatic sequence. Either*
- *every factor $\mathbf{w}[i..i+\ell-1]$ has a string attractor of constant size, and there exists a finite automaton outputting the minimum size given i and ℓ, or*
- *for all $n \geq 1$, the minimum size string attractor for the length-n prefix $\mathbf{w}[0..n-1]$ grows as $\Theta(\log n)$,*

and we can decide which is the case for $\mathbf{w}$*.*

For more about `Walnut` and its applications in combinatorics on words, see my forthcoming book [15].

References

1 A. Amir, Y. Aumann, A. Levy, and Y. Roshko. Quasi-distinct parsing and optimal compression methods. In G. Kucherov and E. Ukkonen, editors, *CPM 2009*, volume 5577 of *Lecture Notes in Computer Science*, pages 12–25. Springer-Verlag, 2009.

2 H. Bannai, M. Funakoshi, T. I, D. Köppl, T. Mieno, and T. Nishimoto. A separation of γ and b via Thue–Morse words. In T. Lecroq and H. Touzet, editors, *SPIRE 2021*, volume 12944 of *Lecture Notes in Computer Science*, pages 168–178. Springer-Verlag, 2021.

3 V. Bruyère, G. Hansel, C. Michaux, and R. Villemaire. Logic and p-recognizable sets of integers. *Bull. Belgian Math. Soc.*, 1:191–238, 1994. Corrigendum, *Bull. Belg. Math. Soc.* 1:577, 1994.

4 J. R. Büchi. Weak secord-order arithmetic and finite automata. *Z. Math. Logik Grundlagen Math.*, 6:66–92, 1960. Reprinted in S. Mac Lane and D. Siefkes, eds., *The Collected Works of J. Richard Büchi*, Springer-Verlag, 1990, pp. 398–424.

5 M. Karpinski, W. Rytter, and A. Shinohara. An efficient pattern-matching algorithm for strings with short descriptions. *Nordic J. Computing*, 4:172–186, 1997.

6 D. Kempa and N. Prezza. At the roots of dictionary compression: string attractors. In *STOC'18 Proceedings*, pages 827–840. ACM Press, 2018.

7 T. Kociumaka, G. Navarro, and N. Prezza. Towards a definitive measure of repetitiveness. In Y. Kohayakawa and F. K. Miyazawa, editors, *LATIN 2020*, volume 12118 of *Lecture Notes in Computer Science*, pages 207–219. Springer-Verlag, 2020.

8 K. Kutsukake, T. Matsumoto, Y. Nakashima, S. Inenaga, H. Bannai, and M. Takeda. On repetitiveness measures of Thue-Morse words. In C. Boucher and S. V. Thankachan, editors, *SPIRE 2020*, volume 12303 of *Lecture Notes in Computer Science*, pages 213–220. Springer-Verlag, 2020.

9 S. Mantaci, A. Restivo, G. Romana, G. Rosone, and M. Sciortino. String attractors and combinatorics on words. In *ICTCS 2019*, volume 2504 of *CEUR Workshop Proceedings*, pages 57–71, 2019. Available at `http://ceur-ws.org/Vol-2504/paper8.pdf`.

10 S. Mantaci, A. Restivo, G. Romana, G. Rosone, and M. Sciortino. A combinatorial view on string attractors. *Theoret. Comput. Sci.*, 850:236–248, 2021.

11 H. Mousavi. Automatic theorem proving in `Walnut`. Arxiv preprint arXiv:1603.06017 [cs.FL], 2016. `arXiv:1603.06017`.

12 J. Radoszewski and W. Rytter. On the structure of compacted subword graphs of Thue-Morse words and their applications. *J. Discrete Algorithms*, 11:15–24, 2012.

13 W. Rytter. The structure of subword graphs and suffix trees of Fibonacci words. *Theoret. Comput. Sci.*, 363:211–223, 2006.

14 L. Schaeffer and J. Shallit. String attractors for automatic sequences. Arxiv preprint arXiv:2012.06840 [cs.FL], 2021. `arXiv:2012.06840`.

15 J. Shallit. *The Logical Approach To Automatic Sequences: Exploring Combinatorics on Words with* `Walnut`. Cambridge University Press, 2022. In press.