A generalized pseudostandard prefix with no string attractor of size four

P. Graham
https://orcid.org/0009-0007-8493-1505

October 26, 2026

Abstract

A string attractor of a finite word is a set of positions such that every nonempty factor has an occurrence containing one of them. Binary generalized pseudostandard sequences are built by iterating palindromic and antipalindromic closures along a directive bi-sequence (Δ, Θ), where Δ lists the letters appended and Θ the closures taken, R for palindromic and E for antipalindromic. Dvořáková and Hendrychová (2023) determined the minimal string attractors of the pseudopalindromic prefixes of standard complementary-symmetric Rote sequences, gave attractors of at most three positions for the antipalindromic prefixes of pseudostandard sequences, and conjectured that the pseudopalindromic prefixes of every binary generalized pseudostandard sequence have a minimal string attractor of size at most four. We show that this bound fails at length 36: for Δ beginning 00101 and Θ beginning $EEEER$, the fifth word of the recurrence, the palindromic prefix 010110010101100101101001101010011010, has smallest attractor size five. The lower bound is an elementary argument with seven factors; the upper bound is an explicit five-position attractor. Both bounds are also confirmed by computer. A second example has length 43.

The proofs, the certificate checkers and the text were produced by AI models (from OpenAI and Anthropic), working under the author's direction.

Status: Computer-assisted. This work has not yet been reviewed by experts.

1 Introduction

A *string attractor* of a finite word $w = w_0 w_1 \cdots w_{n-1}$, a notion due to Kempa and Prezza [8], is a set $\Gamma \subseteq \{0, 1, \ldots, n-1\}$ of positions such that every nonempty factor of w has an occurrence $w_i w_{i+1} \cdots w_{j-1}$ with $\{i, \ldots, j-1\} \cap \Gamma \neq \emptyset$. For example, $\{1, 3\}$ is an attractor of `010010` [5, Section 2]. Positions are numbered from 0, as in [5].

Generalized pseudostandard sequences, introduced by de Luca and De Luca [3], are built by iterated closures. For a binary word v, let $R(v)$ be its reversal and $E(v)$ its reversal with 0 and 1 exchanged; v^R and v^E are the shortest words with prefix v fixed by R and by E. A directive bi-sequence (Δ, Θ), with $\Delta = \delta_1 \delta_2 \cdots$ over $\{0, 1\}$ and $\Theta = \vartheta_1 \vartheta_2 \cdots$ over $\{E, R\}$, gives the words $w_0 = \varepsilon$ (the empty word) and $w_{n+1} = (w_n \delta_{n+1})^{\vartheta_{n+1}}$, each a prefix of the next. Their limit is the sequence $\mathbf{u}(\Delta, \Theta)$, and each w_n is a pseudopalindromic prefix of it, that is, a prefix fixed by R or by E (Section 2 quotes the definitions).

Dvořáková and Hendrychová [5] conjectured in August 2023, from computer experiments, that every pseudopalindromic prefix of every such sequence has an attractor of at most four positions (Conjecture 2). The bound fails at length 36.

Theorem 1. *Let* (Δ, Θ) *be any directive bi-sequence with* Δ *beginning* 00101 *and* Θ *beginning* $EEEER$*. The fifth word* w_5 *of the recurrence is the palindrome*

$$w_5 = 0101100101011001011010011010010011010$$

of length 36*, a pseudopalindromic prefix of* $\mathbf{u}(\Delta, \Theta)$*. It has no string attractor of size four or less, and* $\Gamma = \{3, 7, 15, 21, 25\}$ *is a string attractor of it. So its smallest string attractor has size five, and Conjecture 2 is false.*

Figure 1 shows w_5 with the attractor Γ.

The smallest size of an attractor measures how repetitive a word is: Kempa and Prezza showed that it is at most the size of the representations built by dictionary compressors, such as the number of phrases of the LZ77 parse [8]. The conjecture extends results of [5] and earlier work. For a palindromic prefix of a standard Sturmian sequence the minimal attractors have one position per letter that occurs [4]. For an antipalindromic prefix v of a pseudostandard sequence ($\Theta = E^\omega$, the letter E repeated forever) Dvořáková and Hendrychová gave an attractor of at most three positions; for $v = w_n$ with $n \geq 2$ and Δ beginning with 0, it has minimum size exactly when Δ does not begin with 01^{n-1} [5, Theorem 8]. For a pseudopalindromic prefix of a standard complementary-symmetric Rote sequence, a subclass of the generalized pseudostandard sequences found by Blondin Massé, Paquin, Tremblay and Vuillon [2], they proved that the minimal attractors have one position per letter that occurs [5, Theorem 19]. The Thue–Morse sequence is also generalized pseudostandard, and its pseudopalindromic prefixes of length 2^n need four positions once $n \geq 4$ [9]; all their smallest attractors were recently described [1, 6]. So four would have been best possible.

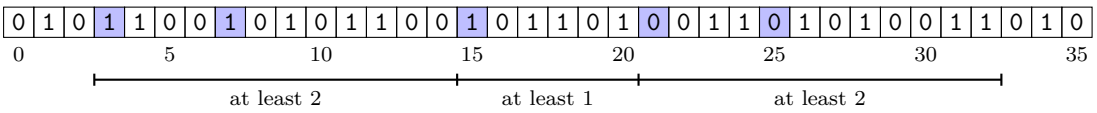


Figure 1: The prefix w_5, positions 0 to 35, with the attractor $\Gamma = \{3, 7, 15, 21, 25\}$ shaded. Below it are the three disjoint regions $[3, 14]$, $[15, 20]$ and $[21, 32]$ of the lower-bound proof, with the number of positions each must contain.

What is new, and the hard step

Dvořáková and Hendrychová's conjecture rests on their computer experiments. The subclass results—Dvořáková's standard Sturmian result [4], restated and re-proved in [5], and the pseudostandard and standard complementary-symmetric Rote results of Dvořáková and Hendrychová [5]—are untouched by Theorem 1. The word w_5 has ten distinct factors of length four, while a Sturmian sequence has five and a Rote sequence eight [10], so w_5 is a factor of neither; and w_5 is a palindrome, not one of the antipalindromic prefixes that [5, Theorem 8] describes. What this note adds is a prefix outside these classes that needs five positions. The upper bound is a direct check. The hard step is the lower bound: that no four positions form an attractor. A priori this means ruling out every set of four positions; we replace that by seven factors whose occurrences force five positions (Section 4).

Proof overview and map

Section 2 quotes the definitions and the conjecture. Section 3 computes the five closures that produce w_5, so that w_5 is a pseudopalindromic prefix exactly in the sense of the conjecture. Section 4

proves the lower bound: an attractor must meet the union of the occurrences of each factor, and seven factors, with only one or two occurrences each, force at least two positions in $[3, 14]$, one in $[15, 20]$ and two in $[21, 32]$. Section 5 proves that Γ is an attractor, by reducing the check to the factors that occur inside a gap of Γ, the end gaps included (listed in Appendix A). Section 6 gives a second example, of length 43, and Section 7 says what remains open. Appendix B describes the computer checks. The proof of Theorem 1 is complete without them; the second example rests on them.

2 Definitions and the conjecture

This section quotes the definitions of [5] (arXiv version 5, which is the published version) word for word, so that the example can be checked against them. Section 2 of [5] defines attractors: "A string attractor (or attractor for short) of a word $w = w_0 w_1 \cdots w_{n-1}$, where $w_i \in \mathcal{A}$, is a set $\Gamma \subset \{0, 1, \ldots, n-1\}$ such that every non-empty factor of w has an occurrence in w containing at least one element of Γ." Here an occurrence of a factor $w_i w_{i+1} \cdots w_{j-1}$ is the set of positions $\{i, i+1, \ldots, j-1\}$. Section 3 of [5] defines, after its Definition 1 of R and E:

> "Consider $w \in \{0,1\}^*$. Then w^R is the shortest palindrome having w as prefix and it is called the palindromic closure of w. Similarly, w^E is the shortest antipalindrome having w as prefix and it is called the antipalindromic closure of w."
>
> "Let $w \in \{0,1\}^+$. It follows immediately from the definition that $w^R = vxR(v)$, where $w = vx$ and x is the longest palindromic suffix of w. Similarly, $w^E = vyE(v)$, where y is the longest antipalindromic suffix of w (possibly empty)."

Definition 3 of [5] reads:

> "**Definition 3.** Let $\Delta = \delta_1 \delta_2 \cdots$ and $\Theta = \vartheta_1 \vartheta_2 \cdots$, where $\delta_i \in \{0,1\}$ and $\vartheta_i \in \{E, R\}$ for all $i \in \mathbb{N}$, $i \geq 1$. The sequence $\mathbf{u}(\Delta, \Theta)$, called generalized pseudostandard sequence, is the sequence having prefixes w_n obtained from the recurrence relation
>
> $$w_{n+1} = (w_n \delta_{n+1})^{\vartheta_{n+1}}, \qquad w_0 = \varepsilon.$$
>
> The bi-sequence (Δ, Θ) is called the directive bi-sequence of the word $\mathbf{u}(\Delta, \Theta)$."

The conjecture is not numbered in [5]. It is stated in the abstract and in Section 7 (Open problems); we number it here for reference.

Conjecture 2 (Dvořáková and Hendrychová [5])**.** *Abstract: "We conjecture that the pseudopalindromic prefixes of any binary generalized pseudostandard sequence have a minimal string attractor of size at most four." Section 7: "Based on computer experiments, we conjecture that the size of minimal string attractors of pseudopalindromic prefixes of binary generalized pseudostandard sequences is of size at most four."*

In [5] a *minimal* attractor is one of minimum size (Section 2 there), so the conjecture says that every such prefix has an attractor of at most four positions. Definition 3 places no condition on (Δ, Θ), and neither statement of the conjecture adds one.

3 The prefix w_5

This section shows that w_5 is the fifth word of the recurrence in Definition 3. Write $\vartheta(v)$ for $R(v)$ or $E(v)$. By the remark quoted above, if the seed $w_{n-1}\delta_n$ is written as vx with x its longest ϑ_n-fixed suffix, then $w_n = vx\,\vartheta_n(v)$. Table 1 lists the five steps for $\Delta = 00101\cdots$ and $\Theta = EEEER\cdots$. In steps 1 and 3 no nonempty suffix of the seed is fixed by E (a single letter never is), so x is empty and w_n has twice the length of the seed.

n	δ_n	ϑ_n	seed $w_{n-1}\delta_n = vx$	x	$\vartheta_n(v)$	$\lvert w_n\rvert$
1	0	E	0	ε	1	2
2	0	E	0|10	10	1	4
3	1	E	01011	ε	00101	10
4	0	E	0101100|1010	1010	1100101	18
5	1	R	01011001010110010|11	11	01001101010011010	36

Table 1: The closures producing w_5. A bar separates v from x in the seed, and $\lvert w_n\rvert$ is the length of w_n.

The words are

$$w_1 = 01, \quad w_2 = 0101, \quad w_3 = 0101100101, \quad w_4 = 010110010101100101,$$

and w_5 is the word of Theorem 1; it equals its reversal.

Lemma 3. *For every directive bi-sequence (Δ, Θ) with Δ beginning 00101 and Θ beginning EEEER, the word w_5 of Definition 3 is a pseudopalindromic prefix of the generalized pseudostandard sequence $\mathbf{u}(\Delta, \Theta)$.*

Proof. The words $w_1, \ldots, w_5$ depend only on the first five letters of Δ and Θ, and Table 1 computes them. Each w_{n+1} has the word $w_n\delta_{n+1}$, one letter longer than w_n, as a prefix, so the w_n are prefixes of one another with strictly increasing lengths, and $\mathbf{u}(\Delta, \Theta)$ is the infinite sequence they define. Thus w_5 is a prefix of it, and a palindrome. □

Counted by length, w_5 is the seventh nonempty pseudopalindromic prefix of $\mathbf{u}(\Delta, \Theta)$: the prefixes of lengths $1, 2, 3, 4, 10, 18$ and 36 are fixed by R or E. The conjecture covers it under either reading, as a word w_n of Definition 3 or as a pseudopalindromic prefix.

Remark 4. The conjecture asks for no further property of $\mathbf{u}(\Delta, \Theta)$, but the example also lives in an aperiodic sequence. Take $\Delta = 00101(01)^\omega$ and $\Theta = EEEERR^\omega$. Florian and Dvořáková [7] showed (see [5, Theorem 16]) that $\mathbf{u}(\Delta, \Theta)$ is eventually periodic exactly when some bijection $f\colon \{E, R\} \to \{0, 1\}$ satisfies $f(\vartheta_n) = \delta_{n+1}$ for all large n. Here $\vartheta_n = R$ for $n \geq 5$ while δ_{n+1} alternates, so no such f exists.

4 No four positions suffice

This section proves the lower bound of Theorem 1, the step that refutes the conjecture. For a factor f of w_5, let $C(f)$ be the set of positions lying in at least one occurrence of f. A set Γ attracts f exactly when it meets $C(f)$. In blocks of six, starting at positions $0, 6, 12, 18, 24, 30$,

$$w_5 = 010110\ 010101\ 100101\ 101001\ 101010\ 011010.$$

Lemma 5. *The following seven factors of w_5 occur exactly at the listed starting positions, so their position sets are as shown.*

region	factor f	starts	$C(f)$
left	$f_1 =$ `001010`	5	$[5, 10]$
left	$f_2 =$ `101011`	7	$[7, 12]$
left	$f_3 =$ `1100`	3, 11	$[3, 6] \cup [11, 14]$
middle	$f_4 =$ `101101`	15	$[15, 20]$
right	$f_5 =$ `010100`	25	$[25, 30]$
right	$f_6 =$ `110101`	23	$[23, 28]$
right	$f_7 =$ `0011`	21, 29	$[21, 24] \cup [29, 32]$

Proof. Each is found by reading w_5 in the display above. For instance `1100` needs two consecutive ones followed by two zeros; the only `11`'s start at $3, 11, 17, 23, 31$, and only those at 3 and 11 are followed by `00`. □

Proposition 6. *Every string attractor of w_5 has at least five positions.*

Proof. Let Γ be an attractor; it meets each of the seven sets of Lemma 5. The three left sets lie in $[3, 14]$. Their common intersection is empty, since $C(f_1) \cap C(f_2) = [7, 10]$ is disjoint from $C(f_3)$. So no single position meets all three, and Γ has at least two positions in $[3, 14]$. Likewise the right sets lie in $[21, 32]$, and $C(f_5) \cap C(f_6) = [25, 28]$ is disjoint from $C(f_7)$, so Γ has at least two positions in $[21, 32]$. Finally Γ meets $C(f_4) = [15, 20]$. The three intervals $[3, 14]$, $[15, 20]$ and $[21, 32]$ are disjoint, so $|\Gamma| \geq 2 + 1 + 2 = 5$. □

Faro, Marino and Pavone [6] use the same kind of certificate, a few factors whose occurrence sets force the size of every attractor, for the Thue–Morse words. The proposition is also confirmed by computer (Appendix B): no set of four positions is an attractor, and any set of fewer positions would be contained in one.

5 Five positions suffice

This section proves that $\Gamma = \{3, 7, 15, 21, 25\}$ is an attractor of w_5, which gives the upper bound. The positions of Γ split the rest of the word into the *gaps* $[0, 2]$, $[4, 6]$, $[8, 14]$, $[16, 20]$, $[22, 24]$ and $[26, 35]$.

Lemma 7. *A set Γ of positions of a word is an attractor if and only if every factor of every gap word (the factor occupying a gap) has an occurrence that contains a position of Γ.*

Proof. An occurrence that contains no position of Γ lies inside one gap. So a factor all of whose occurrences miss Γ is a factor of a gap word. The converse is the definition. □

Proposition 8. *The set $\Gamma = \{3, 7, 15, 21, 25\}$ is a string attractor of w_5.*

Proof. The six gap words are `010`, `100`, `0101100`, `01101`, `011` and `1010011010`. Together they have 48 distinct factors. Table 3 in Appendix A gives, for each, a starting position of an occurrence and a position of Γ that this occurrence contains. Lemma 7 completes the proof. □

For example, the longest gap word `1010011010` occupies $[26, 35]$ and also occurs at $[18, 27]$, which contains 21 and 25.

6 A second example

This section records a second prefix that needs five positions; its lower bound is checked only by computer. Take Δ beginning 001001 and Θ beginning $RRREER$. Table 2 lists the six closures, in the notation of Table 1. They give $w_1 = \mathtt{0}$, $w_2 = \mathtt{00}$, $w_3 = \mathtt{00100}$, then words of lengths 12 and 24, and finally the palindrome of length 43

$$w_6 = \mathtt{00100011101100100011101110001001101110000100}.$$

n	δ_n	ϑ_n	seed $w_{n-1}\delta_n = vx$	x	$\vartheta_n(v)$	$\lvert w_n\rvert$
1	0	R	`|0`	`0`	ε	1
2	0	R	`|00`	`00`	ε	2
3	1	R	`00|1`	`1`	`00`	5
4	0	E	`001000`	ε	`111011`	12
5	0	E	`00100011101|10`	`10`	`01000111011`	24
6	1	R	`001000111011001000|1110111`	`1110111`	`000100110111000100`	43

Table 2: The closures producing w_6. A bar separates v from x in the seed.

Proposition 9. *The smallest string attractor of w_6 has size five; $\{3, 11, 20, 25, 30\}$ is one.*

The proof is by computer (Appendix B). We know no short argument like Lemma 5 for w_6.

7 What remains open

Dvořáková and Hendrychová based the conjecture on computer experiments, and the bound four is attained by the Thue–Morse sequence [9]; Theorem 1 shows that it is exceeded already at length 36. The standard Sturmian result of Dvořáková [4], restated and re-proved in [5], and the pseudo-standard and standard complementary-symmetric Rote results of Dvořáková and Hendrychová [5] are untouched, since Theorem 1 concerns a directive bi-sequence outside those classes. We do not claim that w_5 is the shortest counterexample. Whether some constant bounds the attractor sizes of all these prefixes, and if so the best constant, remains open.

Appendix A The 48 factors of the gap words

Table 3 lists every distinct factor of the six gap words of Section 5, by length, with the start s of an occurrence in w_5 and a position $p \in \Gamma$ with $s \le p \le s + |f| - 1$. The same table is the file `data/upper-bound.json` in the result's folder.

Appendix B Computer checks

Two programs, written independently and using only the Python standard library, check the claims from the definitions. They are in the folder that accompanies this paper, whose `checks/README.md` explains how to run them.

The first, `checks/check_exhaustive.py`, rebuilds each closure as the shortest word fixed by R or E with the given prefix, trying the lengths in increasing order. It checks that the listed five positions attract every factor, and tests every set of four positions: none of the $\binom{36}{4} = 58{,}905$ sets for

factor	s	p	factor	s	p	factor	s	p
0	21	21	1001	4	7	010110	0	3
1	3	3	1010	7	7	011010	16	21
00	21	21	1011	1	3	100110	20	21
01	2	3	1100	3	3	101001	18	21
10	7	7	1101	23	25	101100	1	3
11	3	3	00110	21	21	0011010	21	21
001	5	7	01001	19	21	0100110	19	21
010	6	7	01011	0	3	0101100	0	3
011	2	3	01100	2	3	1001101	20	21
100	20	21	01101	22	25	1010011	18	21
101	1	3	10011	20	21	01001101	19	21
110	3	3	10100	18	21	10011010	20	21
0011	21	21	10110	1	3	10100110	18	21
0100	19	21	11010	17	21	010011010	19	21
0101	0	3	001101	21	21	101001101	18	21
0110	2	3	010011	19	21	1010011010	18	21

Table 3: Each factor of a gap word, with an occurrence starting at s that contains the position p of Γ.

w_5 and none of the $\binom{43}{4} = 123{,}410$ sets for w_6 is an attractor. Adding positions to an attractor gives an attractor, and a word of length at least four has a set of four positions containing any smaller set, so sets of fewer than four positions are excluded too. As controls it reproduces Example 2 and Example 4 of [5] and the attractor $\{1, 3\}$ of 010010.

The second, `checks/check_proof.py`, rebuilds each closure by the longest-suffix formula and confirms by trying every shorter extension that no shorter fixed word exists. It checks Lemma 5 and the proof of Proposition 6 (complete occurrence lists, empty intersections, disjoint regions), checks Table 3 against the gaps of Γ, and computes the smallest attractor size of w_5 and w_6 by a branch-and-bound search for a smallest set of positions meeting every $C(f)$. Both programs reject corrupted inputs: a flipped letter, a changed closure type, five positions that are not an attractor, a Thue–Morse prefix (which has an attractor of size four), and, for the second program, a missing occurrence or factor in the obstruction or the table. The whole check takes a few seconds.

References

[1] M. Banbara, H. Bannai, P. Guo, D. Köppl, T. Mieno and Y. Okamoto, *The smallest string attractors of Fibonacci, period-doubling, and Thue–Morse words*, arXiv:2602.16152, 2026.

[2] A. Blondin Massé, G. Paquin, H. Tremblay and L. Vuillon, *On generalized pseudostandard words over binary alphabets*, J. Integer Seq. **16** (2013), Article 13.2.11.

[3] A. de Luca and A. De Luca, *Pseudopalindrome closure operators in free monoids*, Theoret. Comput. Sci. **362** (2006), 282–300.

[4] L. Dvořáková, *String attractors of episturmian sequences*, Theoret. Comput. Sci. **986** (2024), 114341.

[5] L. Dvořáková and V. Hendrychová, *String attractors of Rote sequences*, Discrete Math. Theor. Comput. Sci. **26**:3 (2024), #2, doi:10.46298/dmtcs.12385, arXiv:2308.00850, first posted August 1, 2023.

[6] S. Faro, F. P. Marino and A. Pavone, *Smallest string attractors and minimal coverage certificates of Thue–Morse words*, arXiv:2610.10950, 2026.

[7] J. Florian and L. Dvořáková, *On periodicity of generalized pseudostandard words*, Electron. J. Combin. **23** (2016), P1.2.

[8] D. Kempa and N. Prezza, *At the roots of dictionary compression: string attractors*, in Proc. 50th ACM Symposium on Theory of Computing (STOC 2018), 827–840.

[9] K. Kutsukake, T. Matsumoto, Y. Nakashima, S. Inenaga, H. Bannai and M. Takeda, *On repetitiveness measures of Thue–Morse words*, in String Processing and Information Retrieval (SPIRE 2020), Lecture Notes in Comput. Sci. **12303**, Springer, 2020, 213–220.

[10] G. Rote, *Sequences with subword complexity* $2n$, J. Number Theory **46** (1994), 196–213.