Hairpin completion with bounded stem-loop

Szilárd Zsolt Fazekas[1⋆], Robert Mercaş[2⋆⋆], and Kayoko Shikishima-Tsuji[3]

[1] Department of Mathematics, Kyoto Sangyo University,
Motoyama, Kamigamo, Kita-Ku Kyoto 603-8555, Japan,
szilard.fazekas@gmail.com
[2] Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik,
PSF 4120,D-39016 Magdeburg, Germany,
robertmercas@gmail.com
[3] Tenri University, 1050 Somanouchi Tenri 632-8510, Japan,
tsuji@sta.tenri-u.ac.jp

Abstract. Pseudopalindromes are words that are fixed points for some antimorphic involution. In this paper we discuss a newer word operation, that of pseudopalindromic completion, in which symbols are added to either side of the word such that the new obtained words are pseudopalindromes. This notion represents a particular type of hairpin completion, where the length of the hairpin is at most one. We give precise descriptions of regular languages that are closed under this operation and show that the regularity of the closure under the operation is decidable.

1 Introduction and preliminaries

Palindromes are sequences which read the same starting from either end. Besides their importance in combinatorial studies of strings, mirrored complementary sequences occur frequently in DNA and are often found at functionally interesting locations such as replication origins or operator sites. Several operations on words were introduced which are either directly motivated by the biological phenomenon called stem-loop completion, or are very similar in nature to it. The mathematical hairpin concept introduced in [17] is a word in which some suffix is the mirrored complement of a middle factor of the word. The hairpin completion operation, which extends such a word into a pseudopalindrome with a non-matching part in the middle was thoroughly investigated in [1, 4, 8, 14–16]. Most basic algorithmic questions about hairpin completion have been answered ([1, 4]) with a noteworthy exception: given a word, can we decide whether the iterated application of the operation leads to a regular language? For the so called bounded hairpin completion [7], even the latter problem is settled [10].

Another operation related to our topic is iterated palindromic closure, which was first introduced in the study of the Sturmian words [2], and later generalized to pseudopalindromes [3]. This operator allows one to construct words with infinitely many pseudopalindromic prefixes, called pseudostandard words.

⋆ Work supported by *Japanese Society for the Promotion of Science* under no. *P10827*
⋆⋆ Work supported by *Alexander von Humboldt Foundation*

In [12] the authors propose the study of palindromic completion of a word, which considers all possible ways of extending the word into a palindrome. This operation, of course, produces an infinite set from any starting word.

The operation studied here, is (pseudo)palindromic completion. It differs from palindromic completion ([12]) in that we require the word to have a pseudopalindromic prefix or suffix in order to be completed. The (iterated) palindromic closure ([2]) considers the unique shortest word which completes the starting word into a (pseudo)palindrome, whereas we take all possible extensions. The subject of this work is closest in nature to the first operation, in fact it is a rather restricted form of it (we do not allow for non-matching middles), and the questions asked are also a subset of problems considered for hairpin completion; since in the biological phenomenon serving as inspiration, the hairpin's length in the case of stable bindings is limited (approx. 4-8 base-pairs) it is natural to consider completions with bounded middle part. Furthermore, as we will see, this restriction allows us to state decidability results, which remain open for hairpin completion as mentioned above.

After presenting the notions and results needed for our treatise, in Section 2 we state some simple one-step completion results. In Section 3 we gradually build the characterization of regular languages which stay regular under the iterated application of completion. Section 4 is a collection of algorithmic results on this operation: membership problem for the iterated completion of a word, decision methods telling whether the regularity of the iterated completion is preserved.

We assume the reader to be familiar with basic concepts as alphabet, word, language and regular expression (for more details see [5]) and end this Section with some definitions regarding combinatorics on words and formal languages.

The length of a finite word w is the number of not necessarily distinct symbols it consists of and is written $|w|$. The ith symbol we denote by $w[i]$ and by $w[i \dots j]$ we refer to the part of the word starting at ith and ending at jth position.

Words together with the operation of concatenation form a free monoid, which is usually denoted by Σ^* for an alphabet Σ. Repeated concatenation of a word w with itself is denoted by w^i for natural numbers i.

A word u is a *prefix* of w if there exists an $i \leq |w|$ such that $u = w[1 \dots i]$. We denote this by $u \leq_p w$. If $i < |w|$, then the prefix is called *proper*. Suffixes are the corresponding concept reading from the back of the word to the front. A word w has a positive integer k as a *period* if for all i, j such that $i \equiv j (\mathrm{mod} k)$ we have $w[i] = w[j]$, whenever both $w[i]$ and $w[j]$ are defined.

The central concept to this work is *palindromicity* in the general sense. First off, for a word $w \in \Sigma^*$ by w^R we denote its reversal, that is $w[|w| \dots 1]$. If $w = w^R$, the word is called a palindrome. Let $\mathcal{P}al(L) = \mathcal{P}al \cap L$ be the set of all palindromes of a language $L \subseteq \Sigma^*$, where $\mathcal{P}al$ is the language of all palindromes over Σ.

We can generalise this definition by allowing the "two" sides of the words to be "complementary" to each other's reverse. In formulae, let θ be an antimorphic involution, i.e. $\theta : \Sigma^* \to \Sigma^*$ is a function, such that $\theta(\theta(a)) = a$ for all $a \in \Sigma$, and $\theta(uv) = \theta(v)\theta(u)$ for all $u, v \in \Sigma^+$. Then, w is a (θ-)pseudopalindrome if

$w = \theta(w)$. To make notation cleaner, we write $\overline{u}$ for $\theta(u)$, when θ is understood. The language of all pseudopalindromes, when the alphabet and θ are fixed, is $\mathcal{P}sepal$. Note that this is a linear context-free language, just like $\mathcal{P}al$.

It is worth noting that the primitive root of every palindrome is a palindrome.

Trivially, palindromes $p = aqa^R$ with q palindrome have palindromic prefixes λ, a and aqa^R. Hence when we say a palindrome has a non-trivial palindromic prefix (suffix), we mean it has a proper prefix (suffix) of length at least two which is a palindrome. This notion is extended to pseudopalindromes as well.

Definition 1. Let θ be an antimorphic involution. For a factorization uv of some word w, where $v \notin \Sigma \cup \{\lambda\}$ (respectively, $u \notin \Sigma \cup \{\lambda\}$) is a ($\theta$-)pseudopalindrome, $uv\overline{u}$ (respectively, $\overline{v}uv$) is in the right(left) (θ)-*completion* (completion, when θ is clear from context) of w. We say that w' is in the completion of w if it is either in the right or left completion of w. We denote this relation by $w \ltimes w'$. The reflexive, transitive closure of $\ltimes$ is the *iterated completion*, in notation $\ltimes^*$, where for two words w and w' we say $w \ltimes^* w'$ if $w = w'$ or there exist words $v_1, \ldots, v_n$ with $v_1 = w$, $v_n = w'$ and $v_i \ltimes v_{i+1}$ for $1 \leq i \leq n-1$.

Definition 2. For a language L, we let $L = L^{\ltimes_0}$ and for $n > 0$ we let $L^{\ltimes_n}$ be the completion of $L^{\ltimes_{n-1}}$, i.e., $L^{\ltimes_n} = \{w \mid \exists u \in L^{\ltimes_{n-1}} : u \ltimes w\}$. Also, we say $L^{\ltimes_*}$ is the iterated pseudopalindromic completion of L, i.e., $L^{\ltimes_*} = \bigcup_{n \geq 0} L^{\ltimes_n}$.

For a singleton language $L = \{w\}$, let $w^{\ltimes_n}$ denote $L^{\ltimes_n}$, i.e., the nth completion of the word w. Moreover, in what follows we fix some literal antimorphic involution θ, hence do not explicitly mention it in the notation.

The following lemma and theorem will appear frequently in our proofs:

Lemma 1. *[Regular pumping lemma] For every regular language L there exists an integer k_L such that every word $w \in L$ longer than k_L, has a factorization $w = w_1w_2w_3$ such that $w_2 \neq \lambda$, $|w_1w_2| \leq k_L$ and $w_1w_2^*w_3 \subseteq L$.*

Theorem 1. *[Fine and Wilf] If two non-empty words p^i and q^j share a prefix of length $|p| + |q|$, then there exists a word r such that $p, q \in r^+$.*

2 Pseudopalindromic Regular Languages

A first observation we make is that a word's pseudopalindromic completion is a finite set, since it always has finitely many pseudopalindromic prefixes or suffixes.

In order to see that the class of regular languages is not closed under pseudopalindromic completion, consider the language $L = aa^+\overline{a}$. After one pseudopalindromic completion step we get $L^{\ltimes_1} = \{a^n\overline{a}^n \mid n \geq 2\}$, which is a non-regular context-free language. This actually settles (negatively) the question whether whenever the iterated completion of a language is non-regular it is also non-context-free.

Lemma 2. *The language $w^{\ltimes_*}$ is infinite iff the word w has both non-trivial pseudopalindromic prefixes and suffixes. Then $w^{\ltimes_i} \subsetneq w^{\ltimes_{i+1}}$ for all $i \geq 1$.*

Proof. The first part of the result is a case analysis result, while the second comes from the fact that the length increases with each iteration. □

Lemma 3. *For pseudopalindromes, the right and left completion steps are equal.*

Proof. For a word to have a right completion, it needs to have a decomposition $uvw\overline{v}$, where $w \in \Sigma \cup \{\lambda\}$ and $v \neq \lambda$. Then, $uvw\overline{v} \ltimes uvw\overline{v}\,\overline{u}$. Since the starting word is a pseudopalindrome, $uvw\overline{v} = \overline{uvw\overline{v}} = v\overline{w}\,\overline{v}\,\overline{u}$, and a left completion gives us $uv\overline{w}\,\overline{v}\,\overline{u}$. Since when $|w| = 1$ we have $w = \overline{w}$, the conclusion follows. □

Hence, whenever considering several completion steps for some pseudopalindromic language L, it is enough to consider either the right or the left completion. Similar to the palindromic languages characterization in [6]:

Theorem 2. *A regular language $L \subseteq \Sigma^*$ is pseudopalindromic, iff it is a union of finitely many languages of the form $L_p = \{p\}$ or $L_{r,s,q} = qr(sr)^*q^R$ where p, r and s are pseudopalindromes, and q is an arbitrary word.*

Proof. For any suitably long word $w \in L$, according to Lemma 1, we have a factorization $w = uvz$ with $0 < |uv| \leq n$ and $v \neq \lambda$, such that $uv^iz \in L$ for any $i \geq 0$ and some language-specific constant n. W.l.o.g., we assume $|u| \leq |z|$, i.e., for big enough i, the fact $uv^iz \in L$ means $z = x\overline{u}$ for some $x \in \Sigma^*$ with v^ix being a pseudopalindrome. This gives us $x = v_1\overline{v}^j$, where $\overline{v} = v_2v_1$ and $j \geq 0$. Again, if i was great enough, we instantly get $v = v_1v_2$ and thus $\overline{v} = \overline{v_2}\,\overline{v_1}$. From $v_2v_1 = \overline{v_2}\,\overline{v_1}$ we get that v_1 and v_2 are pseudopalindromes and, hence, w can be written as $uv_1(v_2v_1)^{j+1}\overline{u}$. According to Lemma 1 a similar decomposition exists for all words longer than n. Since all parts of the decomposition, u, v_1 and v_2 are shorter than n, finitely many such triplets exist. □

3 Iterated Pseudopalindromic Completion

W.l.o.g, we assume that all languages investigated in the case of iterated completion have only words longer than two. The case of pseudopalindromic completion on unary alphabets is not difficult to prove; even for arbitrary unary languages the iterated pseudopalindromic completion is regular:

Proposition 1. *The class of unary regular languages is closed under pseudopalindromic completion. Furthermore, the iterated pseudopalindromic completion of any unary language is regular.*

Proof. We know that all unary regular languages are expressed as a finite union of languages of the form $\{a^k(a^n)^* \mid k, n \text{ are some non-negative integers}\}$. Since for unary words to be pseudopalindromes we have $\overline{a} = a$, a one step pseudopalindromic completion of each word a^m gives the language $\{a^\ell \mid \ell < 2m\}$ and the first part of our result. For arbitrary unary language, after the iterated completion we get the language $\{a^ja^* \mid j \text{ is the minimum integer among all } m\text{'s }\}$. □

Next let us investigate what happens in the singleton languages case.

Proposition 2. *The class of iterated pseudopalindromic completion of singletons is incomparable with the class of regular languages.*

Proof. To show that regular languages are obtained take the word $a\overline{a}a$. It is not difficult to check that the language obtained is $\{a\overline{a}a\} \cup \{(a\overline{a})^n, (\overline{a}a)^n \mid n \geq 2\}$. Since all languages are regular, so is their union.

To see we not always get regular, nay, non-context-free, consider the word $u = a^3ba^3$ and θ just the reverse function. A one step completion gives us $\{a^3ba^3ba^3, a^3ba^4ba^3\}$. From $\{a^3b(a^4b)^na^3 \mid n \geq 1\}$ we get $\{a^3b(a^4b)^ma^3 \mid 1 < n+1 \leq m \leq 2n-1\}$ and $\{a^3b(a^4b)^na^3b(a^4b)^na^3 \mid n \geq 1\}$. The latter's completion includes $L = \{a^3b(a^4b)^na^3b(a^4b)^ma^3b(a^4b)^na^3 \mid 1 \leq n \leq m \leq 2n+1\}$. Actually, $L = u^{\ltimes *} \cap a^3(b(a^4b)^+a^3)^3$ and is easily shown to be non-context-free. □

Lemma 4. *If v is a non-trivial pseudopalindromic prefix or suffix of some other pseudopalindrome u, there always exist pseudopalindromes $x \neq \lambda$ and y, such that $v, w \in x(yx)^*$. Moreover, for two pseudopalindromes $v = p_1(q_1p_1)^{i_1}$ and $u = p_2(q_2p_2)^{i_2}$, where $i_1, i_2 > 2$, $|p_1|, |p_2| > 1$ and $2|v| > |u|$, and $v \leq_p u$, there exist pseudopalindromes p, q, such that $p_j(q_jp_j)^+ \subseteq p(qp)^+$, $j \in \{1, 2\}$.*

Proof. The first statement follows from [9, Proposition 5 (2) and Lemma 5 (2)]. Now let us see the second statement. By our assumptions, we have $p_2(q_2p_2)^{i_2} > \frac{|p_1(q_1p_1)^{i_1}|}{2}$. If $|p_2q_2| \geq |p_1q_1|$, then we can apply Theorem 1 and get that p_1q_1 and p_2q_2 have the same primitive root r. If $|p_2q_2| < |p_1q_1|$, then we have two cases. If $|p_2(q_2p_2)^{i_2}| \geq |(p_1q_1)^2|$, then Fine and Wilf applies directly giving that p_1q_1 and p_2q_2 have the same primitive root r. From $|(p_1q_1)^2| > |p_2(q_2p_2)^{i_2}| > |p_1q_1p_1| + \frac{|q_1|}{2}$, either $|p_2q_2| \leq |p_1| + \frac{|q_1|}{2}$, or $|(q_2p_2)^2| > |p_1q_1p_1|$, and hence, $|p_2(q_2p_2)^{i_2}| = |(q_2p_2)^2| + |p_2q_2p_2| > |p_1q_1p_1| + |p_2q_2p_2| > |p_1q_1| + |p_2q_2|$. In both cases, we apply Theorem 1 to the same end.

Then, there exist pseudopalindromes p, q such that $r = pq$ is primitive and $p_1 = p(qp)^{m_1}$, $q_1 = q(pq)^{n_1}$, for some $m_1, n_1 \geq 0$, so $p_1(q_1p_1)^+ \in p(qp)^+$. Since $v \leq_p u$ and both are pseudopalindromes, $v \leq_s u$ and u ends in $p((qp)^{m_1+n_1+1})^{i_1}$. But u also ends in $p_2(q_2p_2)^2$, so by the above argument, $p_2(q_2p_2)^+ \subseteq p(qp)^+$. □

Proposition 3. *For all words of the form $w = up(qp)^n\overline{u}$, where p and q are pseudopalindromes and u is a suffix of pq, there exist pseudopalindromes p', q' such that $w = p'(q'p')^m$ with $n \leq m \leq n+2$.*

Proof. Depending on the lengths of u and q we distinguish the following cases:
1. $|u| \leq \frac{|q|}{2}$ - in this case $q = \overline{u}xu$, for some (possibly empty) pseudopalindrome x. Thus, w can be written as $up(\overline{u}xup)^n\overline{u} = up\overline{u}(x.up\overline{u})^n$.
2. $\frac{|q|}{2} < |u| \leq |q|$ - in this case the prefix u and the suffix $\overline{u}$ overlap in q, i.e., $q = xyxyx$ for some pseudopalindromes x and y, where $u = xyx$. Thus, $w = xyxp(xyxyxp)^nxyx = x(yxpxy.x)^{n+1}$ so we can set $p' = x$ and $q' = yxpxy$.
3. $|q| < |u|$ - in this case $u = xq$ for some suffix x of p. Thus, $w = xqp(qp)^nq\overline{x} = xq(pq)^{n+1}\overline{x}$ with x a suffix of p, which brings us back to cases 1 or 2 (if the latter, the exponent increases by one yet again). □

Proposition 4. *Let $u_ip_i(q_ip_i)^{k_i}\overline{u_i}$ with $1 \leq i \leq n$ be a sequence of pseudopalindromes with $u_ip_i(q_ip_i)^{k_i}\overline{u_i} \ltimes u_{i+1}p_{i+1}(q_{i+1}p_{i+1})^{k_{i+1}}\overline{u_{i+1}}$, where p_i, q_i are pseudopalindromes and $u_1 = u_n$, $p_1 = p_n$ and $q_1 = q_n$. There exist pseudopalindromes p, q and positive integers t_i with $1 \leq i \leq n$, such that $u_ip_i(q_ip_i)^{k_i}\overline{u_i} = p(qp)^{t_i}$.*

Proof. Since $w \ltimes^* w'$ implies $w \leq_p w'$, we get $\overline{u_1} \leq_p (q_1p_1)^{k_n-k_1}$. Then, there exist words u and v with $uv = q_1p_1$ and some $t \geq 0$, such that we can write $\overline{u_1} = (q_1p_1)^tu$, hence $u_1 = \overline{u}(p_1q_1)^t$. But, $p_1q_1 = \overline{p_1}\,\overline{q_1} = \overline{q_1p_1} = \overline{uv} = \overline{v}\,\overline{u}$, therefore $u_1p_1(q_1p_1)^{k_1}\overline{u_1} = \overline{u}(\overline{v}\,\overline{u})^t(\overline{v}\,\overline{u})^{k_1}p_1(q_1p_1)^tu = \overline{u}(p_1q_1)^{2t+k_1}p_1u$ and also $u_1p_1(q_1p_1)^{k_n}\overline{u_1} = \overline{u}(p_1q_1)^{2t+k_n}p_1u$. Taking this further gives us that for every i with $1 \leq i \leq n$ there exists a $t_i > 0$ and a suffix x_i of p_iq_i such that $x_ip_i(q_ip_i)^{k_i}\overline{x_i} \ltimes^* x_ip_i(q_ip_i)^{k_i+t_i}\overline{x_i}$. Now we can apply Proposition 3, which gives us that these are all words of the form $p(qp)^+$ and Lemma 4 makes sure that one can find a unique pair p, q to express all of the words. □

Theorem 3. *The iterated pseudopalindromic completion of a word w is regular iff w has at most one pseudopalindromic prefix or one suffix, or for all words $w' \in w^{\ltimes_1}$ there exist unique pseudopalindromes p and q with $|p| \geq 2$, such that:*

- *$w' \in p(qp)^+$*
- *w' has no pseudopalindromic prefixes except for the words in $p(qp)^*$.*

Proof. Due to Lemma 3, for $w^{\ltimes_*}$ we need only consider the finite union of all one sided iterated pseudopalindromic completion of words $w' \in w^{\ltimes_1}$.
(IF) For this direction the result is easily obtained, since, at each completion step, from some word of form $p(qp)^n$ with $n \geq 1$ we get all words $p(qp)^n, \ldots, p(qp)^{2n}$, for $n \geq 1$. Thus, the final result is a finite union of regular languages.
(ONLY IF) Now assume that $w^{\ltimes_*}$, the iterated pseudopalindromic completion of some word w, is regular. The first case is trivial. For the second, following Theorem 2, $w^{\ltimes_*}$ is the union of some finite language $\{p \mid p \text{ pseudopalindrome}\}$ and some finite union of languages $\{qr(sr)^*\overline{q} \mid r, s \in \Sigma^* \text{ pseudopalindromes}\}$.

We neglect the case of the finite language $\{p \mid p \text{ pseudopalindrome}\}$, since this, according to Proposition 2 would contain just elements of $w^{\ltimes_1}$ that cannot be extended further on, and consider from $w^{\ltimes_*}$ only the finite union of languages of form $\{qr(sr)^*\overline{q} \mid q, r, s \in \Sigma^* \text{ and } r, s \text{ pseudopalindromes}\}$.

Following Dirichlet's principle for the finiteness of variables q with the help of the pigeon hole principle, we get that for some big enough integer k_1 and some i_1, we have that $qr(sr)^{k_1}\overline{q} \ltimes^* qr(sr)^{k_1+i_1}\overline{q}$. We can apply Proposition 4 and get some pseudopalindromes u, v, such that $qr(sr)^*\overline{q} \subset u(vu)^*$. Moreover, from the same Proposition we have that all the intermediate pseudopalindromic completion steps are in the language $qr(sr)^*\overline{q}$, hence, in $u(vu)^+$. Now we know there exist at most finitely many pairs of pseudopalindromes u, v, such that $w' \in u(vu)^+$. Suppose that exist n pairs of pseudopalindromes (u_i, v_i) such that $w' \in u_i(v_iu_i)^+$ with $u_i \neq u_j$ and $|u_i| \geq 2$, for $1 \leq i, j \leq n$, $i \neq j$. If $|u_1v_1| = |u_2v_2|$, then $|u_1| = |u_2|$ and since they are suffixes of the same word, $u_1 = u_2$ and, hence, $v_1 = v_2$, which is a contradiction. Therefore, w.l.o.g, we may assume $|u_1v_1| > |u_2v_2|$. In this case, $u_2v_2u_2$ is a pseudopalindromic

prefix of $u_1v_1u_1$, and Lemma 4 gives us $u_1v_1u_1, u_2v_2u_2 \in x_1(y_1x_1)^+$ for some pseudopalindromes x_1 and y_1. Repeating the argument for all the pairs (x_i, y_i) and (u_{i+2}, v_{i+2}), we can conclude the proof. □

What happens in the case of regular languages? We already know that the one step pseudopalindromic completion is not closed to regularity.

Proposition 5. *Iterated pseudopalindromic completion of a regular language is not necessarily context-free.*

Proof. Indeed, for this consider the language $L = \{aa^nba \mid n \geq 1\}$ and take θ to be just the reverse function. A closer look at the iterated pseudopalindromic completion of L, shows that the language obtained is $L^{\ltimes *} \subset L \cup L'$, where $L' \subset \{(\prod_{i\geq 1} a^{n_i}b)a^{n_1} \mid n_1 \leq n_i \leq 2n_1 - 2 \text{ for all } i\}$. Employing the context-free languages pumping lemma we get that $L^{\ltimes *} \cap a^+ba^+ba^+$ is non-context-free. The closure under intersection with regular languages gives us the result. □

Proposition 6. *Let $p, q, u \in \Sigma^*$ with p, q pseudopalindromes. If all pseudopalindromic prefixes of $upqp\overline{u}$ are trivial, then for any $i \geq 0$ so are those of $up(qp)^i\overline{u}$.*

Proof. Suppose p' is the shortest non-trivial pseudopalindromic prefix of any word $up(qp)^k\overline{u}$, $k \geq 0$. Since p' is not a prefix of $upqp\overline{u}$, the length of up is less than the length of p', hence, we have $p' = up(qp)^ix$, for some $i \leq k$ and word x which is a prefix of q, qp or $\overline{u}$. If x is a prefix of q, then $\overline{x}px$ is a suffix of p', hence, a non-trivial pseudopalindromic prefix of p', and, therefore, p' is not the shortest. If x is a prefix of qp, but not of q, then $x = qx'$ and $\overline{x'}(qp)^iqx'$ is a pseudopalindromic suffix, hence, prefix of p', contradicting our assumption. Similarly, if x is a prefix of $\overline{u}$, then $\overline{x}p(qp)^ix$ is a shorter non-trivial pseudopalindromic prefix than p' itself. □

By [2, Lemma 3] the following is straightforward:

Lemma 5. *A pseudopalindrome w has period $p < |w|$ iff it has a pseudopalindromic prefix of length $|w| - p$.*

Theorem 4. *For a regular language L, its iterated pseudopalindromic completion $L^{\ltimes *}$ is regular iff L can be written as the union of disjoint regular languages L', L'', and L''', where*

- $L' = L'^{\ltimes 1} = \{w \in L \mid w^{\ltimes *} \subseteq L\}$;
- $L'' = \{w \in L \mid w^{\ltimes 1} = w^{\ltimes *} \not\subseteq L\}$ *and all words of L'' are prefixes*[4] *(suffixes) of words in the finite union of languages of the form $up(qp)^*\overline{u}$, where $upqp\overline{u}$ has only trivial pseudopalindromic prefixes and p, q are pseudopalindromes;*

[4] Note, that the prefixes have to be at least $|up| + \lceil \frac{|q|}{2} \rceil + 1$ and $|p_i| + \lceil \frac{|q_i|}{2} \rceil + 1$ long, respectively, because the shorter ones do not extend beyond one step completion when pq (and p_iq_i, respectively) is primitive. This does not make a difference for the characterization, only for the decision process.

– $L''' = \{w \in L \mid \{w\} \cup w^{\ltimes_1} \neq w^{\ltimes_*} \not\subseteq L\}$ *and all words of* L''' *are prefixes*[4] *(suffixes) of words in* $\bigcup_{i=1}^{m} p_i(q_ip_i)^+$*, where* $m \geq 0$ *is an integer depending on* L *and* p_i*,* q_i *are pseudopalindromes such that* p_iq_i *have only one non-trivial pseudopalindromic prefix.*

Proof. (IF) This direction is immediate since L is a union of regular languages. (ONLY IF) Clearly, any language $L \subset \Sigma^*$ can be written as a union of three disjoint languages where one of them (L') contains the words which have neither non-trivial pseudopalindromic prefixes nor suffixes or their iterated pseudopalindromic completion is included in L, another (L'') has all the words which have either non-trivial prefixes or suffixes, and the third one (L''') contains the words which can be extended in both directions by pseudopalindromic completion. If $L^{\ltimes_*}$ and two of the other languages are regular, then the third one is, as well.

Here, we assume that $L^{\ltimes_*}$ is regular, hence $L^{\ltimes_*} \setminus L$ is regular, too. Moreover, $L^{\ltimes_*} \setminus L$ is a pseudopalindromic language, since all of its words are the result of pseudopalindromic completion. From Theorem 2 it follows that there exists a finite set of words x_i, r_i, s_i, where $i \in \{1, \ldots, n\}$ and r_i, s_i are pseudopalindromes, such that the words in $L^{\ltimes_*} \setminus L$ are elements of $x_ir_i(s_ir_i)^*\overline{x_i}$ with $1 \leq i \leq n$.

First we identify L'''. For each j, using once more the pigeon hole principle, it must that there exist big enough integers k_1 and k_2 with $x_jr_j(s_jr_j)^{k_1}\overline{x_j} \ltimes^* x_jr_j(s_jr_j)^{k_2}\overline{x_j}$, or we have $x_jr_j(s_jr_j)^{k_j}\overline{x_j} \ltimes^* x_ir_i(s_ir_i)^{k_1}\overline{x_i} \ltimes^* x_ir_i(s_ir_i)^{k_2}\overline{x_i}$ for some $i \neq j$ and k_j. In the first case we apply Proposition 4 and get that there exist pseudopalindromes $p \neq \lambda$ and q such that $x_jr_j(s_jr_j)^{k_i}\overline{x_j} \in p(qp)^+$, for $i \in \{1, 2\}$, and all intermediary words $x_jr_j(s_jr_j)^{k_j}\overline{x_j}$ are also in $p(qp)^+$. In the second case we apply Proposition 4 to the second relation. Then by Lemma 4 and Proposition 4 we get that all three words are in $p(qp)^+$, for suitable p, q. Also, pq has no non-trivial pseudopalindromic prefixes except for p, otherwise by Theorem 3 its iterated completion leads to non-regular languages. After finding these finitely many (say, m) pairs p_k, q_k, the language of all prefixes of $\bigcup_{k=1}^{m} p_k(q_kp_k)^+$ is a regular language, hence, its intersection with L is also regular.

We know that $L^{\ltimes_*} \setminus L'''^{\ltimes_*} = L'^{\ltimes_*} \cup L''^{\ltimes_*}$ is regular, therefore $L_{\mathbf{diff}} = (L'^{\ltimes_*} \cup L''^{\ltimes_*}) \setminus L \subset L''^{\ltimes_*}$ is a pseudopalindromic regular language. Again, from Theorem 2 we know that $L_{\mathbf{diff}}$ can be written as the finite union of languages of the form $up(qp)^*\overline{u}$. Clearly then, all words in L'' are prefixes of some word in $up(qp)^*\overline{u}$. Since by definition $L''^{\ltimes_1} = L''^{\ltimes_*}$, the words in $up(qp)^*\overline{u} \cap L_{\mathbf{diff}}$ have no non-trivial pseudopalindromic prefixes, hence, by Proposition 6 we have that $upqp\overline{u}$ does not either. Let L'' be the finite union of the languages $\mathbf{Pref}(up(qp)^+) \cap L$, where $\mathbf{Pref}(A)$ is the language of all prefixes of A. This way, L'' is regular and since from it we obtain $L_{\mathbf{diff}}$ by pseudopalindromic completion, it meets the requirements. All that is left is to assign $L' = (L \setminus L''') \setminus L''$, which is regular and all its words either have only trivial pseudopalindromic prefixes or suffixes, or their pseudopalindromic completion is already in L. □

As a consequence of Theorems 2 and 4, the following result is obtained:

Corollary 1. *If for some regular language* L *we have that* $L^{\ltimes_*}$ *is regular, then for any integer* $n \geq 1$ *we have that* $L^{\ltimes_n}$ *is regular.*

4 Decidability questions

We conclude this paper with some complexity results, which build on the previously obtained characterizations.

While in the classical hairpin completion case the extension of a word is both to the right and the left of the word, here, due to the pseudopalindromicity property the two extensions are identical making the problem simpler. The membership problem for the one step pseudopalindromic completion of a word is trivial as one has to check for the shorter word if it is a prefix while its θ image is a suffix of the longer one, or vice-versa, and these two occurrences overlap. Obviously, the time needed for this is linear. A more interesting problem is that of membership for the iterated pseudopalindromic completion; in this setting the problem is decidable, and solvable in quadratic time.

Lemma 6. *If u, v are pseudopalindromes with u prefix of v and $|u| > \lceil |v|/2 \rceil$, then $u \ltimes v$.*

Proof. The result is an immediate consequence of Lemma 4. □

Proposition 7. *For two pseudopalindromes u, v, we have $u \ltimes^* v$ iff u is a prefix of v and for every prefix w of v with length greater than u, w has as prefix a non-trivial pseudopalindrome of length greater than $\lceil |w|/2 \rceil$.*

Proof. In other words for pseudopalindromes u and v, we say that v can be obtained from u iff u is a prefix of v and for any pseudopalindromic prefixes of v they all have as prefix some pseudopalindrome of length greater than half theirs. (ONLY IF) Since starting with the pseudopalindrome u we have after some completions steps u as prefix and suffix. Moreover, after each step the pseudopalindrome we do the completion on is both prefix and suffix of the new word. (IF) In order for v to be part of the iterated pseudopalindromic completion of a word it must be the case that second of the properties holds. Since v starts with u and the second property holds, with the help of Lemma 6 we get that v is in the language given by the iterated pseudopalindromic completion of u. □

Theorem 5. *One can decide in linear time if for two words u and v, where v is a pseudopalindrome of length n greater than $|u|$, we have $u \ltimes^* v$.*

Proof. By Proposition 7, it suffices to check two things: if the pseudopalindromic completion of u contains some prefix of v, which is done in linear time, and then whether all pseudopalindromic prefixes of v have as prefix a pseudopalindrome of length more than half of theirs. Identifying all pseudopalindromic prefixes of v of length greater than that of w is easily done in $\mathcal{O}(n)$ using a slight modification of the algorithm from [13]. Next, looking at the lengths of all elements in this set, we check that the difference between no two consecutive ones is double the smallest of them; again linear time is enough to do this and we conclude. □

As previously mentioned, one can identify in time $\mathcal{O}(n)$ all pseudopalindromic prefixes of some word v of length n. From those, one can efficiently compute

the pseudopalindromic completion distance between two given words u and v. We start with the longest element of $u^{\ltimes_1}$, and in each step choose v's longest pseudopalindromic prefix which is shorter than twice the length of the current one. The greedy technique ensures optimality with the help of Proposition 7, while Lemma 6 proves the correctness of each step, therefore:

Theorem 6. *Given a word u and a pseudopalindrome v of length $n > |u|$, one can compute in linear time the minimum number of pseudopalindromic completion iterations needed in order to get from u to v, when possible.*

Let us now look at the regular closure property related to this operation.

Theorem 7. *For some word w of length n, it is decidable in $\mathcal{O}(n^2)$ whether its iterated pseudopalindromic completion $w^{\ltimes_*}$ is regular.*

Proof. For each of the finitely many w' (the number is, of course, linear in $|w|$), with $w \ltimes w'$, consider the following procedure. In linear time one can find all periods of w'. Let $n = p_i q_i + r_i$, where p_i are all periods of w', with $r_i < p_i$. Taking r' to be the smallest of r_i, according to Lemma 5, it is left to check if there exists a unique pseudopalindrome v, such that for all $r_j > r'$, we have $w'[1 \dots r_j] \in w'[1 \dots r'](vw'[1 \dots r'])^*$. Since deciding whether a word is pseudopalindrome is done in $\mathcal{O}(n)$, the result is concluded. □

In what follows, a deterministic finite automaton (DFA) is defined by a quintuple $\langle Q, \Sigma, q_0, \sigma, F \rangle$, where Q is the set of states, q_0 the initial state, Σ the input alphabet, σ the transition function and F the set of final states. For details on finite automata and closure properties, see [5]. For the next results we suppose - w.l.o.g, as the algorithm given here is intractable even for DFAs - that L is presented to us as a DFA as above, with $|Q| = n$.

Theorem 8. *Given a regular language L, it is decidable whether $L = L^{\ltimes_*}$.*

Proof. If $L \neq L^{\ltimes_*}$, then there exist some non-empty word u and pseudopalindrome p of length at least two, such that $up \in L$, but $up\overline{u} \notin L$. Let us suppose that u is the shortest such word. We show that, should u exist we can find it after finitely many steps. Let L_{ul} denote the language $\{w \mid \sigma(q_0, w) = \sigma(q_0, u)\}$. Define the set of final states reachable by a pseudopalindrome after first reading u, as $F_u = \{q \in F \mid \exists w$ pseudopalindrome with $\sigma(q_0, uw) = q\}$, and the language accepted starting from such a state $L_{ur} = \{w \mid \exists p \in F_u, q \in F : \sigma(p, w) = q\}$.

Then, u is the shortest word in $L_{ul} \setminus L_{ur}^{\theta} = L_{ul} \cap (\Sigma^* \setminus L_{ur}^{\theta})$, where $L^{\theta} = \{\theta(w) | w \in L\}$ is the θ image of L. Note that the languages L_{ul} and L_{ur} depend only on the state to which our supposed u takes the automaton, therefore all possibilities can be accounted for by considering all states of the automaton. The number of states of the automaton $L_{ul} \setminus L_{ur}^{\theta}$ is unfortunately quite high, hence so is the length up to which we have to check all words whether they are u:
– the automaton accepting L_{ul} has at most n states;
– for L_{ur} we get a NFA of at most n states, so at most 2^n states for the DFA;
– reversal and determinisation of the L_{ur} automaton takes it up to 2^{2^n} states;

– $L_{ul} \cap (\Sigma \setminus L^{\theta}_{ur})$ results in an automaton with at most $n2^{2^n}$ states and the shortest word accepted by it being at most as long as the number of states.

Thus, for all words u with $|u| \leq n2^{2^n}$, we have to check $((u \cdot \mathcal{P}sepal) \cap L)\overline{u} \setminus L = \emptyset$. If for at least one the set is not empty, we answer NO, otherwise YES. □

Theorem 9. *Given a regular language L, it is decidable whether $L^{\ltimes *}$ is regular. If the answer is YES, we can construct an automaton accepting $L^{\ltimes *}$.*

Proof. The outline of the decision procedure, based on the description of $L^{\ltimes *}$ given in Theorem 4, is as follows: first we identify the words p_i, q_i forming L''', if any exist. Then we construct a DFA which accepts $L' \cup L'' = L \setminus L'''$. In the resulting automaton we check for the words u_k, p_k and q_k - if any - which form L'' and construct the automaton for $L' = (L \setminus L''') \setminus L''$. Last, we check whether $L' = L'^{\ltimes *}$, that is $L' = L'^{\ltimes 1}$, with the help of Theorem 8. If yes, then $L^{\ltimes *}$ is regular, otherwise it is not.

The automata for the intermediary steps are computable using well-known algorithms (see [5]). What we have to show, is that the words u_k, p_k, q_k can be found, given an automaton. First, we check every cycle of length at most N_L in the automaton, where N_L is a constant computable from the representation of L (for the argument on N_L see the last part of the proof). This can be easily done by a depth-first search. If the label of the cycle can be written as pq for some pseudopalindromes $p \neq \lambda$ and q, then we check all paths w of length at most N_L, which lead to the cycle from the initial state and all paths v of length at most N_L, going from the cycle to a final state. If there exist pseudopalindromes $x \neq \lambda$ and y such that xy is a cyclic shift of pq and $wpqv$ is a prefix or suffix of a word in $x(yx)^+$, then we identified a pair p_i, q_i for L'''. If there exist pseudopalindromes $x \neq \lambda$ and y, and some word u, such that xy is a cyclic shift of pq and $wpqv = ux(yx)^i$ for some $i \geq 1$, then we identified a triple u_k, p_k, q_k for L''. After finding all pairs p, q for L''', we construct for each of them the automaton accepting $L \setminus L_{pq}$, where L_{pq} is the set of prefixes of $p(qp)^+$ longer than $|p| + \lceil \frac{|q|}{2} \rceil + 1$. The language we get finally is $L' \cup L''$. Afterwards we subtract, for each triple u, p, q forming L'', the language of prefixes of $up(qp)^+u^R$ which are longer that $|up| + \lceil \frac{|q|}{2} \rceil + 1$. The resulting language is our candidate for L'. As mentioned above, if $L' = L'^{\ltimes 1}$, output YES, otherwise NO.

We end the proof by showing that N_L is computable from the presentation of L, as it is the number of states of a newly constructed automaton.

If $L^{\ltimes *}$ is regular, then so is $L^{\ltimes 1}$, by Corollary 1. If $L^{\ltimes 1}$ is regular, then Theorem 2 applies to $L^{\ltimes *} \setminus L$ and gives us that it can be written as the finite union of languages of the form $xr(sr)^*\overline{x}$, with r, s pseudopalindromes.

For every state $p \in Q$, let us define the languages $\text{LEFT}_p = \{u \mid \sigma(q_0, u) = p\}$ and $\text{RIGHT}_p = \{u \mid \exists q \in F : \sigma(p, u) = q\}$. For every pair of states $p \in Q$, $q \in F$, let L_{pq} denote the language $\text{LEFT}_p \setminus \theta(\text{RIGHT}_q)$, when $\sigma(p, w) = q$ for some pseudopalindrome $w \notin \Sigma \cup \{\lambda\}$, and $L_{pq} = \emptyset$, otherwise. Now, the language

$$L_c = \bigcup_{p,q \in Q} L_{pq}$$

is regular, as it is the finite union of regular languages. Also, every word in L_c is the prefix of a word in one of the finitely many languages $xr(sr)^*\overline{x}$ mentioned above. If L_c is infinite, then by Lemma 1 and Theorem 1 we get that the label of every cycle in the automaton accepting L_c is of the form w^k, where w is a cyclic shift of pq and $k \geq 1$. Hence, the same holds for cycles of length at most m, where m is the number of states of the automaton accepting L_c. On the other hand, suppose there is a pair r_1, s_1, such that all cycles which are cyclic shifts of $(r_1 s_1)^k$ for some $k \geq 1$ are longer than m. Then, again by pumping lemma and pigeon hole principle, we get that $r_1 s_1$ is the cyclic shift of some other pair r_2, s_2, where $|r_2 s_2| \leq m$. Hence, we conclude that by checking all cycles of length at most m of the automaton accepting L_c we discover the pairs r, s from the characterization in Theorem 2. The automaton accepting L_c can be constructed, given L, and m is computed by counting the states, hence take N_L to be m. □

References

1. D. Cheptea, C. Martín-Vide, and V. Mitrana. A new operation on words suggested by DNA biochemistry: Hairpin completion. *Trans. Comput.*, pages 216–228, 2006.
2. A. de Luca. Sturmian words: Structure, combinatorics, and their arithmetics. *Theor. Comput. Sci.*, 183(1):45–82, 1997.
3. A. de Luca and A. De Luca. Pseudopalindrome closure operators in free monoids. *Theor. Comput. Sci.*, 362(1-3):282–300, 2006.
4. V. Diekert, S. Kopecki, and V. Mitrana. On the hairpin completion of regular languages. In *ICTAC*, vol. 5684 of *Lect. Notes Comput. Sci.*, 170–184, 2009.
5. M. A. Harrison. *Introduction to Formal Language Theory.* Addison-Wesley, Reading, Massachusetts, 1978.
6. S. Horváth, J. Karhumäki, and J. Kleijn. Results concerning palindromicity. *J. Inf. Process. Cybern.*, 23:441–451, 1987.
7. Masami Ito, Peter Leupold, Florin Manea, and Victor Mitrana. Bounded hairpin completion. *Inf. Comput.*, 209(3):471–485, 2011.
8. L. Kari, S. Kopecki, and S. Seki. Iterated hairpin completions of non-crossing words. In *SOFSEM*, vol. 7147 of *Lect. Notes Comput. Sci.*, 337–348, 2012.
9. L. Kari and K. Mahalingam. Watson–Crick palindromes in DNA computing. *Nat. Comput.*, 9(2):297–316, 2010.
10. S. Kopecki. On iterated hairpin completion. *Theor. Comput. Sci.*, 412(29):3629–3638, 2011.
11. M. Lothaire. *Combinatorics on Words.* Cambridge University Press, 1997.
12. K. Mahalingam and K. G. Subramanian. Palindromic completion of a word. In *BIC-TA*, 1459–1465, IEEE, 2010.
13. G. Manacher. A new linear-time "on-line" algorithm for finding the smallest initial palindrome of a string. *Journal of the ACM*, 22(3):346–351, 1975.
14. F. Manea, C. Martín-Vide, and V. Mitrana. On some algorithmic problems regarding the hairpin completion. *Discrete Appl. Math.*, 157(9):2143–2152, 2009.
15. F. Manea and V. Mitrana. Hairpin completion versus hairpin reduction. In *CiE*, vol. 4497 of *Lect. Notes Comput. Sci.*, 532–541, 2007.
16. F. Manea, V. Mitrana, and T. Yokomori. Some remarks on the hairpin completion. *Int. J. Found. Comput. Sci.*, 21(5):859–872, 2010.
17. G. Paun, G. Rozenberg, and T. Yokomori. Hairpin languages. *Int. J. Found. Comput. Sci.*, pages 837–847, 2001.