On the Iterated Hairpin Completion

Steffen Kopecki

Institute for Formal Methods in Computer Science
Division of Theoretical Computer Science

University of Stuttgart
Germany

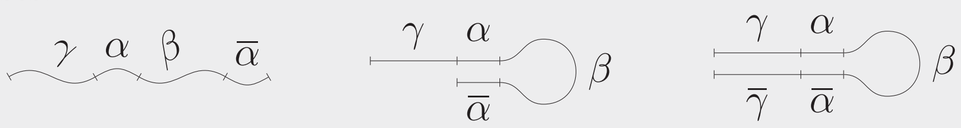
Abstract

The hairpin completion is a natural operation on formal languages which has been inspired by biochemistry and DNA-computing. In our work we solve two open problems from literature.

1. A natural variant of the hairpin completion is the bounded hairpin completion. It was unknown whether the iterated bounded hairpin completion of a regular language remains regular. We prove that this is indeed the case. Actually we derive a more general result. We prove that each language class which is closed under some basic operations is also closed under iterated bounded hairpin completion.
2. It is known that the iterated hairpin completion of a singleton (or finite language) is always context-sensitive, but it was unknown whether the iterated hairpin completion of a singleton is always regular, see [2]. We show that it might even be not context-free.

Biochemical Background

A *DNA strand* is a polymer composed of nucleotides which differ from each other by their bases A (adenine), C (cytosine), G (guanine), and T (thymine). A strand can be seen as a finite sequence of bases. By *Watson-Crick base pairing* two base sequences can bind to each other if they are pairwise complementary where A is complementary to T and C to G. The hairpin completion is best explained by the following figure. By the base sequence $\overline{\alpha}$ we mean to read α from right to left and complement its bases.



(a) strand (b) hairpin (c) hairpin completion

During chemical processes a strand which contains a sequence α and ends on the complementary sequence $\overline{\alpha}$ (a) can form an intramolecular base-pairing which is known as *hairpin* (b). By complementing the unbound sequence γ, the *hairpin completion* (c) arises.

On an abstract level a strand can be seen as a word and a (possibly infinite) set of strands corresponds to a formal language. The hairpin completion of a formal language was defined first in [1].

The Hairpin Completion of a Formal Language

Let Σ be an alphabet with an involution $\overline{\ } : \Sigma \to \Sigma$ (i.e., $\forall a \in \Sigma\colon \overline{\overline{a}} = a$). For a word $w = a_1 \cdots a_n$ let $\overline{w} = \overline{a_n} \cdots \overline{a_1}$.

If a word $w \in \Sigma^*$ has a factorization $w = \gamma\alpha\beta\overline{\alpha}$, then $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$ is called a *(right) hairpin completion* of w. Since a hairpin in biochemistry is stable only if α is long enough, we fix a constant $k \geq 1$ and ask $|\alpha| = k$. Symmetrically, if $w = \alpha\beta\overline{\alpha}\overline{\gamma}$ with $|\alpha| = k$, then $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$ is called a *(left) hairpin completion* of w.

The *hairpin completion* of a formal language $L \subseteq \Sigma^*$ is the union of all hairpin completions of all words in L:

$$\mathcal{H}_k(L) = \{\gamma\alpha\beta\overline{\alpha}\overline{\gamma} \mid |\alpha| = k \wedge (\gamma\alpha\beta\overline{\alpha} \in L \vee \alpha\beta\overline{\alpha}\overline{\gamma} \in L)\}.$$

For the *bounded hairpin completion* we assume that the length of the factor γ meets a length constraint:

$$\mathcal{H}_{k,\ell}(L) = \{\gamma\alpha\beta\overline{\alpha}\overline{\gamma} \mid |\alpha| = k \wedge |\gamma| \leq \ell \wedge (\gamma\alpha\beta\overline{\alpha} \in L \vee \alpha\beta\overline{\alpha}\overline{\gamma} \in L)\}.$$

Our focus is on the iterated versions of these two operations. Let

$$\mathcal{H}_k^0(L) = L, \qquad \mathcal{H}_k^i(L) = \mathcal{H}_k(\mathcal{H}_k^{i-1}(L)) \quad \text{for } i \geq 0.$$

The *iterated hairpin completion* is defined as

$$\mathcal{H}_k^*(L) = \bigcup_{i \geq 0} \mathcal{H}_k^i(L).$$

The *iterated bounded hairpin completion* $\mathcal{H}_{k,\ell}^*(L)$ is defined analogously.

The Iterated Bounded Hairpin Completion

In [3] some closure properties of the iterated bounded hairpin completion have been proved. E.g., the classes of context-free, context-sensitive, and recursively enumerable languages are closed under iterated bounded hairpin completion. But one interesting question has has been left unanswered: Is the class of regular languages closed under iterated bounded hairpin completion? We were able to answer this question positively, and we even provide a more general result.

Theorem. *Every language class which is closed under union, intersection with regular sets, and concatenation with regular sets is also closed under iterated bounded hairpin completion.*

The proof can be found in the technical report [4].

Our theorem yields a new representation for $\mathcal{H}_{k,\ell}^*(L)$ which uses L and the operations union, intersection with regular sets, and concatenation with regular sets.

In case L is a regular language this representation leads to a regular expression. Written down the length of the representation is more than exponential with respect to the bound ℓ, but it is linear in the size of a regular expression accepting L.

The Iterated Hairpin Completion of Singletons

The class of iterated hairpin completions of singletons

$$HCS_k = \{\mathcal{H}_k^*(\{w\}) \mid w \in \Sigma^*\}$$

has been investigated in [2]. HCS_k is included in the class of context-sensitive languages because context-sensitive languages are closed under iterated hairpin completion, see [1]. However, the problem if HCS_k contains non-regular or non-context-free languages has been unsolved. We solve this problem by stating a singleton whose iterated hairpin completion is not context-free.

Theorem. *The iterated hairpin completion of a singleton is context-sensitive and it is not in general context-free.*

Let $\Sigma = \{a, \bar{a}, b, \bar{b}, c, \bar{c}\}$ and $\alpha = a^k$. Consider

$$w = \alpha b \alpha \bar{\alpha} \alpha c \bar{\alpha}.$$

We claim that $\mathcal{H}_k^*(\{w\})$ is not context-free.

Since context-free languages are closed under intersection with regular sets, it suffices to show that the intersection $\mathcal{H}_k^*(\{w\}) \cap \mathcal{R}$ is not context-free for some regular set $\mathcal{R}$. Let $u = \bar{b}\bar{\alpha}$, let $v = \alpha\bar{\alpha}\bar{b}\bar{\alpha}$, and define

$$\mathcal{R} = wu^+v\bar{u}^+\overline{w}\bar{u}^+\overline{w}.$$

In our paper we prove

$$\mathcal{H}_k^*(\{w\}) \cap \mathcal{R} = \{wu^nv\bar{u}^n\overline{w}\bar{u}^n\overline{w} \mid n \geq 1\}$$

which is a non-context-free language.

For details see the technical report [4].

Some References

[1] D. Cheptea, C. Martin-Vide, and V. Mitrana.
A new operation on words suggested by DNA biochemistry: Hairpin completion.
Trans. Comp., pages 216–228, 2006.

[2] F. Manea, V. Mitrana, and T. Yokomori.
Some remarks on the hairpin completion.
In *12th International Conference AFL 2008 Proceedings*, pages 302–312, 2008.

[3] M. Ito, P. Leupold, and V. Mitrana.
Bounded hairpin completion.
In *LATA 2009*, pages 434–445, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] S. Kopecki.
On the Iterated Hairpin Completion.
Technical Report Computer Science 2010/02, University of Stuttgart, May 2010.
http://www2.informatik.uni-stuttgart.de/cgi-bin/NCSTRL/NCSTRL_view.pl?id=TR-2010-02&engl=1

http://www.fmi.uni-stuttgart.de/ti/ kopecki@fmi.uni-stuttgart.de