On Hairpin-Free Words and Languages

L. Kari[1], S. Konstantinidis[2] P. Sosík[3,4], G. Thierrin[5]

[1]Department of Computer Science, The University of Western Ontario, London, ON, Canada, N6A 5B7, lila@csd.uwo.ca

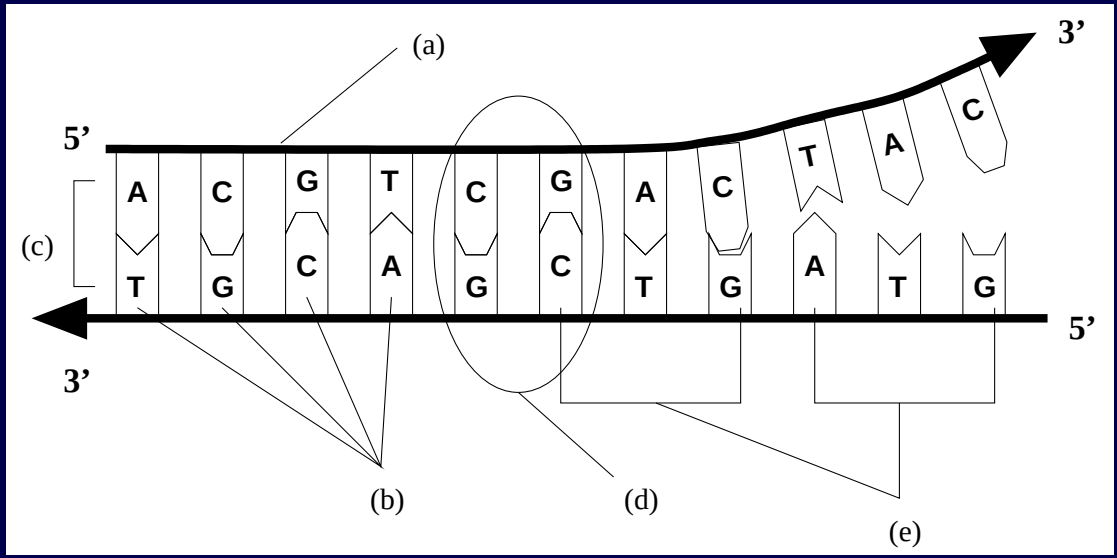
[2]Dept. of Mathematics and Computing Science, Saint Mary's University, Halifax, Nova Scotia, B3H 3C3 Canada, s.konstantinidis@stmarys.ca

[3]Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Madrid, Spain,

[4]Institute of Computer Science, Silesian University, Opava, Czech Republic, petr.sosik@fpf.slu.cz

[5]Department of Mathematics, The University of Western Ontario, London, ON, Canada, N6A 5B7, thierrin@uwo.ca

A DNA molecule



A segment of a double-stranded DNA molecule. (a) Sugar-phosphate backbone, with a 5' $\rightarrow$ 3' orientation; (b) nucleotides; (c) bindings; (d) Watson-Crick complementarity principle; (e) triples of nucleotides: *codons.*

Formal DNA description

- Let X be an alphabet, an involution θ of X^* is a map such that $\theta(\theta(w)) = w$ for all $w \in X^*$.

Formal DNA description

- Let X be an alphabet, an involution θ of X^* is a map such that $\theta(\theta(w)) = w$ for all $w \in X^*$.
- The DNA alphabet $\Delta = \{A, C, T, G\}$.

Formal DNA description

- Let X be an alphabet, an involution θ of X^* is a map such that $\theta(\theta(w)) = w$ for all $w \in X^*$.
- The DNA alphabet $\Delta = \{A, C, T, G\}$.
- The simplest involution: the identity function ϵ.

Formal DNA description

- Let X be an alphabet, an involution θ of X^* is a map such that $\theta(\theta(w)) = w$ for all $w \in X^*$.
- The DNA alphabet $\Delta = \{A, C, T, G\}$.
- The simplest involution: the identity function ϵ.
- The mirror involution is denoted by μ.

Formal DNA description

- Let X be an alphabet, an involution θ of X^* is a map such that $\theta(\theta(w)) = w$ for all $w \in X^*$.
- The DNA alphabet $\Delta = \{A, C, T, G\}$.
- The simplest involution: the identity function ϵ.
- The mirror involution is denoted by μ.
- The DNA complementarity involution γ is a morphism given by $\gamma(A) = T$, $\gamma(T) = A$, $\gamma(C) = G$, $\gamma(G) = C$.
- Example: $\epsilon(ACGCTG) = ACGCTG = \mu(GTCGCA) = \gamma(TGCGAC)$.

Formal DNA description

- Let X be an alphabet, an involution θ of X^* is a map such that $\theta(\theta(w)) = w$ for all $w \in X^*$.
- The DNA alphabet $\Delta = \{A, C, T, G\}$.
- The simplest involution: the identity function ϵ.
- The mirror involution is denoted by μ.
- The DNA complementarity involution γ is a morphism given by $\gamma(A) = T$, $\gamma(T) = A$, $\gamma(C) = G$, $\gamma(G) = C$.
- Example: $\epsilon(ACGCTG) = ACGCTG = \mu(GTCGCA) = \gamma(TGCGAC)$.
- The Watson-Crick involution $\tau = \mu\gamma$ corresponds to the DNA bond formation.

Hairpin secondary structures

- In some DNA computing models hairpins are undesirable — e.g. the Adleman's experiment.

Hairpin secondary structures

- In some DNA computing models hairpins are undesirable — e.g. the Adleman's experiment.
- However, hairpins play an important role in insertion/deletion operations.

Hairpin secondary structures

- In some DNA computing models hairpins are undesirable — e.g. the Adleman's experiment.
- However, hairpins play an important role in insertion/deletion operations.
- Hairpins are the main tool used in the Whiplash PCR computing techniques.

Hairpin secondary structures

- In some DNA computing models hairpins are undesirable — e.g. the Adleman's experiment.
- However, hairpins play an important role in insertion/deletion operations.
- Hairpins are the main tool used in the Whiplash PCR computing techniques.
- Hairpins serve as a binary information medium for DNA RAM.

Hairpin secondary structures

- In some DNA computing models hairpins are undesirable — e.g. the Adleman's experiment.
- However, hairpins play an important role in insertion/deletion operations.
- Hairpins are the main tool used in the Whiplash PCR computing techniques.
- Hairpins serve as a binary information medium for DNA RAM.
- Hairpins are a basic component of "smart drugs."

Formalization of hairpins

Definition 1 *Let θ be a morphic or antimorphic involution of X^*, let $k \geq 1$, then a word $u \in X^*$ is said to be θ-k-hairpin-free or simply hp(θ,k)-free if $u = xvy\theta(v)z$ for some $x, v, y, z \in X^*$ implies $|v| < k$.*

Observe: if $v = 1$, the empty word, $\theta = \epsilon$, the identity involution, and $k = 1$, then a word u is hp(ϵ,1)-free iff it is square-free.

Formalization of hairpins

Definition 1 *Let θ be a morphic or antimorphic involution of X^*, let $k \geq 1$, then a word $u \in X^*$ is said to be θ-k-hairpin-free or simply hp(θ,k)-free if $u = xvy\theta(v)z$ for some $x, v, y, z \in X^*$ implies $|v| < k$.*

Observe: if $v = 1$, the empty word, $\theta = \epsilon$, the identity involution, and $k = 1$, then a word u is hp(ϵ,1)-free iff it is square-free.

Definition 2 *Denote by $hpf(\theta, k)$ the set of all hp(θ,k)-free words in X^*. The complement of $hpf(\theta, k)$ is $hp(\theta, k) = X^* - hpf(\theta, k)$.*

Formalization of hairpins

Definition 1 *Let θ be a morphic or antimorphic involution of X^*, let $k \geq 1$, then a word $u \in X^*$ is said to be θ-k-hairpin-free or simply hp(θ,k)-free if $u = xvy\theta(v)z$ for some $x, v, y, z \in X^*$ implies $|v| < k$.*

Observe: if $v = 1$, the empty word, $\theta = \epsilon$, the identity involution, and $k = 1$, then a word u is hp(ϵ,1)-free iff it is square-free.

Definition 2 *Denote by $hpf(\theta, k)$ the set of all hp(θ,k)-free words in X^*. The complement of $hpf(\theta, k)$ is $hp(\theta, k) = X^* - hpf(\theta, k)$.*

Definition 3 *A language L is called θ-k-hairpin-free or simply hp(θ, k)-free if $L \subseteq hpf(\theta, k)$.*

Observe: L is hp(θ, k)-free *iff* $X^*vX^*\theta(v)X^* \cap L = \emptyset$ for all $|v| \geq k$.

Examples

1. Let $X = \{a, b\}$ with $\theta(a) = b, \theta(b) = a$. Then:

$$hpf(\theta, 1) = a^* \cup b^*$$

Observe: product of $\mathrm{hp}(\theta, 1)$-free words is not a $\mathrm{hp}(\theta, 1)$-free word.

Examples

1. Let $X = \{a, b\}$ with $\theta(a) = b, \theta(b) = a$. Then:

$$hpf(\theta, 1) = a^* \cup b^*$$

Observe: product of $\mathrm{hp}(\theta, 1)$-free words is not a $\mathrm{hp}(\theta, 1)$-free word.

2. If $\theta = \gamma$ is the DNA complementary involution over Δ^*, then:

$$hpf(\theta, 1) = \{A, C\}^* \cup \{A, G\}^* \cup \{T, C\}^* \cup \{T, G\}^*$$

Examples

1. Let $X = \{a, b\}$ with $\theta(a) = b, \theta(b) = a$. Then:

$$hpf(\theta, 1) = a^* \cup b^*$$

Observe: product of $\mathrm{hp}(\theta, 1)$-free words is not a $\mathrm{hp}(\theta, 1)$-free word.

2. If $\theta = \gamma$ is the DNA complementary involution over Δ^*, then:

$$hpf(\theta, 1) = \{A, C\}^* \cup \{A, G\}^* \cup \{T, C\}^* \cup \{T, G\}^*$$

3. Let $\theta = \mu$ be the mirror involution and let $X = \{a, b\}$, then:

$$hpf(\theta, 1) = \{1, a, b, ab, ba\}$$

Observe: if $\theta = \mu$, then $hpf(\theta, 1)$ is always finite.

Properties of hp$(\theta, 1)$-free languages

Definition 4 *(i) A language L is* $\leq_e$-convex *if $u \leq_e w \leq_e v$, $u, v \in L$ implies $w \in L$.* *(ii) L is* right $\leq_e$-convex *if $u \leq_e w$, $u \in L$ implies $w \in L$.*

Properties of hp$(\theta, 1)$-free languages

Definition 4 *(i) A language L is $\leq_e$-convex if $u \leq_e w \leq_e v$, $u, v \in L$ implies $w \in L$.* *(ii) L is right $\leq_e$-convex if $u \leq_e w$, $u \in L$ implies $w \in L$.*

Proposition 5 *The language $hp(\theta, 1)$ is right $\leq_e$ -convex.*

Properties of hp$(\theta, 1)$-free languages

Definition 4 *(i) A language L is $\leq_e$-convex if $u \leq_e w \leq_e v$, $u, v \in L$ implies $w \in L$.* *(ii) L is right $\leq_e$-convex if $u \leq_e w$, $u \in L$ implies $w \in L$.*

Proposition 5 *The language $hp(\theta, 1)$ is right $\leq_e$ -convex.*

Definition 6 *Let H be a nonempty subset of X^+.*

(i) We denote $S(H) = \{w \in X^* | u \leq_e w, u \in H\}$.

(ii) H is called a hypercode over X^* iff *$x \leq_e y$ and $x, y \in H$ imply $x = y$.*

Properties of hp$(\theta, 1)$-free languages

Definition 4 *(i) A language L is $\leq_e$-convex if $u \leq_e w \leq_e v$, $u, v \in L$ implies $w \in L$.* *(ii) L is right $\leq_e$-convex if $u \leq_e w$, $u \in L$ implies $w \in L$.*

Proposition 5 *The language $hp(\theta, 1)$ is right $\leq_e$ -convex.*

Definition 6 *Let H be a nonempty subset of X^+.*

(i) We denote $S(H) = \{w \in X^* | u \leq_e w, u \in H\}$.

(ii) H is called a hypercode over X^* iff *$x \leq_e y$ and $x, y \in H$ imply $x = y$.*

Proposition 7 *Let θ be an involution. Then there exists a unique hypercode H such that $hp(\theta, 1) = S(H)$.*

Properties of hp(θ, k)-free languages

Proposition 8 *The languages* $hp(\theta, k)$ *and* $hpf(\theta, k)$, $k \geq 1$, *are regular.*

Properties of hp(θ, k)-free languages

Proposition 8 *The languages* $hp(\theta, k)$ *and* $hpf(\theta, k)$, $k \geq 1$, *are regular.*

Proposition 9 *Let* X *be a binary alphabet. For every word* $w \in X^*$ *in* $hpf(\mu, 4)$ *we have that* $|w| \leq 31$. *Moreover the following word of length* 31 *is in* $hpf(\mu, 4)$

$$a^7ba^3bababab^2ab^2a^2b^7.$$

Properties of hp(θ, k)-free languages

Proposition 8 *The languages* $hp(\theta, k)$ *and* $hpf(\theta, k)$, $k \geq 1$, *are regular.*

Proposition 9 *Let* X *be a binary alphabet. For every word* $w \in X^*$ *in* $hpf(\mu, 4)$ *we have that* $|w| \leq 31$. *Moreover the following word of length* 31 *is in* $hpf(\mu, 4)$

$$a^7ba^3bababab^2ab^2a^2b^7.$$

Proposition 10 *Consider a binary alphabet* X. *Then* $hpf(\mu, k)$ *is finite if and only if* $k \leq 4$.

Properties of hp(θ, k)-free languages

Proposition 8 *The languages* $hp(\theta, k)$ *and* $hpf(\theta, k)$, $k \geq 1$, *are regular.*

Proposition 9 *Let* X *be a binary alphabet. For every word* $w \in X^*$ *in* $hpf(\mu, 4)$ *we have that* $|w| \leq 31$. *Moreover the following word of length* 31 *is in* $hpf(\mu, 4)$

$$a^7ba^3bababab^2ab^2a^2b^7.$$

Proposition 10 *Consider a binary alphabet* X. *Then* $hpf(\mu, k)$ *is finite if and only if* $k \leq 4$.

Proposition 11 *Let* θ *be an involution. The language* $hpf(\theta, k)$ *over a non-singleton alphabet* X *is finite if and only if one of the following holds:*
(a) $\theta = \epsilon$, *the identity involution;*
(b) $\theta = \mu$, *the mirror involution, and either* $k = 1$ *or* $|X| = 2$ *and* $k \leq 4$.

Decidability of hp(θ, k)-freedom

Corollary 12 *The following problem is decidable in linear time w.r.t.* $|M|$*:*

Input: *An NFA* M*.*

Output: *Yes/No depending on whether* $L(M)$ *is* $hp(\theta, k)$*-free.*

Decidability of hp(θ, k)-freedom

Corollary 12 *The following problem is decidable in linear time w.r.t.* $|M|$*:*

Input: *An NFA* M*.*

Output: *Yes/No depending on whether* $L(M)$ *is* $hp(\theta, k)$*-free.*

Corollary 13 *The following problem is decidable in linear time w.r.t.* $|M_1| \cdot |M_2|$*:*

Input: *A DFA* M_1 *such that* $L(M_1)$ *is* $hp(\theta, k)$*-free, and a NFA* M_2*.*

Output: *Yes/No depending on whether there is a word* $w \in L(M_2) - L(M_1)$ *such that* $L(M_1) \cup \{w\}$ *is* $hp(\theta, k)$*-free.*

The context-free case

Corollary 14 *The following problem is decidable in cubic time w.r.t.* $|M|$*:*

Input: *A PDA* M*.*

Output: *Yes/No depending on whether* $L(M)$ *is* $hp(\theta, k)$*-free.*

The context-free case

Corollary 14 *The following problem is decidable in cubic time w.r.t.* $|M|$*:*

Input: *A PDA* M*.*

Output: *Yes/No depending on whether* $L(M)$ *is* $hp(\theta, k)$*-free.*

Corollary 15 *The following problem is decidable in cubic time w.r.t.* $|M_1| \cdot |M_2|$*:*

Input: *A DPDA* M_1 *such that* $L(M_1)$ *is* $hp(\theta, k)$*-free, and a NFA* M_2*.*

Output: *Yes/No depending on whether there is a word* $w \in L(M_2) - L(M_1)$ *such that* $L(M_1) \cup \{w\}$ *is* $hp(\theta, k)$*-free.*

Descriptional complexity issues

Proposition 16 *The number of states of a minimal NFA accepting the language* $hp(\theta, k)$, $k \geq 1$, *over an alphabet* X *of the cardinality* $\ell > 1$, *is between* ℓ^k *and* $3\ell^k$, *its size is at most* $3(\ell^k + \ell^{k+1})$.

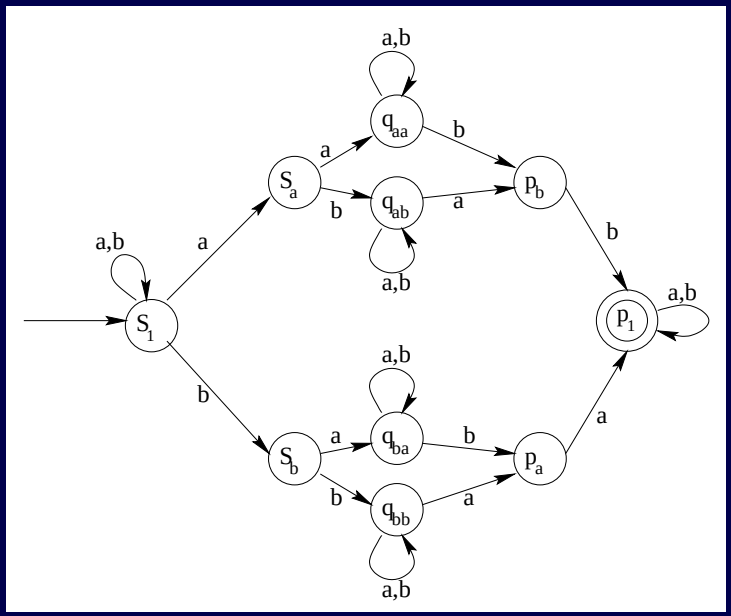


Figure 1: An example of an NFA accepting the language $hp(\theta, 2)$.

Proposition 17 *Assume that there are distinct letters* $a, b \in X$ *such that* $a = \theta(b)$. *Then the number of states of a minimal NFA accepting* $hpf(\theta, k)$, $k \geq 1$, *over an alphabet* X *with the cardinality* ℓ, *is at least* $2^{(\ell-2)^k/2}$.

Proposition 17 *Assume that there are distinct letters* $a, b \in X$ *such that* $a = \theta(b)$. *Then the number of states of a minimal NFA accepting* $hpf(\theta, k)$, $k \geq 1$, *over an alphabet* X *with the cardinality* ℓ, *is at least* $2^{(\ell-2)^k/2}$.

Corollary 18 *Let* X *be an alphabet such that* $|X| = \ell$, $\ell \geq 2$. *Let there be distinct letters* $a, b \in X$ *such that* $a = \theta(b)$. *Then the number of states of a minimal DFA over the alphabet* X, *accepting either* $hp(\theta, k)$ *or* $hpf(\theta, k)$, $k \geq 1$, *is between* $2^{(\ell-2)^k/2}$ *and* $2^{3\ell^k}$.

Proposition 17 *Assume that there are distinct letters $a, b \in X$ such that $a = \theta(b)$. Then the number of states of a minimal NFA accepting $hpf(\theta, k)$, $k \geq 1$, over an alphabet X with the cardinality ℓ, is at least $2^{(\ell-2)^k/2}$.*

Corollary 18 *Let X be an alphabet such that $|X| = \ell$, $\ell \geq 2$. Let there be distinct letters $a, b \in X$ such that $a = \theta(b)$. Then the number of states of a minimal DFA over the alphabet X, accepting either $hp(\theta, k)$ or $hpf(\theta, k)$, $k \geq 1$, is between $2^{(\ell-2)^k/2}$ and $2^{3\ell^k}$.*

Corollary 19 *Consider the DNA alphabet $\Delta = \{A, C, T, G\}$ and the Watson-Crick involution τ.*

(i) The size of a minimal NFA accepting $hp(\tau, k)$ is at most $15 \cdot 4^k$. The number of its states is between 4^k and $3 \cdot 4^k$.

(ii) The number of states of either a minimal DFA or an NFA accepting $hpf(\tau, k)$ is between $2^{2^{k-1}}$ and $2^{3 \cdot 2^{2k}}$.

Thank you!