Word periods under involution[1]

This talk addresses questions about unbordered words, local and global periods generalized by considering involutions. Apart from general interest, this topic draws its motivation from combinatorial questions in computational biology and DNA computing. The complementation of a single stranded DNA, understood as a word over the alphabet $\{A, C, G, T\}$, constitutes what we call an antimorphic involution where A is mapped to T and C to G and the order of letters is reversed; see also the work of Lila Kari et.al. [2, 5, 6]. The presented material results from work in progress.

Let A denote an alphabet and A^* the free monoid of all finite words over A. Let $w \in A^*$ denote a word over A. Let $|w|$ denote the length of w. Let θ denote an involution on A^*, that is, $\theta(\theta(w)) = w$. Then θ is called morphic, if $\theta(uv) = \theta(u)\theta(v)$, and antimorphic, if $\theta(uv) = \theta(v)\theta(u)$. A word w is called primitive, if $w = u^i$ implies $i = 1$ for any $u \in A^*$. The primitive root of w is the shortest word u such that $w = u^i$ for some $i \geq 1$. A word w is called θ-primitive, if $w \in \{u, \theta(u)\}^i$ implies $i = 1$ for any $u \in A^*$. The θ-primitive root of w is the shortest word u such that $w \in \{u, \theta(u)\}^i$ for some $i \geq 1$. If w is a prefix of u^ω for some u then $|u|$ is called (global) period of w. Let $\Pi(w)$ denote the set of all periods of w, and let $\pi(w)$ denote the shortest period of w.

The notion of periodicity under involution can be defined in several ways. If w is a prefix of $\{u, \theta(u)\}^\omega$ for some u then $|u|$ is called θ-period of w. Let $\Pi_\theta(w)$ denote the set of all θ-periods of w, and let $\pi_\theta(w)$ denote the shortest θ-period of w. If w is a prefix of $(u\,\theta(u))^\omega$ for some u then $|u|$ is called alternating θ-period of w. Let $\Pi_\theta^{alt}(w)$ denote the set of all alternating θ-periods of w, and let $\pi_\theta^{alt}(w)$ denote the shortest alternating θ-period of w.

Another θ-generalization of the notion of periodicity is the following. We say that a natural p is called weak θ-period of w, if $w_{[i]} = w_{[i+p]}$ or $w_{[i]} = \theta(w_{[i+p]})$ for all $1 \leq i \leq |w| - p$ where $w_{[j]}$ denotes the jth letter of w. Let $\Pi_\theta^{weak}(w)$ denote the set of all weak θ-periods of w. However, the following result shows that weak θ-periods do not seem to imply anything different than ordinary periods.

Theorem 1 *Let θ be an involution on A^* and $w \in A^*$. Let ψ be a suitable substitution for θ. Then $\Pi_\theta^{weak}(w) = \Pi(\psi(w))$.*

Where a suitable substitution ψ for θ is a substitution with the following properties: Let A' be an alphabet such that (1) $a \in A'$ or $\theta(a) \in A'$ for all $a \in A$ and (2) $a \in A'$ and $a \neq \theta(a)$ implies $\theta(a) \notin A'$. Then ψ is a substitution where $\psi(a) = a$, if $a \in A'$, and $\psi(a) = \theta(a)$, if $a \notin A'$. We consider only (alternating) θ-periods in the following.

A very natural question regarding periods of words is the effects caused by overlaps. The classical result by Fine and Wilf [4] considers periods without involutions.

[1]by Dirk Nowotka jointly with Bastian Bischoff

Theorem 2 ([4]) *Let $p, q \in \Pi(w)$ for some word w. If $|w| \geq p + q - \gcd\{p, q\}$ then $\gcd\{p, q\} \in \Pi(w)$.*

Czeizler et.al. consider in [2] the general antimorphic case and state:

Theorem 3 ([2]) *Let θ be an anti-morphic involution on A^* and $w \in A^*$ and $p, q \in \Pi_\theta(w)$ where $p > q$. Let u and v be prefixes of w of length p and q, respectively. If $|w| \geq 2p + q - \gcd\{p, q\}$ then u and v have the same θ-primitive root, that is they have a common period not longer than q.*

We add the following for the morphic case.

Theorem 4 *Let θ be a morphic involution on A^* and $w \in A^*$.*
If $|w| \geq p + q - \gcd\{p, q\}$ with $p, q \in \Pi_\theta(w)$ then $\gcd\{p, q\} \in \Pi_\theta(w)$.
If $|w| \geq p + q$ with $p, q \in \Pi_\theta^{alt}(w)$ then $\gcd\{p, q\} \in \Pi_\theta^{alt}(w)$.

All bounds given so far are tight. The only open case is the one for alternating θ-periods where θ is antimorphic.

Theorem 5 *Let θ be an antimorphic involution on A^* and $w \in A^*$.*
If $|w| \geq p + q$ with $p, q \in \Pi_\theta^{alt}(w)$ then $\gcd\{p, q\} \in \Pi_\theta^{alt}(w)$.

This bound however could not shown to be tight. We conjecture that the alternating antimorphic case has actually the bound $|w| \geq p + q - \gcd\{p, q\}$. This conjecture is still open.

The Critical Factorization theorem (CFT) is fundamental in the investigation of local periods. Let $u \sim_p v$ denote the fact that either u is a prefix of v or vice versa. Similarly, we write $u \sim_s v$, if u is a suffix of v or vice versa. Consider a factorization $w = uv$, then x is called a repetition word for this factorization of w, if $u \sim_s x$ and $v \sim_p x$. The length of x is called local period for the factorization uv. The smallest local period is denoted by $\pi(u, v)$. It is straightforward to see that $\pi(u, v) \leq \pi(w)$. A factorization is called critical if $\pi(u, v) = \pi(w)$. The critical factorization theorem was developed in several papers, see [7, 1, 3], and can be stated as follows.

Theorem 6 (CFT) *Among any $\pi(w)$ many consecutive factorizations uv of a word w exists at least one that is critical, that is, where $\pi(u, v) = \pi(w)$.*

It is a natural generalization to consider local periods under an involution as we did for the global periods. Given a factorization uv of w, we call x a θ-repetition word, if $u \sim_s x$ and $v \sim_p \theta(x)$. The length of x is called local θ-period for the factorization uv. The smallest local θ-period is denoted by $\pi_\theta(u, v)$. However, this notion does not yield a structural property like the CFT, neither when $\pi_\theta(u, v)$ is related to the ordinary nor the θ-period of w, and neither for the morphic nor antimorphic case, and neither for the alternating nor non-alternating case. Let the following propositions exemplify this for the case of alternating θ-periods where θ is morphic.

Proposition 7 *Let $|A| \geq 3$ and θ be a morphic involution (not the identity) on A^*. Then there exists for all $p \geq 1$ a word w such that $p = \pi_\theta^{alt}(w)$ and $p = \pi_\theta(u, v)$ for all factorizations uv of w.*

Proposition 8 *Let θ be a morphic involution (not the identity) on A^*. Then there exists for all $p \neq 3$ a word w such that $p = \pi_\theta^{alt}(w)$ and $\pi_\theta(u,v) \in \{1,2\}$ for all factorizations uv of w.*

The inability to directly transfer a strong result on local periods to the case of local θ-periods suggests a deeper investigation of the local θ-periodic structure of words. Unbordered factors are an obvious subject of interest here. A word is bordered if there exists a proper prefix that is also a suffix. Otherwise, a word is called unbordered. Let $\tau(w)$ denote the maximal length of unbordered factors in w. The following are straightforward observations: $\tau(w) \leq \pi(w)$, the shortest border of a bordered word is unbordered, a shortest local repetition word is unbordered. Moreover, a short argument using Lyndon words establishes that, if $|w| \geq 2\pi(w) - 1$ then $\tau(w) = \pi(w)$. A more involved proof shows that, if $|w| \geq 7/3\,\tau(w)$ then $\tau(w) = \pi(w)$ (see the abstract by Štěpán Holub on page **??**). How does that property of unbordered factors relate to the θ-bordered case? A word w is called θ-bordered, if w has a proper prefix u such that $\theta(u)$ is a suffix of w. Otherwise, w is called θ-unbordered. Consider the morphic case and alternating θ-periods. The following sequence of words has a ration of word length to the maximal length of unbordered factors that approaches 3 and in the limit yet $\tau_\theta(w_i) \neq \pi_\theta^{alt}(w)$ thereby showing that the $7/3\,\tau_\theta(w)$ bound does not hold. Let

$$w_i = (ab)^i abb(ab)^i aab(ab)^i a$$

for any $i \geq 2$, and let $\theta(a) = b$ and $\theta(b) = a$. Then $|w_i| = 6(i+1)+1$ and $\tau_\theta(w_i) = 2i+4$ and $\pi_\theta^{alt}(w_i) = 4i+5$. We have $|w_i| \geq 7/3\,\tau_\theta(w_i)$ for all $i \geq 2$ and $\lim_{i\to\infty} |w_i|/\tau_\theta(w_i) = 3$. We conjecture that this example is the best possible.

Conjecture 9 *Let θ be a morphic involution on A^* and $w \in A^*$.*
If $|w| \geq 3\tau_\theta(w)$ then $\tau_\theta(w) = \pi_\theta^{alt}(w)$.

References

[1] Y. Césari and M. Vincent, *Une caractérisation des mots périodiques*, C. R. Acad. Sci. Paris Sér. A **286** (1978), 1175–1177.

[2] E. Czeizler, L. Kari, and S. Seki, *On a special class of primitive words*, Theoret. Comput. Sci. **411** (2010), no. 3, 617–630.

[3] J.-P. Duval, *Périodes et répétitions des mots du monoïde libre*, Theoret. Comput. Sci. **9** (1979), no. 1, 17–26.

[4] N. J. Fine and H. S. Wilf, *Uniqueness theorem for periodic functions*, Proc. Amer. Math. Soc. **16** (1965), 109–114.

[5] L. Kari and K. Mahalingam, *Involutively bordered words*, Internat. J. Found. Comput. Sci. **18** (2007), no. 5, 1089–1106.

[6] ———, *Watson–Crick palindromes in DNA computing*, Natural Computing **9** (2010), 297–316.

[7] M.-P. Schützenberger, *A property of finitely generated submonoids of free monoids*, Algebraic theory of semigroups (Proc. Sixth Algebraic Conf., Szeged, 1976) (Amsterdam), Colloq. Math. Soc. János Bolyai, vol. 20, North-Holland, 1979, pp. 545–576.