Periods and binary words

Vesa Halava* Tero Harju† Lucian Ilie‡

Abstract

We give an elementary short proof for a well known theorem of Guibas and Odlyzko stating that the sets of periods of words are independent of the alphabet size. As a consequence of our constructive proof, we obtain a linear time algorithm which, given a word, computes a binary one with the same periods. We give also a very short proof for the famous Fine-Wilf periodicity lemma.

Keywords: word, period, binary image, binary word, Fine-Wilf lemma

1. Introduction and basic definitions

Let A be a finite alphabet of at least two letters and A^* the set of all words over A; the empty word is denoted ε. For $w \in A^*$, $|w|$ denotes the length of w and w_i its ith letter. Therefore each word $w \in A^*$ has the form $w = w_1 w_2 \cdots w_{|w|}$. An integer p, with $1 \leq p \leq |w|$, is called a *period* of w if $w_i = w_{i+p}$, for all $1 \leq i \leq |w| - p$. The set of all periods of w is denoted by $\mathcal{P}(w)$. Notice that $\mathcal{P}(w) = \{|w|\}$ if and only if w is *unbordered*, that is, $w = uvu$ implies $u = \varepsilon$.

The notion of period of a word is very central in the theory of combinatorics on words. There are many beautiful results on periods of words. Among them is the theorem of Guibas and Odlyzko which states that the sets of periods of words are independent of the alphabet size. (Unary alphabets are, of course, out of discussion.) Put otherwise, it says that, for every word w, there exists a binary one, say w', such that $\mathcal{P}(w') = \mathcal{P}(w)$; w' will be called a *binary image* of w.

The proof given by [GuOd] to this unexpected result uses properties of the correlation and is somewhat complicated. In this note, we give an elementary short proof for this theorem. As the proof is constructive, we give also a fast algorithm which computes a binary image of a given word. The algorithm runs in linear time, so it is optimal. We shall give also a very short proof (the shortest to our

*Turku Centre for Computer Science, FIN-20520, Turku, Finland, e-mail: vehalava@cs.utu.fi.

†Department of Mathematics, University of Turku, FIN-20014, Turku, Finland, e-mail: harju@utu.fi.

‡Turku Centre for Computer Science, FIN-20520, Turku, Finland, e-mail: lucili@cs.utu.fi. Research supported by the Academy of Finland, Project 137358. On leave of absence from Faculty of Mathematics, University of Bucharest, Str. Academiei 14, R-70109 Bucharest, Romania.

knowledge) for the famous Fine-Wilf periodicity lemma in its form for words, cf [ChKa] (Theorem 6.1) or [Lo] (Proposition 1.3.5). (Initially, Fine and Wilf proved the result for real functions, cf. [FiWi].)

For a word w, the powers of w are defined inductively by $w^0 = \varepsilon$ and, for any $n \geq 1$, $w^n = ww^{n-1}$. A word u is *primitive* if there is no word v such that $u = v^k$, where $k \geq 2$. For two words u, v, we say that u is a *prefix* of v if $v = ux$, for some $x \in A^*$; u is a *factor* of v if $v = xuy$, for some $x, y \in A^*$.

For basic notions and results on words we refer to [ChKa] and [Lo].

2. Properties of words and periods

In this section we give first the announced proof of Fine-Wilf lemma (Lemma 1) and then prove some properties of words and periods needed in the proof of the main theorem in the next section.

Lemma 1. *If a word w has periods p and q and $|w| \geq p + q - \gcd(p, q)$, then w has also period $d = \gcd(p, q)$.*

Proof. For a fixed d, by induction on $p + q$. The first step is trivial. Suppose the statement holds for all integers smaller than $p + q$. Assume $p > q$ and put $w = uv$, where $|u| = p - d$. Now, for any $1 \leq i \leq q - d$, we have $u_i = w_i = w_{i+p} = w_{i+p-q} = u_{i+p-q}$, and so u has period $p - q$. Since u has also period q and $\gcd(p - q, q) = d$, the inductive hypothesis shows that u has period d. Now, $|u| \geq q$ implies that the prefix of length q of w has period d. Since w has period q and $d|q$, it follows that w has period d, too. ■

Next lemma gives us the structure of the set of periods. We call the least $p \in \mathcal{P}(w)$, the *minimum period* of w. Notice that, for any $w, \mathcal{P}(w) \neq \emptyset$, since at least $|w| \in \mathcal{P}(w)$. Hence the minimum period is well defined.

Lemma 2. *Let $w \in A^*$ and $p \in \mathcal{P}(w)$ be the minimum period of w. Then, for any $q \in \mathcal{P}(w)$ with $q \leq |w| - p$, q is a multiple of p.*

Proof. Since $p + q \leq |w|$, we get by Lemma 1 that $\gcd(p, q) \in \mathcal{P}(w)$. As p is the minimum period, we must have $p = \gcd(p, q)$, and so $p|q$. ■

As a corollary, we obtain that if the minimum period satisfies $p \leq |w|/2$, then the set of periods can be partitioned into two sets, the first one including the minimum period p and all of its multiples and the second one including all the periods for which $q > |w| - p$.

Lemma 3. *For any word $w \in \{0, 1\}^*$, $w0$ or $w1$ is primitive.*

Proof. For $w = \varepsilon$ both $w0$ and $w1$ are primitive. Let then $|w| > 0$, and assume that $w0 = v^k, w1 = u^\ell$, for some primitive words u, v and integers $k, \ell \geq 2$. Both $|v|$ and $|u|$ are periods of w, and, since $k, \ell \geq 2$, $|w| = k|v| - 1 = \ell|u| - 1 \geq 2\max\{|v|, |u|\} - 1 \geq |v| + |u| - 1$. By Lemma 1, also $d = \gcd(|v|, |u|)$ is a period of

w. However, d divides $|v|$ and $|u|$, and since v and u are primitive, we conclude that $|v| = d = |u|$. Now u and v are prefixes of w, and thus $v = u$, which contradicts the fact that v and u end with different letters, 0 and 1, respectively. ■

3. Main theorem

We prove in this section our main theorem (Theorem 1). For a given word, we recursively construct a binary image of it. The recursion has two different cases which, for clarity, we study separately in Lemmata 5 and 6.

In the following three lemmata, let $w \in A^*$ be a word with the minimum period p. Then there are words $u, v \in A^*$ (possibly $u = \varepsilon$) such that

$$w = (uv)^k u \quad \text{where } p = |uv|,\ v \neq \varepsilon \text{ and } k \geq 1. \tag{1}$$

Lemma 4. *Let $w \in A^*$ be as in (1) with $k \geq 2$, and let q be such that $|w| - p < q < |w|$. Put $q = (k-1)p + r$, where $|u| < r < |u| + p$. Then $q \in \mathcal{P}(w)$ if and only if $r \in \mathcal{P}(uvu)$.*

Proof. For any $0 < i < |w| - q = p + |u| - r$, we have $w_i = (uvu)_i$ and $w_{i+q} = (uvu)_{i+r}$. Hence $w_i = w_{i+q}$ if and only if $(uvu)_i = (uvu)_{i+r}$ which means exactly that $q \in \mathcal{P}(w)$ if and only if $r \in \mathcal{P}(uvu)$, as claimed. ■

Lemma 5. *Let $w \in A^*$ be as in (1) with $k \geq 1$. Suppose that $u'v'u'$ is a binary image of uvu, where $|u'v'| = |uv|$. Then $\mathcal{P}(w) = \mathcal{P}(w')$ for the binary word $w' = (u'v')^k u'$.*

Proof. The case $k = 1$ is trivial. Assume thus that $k \geq 2$. Notice that $p = |uv| = |u'v'| \in \mathcal{P}(w')$.

First, consider q with $q \leq |w| - p$. Assume $q \in \mathcal{P}(w')$. Now $|w'| = |w| \geq p + q$, and thus, by Lemma 1, $d = \gcd(p, q) \in \mathcal{P}(w')$. Since $d \leq p = |u'v'|$, also $d \in \mathcal{P}(u'v'u') = \mathcal{P}(uvu)$, and so $d|p$ implies $d \in \mathcal{P}(w)$. By the minimality of p, we have $d = p$, and therefore $p|q$, which implies $q \in \mathcal{P}(w)$. On the other hand, if $q \in \mathcal{P}(w)$, then Lemma 2 gives that $p|q$, and therefore $q \in \mathcal{P}(w')$, since $p \in \mathcal{P}(w')$.

Let then $|w| - p < q < |w|$, and put $q = (k-1)p + r$, where $|u| < r < p + |u|$. Then, by Lemma 4, $q \in \mathcal{P}(w)$ if and only if $r \in \mathcal{P}(uvu) = \mathcal{P}(u'v'u')$ which, also by Lemma 4, is equivalent with $q \in \mathcal{P}(w')$. This completes the proof. ■

Lemma 6. *Let $w = uvu \in A^*$ be as in (1) with $k = 1$. Let also $u' \in \{0,1\}^*$ be a binary image of u and assume that u' begins with letter 0. For $a \in \{0,1\}$ such that $u'1^{|v|-1}a$ is primitive, $\mathcal{P}(w) = \mathcal{P}(w')$, for the binary word $w' = u'1^{|v|-1}au'$.*

Proof. Clearly, $\mathcal{P}(w) \subseteq \mathcal{P}(w')$, since u' is a binary image of u and all periods q of w satisfy $q \geq p = |uv| = |u'1^{|v|-1}a|$. Assume then that there is $q \in \mathcal{P}(w') - \mathcal{P}(w)$ and also that q is minimal with this property. Clearly, either $q < |u'|$ or $|u'| + |v| - 1 \leq q < |w|$, since u' does not begin with 1.

If $q < |u'|$, then, by the minimality of q, it is the minimum period of w', and Lemma 2 implies $q|p$, and so $u'1^{|v|-1}a$ is not primitive, a contradiction. If $q = |u'| + |v| - 1$, then $a = 0$, in which case we get $u'1 = 0u'$, which is impossible. Therefore $q > p = |uv|$. Put $q = p + r, r > 0$. Then, clearly, r is a period of u' and hence of u. But this implies $q \in \mathcal{P}(w)$, a contradiction. The proof is completed. ■

Theorem 1. *For any alphabet A and any word $w \in A^*$, there exists a word $w' \in \{0,1\}^*$ such that $\mathcal{P}(w') = \mathcal{P}(w)$.*

Proof. By induction on $|w|$. For $|w| \leq 2$, the claim is clear.

Assume that the claim holds for all words of length less than or equal to $n \geq 2$. Let $w \in A^*$ be as in (1) with $|w| = n + 1$.

For $k \geq 2$ we have $|uvu| \leq n$ and, by the inductive hypothesis, there exists a binary image $u'v'u'$ of uvu such that $|u'v'| = |uv|$. Now, by Lemma 5, $(u'v')^k u'$ is a binary image of w.

Consider next the case $k = 1$. As $v \neq \varepsilon$, we have $|u| \leq n$ and thus, by the inductive hypothesis, there exists a binary image u' of u. If $u' = \varepsilon$ then $w' = 01^{|v|-1}$ is clearly a binary image of $w = v$, since, in this case, $\mathcal{P}(w) = \{|w|\}$. Otherwise, assume that u' begins with the letter 0. By Lemma 3, there exists $a \in \{0,1\}$ such that $u'1^{|v|-1}a$ is primitive. By Lemma 6, the word $w' = u'1^{|v|-1}au'$ is a binary image of w. The theorem is proved. ■

4. The algorithm

From the proof of Theorem 1, we obtain a recursive algorithm for constructing a binary image of a given word w, denoted below by $\mathrm{Bin}(w)$.

Bin(w)

1. Find the minimum period p of w. If $p = |w|$, then output $\mathrm{Bin}(w) = 01^{|w|-1}$.
2. Find $u, v \in A^*$ and $k \geq 1$ such that $w = (uv)^k u$, where $v \neq \varepsilon$ and $|uv| = p$.
3. If $k \geq 2$, then compute $\mathrm{Bin}(uvu) = u'v'u'$, $|u'v'| = |uv|$, and output $\mathrm{Bin}(w) = (u'v')^k u'$.
4. Compute $\mathrm{Bin}(u) = u'$.
5. Find $a \in \{0,1\}$ such that the word $u'1^{|v|-1}a$ is primitive and then output $\mathrm{Bin}(w) = u'1^{|v|-1}au'$.

The correctness follows from the proof of Theorem 1. We finally consider the complexity of the algorithm.

Theorem 2. *The algorithm* Bin *runs in linear time and therefore is optimal.*

Proof. The algorithm is recursive, so let us compute the complexity of a single call of the procedure Bin, say $f(m)$, where m is the length of the current word for this call, say x.

Consider first Step 1. A linear pattern matching algorithm M (see e.g. [CrRy]) can be easily adapted to compute the minimum period of x as follows. Assume that, given two words (u, v), the algorithm finds the leftmost occurrence, if any, of u as a factor of v. The comparisons done by the algorithm are of the type $a \stackrel{?}{=} b$, for letters a and b. We consider a new letter (wild card) # which passes the test $\# \stackrel{?}{=} a$, for any letter a. Put $x = ax'$, where a is a letter. Then we run the algorithm M on the inputs $(x, x'\#^{|x|})$. Clearly, an integer $p, 1 \leq p \leq |x|$, is a period of x if and only if x is a factor starting at position $p-1$ of $x'\#^{|x|}$. Therefore, the leftmost occurrence of x as a factor of $x'\#^{|x|}$ (which always exists) gives the minimum period of x. Consequently, Step 1 can be performed in linear time.

Step 2 is performed in linear time, since we know that $p = |uv|$ from Step 1. Step 3 is obviously performed in linear time. At Step 5 we have to test which of the words $u'1^{|v|-1}0$ and $u'1^{|v|}$ is primitive. It is well known that primitivity can be tested in linear time. Indeed, a word z is primitive if and only if z is not a proper factor of z^2, that is, $z^2 = xzy$ implies that either $x = \varepsilon$ or $y = \varepsilon$; see e.g. [ChKa].

Consequently, we have shown so far that a single call of Bin requires $f(m) = \mathcal{O}(m)$ steps, where m is the length of the current word. More precisely, there is a constant c such that $f(m) \leq cm$, for any $m \geq 0$.

To calculate the time required for the whole algorithm on an input w of length n, we first see how fast the length of the current word decreases from a call to the next one. Consider x and y the current words for two consecutive calls of Bin, respectively. We have that either $x = (uv)^k u, y = uvu, k \geq 2$ (if Bin(y) is called at Step 3 in Bin(x)), or $x = uvu, y = u$ (if Bin(y) is called at Step 4 in Bin(x)). In either case, $|y| \leq \frac{2}{3}|x|$.

Therefore, the time required by the algorithm to compute Bin(w) is at most

$$\sum_{i \geq 0} f([(\tfrac{2}{3})^i n]) \leq \sum_{i \geq 0} c[(\tfrac{2}{3})^i n] \leq 3cn,$$

hence it is linear, as claimed. Finally, it is clear that the algorithm is optimal, as the problem requires at least linear time. ■

References

[ChKa] Choffrut, C., and Karhumäki, J., Combinatorics of Words, in G. Rozenberg, A. Salomaa, eds., *Handbook of Formal Languages, Vol. 1* (Springer-Verlag, Berlin, Heidelberg, 1997) 329 – 438.

[CrRy] Crochemore, M., and Rytter, W., *Text Algorithms* (Oxford University Press, 1994).

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

[GuOd] Guibas, L. J., and Odlyzko, A. M., Periods in strings, *J. Combin. Theory, Ser A*, **30**(1) (1981) 19 – 42.

[Lo] Lothaire, M., *Combinatorics on Words* (Addison-Wesley, Reading, MA., 1983).