Conjugacy on partial words

By: Francine Blanchet-Sadri and D.K. Luhmann

F. Blanchet-Sadri and D.K. Luhmann, "Conjugacy on Partial Words." Theoretical Computer Science, Vol. 289, No. 1, 2002, pp 297-312.

Abstract:
The study of the combinatorial properties of strings of symbols from a finite alphabet (also referred to as words) is profoundly connected to numerous fields such as biology, computer science, mathematics, and physics. In this paper, we examine to which extent some fundamental combinatorial properties of words, such as conjugacy, remain true for *partial* words. The motivation behind the notion of a partial word is the comparison of two genes (alignment of two such strings can be viewed as a construction of two partial words that are said to be compatible). This study on partial words was initiated by Berstel and Boasson.

Article:
1. Introduction
The study of the combinatorial properties of strings of symbols from a finite alphabet is profoundly connected to numerous fields such as biology, computer science, mathematics, and physics. The symbols from the alphabet are also referred to as *letters* and the strings as *words*. The stimulus for recent works on combinatorics of finite words is the study of molecules such as DNA that play a central role in molecular biology [1–3, 5–10]. The genetic information in almost all organisms is carried by molecules of DNA. A DNA molecule is a quite long but finite string of nucleotides of four possible types: *a* (for adenine), *c* (for cytosine), *g* (for guanine), and *t* (for thymine). A protein is a string of amino acids of 20 possible types. The set of the 64 3-letter words over the alphabet $\{a, c, g, t\}$ is the set $\{aaa, aac, \ldots, ttt\}$ whose elements are referred to as *codons*. Every codon codes a uniquely determined amino acid except for the codons *taa*, *tag*, and *tga*, hence, every codon except *taa*, *tag*, and *tga* is referred to as a coding codon. Several codons may code the same amino acid. A (protein coding) *gene* is a string of nucleotides of the form $u = xy_1y_2 \ldots y_ny_{n+1}z$ where $y_1 = atg$, $y_{n+1} \in \{taa, tag, tga\}$; $y_2, \ldots, y_n$ are coding codons, and y_1 is the first occurrence of *atg* in *u*. Proteins are synthesized from genes in the transcription – translation process which takes as input a gene like *u* above and which outputs the string of *n* amino acids corresponding to $y_1y_2 \ldots y_n$.

Partial words appear in comparing genes. Indeed, alignment of two strings can be viewed as a construction of two *partial words* that are *compatible* in a sense that will be described in Section 3. More precisely, a word of length *n* over a finite alphabet *A* is a map from $\{1, \ldots, n\}$ into *A* while a *partial* word of length *n* over *A* is a partial map from $\{1, \ldots, n\}$ into *A*. In the latter case, elements of $\{1, \ldots, n\}$ without an image are called holes (a word is just a partial word without holes). In this paper, we extend some fundamental combinatorial properties of words to partial words with an arbitrary number of holes. This study was initiated by Berstel and Boasson [2]. In particular, in Section 4, we extend results which were proved for partial words with a single hole to partial words with an arbitrary number of holes. The definition of *special* partial word is crucial for these extensions. In Section 5, we extend the important combinatorial property of conjugacy of words to partial words with an arbitrary number of holes by answering a question that was raised.

2. Preliminaries on words
This section is devoted to reviewing basic concepts on words. For more information on the matters discussed here, see the book by Lothaire [11] or the Handbook of Formal Languages (Vol. 1, Chapter 6 by Choffrut and

Karhumäki) [4].

Let A be a nonempty finite set, or an *alphabet*. Elements of A are called *letters* and finite sequences of letters of A are called (finite) *words* over A. The unique sequence of length 0, denoted by ε, is called the *empty word*. The set of all words over A of finite length (greater than or equal to 0) is denoted by A^*. It is a monoid under the associative operation of concatenation or product of words (ε serves as identity) and is referred to as the *free monoid* generated by A. Similarly, the set of all nonempty words over A is denoted by A^+. It is a semigroup under the operation of concatenation of words and is referred to as the *free semigroup* generated by A. The set of all words over A of length n is denoted by A^n ($A^* = \bigcup_{n\geq \text{or}=0} A^n$ and $A^+ = \bigcup_{n>0} A^n$).

A word of length n over A can be defined by a map u: $\{1, \ldots, n\} \rightarrow A$ where $u(i) = a_i$ with $a_i \in A$ and is usually represented as $u = a_1a_2 \ldots a_n$. The length of u or n is denoted by $|u|$.

For any word u and $n \geq 0$, the n-power u^n is defined as $u^0 = \varepsilon$, $u^n = uu^{n-1}$ for $n > 0$. A word u is a *factor* of v if there exist words x and y such that $v = xuy$. The factor u is called proper if $u \neq v$. If $x = \varepsilon$ (respectively, $y = \varepsilon$), then u is called a *prefix* (respectively, *suffix*) of v.

We now state some well-known combinatorial properties of words.

The following important property, usually referred to as the *equidivisibility property* or *lemma of Levi*, holds.

Lemma 1. *Let u,v,x,y be words such that $ux = vy$.*

- *If $|u| \geqslant |v|$, then there exists a word z such that $u = vz$ and $y = zx$.*
- *If $|u| \leqslant |v|$, then there exists a word z such that $v = uz$ and $x = zy$.*

The following two properties on words u and v are equivalent to u and v being powers of the same word.

Theorem 1. *Let u and v be words. Then $u^k = v^\ell$ for some integers k and ℓ if and only if there exists a word w such that $u = w^m$ and $v = w^n$ for some integers m and n.*

Theorem 2. *Let u and v be words. Then $uv = vu$ if and only if there exists a word w such that $u = w^m$ and $v = w^n$ for some integers m and n.*

We end this section with the important combinatorial property of conjugacy.

Definition 1. Two words u and v are called conjugate if there exist words x and y such that $u = xy$ and $v = yx$.

Theorem 3. *Two nonempty words u and v are conjugate if and only if there exists a word z such that $uz = zv$. Moreover; in this case there exist words x; y such that $u = xy$; $v = yx$; and $z = x(yx)^n$ for some $n \geq 0$.*

3. Preliminaries on partial words

In this section, we give a brief overview of partial words [2]. Throughout this section, we let A be a finite alphabet.

A *partial word* u of length n over A is a partial map u: $\{1, \ldots, n\} \rightarrow A$. If $1 \leq i \leq n$; then i belongs to the *domain* of u (denoted by *Domain*(u)) in the case where $u(i)$ is defined and i belongs to the *set of holes* of u (denoted by *Hole*(u)) otherwise. A word over A is a partial word over A with an empty set of holes (we will sometimes refer to words as *full* words).

Let u be a partial word of length n over A. The *companion* of u (denoted by u_o) is the map $u_o : \{1, \ldots, n\} \rightarrow A \cup \{o\}$ defined by

$$u_o(i) = \begin{cases} u(i) & \text{if } i \in \mathit{Domain}(u), \\ o & \text{otherwise.} \end{cases}$$

The bijectivity of the map $u \longmapsto u_o$ allows us to define for partial words concepts such as concatenation in a trivial way. The symbol *o* is viewed as a "do not know" symbol and not as a "do not care" symbol as in pattern matching. The word $u_o = abbobocbb$ is the companion of the partial word u of length 9 where $Domain(u) = \{1, 2,3,5,7,8,9\}$ and $Hole(u) = \{4;6\}$.

Let u be a partial word over A. A period of u is a positive integer k such that $u(i) = u(j)$ whenever $i, j \in Domain(u)$ and $i \equiv j \pmod k$ (in other words, k divides $j - i$). In such a case, we call u k-periodic. For example, the partial word with companion *abboboabb* is 3-periodic.

Let u and v be two partial words of length n. The partial word u is said to be contained in the partial word v (denoted by $u \subset v$) if $Domain(u) \subset Domain(v)$ and $u(i) = v(i)$ for all $i \in Domain(u)$. The partial words u and v are called *compatible* (denoted by $u \uparrow v$) if there exists a partial word w such that $u \subset w$ and $v \subset w$ (in which case we define $u \vee v$ by $u \subset u \vee v$ and $v \subset u \vee v$ and $Domain(u \vee v) = Domain(u) \cup Domain(v)$). As an example, $u_o = abaooa$ and $v_o = aooboa$ are the companions of two partial words u and v that are compatible and $(u \vee v)_o = ababoa$.

The following rules are useful for computing with partial words.

- *Multiplication*: If $u \uparrow v$ and $x \uparrow y$, then $ux \uparrow vy$.
- *Simplification*: If $ux \uparrow vy$ and $|u| = |v|$, then $u \uparrow v$ and $x \uparrow y$.
- *Weakening*: If $u \uparrow v$ and $w \subset u$, then $w \uparrow v$.

Lemma 1's version for partial words can be stated as follows.

Lemma 2 (Berstel and Boasson [2]). *Let u, v, x, y be partial words such that $ux \uparrow vy$.*
- *If $|u| \geqslant |v|$, then there exists partial words w, z such that $u = wz$, $v \uparrow w$, and $y \uparrow zx$.*
- *If $|u| \leqslant |v|$, then there exists partial words w, z such that $v = wz$, $u \uparrow w$, and $x \uparrow zy$.*

The following lemma was used to prove Theorem 6 that follows.

Lemma 3 (Berstel and Boasson [2]). *Let u and v be two words and let w be a partial word with only one hole. If $w \subset uv$ and $w \subset vu$; then $uv = vu$.*

Theorem 1's version for partial words can be stated as follows.

Theorem 4 (Berstel and Boasson [2]). *Let u and v be partial words. Then $u^k \uparrow v^\ell$ for some integers k and ℓ if and only if there exists a word w such that $u \subset w^m$ and $v \subset w^n$ for some integers m and n.*

The following two results relate to extending Theorem 2 to partial words.

Theorem 5 (Berstel and Boasson [2]). *Let u and v be partial words. If there exists a word w such that $u \subset w^m$ and $v \subset w^n$ for some integers m and n; then $uv \uparrow vu$.*

Theorem 6 (Berstel and Boasson [2]). *Let u and v be partial words such that uv has at most one hole. If $uv \uparrow vu$; then there exists a word w such that $u \subset w^m$ and $v \subset w^n$ for some integers m and n.*

As stated in Ref. [2], Theorem 6 is false if uv has two holes even if $uv \vee vu$ has no hole. Take for example $u_o = obb$ and $v_o = abbo$.

4. Extensions of Berstel and Boasson's results

In this section, we extend Berstel and Boasson's Lemma 3 and Theorem 6 to partial words with an arbitrary number of holes. The concept of a special partial word is crucial for these extensions.

Definition 2. Let k and ℓ be positive integers satisfying $k \le \ell$. For $1 \le i \le k + \ell$, we define the sequence of i relative to k and ℓ as $seq_{k,\ell}(i) = (i_0, i_1, i_2, \ldots, i_n, i_{n+1})$ where

- $i_0 = i = i_{n+1}$,
- for $1 \le j \le n$, $i_j \ne i$,
- for $1 \le j \le n + 1$; i_j is defined as

$$i_j = \begin{cases} i_{j-1} + k & \text{if } i_{j-1} \leqslant \ell, \\ i_{j-1} - \ell & \text{otherwise.} \end{cases}$$

For example, if $k = 4$ and $\ell = 10$, then $seq_{4,10}(1)=(1,5,9,13,3,7,11,1)$ and $seq_{4,10}(6) = (6;10;14;4;8;12;2;6)$.

Definition 3. Let k and ℓ be positive integers satisfying $k \le \ell$ and let w be a partial word of length $k + \ell$. We say that w is $\{k, \ell\}$-special if there exists $1 \le i \le k$ such that $seq_{k,\ell}(i) = (i_0, i_1, i_2, \ldots, i_n, i_{n+1})$ satisfies one of the following conditions:

1. $seq_{k,\ell}(i)$ contains two consecutive positions that are holes of w.
2. $seq_{k,\ell}(i)$ contains two positions that are holes of w while $w_o(i_0)w_o(i_1)w_o(i_2) \ldots w_o(i_{n+1})$ is not 1-periodic.

For example, if $k = 4$ and $\ell = 10$, then

- The partial word u with companion $u_o = aboaaboaabaaoo$ is $\{4; 10\}$-special since $seq_{4,10}(1) = (1,5,9,13,3,7,11,1)$ contains the consecutive positions 13 and 3 which are in $Hole(u) = \{3,7,13,14\}$.
- The partial word v with companion $v_o = aobaaboaabaaoo$ is $\{4;10\}$-special since $seq_{4,10}(1)$ contains the positions 7 and 13 which are in $Hole(v) = \{2,7,13,14\}$ while $v_o(1)v_o(5)v_o(9)v_o(13)v_o(3)v_o(7)v_o(11)v_o(1) = aaaoboaa$ is not 1-periodic.
- The partial word w with companion $w_o = obababobababob$ is not $\{4; 10\}$-special.

We now extend Lemma 3. The following examples show that the assumption that w is not $\{|u|, |v|\}$-special is necessary. As a first example, let $w_o = occaccoccbcc$, $u = bcc$, and $v = accaccbcc$. Here $w \subset uv$ and $w \subset vu$, but $uv \ne vu$ (w is $\{3,9\}$-special). As a second example, let $w_o = aobaaboaabaaoo$, $u = aaba$, and $v = abbaabaaab$. Here $w \subset uv$ and $w \subset vu$, but $uv \ne vu$ (w is $\{4, 10\}$-special).

Lemma 4. *Let w be a partial word and u, v be full words such that $w \subset uv$ and $w \subset vu$. If w is not $\{|u|, |v|\}$-special; then $uv = vu$.*

Proof. Put $|u| = k$ and $|v| = \ell$. Without loss of generality, we can assume that $k \le \ell$. The proof is split into cases which refer to a given position i of w_o. Case 1 treats the situation when $1 \le i \le k$, Case 2 the situation when $k < i \le \ell$, and Case 3 when $\ell < i \le \ell + k$ (Cases 1 and 3 are symmetric). The following diagram pictures the inclusions $w \subset uv$ and $w \subset vu$:

$$\begin{array}{l|lll|lll|lll}
w_o & w_o(1) & \cdots & w_o(k) & w_o(k+1) & \cdots & w_o(\ell) & w_o(\ell+1) & \cdots & w_o(\ell+k) \\
uv & u(1) & \cdots & u(k) & v(1) & \cdots & v(\ell-k) & v(\ell-k+1) & \cdots & v(\ell) \\
vu & v(1) & \cdots & v(k) & v(k+1) & \cdots & v(\ell) & u(1) & \cdots & u(k)
\end{array}$$

Put $\ell = mk + r$ where $0 \le r < k$. We first treat the case where $r = 0$.

Case 1. If $i \in Domain(w)$, then $w(i) \subset u(i)$ and $w(i) \subset v(i)$ and so $u(i) = v(i)$. If $i \in Hole(w)$, then we prove that $u(i) = v(i)$ as follows. We have

$$w_o(i) \subset u(i) \text{ and } w_o(i) \subset v(i),$$

$$w_o(i+k) \subset v(i) \text{ and } w_o(i+k) \subset v(i+k),$$

$$w_o(i+2k) \subset v(i+k) \text{ and } w_o(i+2k) \subset v(i+2k),$$

$$\vdots$$

$$w_o(i+(m-1)k) \subset v(i+(m-2)k) \text{ and } w_o(i+(m-1)k) \subset v(i+(m-1)k),$$

$$w_o(i+mk) \subset v(i+(m-1)k) \text{ and } w_o(i+mk) \subset u(i).$$

Here $w_o(i)w_o(i+k)w_o(i+2k) \ldots w_o(i+mk)w_o(i)$ does not contain consecutive holes and does not contain two holes while not 1-periodic (w is not $\{k,\ell\}$-special). So $u(i) = v(i+(m-1)k) = v(i+(m-2)k) = \cdots = v(i+k) = v(i)$ ($Hole(w)$ does not contain in particular $i+k$, $i+mk$).

Case 2. If $i \in Domain(w)$, then $w(i) \subset v(i-k)$ and $w(i) \subset v(i)$ and so $v(i-k) = v(i)$. If $i \in Hole(w)$, then put $i = nk + s$ where $0 \le s < k$. If $s = 0$, then

$$w_o(nk) \subset v((n-1)k) \text{ and } w_o(nk) \subset v(nk),$$

$$w_o((n+1)k) \subset v(nk) \text{ and } w_o((n+1)k) \subset v((n+1)k),$$

$$w_o((n+2)k) \subset v((n+1)k) \text{ and } w_o((n+2)k) \subset v((n+2)k),$$

$$\vdots$$

$$w_o(mk) \subset v((m-1)k) \text{ and } w_o(mk) \subset v(mk),$$

$$w_o((m+1)k) \subset v(mk) \text{ and } w_o((m+1)k) \subset u(k),$$

$$w_o(k) \subset u(k) \text{ and } w_o(k) \subset v(k),$$

$$w_o(2k) \subset v(k) \text{ and } w_o(2k) \subset v(2k),$$

$$\vdots$$

$$w_o((n-2)k) \subset v((n-3)k) \text{ and } w_o((n-2)k) \subset v((n-2)k),$$

$$w_o((n-1)k) \subset v((n-2)k) \text{ and } w_o((n-1)k) \subset v((n-1)k).$$

Here $w_o(nk)w_o((n+1)k)w_o((n+2)k) \ldots w_o(mk)w_o((m+1)k)w_o(k)w_o(2k) \ldots w_o((n-1)k)w_o(nk)$ does not contain consecutive holes and does not contain two holes while not 1 -periodic. We conclude that $v(i-k) = v((n-1)k) = \cdots = v(2k) = v(k) = u(k) = v(mk) = v((m-1)k) = \cdots = v(nk) = v(i)$ ($Hole(w)$ does not contain $i+k$, $i-k$ in particular). If $s > 0$, then

$$w_o(nk+s) \subset v((n-1)k+s) \text{ and } w_o(nk+s) \subset v(nk+s),$$

$$w_o((n+1)k+s) \subset v(nk+s) \text{ and } w_o((n+1)k+s) \subset v((n+1)k+s),$$

$$w_o((n+2)k+s) \subset v((n+1)k+s) \text{ and } w_o((n+2)k+s) \subset v((n+2)k+s),$$

$$\vdots$$

$$w_o((m-1)k+s) \subset v((m-2)k+s) \text{ and } w_o((m-1)k+s) \subset v((m-1)k+s),$$

$$w_o(mk+s) \subset v((m-1)k+s) \text{ and } w_o(mk+s) \subset u(s),$$

$$w_o(s) \subset u(s) \text{ and } w_o(s) \subset v(s),$$

$$w_o(k+s) \subset v(s) \text{ and } w_o(k+s) \subset v(k+s),$$

$$w_o(2k+s) \subset v(k+s) \text{ and } w_o(2k+s) \subset v(2k+s),$$

$$\vdots$$

$$w_o((n-2)k+s) \subset v((n-3)k+s) \text{ and } w_o((n-2)k+s) \subset v((n-2)k+s),$$

$$w_o((n-1)k+s) \subset v((n-2)k+s) \text{ and } w_o((n-1)k+s) \subset v((n-1)k+s).$$

As in the case when $s = 0$, we conclude that $v(i-k) = v((n-1)k+s) = \cdots = v(k+s) = v(s) = u(s) = v((m-1)k$

$+ s) = v((m-2)k + s) = \cdot\cdot\cdot = v(nk + s) = v(i)$. We now treat the case where $r > 0$.

Case 1. If $i \in Domain(w)$, then we proceed as in the case where $r = 0$. If $i \in Hole(w)$, we consider the cases where $i \leq r$ and $i > r$. If $i \leq r$, then

$$
\begin{aligned}
&w_o(i) \subset u(i) \text{ and } w_o(i) \subset v(i),\\
&w_o(i+k) \subset v(i) \text{ and } w_o(i+k) \subset v(i+k),\\
&w_o(i+2k) \subset v(i+k) \text{ and } w_o(i+2k) \subset v(i+2k),\\
&\vdots\\
&w_o(i+(m-1)k) \subset v(i+(m-2)k) \text{ and } w_o(i+(m-1)k) \subset v(i+(m-1)k),\\
&\qquad w_o(i+mk) \subset v(i+(m-1)k) \text{ and } w_o(i+mk) \subset v(i+mk),\\
&\qquad w_o(i+(m+1)k) \subset v(i+mk) \text{ and } w_o(i+(m+1)k) \subset u(i+k-r),\\
&\qquad w_o(i+k-r) \subset u(i+k-r) \text{ and } w_o(i+k-r) \subset v(i+k-r),\\
&\qquad \vdots
\end{aligned}
$$

If $i>r$, then

$$
\begin{aligned}
&w_o(i) \subset u(i) \text{ and } w_o(i) \subset v(i),\\
&w_o(i+k) \subset v(i) \text{ and } w_o(i+k) \subset v(i+k),\\
&w_o(i+2k) \subset v(i+k) \text{ and } w_o(i+2k) \subset v(i+2k),\\
&\vdots\\
&w_o(i+(m-1)k) \subset v(i+(m-2)k) \text{ and } w_o(i+(m-1)k) \subset v(i+(m-1)k),\\
&w_o(i+mk) \subset v(i+(m-1)k) \text{ and } w_o(i+mk) \subset u(i-r),\\
&w_o(i-r) \subset u(i-r) \text{ and } w_o(i-r) \subset v(i-r),\\
&\vdots
\end{aligned}
$$

Applying the above repeatedly, we can show that $v(i) = u(i)$. More precisely, in the case where $i \leq r$, $seq_{k,\ell}(i) = (i,i + k, \ldots, i + mk, i + (m + 1)k, i + k - r, \ldots, i)$ leads to $v(i) = v(i + k) = \cdots = v(i+(m-1)k) = v(i+mk) = u(i + k - r) = \cdots = u(i)$ since w is not $\{k, \ell\}$-special. Similarly, in the case where $i > r$, $seq_{k,\ell}(i) = (i, i + k, \ldots, i + (m - 1)k, i + mk, i - r, \ldots, i)$ leads to $v(i) = v(i + k) = \cdot\cdot\cdot = v(i + (m-2)k) = v(i + (m-1)k) = u(i-r) = \cdots = u(i)$.

Case 2. If $i \in Domain(w)$, then we proceed as in the case where $r = 0$. If $i \in Hole(w)$, then put $i = nk+s$ where $0 \leq s < k$. If $s = 0$, then

$$
\begin{aligned}
&w_o(nk) \subset v((n-1)k) \text{ and } w_o(nk) \subset v(nk),\\
&w_o((n+1)k) \subset v(nk) \text{ and } w_o((n+1)k) \subset v((n+1)k),\\
&w_o((n+2)k) \subset v((n+1)k) \text{ and } w_o((n+2)k) \subset v((n+2)k),\\
&\vdots\\
&w_o(mk) \subset v((m-1)k) \text{ and } w_o(mk) \subset v(mk),\\
&w_o((m+1)k) \subset v(mk) \text{ and } w_o((m+1)k) \subset u(k-r),\\
&w_o(k-r) \subset u(k-r) \text{ and } w_o(k-r) \subset v(k-r),\\
&\vdots
\end{aligned}
$$

Since $seq_{k,\ell}(i) = (i, (n+1)k, (n+2)k, \ldots, (m+1)k, k-r, 2k-r, 3k-r, \ldots, k, 2k, \ldots, i-k, i)$, we conclude that $v(i) = v(nk) = \cdots = v((m-1)k) = v(mk) = u(k-r) = \cdots = u(k) = v(k) = v(2k) = \cdots = v((n-1)k) = v(i-k)$.

If $s > 0$, then

$$
\begin{aligned}
&w_o(nk+s) \subset v((n-1)k+s) \text{ and } w_o(nk+s) \subset v(nk+s),\\
&w_o((n+1)k+s) \subset v(nk+s) \text{ and } w_o((n+1)k+s) \subset v((n+1)k+s),\\
&w_o((n+2)k+s) \subset v((n+1)k+s) \text{ and } w_o((n+2)k+s) \subset v((n+2)k+s),\\
&\vdots\\
&w_o((m-1)k+s) \subset v((m-2)k+s) \text{ and } w_o((m-1)k+s) \subset v((m-1)k+s).
\end{aligned}
$$

If $s \leqslant r$, we also get

$$
\begin{aligned}
&w_o(mk+s) \subset v((m-1)k+s) \text{ and } w_o(mk+s) \subset v(mk+s),\\
&w_o((m+1)k+s) \subset v(mk+s) \text{ and } w_o((m+1)k+s) \subset u(k-r+s),\\
&w_o(k-r+s) \subset u(k-r+s) \text{ and } w_o(k-r+s) \subset v(k-r+s),\\
&\vdots
\end{aligned}
$$

Since $seq_{k,\ell}(i) = (i, (n+1)k+s, (n+2)k+s, \ldots, (m+1)k+s, k-r+s, 2k-r+s, \ldots, s, k+s, \ldots, i-k, i)$, we conclude that $v(i) = v(nk+s) = \cdots = v((m-1)k+s) = v(mk+s) = u(k-r+s) = \cdots = u(s) = v(s) = v(k+s) = \cdots = v((n-1)k+s) = v(i-k)$. If $s > r$, we also get

$$
\begin{aligned}
&w_o(mk+s) \subset v((m-1)k+s) \text{ and } w_o(mk+s) \subset u(s-r),\\
&w_o(s-r) \subset u(s-r) \text{ and } w_o(s-r) \subset v(s-r),\\
&\vdots
\end{aligned}
$$

Since $seq_{k,\ell}(i) = (i, (n+1)k+s, (n+2)k+s, \ldots, mk+s, s-r, k+s-r, \ldots, s, k+s, \ldots, i-k, i)$, we conclude that $v(i) = v(nk+s) = \cdots = v((m-1)k+s) = u(s-r) = \cdots = u(s) = v(s) = v(k+s) = \cdots = v((n-1)k+s) = v(i-k)$.

We now give an extension of Theorem 6 under the needed assumption that neither uv nor vu is $\{|u|, |v|\}$-special. Refer to the example stated at the end of the preceding section where $u_o = obb$ and $v_o = abbo$. Here $seq_{3,4}(1) = (1, 4, 7, 3, 6, 2, 5, 1)$ which contains the holes 1,7 of uv while $(uv)(1)(uv)(4)(uv)(7)(uv)(3)(uv)(6)(uv)(2)(uv)(5)(uv)(1) = oaobbbbo$ is not 1-periodic implying that uv is $\{3,4\}$-special.

Theorem 7. *Let u and v be partial words: If uv is not $\{|u|, |v|\}$-special and $uv \uparrow vu$; then there exists a word w such that $u \subset w^m$ and $v \subset w^n$ for some integers m and n.*

Proof. Since $uv \uparrow vu$; $uv \subset z$ and $vu \subset z$ for some word z. Put $z = xy$ where $|x| = |u|$ and $|y| = |v|$. We have $uv \subset xy$ and we will show that $uv \subset yx$. And then, by Lemma 4, we will have $xy = yx$ and thus by Theorem 2, a word w such that $x = w^m$ and $y = w^n$ for some integers m and n. We will then have $u \subset w^m$ and $v \subset w^n$ as desired.

Put $|u| = k$ and $|v| = \ell$. Without loss of generality, we can assume that $k \leq \ell$. The proof is split into three cases that refer to a given position i of $(uv)_o$. Case 1 refers to $1 \leq i \leq k$, Case 2 to $k < i \leq \ell$, and Case 3 to $\ell < i \leq \ell + k$. If $i \in Hole(uv)$, then there is nothing to prove. Henceforth, we assume that $i \in Domain(uv)$. The following diagram helps in proving the inclusion $uv \subset yx$:

$$
\begin{array}{c|ccc|ccc|ccc}
(uv)_o & u_o(1) & \ldots & u_o(k) & v_o(1) & \ldots & v_o(\ell-k) & v_o(\ell-k+1) & \ldots & v_o(\ell)\\
yx & y(1) & \ldots & y(k) & y(k+1) & \ldots & y(\ell) & x(1) & \ldots & x(k)
\end{array}
$$

Put $\ell = mk + r$ where $0 \leq r < k$. We first treat the case where $r = 0$.

Case 1: Since $uv \subset xy$ and $vu \subset xy$, we have $u(i) \subset x(i)$ and $v_o(i) \subset y(i)$ and $v_o(i) \subset x(i)$. If $i \in Domain(v)$, we get $u(i) = x(i) = v(i) = y(i)$ as desired. If $i \notin Domain(v)$, then $u(i)v_o(i)v_o(i + k) \dots v_o(i + (m - 1)k)u(i)$ does not contain consecutive holes and does not contain two holes while not 1-periodic (since uv is not $\{k,\ell\}$-special). Since $uv \subset xy$ and $vu \subset xy$, we have

$$u(i) \subset x(i) \text{ and } v_o(i) \subset x(i),$$

$$v_o(i) \subset y(i) \text{ and } v_o(i+k) \subset y(i),$$

$$v_o(i+k) \subset y(i+k) \text{ and } v_o(i+2k) \subset y(i+k),$$

$$v_o(i+2k) \subset y(i+2k) \text{ and } v_o(i+3k) \subset y(i+2k),$$

$$\vdots$$

$$v_o(i+(m-2)k) \subset y(i+(m-2)k) \text{ and } v_o(i+(m-1)k) \subset y(i+(m-2)k),$$

$$v_o(i+(m-1)k) \subset y(i+(m-1)k) \text{ and } u(i) \subset y(i+(m-1)k).$$

We get $y(i) = y(i + k) = \cdot\cdot\cdot = y(i + (m - 1)k) = u(i)$ as desired (in particular, $v_o(i + k)$ is full).

Case 2: Since $uv \subset xy$ and $vu \subset xy$, we have $v(i - k) \subset y(i - k)$ and $v_o(i) \subset y(i)$ and $v_o(i) \subset y(i - k)$. If $i \in Domain(v)$, then $v(i - k) = y(i - k) = v(i) = y(i)$. Otherwise, put $i = nk + s$ where $0 \leq s < k$. If $s = 0$, then

$$v_o(nk) \subset y(nk) \text{ and } v_o((n+1)k) \subset y(nk),$$

$$v_o((n+1)k) \subset y((n+1)k) \text{ and } v_o((n+2)k) \subset y((n+1)k),$$

$$v_o((n+2)k) \subset y((n+2)k) \text{ and } v_o((n+3)k) \subset y((n+2)k),$$

$$\vdots$$

$$v_o((m-1)k) \subset y((m-1)k) \text{ and } v_o(mk) \subset y((m-1)k),$$

$$v_o(mk) \subset y(mk) \text{ and } u_o(k) \subset y(mk),$$

$$u_o(k) \subset x(k) \text{ and } v_o(k) \subset x(k),$$

$$v_o(k) \subset y(k) \text{ and } v_o(2k) \subset y(k),$$

$$\vdots$$

$$v_o((n-3)k) \subset y((n-3)k) \text{ and } v_o((n-2)k) \subset y((n-3)k),$$

$$v_o((n-2)k) \subset y((n-2)k) \text{ and } v_o((n-1)k) \subset y((n-2)k),$$

$$v_o((n-1)k) \subset y((n-1)k) \text{ and } v_o(nk) \subset y((n-1)k.$$

We conclude that $v(i-k) = v((n-1)k) = y((n-1)k) = y((n-2)k) = \cdot\cdot\cdot = y(k) = x(k) = y(mk) = y((m-1)k) = \cdot\cdot\cdot = y((n+1)k) = y(nk) = y(i)$.

If $s > 0$, then

$$v_o(nk+s) \subset y(nk+s) \text{ and } v_o((n+1)k+s) \subset y(nk+s),$$

$$v_o((n+1)k+s) \subset y((n+1)k+s) \text{ and } v_o((n+2)k+s) \subset y((n+1)k+s),$$

$$v_o((n+2)k+s) \subset y((n+2)k+s) \text{ and } v_o((n+3)k+s) \subset y((n+2)k+s),$$

$$\vdots$$

$$v_o((m-2)k+s) \subset y((m-2)k+s) \text{ and } v_o((m-1)k+s) \subset y((m-2)k+s),$$

$$v_o((m-1)k+s) \subset y((m-1)k+s) \text{ and } u_o(s) \subset y((m-1)k+s),$$

$u_o(s) \subset x(s)$ and $v_o(s) \subset x(s)$,

$v_o(s) \subset y(s)$ and $v_o(k+s) \subset y(s)$,

$v_o(k+s) \subset y(k+s)$ and $v_o(2k+s) \subset y(k+s)$,

$\vdots$

$v_o((n-2)k+s) \subset y((n-2)k+s)$ and $v_o((n-1)k+s) \subset y((n-2)k+s)$,

$v_o((n-1)k+s) \subset y((n-1)k+s)$ and $v_o(nk+s) \subset y((n-1)k+s)$.

We conclude that $v(i-k) = v((n-1)k+s) = y((n-1)k+s) = y((n-2)k+s) = \cdots = y(k+s) = y(s) = x(s) = y((m-1)k+s) = \cdots = y(nk+s) = y(i)$.

Case 3: Put $i = \ell + j$ where $1 \leq j \leq k$. Since $uv \subset xy$ and $vu \subset xy$, we have $u_o(j) \subset x(j)$ and $u_o(j) \subset y(\ell - k + j)$ and $v(\ell - k + j) \subset y(\ell - k + j)$. If $j \in Domain(u)$, we get $v(\ell - k + j) = y(\ell - k + j) = u(j) = x(j)$ as desired. If $j \notin Domain(u)$, then $u_o(j)v_o(j)v_o(j+k) \ldots v_o(j+(m-1)k)u_o(j)$ does not contain consecutive holes and does not contain two holes while not 1-periodic (since uv is not $\{k,\ell\}$-special). Since $uv \subset xy$ and $vu \subset xy$,

we have

$u_o(j) \subset x(j)$ and $v_o(j) \subset x(j)$,

$v_o(j) \subset y(j)$ and $v_o(j+k) \subset y(j)$,

$v_o(j+k) \subset y(j+k)$ and $v_o(j+2k) \subset y(j+k)$,

$v_o(j+2k) \subset y(j+2k)$ and $v_o(j+3k) \subset y(j+2k)$,

$\vdots$

$v_o(j+(m-2)k) \subset y(j+(m-2)k)$ and $v_o(j+(m-1)k) \subset y(j+(m-2)k)$,

$v_o(j+(m-1)k) \subset y(j+(m-1)k)$ and $u_o(j) \subset y(j+(m-1)k)$.

We get $x(j) = y(j) = y(j+k) = \ldots = y(j+(m-1)k) = v(j+(m-1)k) = v(\ell - k + j)$ as desired (in particular, $v_o(j)$ and $v_o(j+(m-1)k)$ are full).

We now treat the case where $r > 0$.

Case 1: If i $\in Domain(v)$, we proceed as in the case where $r = 0$. Otherwise, we consider the cases where $i \leq r$ and $i > r$. If $i \leq r$, then

$u(i) \subset x(i)$ and $v_o(i) \subset x(i)$,

$v_o(i) \subset y(i)$ and $v_o(i+k) \subset y(i)$,

$v_o(i+k) \subset y(i+k)$ and $v_o(i+2k) \subset y(i+k)$,

$v_o(i+2k) \subset y(i+2k)$ and $v_o(i+3k) \subset y(i+2k)$,

$\vdots$

$v_o(i+(m-2)k) \subset y(i+(m-2)k)$ and $v_o(i+(m-1)k) \subset y(i+(m-2)k)$,

$v_o(i+(m-1)k) \subset y(i+(m-1)k)$ and $v_o(i+mk) \subset y(i+(m-1)k)$,

$v_o(i+mk) \subset y(i+mk)$ and $u_o(i+k-r) \subset y(i+mk)$,

$u_o(i+k-r) \subset x(i+k-r)$ and $v_o(i+k-r) \subset x(i+k-r)$,

$\vdots$

Here $u(i)v_o(i)v_o(i+k) \ldots v_o(i+mk)u_o(i+k-r) \ldots u(i)$ does not contain consecutive holes and does not contain two holes while not 1-periodic. If $i > r$, then

$$u(i) \subset x(i) \text{ and } v_o(i) \subset x(i),$$

$$v_o(i) \subset y(i) \text{ and } v_o(i+k) \subset y(i),$$

$$v_o(i+k) \subset y(i+k) \text{ and } v_o(i+2k) \subset y(i+k),$$

$$v_o(i+2k) \subset y(i+2k) \text{ and } v_o(i+3k) \subset y(i+2k),$$

$$\vdots$$

$$v_o(i+(m-2)k) \subset y(i+(m-2)k) \text{ and } v_o(i+(m-1)k) \subset y(i+(m-2)k),$$

$$v_o(i+(m-1)k) \subset y(i+(m-1)k) \text{ and } u_o(i-r) \subset y(i+(m-1)k),$$

$$u_o(i-r) \subset x(i-r) \text{ and } v_o(i-r) \subset x(i-r),$$

$$\vdots$$

Here $u(i)v_o(i)v_o(i+k) \ldots v_o(i+(m-1)k)u_o(i-r) \ldots u(i)$ does not contain consecutive holes and does not contain two holes while not 1-periodic. Applying the above repeatedly, if $i \leq r$, then we get $y(i) = y(i+k) = \cdots = y(i+mk) = x(i+k-r) = y(i+k-r) = y(i+2k-r) = \cdots = u(i)$, and if $i > r$, then we get $y(i) = y(i+k) = \cdots = y(i+(m-1)k) = x(i-r) = y(i-r) = y(i+k-r) = \cdots = u(i)$.

Case 2. If $i \in Domain(v)$, we proceed as in the case where $r = 0$. Otherwise, put $i = nk + s$ where $0 \leq s < k$. If $s = 0$, then

$$v_o(nk) \subset y(nk) \text{ and } v_o((n+1)k) \subset y(nk),$$

$$v_o((n+1)k) \subset y((n+1)k) \text{ and } v_o((n+2)k) \subset y((n+1)k),$$

$$v_o((n+2)k) \subset y((n+2)k) \text{ and } v_o((n+3)k) \subset y((n+2)k),$$

$$\vdots$$

$$v_o((m-1)k) \subset y((m-1)k) \text{ and } v_o(mk) \subset y((m-1)k),$$

$$v_o(mk) \subset y(mk) \text{ and } u_o(k-r) \subset y(mk),$$

$$u_o(k-r) \subset x(k-r) \text{ and } v_o(k-r) \subset x(k-r),$$

$$v_o(k-r) \subset y(k-r) \text{ and } v_o(2k-r) \subset y(k-r),$$

$$\vdots$$

$$v_o((n-1)k) \subset y((n-1)k) \text{ and } v_o(nk) \subset y((n-1)k).$$

We conclude that $y(i) = y(nk) = \cdots = y((m-1)k) = y(mk) = x(k-r) = y(k-r) = y(2k-r) = \cdots = y((n-1)k) = v((n-1)k) = v(i-k)$.

If $s > 0$, then

$$v_o(nk+s) \subset y(nk+s) \text{ and } v_o((n+1)k+s) \subset y(nk+s),$$

$$v_o((n+1)k+s) \subset y((n+1)k+s) \text{ and } v_o((n+2)k+s) \subset y((n+1)k+s),$$

$$v_o((n+2)k+s) \subset y((n+2)k+s) \text{ and } v_o((n+3)k+s) \subset y((n+2)k+s),$$

$$\vdots$$

$$v_o((m-2)k+s) \subset y((m-2)k+s) \text{ and } v_o((m-1)k+s) \subset y((m-2)k+s).$$

If $s \leqslant r$, we also get

$$v_o((m-1)k+s) \subset y((m-1)k+s) \text{ and } v_o(mk+s) \subset y((m-1)k+s),$$
$$v_o(mk+s) \subset y(mk+s) \text{ and } u_o(k-r+s) \subset y(mk+s),$$
$$\vdots$$

We conclude that $y(i) = y(nk + s) = \cdots = y((m - 1)k + s) = y(mk + s) = x(k - r + s) = y(k - r + s) = y(2k - r + s) = \cdots = y((n - 1)k + s) = v((n - 1)k + s) = v(i - k)$. If $s > r$, we also get

$$v_o((m-1)k+s) \subset y((m-1)k+s) \text{ and } u_o(s-r) \subset y((m-1)k+s),$$
$$\vdots$$

We conclude that $y(i) = y(nk + s) = \cdots = y((m - 1)k + s) = x(s - r) = y(s - r) = y(s - r + k) = \cdots = y((n - 1)k + s) = v((n - 1)k + s) = v(i - k)$.

Case 3: By putting $i = \ell + j$ where $1 \leq j \leq k$, we can see that this case is symmetric to Case 1.

5. Conjugacy

This section is concerned with conjugacy of partial words. In particular, we extend Theorem 3 to partial words.

Definition 4. Two partial words u and v are called *conjugate* if there exist partial words x and y such that $u \subset xy$ and $v \subset yx$.

Conjugacy on words is known to be an equivalence relation. Conjugacy on partial words is trivially reflexive and symmetric. However, it is not transitive as the following example shows. Consider, $u_o = aobabboa$, $v_o = obooaaoo$, and $w_o = baobbbaa$. By putting $x_o = aob$ and $y_o = abboa$, we get $u \subset xy$ and $v \subset yx$ showing that u and v are conjugate. Similarly, by putting $x'_o = obbbaa$ and $y'_o = ba$, we get $v \subset x'y'$ and $w \subset y'x'$ showing that v and w are conjugate. But we can see that u and w are not conjugate.

We now answer a question that was raised in Ref. [2]. Consider nonempty partial words u, v, z such that $uz \uparrow zv$. Under mild assumptions, do partial words x, y exist such that $u \subset xy$, $v \subset yx$, and $z \subset x(yx)^n$ for some $n \geq 0$? As noted in Ref. [2], this is false even if uzv has only one hole. Consider for example $u = a$, $v = b$, and $z_o = obb$ (here $uz \vee zv = abbb$ is not 1 -periodic). The next theorem shows that under the assumption that $uz \vee zv$ is $|u|$-periodic, such partial words x, y exist.

Theorem 8. *Let u and v be nonempty partial words.*

1: *If u and v are conjugate; then there exists a partial word z such that $uz \uparrow zv$: Moreover; in this case there exist partial words x; y such that $u \subset xy$; $v \subset yx$; and $z \subset x(yx)^n$ for some $n \geq 0$.*

2: *If there exists a partial word z such that $uz \uparrow zv$ and $uz \vee zv$ is $|u|$ -periodic; then there exist partial words x; y such that $u \subset xy$; $v \subset yx$; and $z \subset x(yx)^n$ for some $n \geq 0$.*

Proof. To prove Statement 1, let u, v be nonempty partial words. Suppose that u and v are conjugate and let x, y be partial words such that $u \subset xy$ and $v \subset yx$. Then $ux \subset xyx$ and $xv \subset xyx$ and so for $z = x$ we have $uz \uparrow zv$.

To prove Statement 2, assume that $uz \uparrow zv$ and that $uz \vee zv$ is $|u|$-periodic. Let m be such that $m|u| > |z| \geq (m - 1)|u|$. Put $u = x_1y_1$ and $v = y_2x_2$ where $|x_1| = |x_2|= |z| - (m - 1)|u|$ and $|y_1| = |y_2|$ (here $|u| = |v|$). Put $z = x'_1y'_1x'_2y'_1 \ldots x'_{m-1}y'_{m-1}x'_m$ where $|x'_1| = \cdots = |x'_{m-1}| = |x'_m| = |x_1| = |x_2|$ and $|y'_1| = \cdots = |y'_{m-1}| = |y_1| = |y_2|$. Since $uz \uparrow zv$, we get

$$\uparrow \begin{matrix} x_1 & y_1 & x'_1 & y'_1 & x'_2 & y'_2 & \dots & x'_{m-2} & y'_{m-2} & x'_{m-1} & y'_{m-1} & x'_m \\ x'_1 & y'_1 & x'_2 & y'_2 & x'_3 & y'_3 & \dots & x'_{m-1} & y'_{m-1} & x'_m & y_2 & x_2 \end{matrix}$$

Let $1 \leqslant i \leqslant |x_1|$. The partial word

$$\vee \begin{matrix} (x_1)_o(i) & (x'_1)_o(i) & (x'_2)_o(i) & \dots & (x'_{m-2})_o(i) & (x'_{m-1})_o(i) & (x'_m)_o(i) \\ (x'_1)_o(i) & (x'_2)_o(i) & (x'_3)_o(i) & \dots & (x'_{m-1})_o(i) & (x'_m)_o(i) & (x_2)_o(i) \end{matrix}$$

is 1-periodic, say with letter a_i in $A \cup \{o\}$. Now, let $1 \leqslant i \leqslant |y_1|$. The partial word

$$\vee \begin{matrix} (y_1)_o(i) & (y'_1)_o(i) & (y'_2)_o(i) & \dots & (y'_{m-2})_o(i) & (y'_{m-1})_o(i) \\ (y'_1)_o(i) & (y'_2)_o(i) & (y'_3)_o(i) & \dots & (y'_{m-1})_o(i) & (y_2)_o(i) \end{matrix}$$

is 1 -periodic, say with letter b_i in $A \cup \{o\}$. Create $x_o = a_1 a_2 \dots a_{|x_1|}$ and $y_o = b_1 b_2 \dots b_{|y_1|}$. We conclude that $x_1 \subset x$ and $x_2 \subset x$ and $y_1 \subset y$ and $y_2 \subset y$ and thus, $u = x_1 y_1 \subset xy$ and $v = y_2 x_2 \subset yx$. Moreover, $z = x'_1 y'_1 x'_2 y'_2 \dots x'_{m-1} y'_{m-1} x'_m \subset x(yx)^{m-1}$ and the result follows.

Note that if $uz \uparrow zv$ and z is full, then $uz \vee zv$ is $|u|$-periodic (in the proof of Theorem 8, we get $x'_1 = x'_2 = \cdots = x'_{m-1} = x'_m = x$ and $y'_1 = y'_2 = \cdots = y'_{m-1} = y$ in such a case).

References

[1] D.G. Arqués, C.J. Michel, A possible code in the genetic code, Lecture Notes in Computer Science, Vol. 900, Springer, Berlin, 1995, pp. 640–651.
[2] J. Berstel, L. Boasson, Partial words and a theorem of Fine and Wilf, Theoret. Comput. Sci. 218 (1999) 135–141.
[3] F. Blanchet-Sadri, R.A. Hegstrom, Partial words and a theorem of Fine and Wilf revisited, Theoret. Comput. Sci., to appear.
[4] C. Choffrut, J. Karhumäki, Combinatorics of words, in: G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, Vol. 1 (Chapter 6), Springer, Berlin, 1997, pp. 329–438.
[5] A. Colosimo, A. de Luca, Special factors in biological strings, preprint 97/42, Dipartimento di Matematica, Universiti di Roma "La Sapienza".
[6] K. Culik II, T. Harju, Splicing semigroups of dominoes and DNA, Discrete Appl. Math. 31 (1991) 261–277.
[7] A. de Luca, On the combinatorics of finite words, Theoret. Comput. Sci. 218 (1999) 13–39.
[8] T. Head, Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors, Bull. Math. Biol. 49 (1987) 737–759.
[9] T. Head, G. Pă un, D. Pixton, Language theory and molecular genetics, in: G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, Vol. 2, Springer, Berlin, 1997, pp. 295–360 (Chapter 7).
[10] M.A. Jimenez-Montano, On the syntactic structure of protein sequences and the concept of grammar complexity, Bull. Math. Biol. 46 (1984) 641–659.
[11] M. Lothaire, Combinatorics on Words, Cambridge University Press, Cambridge, UK, 1997.