Periodicity of Degenerate Strings

Pengfei Wang[1]
Joint with Estéban Gabory[2], Eric Rivals[1], Michelle Sweering[2] and Hilde Verbeek[2]

[1]LIRMM, Univ. Montpellier, CNRS, Montpellier, France.

[2]CWI, Amsterdam, The Netherlands.

Prague Stringlogy Conference
Aug.2023

Co-financed by the Connecting Europe Facility of the European Union

Multiple Sequence Alignment (MSA)

```
C A - - A G C G C T A A - - - T T
C - - - A G C C G A A G T - - A T
C A - C A A G T C A A G - - - - T
```

Local Gapless Alignment (LGA)

```
C A A G C G C T A A T T
C A G C C G A A G T A T
C A C A A G T C A A G T
```

Degenerate string from LGA

Local Gapless Alignment

C	A	A	G	C	G	C	T	A	A	T	T
C	A	G	C	C	G	A	A	G	T	A	T
C	A	C	A	A	G	T	C	A	A	G	T

Degenerate String

$$\{\mathtt{C}\}\cdot\{\mathtt{A}\}\cdot\left\{\begin{matrix}\mathtt{A}\\\mathtt{G}\\\mathtt{C}\end{matrix}\right\}\cdot\left\{\begin{matrix}\mathtt{G}\\\mathtt{C}\\\mathtt{A}\end{matrix}\right\}\cdot\left\{\begin{matrix}\mathtt{C}\\\mathtt{C}\\\mathtt{A}\end{matrix}\right\}\cdot\{\mathtt{G}\}\cdot\left\{\begin{matrix}\mathtt{C}\\\mathtt{A}\\\mathtt{T}\end{matrix}\right\}\cdot\left\{\begin{matrix}\mathtt{T}\\\mathtt{A}\\\mathtt{C}\end{matrix}\right\}\cdot\left\{\begin{matrix}\mathtt{A}\\\mathtt{G}\\\mathtt{A}\end{matrix}\right\}\cdot\left\{\begin{matrix}\mathtt{A}\\\mathtt{T}\\\mathtt{A}\end{matrix}\right\}\cdot\left\{\begin{matrix}\mathtt{T}\\\mathtt{A}\\\mathtt{G}\end{matrix}\right\}\cdot\{\mathtt{T}\}$$

Period of classical string

Example: String $u = aabbaa$ has periods 0, 4 and 5

	0	1	2	3	4	5	6	7	8	9	10
u	a	a	b	b	a	a	-	-	-	-	-
u	a	a	b	b	a	a	-	-	-	-	-
	-	a	a	b	b	a	a	-	-	-	-
	-	-	a	a	b	b	a	a	-	-	-
	-	-	-	a	a	b	b	a	a	-	-
	-	-	-	-	a	a	b	b	a	a	-
	-	-	-	-	-	a	a	b	b	a	a

- Period set = $\{0, 4, 5\}$ (and corresponding autocorrelation 100011)

Known results about periods of classical strings

- Characterization of period sets [JCTA, Guibas and Odlyzko,1981].
  - (Auto)correlations
  - Lower bound on the number of period sets of a given length
  - Populations (i.e. how many strings share a given period set)

- The combinatorics of periods [JCTA, Rivals and Rahmann,2003].

- The convergence of the number of period sets for strings of given length [ICALP , Rivals, Sweering and Wang, 2022].

How to define periodicity of DS?

0	1	2	3	4	5	6	7	8
$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	-	-	-
$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	-	-	-
-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	-	
-	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	
-	-	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	
-	-	-	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$

Table of Contents

Table of Contents

Degenerate string

Σ: a finite alphabet of size σ and n an integer.

Definition (Degenerate string)

$\hat{w} = \hat{w}[0 \ldots n-1] \in (\mathcal{P}(\Sigma) \setminus \emptyset)^n$ is a string of length n over $\mathcal{P}(\Sigma) \setminus \emptyset$.

- $\hat{w}[i]$ is called an undetermined symbol
- We say two degenerate strings $\hat{x}$ and $\hat{y}$ of length n match, if for all $i \in \{0, \ldots, n-1\}$ the intersection $\hat{x}[i] \cap \hat{y}[i]$ is non-empty.
- A hollow string $\hat{w}$ is a degenerate string such that $\hat{w}[i] = \emptyset$ for at least one $i \in \{0, \ldots, n-1\}$

Language associated to a DS

Definition (Language)

The language of a degenerate string $\hat{w}$ of length n is set of all classical strings that match with it:

$$\mathcal{L}(\hat{w}) = \{w \in \Sigma^n \mid \forall i \in \{0, \dots, n-1\} \quad w[i] \in \hat{w}[i]\}\,.$$

Example 1

Let $\hat{w} = \begin{Bmatrix} \mathtt{a} \\ \mathtt{b} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \{\mathtt{c}\} \cdot \begin{Bmatrix} \mathtt{a} \\ \mathtt{c} \end{Bmatrix}$.

$$\begin{aligned}\mathcal{L}(\hat{w}) = \{&abbca, abbcc, abcca, abccc, acbca, acbcc, accca, acccc,\\ &bbbca, bbbcc, bbcca, bbccc, bcbca, bcbcc, bccca, bcccc\}\end{aligned}$$

Strong periodicity

Definition (Strong period)

A degenerate string $\hat{w}$ has <u>strong period</u> p if there exists a string $w \in \mathcal{L}(\hat{w})$ with period p.

- To model one specific string, whose letters are not precisely known.
- $P^s(\hat{w})$ denotes the set of strong periods of $\hat{w}$.

Example 2

Let $\hat{w} = \begin{Bmatrix} \mathtt{a} \\ \mathtt{b} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \{\mathtt{c}\} \cdot \begin{Bmatrix} \mathtt{a} \\ \mathtt{c} \end{Bmatrix}$.

$accca \in \mathcal{L}(\hat{w})$, $accca$ has periods $0, 4$. $P^s(\hat{w}) = \{0, 4\}$.

Weak periodicity

Definition (Weak period)

A degenerate string $\hat{w} = \hat{w}[0 \dots n-1]$ has <u>weak period</u> $p \in \{0, \dots, n-1\}$ if and only if $\hat{w}[0 \dots n-p-1]$ matches $\hat{w}[p \dots n-1]$

- Weak periods can model variations in a set of related strings.
- $P^w(\hat{w})$ denotes the set of weak periods of $\hat{w}$.
- Running example: $\hat{w} = \begin{Bmatrix} \mathtt{a} \\ \mathtt{b} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \{\mathtt{c}\} \cdot \begin{Bmatrix} \mathtt{a} \\ \mathtt{c} \end{Bmatrix}$.

Example

	0	1	2	3	4	5	6	7	8
$\hat{w}$	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	-	-	-
$\hat{w}$	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	-	-	-
	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	-	
	-	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	-	
	-	-	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$	
	-	-	-	-	$\{a, b\}$	$\{b, c\}$	$\{b, c\}$	$\{c\}$	$\{a, c\}$

Weak periodicity

Definition (Weak period)

A degenerate string $\hat{w} = \hat{w}[0 \dots n-1]$ has <u>weak period</u> $p \in \{0, 1, \dots, n-1\}$ if and only if $\hat{w}[0 \dots n-p-1]$ matches $\hat{w}[p \dots n-1]$

- Weak periods can model variations in a set of related strings.
- $P^w(\hat{w})$ denotes the set of weak periods of $\hat{w}$.

Example 3

$$\hat{w} = \begin{Bmatrix} \mathtt{a} \\ \mathtt{b} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \begin{Bmatrix} \mathtt{b} \\ \mathtt{c} \end{Bmatrix} \cdot \{\mathtt{c}\} \cdot \begin{Bmatrix} \mathtt{a} \\ \mathtt{c} \end{Bmatrix}.$$

Then $P^w(\hat{w}) = \{0, 1, 2, 4\}$,

Definition (Medium period)

A degenerate string $\hat{w} = \hat{w}[0 n-1]$ has medium period $p \in \{0, 1, \ldots, n-1\}$ if and only if for any $0 \leq i, j \leq n-1$ such that $i \equiv j \pmod{p}$ we have $\hat{w}[i] \cap \hat{w}[j] \neq \emptyset$.

- $P^m(\hat{w})$ denotes the set of medium periods of $\hat{w}$.
- $P^w(\hat{w}) = \{0, 1, 2, 4\} \Longrightarrow P^m(\hat{w}) = \{0, 2, 4\}$
- For each degenerate string, $P^s \subseteq P^m \subseteq P^w$.
- The sets of weak, medium and strong period sets of <u>all</u> degenerate strings of length n are denoted by Ω_n^w, Ω_n^m, and Ω_n^s.

Table of Contents

Theorem (Characterization of PS)

Let $P^s \subseteq P^m \subseteq P^w \subseteq \{0,\dots,n-1\}$. Then P^w, P^m and P^s are respectively the weak, medium, and strong period sets of some non-hollow degenerate string $\hat{w}$ of length n if and only if

1. (I) $0 \in P^s$,
2. (II) $\forall p \in P^w$ with $p \geq n/2 \implies p \in P^s$,
3. (III) $p \in P^m$ iff $\forall k \in \mathbb{N}$ with $kp \in \{0,\dots,n-1\} \implies kp \in P^w$
4. (IV) $p \in P^s$ iff $\forall k \in \mathbb{N}$ with $kp \in \{0,\dots,n-1\} \implies kp \in P^s$.

Proof of characterization of periodicities, part 1

Given a non-hollow degenerate string $\hat{w}$.

Proof: Necessity ($\Rightarrow$).

1. Every classical string has period 0, so take any string from $\mathcal{L}(\hat{w})$.
2. $\hat{w} = \hat{w}[0 \ldots n-p-1]\hat{w}[n-p \ldots p-1]\hat{w}[p \ldots n-1]$ with $\hat{w}[0 \ldots n-p-1] \cap \hat{w}[p \ldots n-1] \neq \emptyset$
   $\Longrightarrow \exists$ string $w = w[0 \ldots n-p-1]w[n-p \ldots p-1]w[p \ldots n-1]$:
   - $w[0 \ldots n-p-1] = w[p \ldots n-1] \in \hat{w}[0 \ldots n-p-1] \cap \hat{w}[p \ldots n-1]$
   - and $w[n-p \ldots p-1] \in \hat{w}[n-p \ldots p-1]$.
3. This is the definition of medium periods.
4. If a classical string $w \in \mathcal{L}(\hat{w})$ has period p, then it also has period kp for $k \in \mathbb{N}$, giving $\hat{w}$ strong period kp as well.

$\square$

Proof of characterization of periodicities, part 2

Proof: Sufficiency ($\Leftarrow$).

Sufficiency ($\Leftarrow$): Proven by constructing $\hat{w}$ given the period sets:

$$\hat{w}[i] = \begin{cases} \{\mathtt{a},\mathtt{b}\} & \text{if } i = 0 \\ \{\mathtt{a},\mathtt{c}\} & \text{if } i \in P^s \setminus \{0\} \\ \{\mathtt{b},\mathtt{c}\} & \text{if } i \in P^w \setminus P^s \\ \{\mathtt{c}\} & \text{otherwise} \end{cases}$$

- $\hat{w}[p] \neq \mathtt{c} \iff p \in (P^s \setminus \{0\}) \cup (P^w \setminus P^s) \iff p \in P^w \setminus \{0\}$.
- For every $p \in P^m$, every multiple is a weak period with symbol $\{\mathtt{b},\mathtt{c}\}$.
- For $p \in P^s \setminus \{0\}$, there exists a string $w \in \mathcal{L}(\hat{w})$ such that:

$$w[i] = \begin{cases} \mathtt{a} & \text{if } p \mid i \\ \mathtt{c} & \text{otherwise.} \end{cases}$$

$\square$

Proof of characterization of periodicities, part 3

Proof.

...hence $\hat{w}$ has strong period p. If $p \notin P^s$, then either

- $p \notin P^w$, or
- $p \in P^w \setminus P^s \iff \exists k \in \mathbb{N}$ such that $kp \in [\frac{n}{2}, n-1], kp \in P^s \iff$

$$\hat{w}[0] \cap \hat{w}[p] \cap \hat{w}[kp] = \{\mathtt{a}, \mathtt{b}\} \cap \{\mathtt{b}, \mathtt{c}\} \cap \{\mathtt{a}, \mathtt{c}\} = \emptyset.$$

□

Corollary

Because the constructed string $\hat{w}$ has alphabet size 3, period sets in general are independent of their alphabet Σ when $|\Sigma| \geq 3$.

Table of Contents

Primitive sets and relation to period sets

- By the characterization theorem, we notice that P^m, P^s are multiplicative subsets of $[0, \ldots, n-1]$ and hence $\Omega_n^m = \Omega_n^s$.
- For $P \subseteq [0, n-1]$ we write $\langle P \rangle = \{kp \in [0, n-1] \mid p \in P,\ k \in \mathbb{N}\}$. We say that P is a generator of $\langle P \rangle$.

Definition (Primitive set)

A set P of integers is primitive if it does not contain a pair $i \neq j$ such that i divides j or j divides i.

Lemma

For any $P \subseteq [0, n-1]$, there exists a unique P_{prim} such that P_{prim} is a primitive set and the minimum generator of P. Hence, $P \mapsto P_{prim}$ is a one-to-one mapping between primitive subsets and multiplicative subsets of $[0, n-1]$.

Define $Q(n)$ as the number of primitive sets with greatest element at most n.

Counting the number of period sets

Theorem (Counting and convergence)

1. *The number of weak period sets is:* $|\Omega_n^w| = 2^{n-1}$.
2. *For medium/strong period sets: for any* $\varepsilon > 0$*, we have*

$$|\Omega_n^m| = |\Omega_n^s| = \alpha^{n\left(1+O\left(\exp\left((-1+\varepsilon)\sqrt{\log n \log\log n}\right)\right)\right)}.$$

Proof.

1. The set of weak period sets "corresponds to" the power set of $\{1,\ldots,n-1\}$.
2. We applied the knowledge from number theory and $|\Omega_n^m| = |\Omega_n^s| = Q(n-1)$.

□

Conclusion

- We provided three notions of periodicity (weak, medium, and strong) for degenerate strings.
- We characterized the period sets for each, exhibiting necessary and sufficient conditions.
- We counted the number of period sets for strings of length n for each type and studied its convergence using recent results from number theory.
- We investigated the structure of the families of period sets. i.e. Proved that all the types of period sets form lattices under set intersection and set union.
- We computed how many degenerate strings share a given period set (i.e. its population) using graph theory.

Future work

1. Investigate the combinatorial part of periodicity of languages,.i.e, any set of strings.

2. Study the algorithmic aspect of periodicity of languages,e.g.,Improve the complexity to determine period sets of degenerate string.

3. Apply our notions of periodicity of degenerate string to pattern matching, or string-graph matching problem.

Funding and acknowledgements

*This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No 872539 and No 956229, from the Netherlands Organisation for Scientific Research (NWO) through Gravitation-grant NETWORKS-024.002.003 and from the Constance van Eeden PhD Fellowship. Moreover, we would like to thank Solon P. Pissis for his helpful advice and suggestions.

Thank you for your attention!

Questions?

References I

Guibas, L. and Odlyzko, A. (1981).
Periods in strings.
Journal of Combinatorial Theory, Series A, 30:19–43.

Rivals, E. and Rahmann, S. (2003).
Combinatorics of periods in strings.
Journal of Combinatorial Theory, Series A, 104(1):95–113.

Rivals, E., Sweering, M., and Wang, P. (2022).
Convergence of the number of period sets in strings.