Combinatorics of period sets

Eric Rivals[1] and Michelle Sweering[2] and **Pengfei Wang**[1]

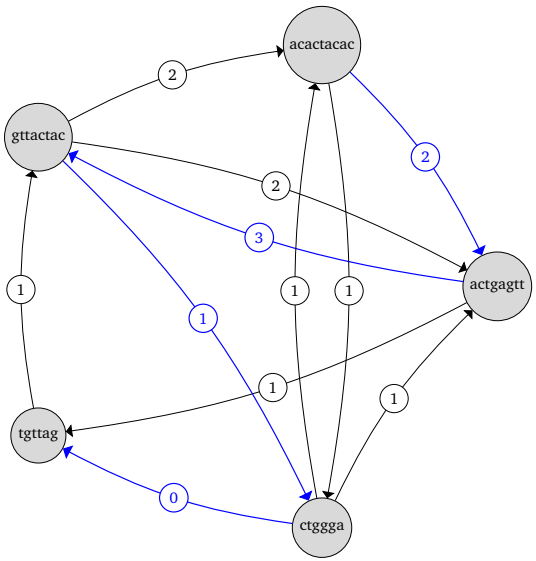
[1]LIRMM, Univ.Montpellier, CNRS, Montpellier, France.

[2]CWI, Amsterdam, The Netherlands.

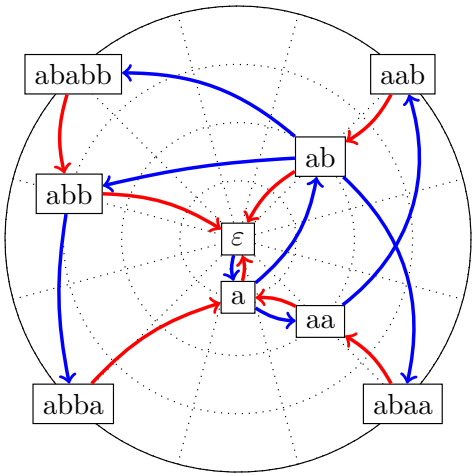
Séquences en Bioinformatique, Informatique et Mathématiques
Nov.2022

Data structures in computational pan-genomics

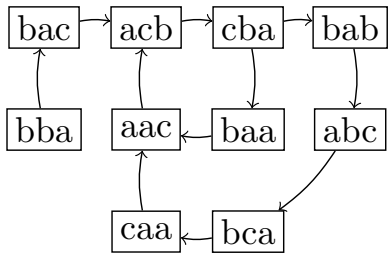
Overlap Graph (OG)



Hierarchical OG

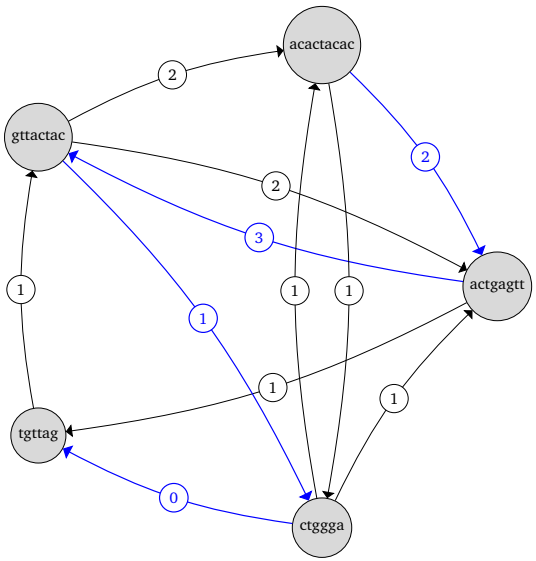


De Bruijn Graph

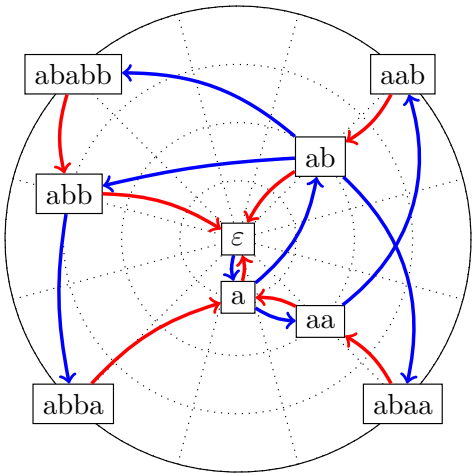


Data structures in computational pan-genomics

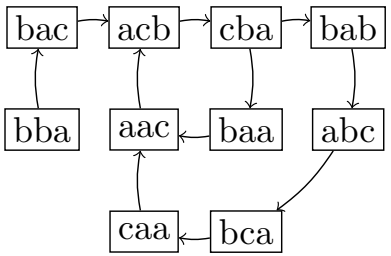
Overlap Graph (OG)



Hierarchical OG



De Bruijn Graph



Question:

- How many combinations of overlaps do exist?

Overlap of one string

Ex: The overlap of $u = aabbaa$, "combination of overlaps" = $\{0, 4, 5\}$

	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

Overlap of one string

Ex: The overlap of $u = aabbaa$, "period set" = $\{0, 4, 5\}$

	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

Question:

- ~~How many combinations of overlaps do exist?~~
- How many distinct period sets do exist for strings of a given length?

Table of Contents

Applications

1. Absence of probability
   The absence probability of a word in random text depends on the autocorrelation polynomial
   - Vocabulary statistics:
     number of missing words in a random text
   - Test procedure for Random Number Generators

2. Sequence or text comparison using k-mers
   - number of common words between random texts
     - for natural language texts.
     - for biological sequences
   - sequence complexity

Table of Contents

Period and Border

Σ: a finite alphabet of size σ and n an integer.
Σ^n: the set of strings of length n over Σ.

Definition Border

A border of a string u is any substring that is both a prefix and a suffix of u.

Definition Period

Let $u = u[0 \dots n-1]$ a string of Σ^n, p an integer such that $p < n$. p is a period of u iff:

$$\forall\, 0 \leq i \leq n-p-1 \ : u[i] = u[i+p]$$

Denote $P(u)$ the period set of u. $P(u) \subseteq [0, n-1]$. The smallest non-trivial period is called basic period.

Examples

- Ex 1: $n := 11$, $u :=$abracadabra has periods $0, 7, 10$

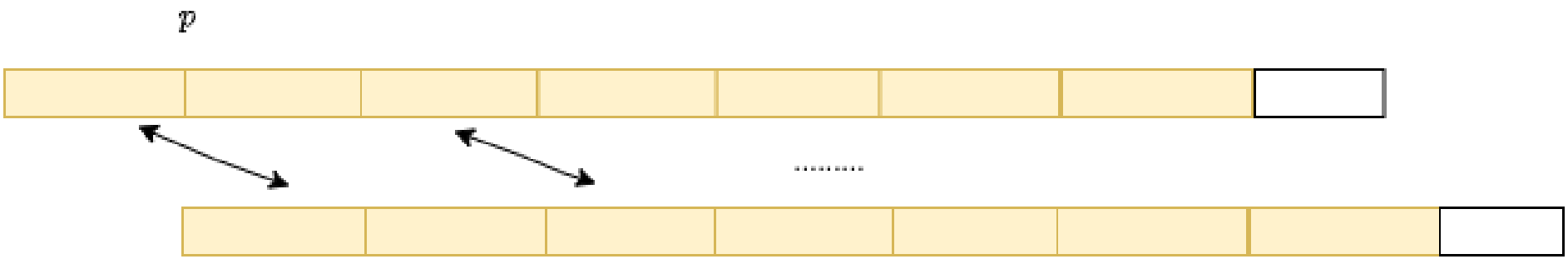
i :	0	1	2	3	4	5	6	7	8	9	10
u :	a	b	r	a	c	a	d	a	b	r	a
	a	b	r	a	-	-	-	a	b	r	a
	a	-	-	-	-	-	-	-	-	-	a

- Ex 2: $n := 9$, $u :=$ abababababa has period set $\{0, 2, 4, 6, 8\}$

Properties of periods

Property 1

Multiple of a period: If p is a period of u, then kp are also periods of u, with $kp \leq |u|$.



- Ex: $n := 9$, $u :=$ abababababa has period set $\{0, 2, 4, 6, 8\}$
- Corollary:
  $u = vvv \ldots vv'$, where $|v| = p$, v' is prefix of v

Properties of periods

Theorem 1– Fine and Wilf Theorem [1]

If p, q are periods of u and $p + q \leq |u| + gcd(p, q)$, then $gcd(p, q)$ is period of u.

Corollary:

If p is the basic period of u, q is another period, then either $p|q$ or $q > n - p$.

Forward propagation rule

Property 2

Forward propagation rule: String $u = u[0 \ldots n-1]$ of length n satisfies the forward propagation rule(FPR) if, whenever $p, q \in P(u)$, with $p < q$, we have:

$$t \in P(u) \text{ if } t = q + i(q-p), i = 0, 1, 2, \ldots, \text{and } t \in [p, n).$$

Ex 1: $n := 12$, $u :=$ abababbababa. Period set P: = $\{0,7,9,11\}$. FPR:

$$11 = 9 + 1(9-7)$$

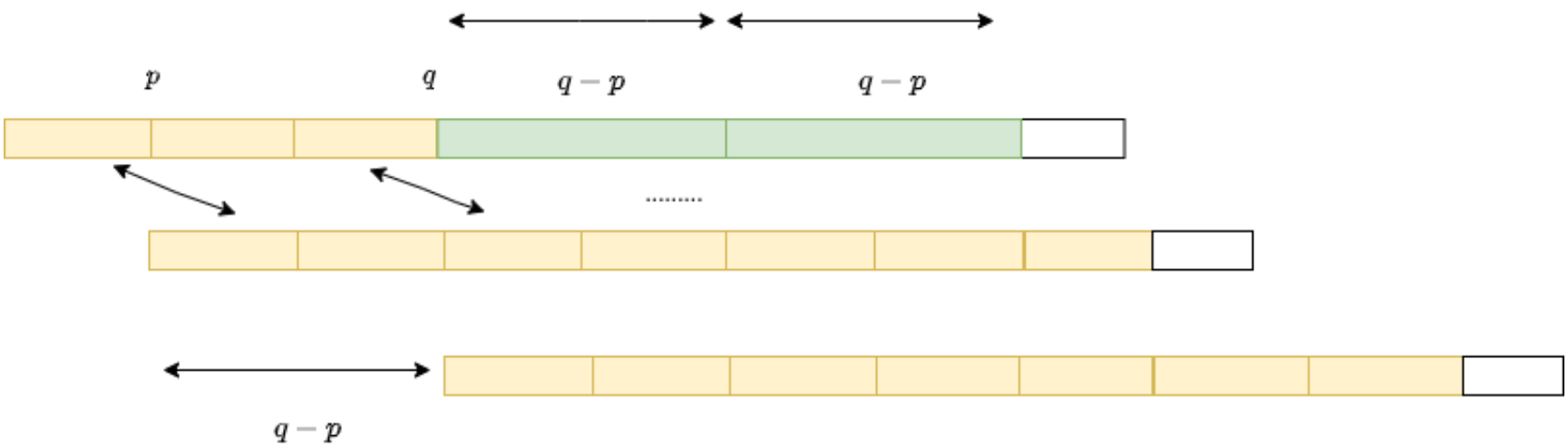
Ex 2: $n := 9$, $u :=$ abababababa. Period set P: = $\{0,2,4,6,8\}$. FPR:

$$4 = 2 + 1(2-0);$$
$$6 = 2 + 2(2-0);$$
$$8 = 2 + 3(2-0).$$

Forward propagation rule

String $u = u[0 \ldots n-1]$ of length n satisfies the forward propagation rule(FPR) if, whenever $p, q \in P(u)$, with $p < q$, we have:

$$t \in P(u) \text{ if } t = q + i(q-p), i = 0, 1, 2, \ldots, \text{and } t \in [p, n).$$



Period set

- Denote $P(u)$ the period set of u. $P(u) \subseteq [0, n-1]$.
- Let $\Gamma(n)$ denote the set of all period sets for strings of length n.

$$\Gamma(n) := \{P(u) | \exists u \in \Sigma^n\}$$

- when $n = 4$. $\Gamma(4)$ = $\{\{0,1,2,3\},\{0,2\},\{0,3\},\{0\}\}$
- Not all subsets from $[0, n-1]$ are period sets.
- Different strings may have the same period set. For example $u = abbbbabb, v = aabaaaab$, they both have period set: $\{0, 5\}$

Table of Contents

Irreducible Period Set

Irreducible Period Set (IPS). From the period set remove redundant periods that are induced by others using forward propagation rule.

Theorem 2– [3]

The mapping between period sets and Irreducible Period Sets(IPSs) is one-to-one.

Denote:
Λ_n = the number of IPSs of strings with length n

Corollary:

$$|\Gamma_n| = |\Lambda_n|$$

.

Examples of IPS

- Ex 1: $n := 11$, $u :=$abracadabra

$$\begin{aligned} \mathrm{P(u)}: &= \{0, 7, 10\}, \\ \mathrm{IPS}: &= \{0, 7, 10\}. \end{aligned}$$

- Ex 2: $n := 12$, $u :=$ abababbababa

$$\begin{aligned} \mathrm{P(u)}: &= \{0, 7, 9, 11\}, \\ \mathrm{IPS}: &= \{0, 7, 9\}. \\ \mathrm{FPR}: 11 &= 9 + 1(9 - 7) \end{aligned}$$

- Ex 3: $n := 9$, $u :=$ ababababa

$$\begin{aligned} \mathrm{P(u)}: &= \{0, 2, 4, 6, 8\}, \\ \mathrm{IPS}: &= \{0, 2\}. \end{aligned}$$

Known bounds on κ_n (nb period sets)

Notation: κ_n is the cardinality of

Theorem 3– Lower & upper bounds [2]

$$\frac{1}{2\ln 2} + o(1) \leq \frac{\ln \kappa_n}{\ln^2(n)} \leq \frac{1}{2\ln(3/2)} + o(1)$$

Conjecture:

$$\lim_{n\to\infty} \frac{\ln \kappa_n}{\ln^2(n)} = \frac{1}{2\ln 2}$$

Lower bound [3]

$$\frac{\ln \kappa_n}{\ln^2(n)} \geq \frac{1}{2\ln 2}\left(1 - \frac{\ln\ln n}{\ln n}\right)^2 + \frac{0.4139}{\ln n} - \frac{1.47123 \ln\ln n}{\ln n} + o\left(\frac{1}{\ln^2(n)}\right)$$

Improved upper bound on κ_n

Theorem 4– Improved upper bound for κ_n

$$\frac{\ln(\kappa_n)}{\ln^2(n)} \leq \frac{1}{2\ln(2)} + \frac{3}{2\ln(2)\ln(n)} \quad \forall n \in \mathbb{N}_{\geq 2}.$$

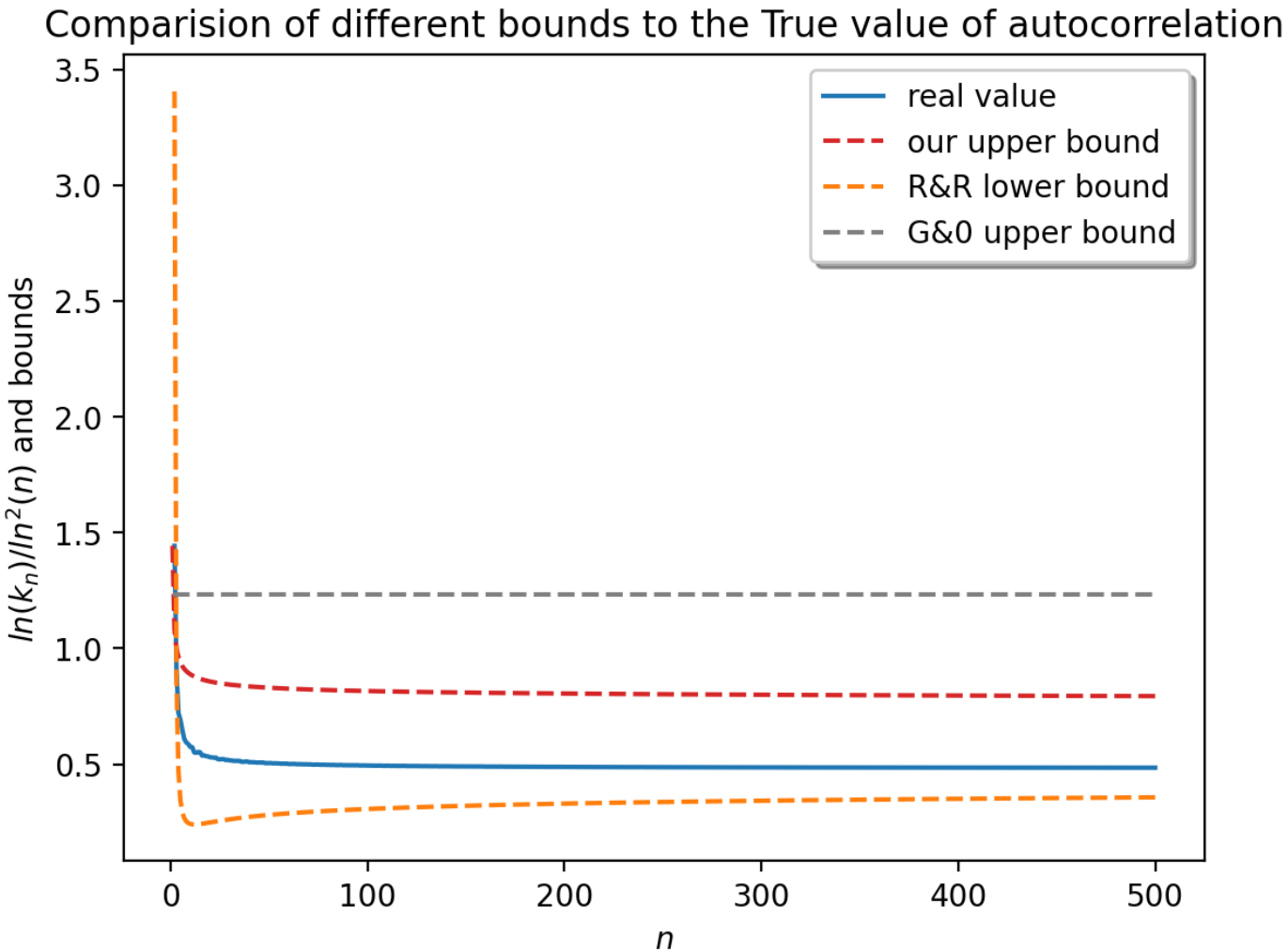
Asymptotically

$$\frac{\ln \kappa_n}{\ln^2(n)} \leq \frac{1}{2\ln 2} + o(1)$$

Proof idea:

- calculate the number of IPSs instead of the whole period sets.
- investigate the distance between deducible and irreducible periods.

Bounds and true values of $\ln \kappa_n / \ln^2(n)$



True values from A005434 of the Encyclopedia of Integer Sequences

Proof of improved upper bound (1)

Denote $IPS = \{0 = a_0 < \ldots < a_i < \ldots < a_k < n, \quad k \in \mathbb{Z}^+, k \in [1, n)\}$.

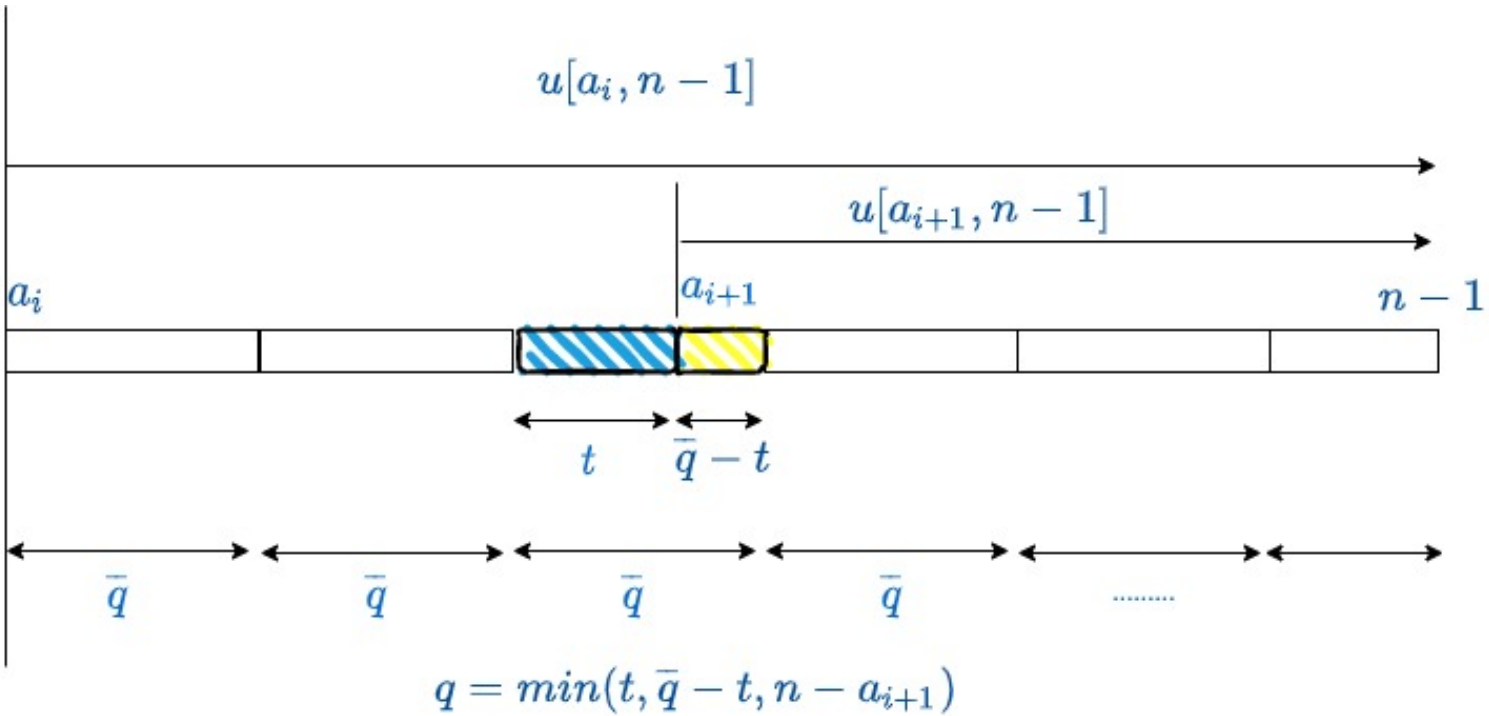
Lemma 1

For all $0 \leq i \leq k$, the suffix $u[a_i \ldots n-1]$ has a period q such that:

1. $1 \leq q \leq n/2^i$, and
2. q can be deduced from $a_0, \ldots, a_i$ using the forward propagation rule.

- Idea: See figure next page.

Figure of Lemma 1



Hints:

- suffix $u[a_i, n-1]$ has the similar period structure with u.
- If $x + y \leq a$, then $\min(x, y) \leq \frac{a}{2}$

Proof of improved upper bound (1)

Lemma 1

For all $0 \leq i \leq k$, the suffix $u[a_i \ldots n-1]$ has a period q such that:

1. $1 \leq q \leq n/2^i$, and
2. q can be deduced from $a_0, \ldots, a_i$ using the forward propagation rule.

Lemma 2

The number of irreducible periods, k satisfies: $k \leq \log_2(n)$.

- Idea: By Lemma 1.

Proof of improved upper bound (2)

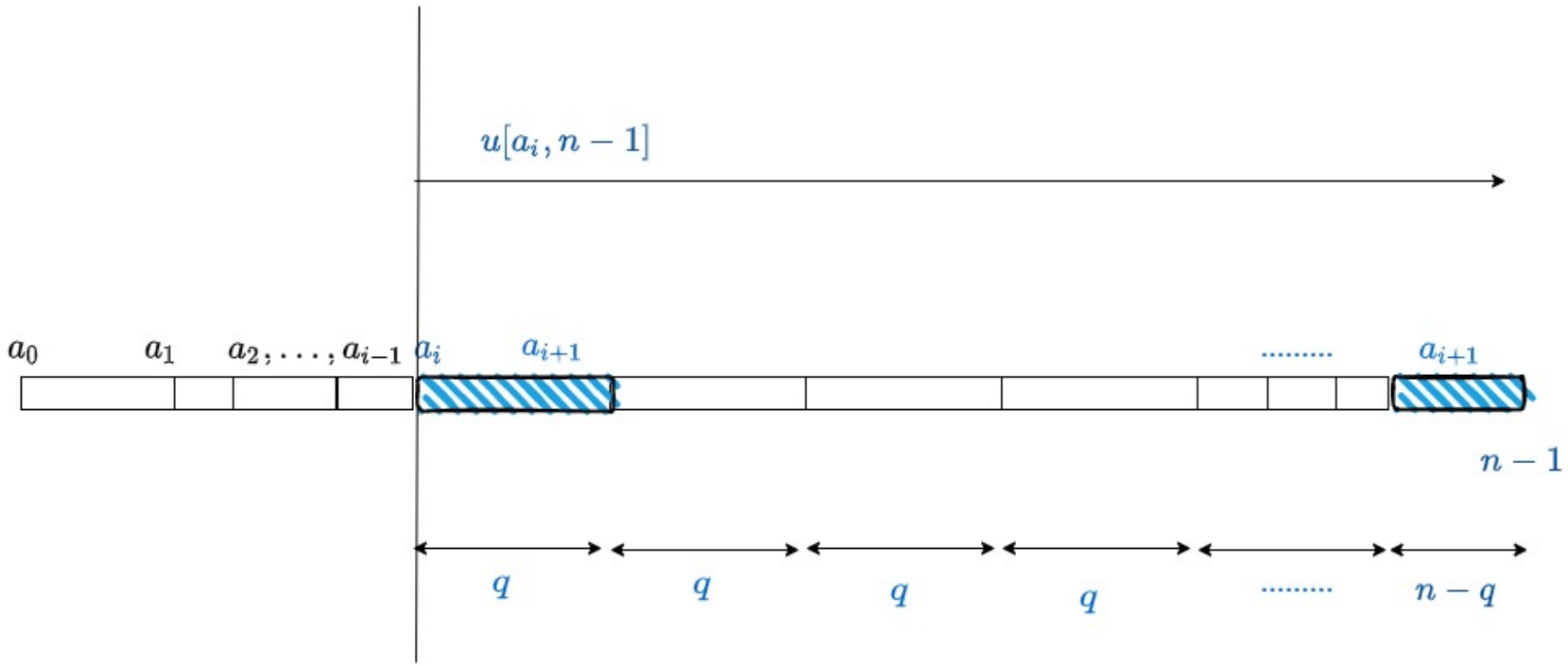
Lemma 3

Let $0 \leq i < k-1$. There are at most

- $\left(\frac{n}{2^{i-1}} - 1\right)$ possibilities for a_{i+1} given $a_0, \ldots, a_i$.

- Idea: See figure next page.

Figure of Lemma 3



- Hint: Use corollary of Fine and Wilf Theorem

Proof of improved upper bound (2)

Lemma 3

Let $0 \leq i < k-1$. There are at most

- $\left(\frac{n}{2^{i-1}} - 1\right)$ possibilities for a_{i+1} given $a_0, \ldots, a_i$.

We know that $\kappa_n = |\Gamma_n| = |\Lambda_n|$

Upper bound on number of IPS

$$\kappa_n = |\Lambda_n| \leq \sum_{k=1}^{\log_2(n)} \prod_{i=0}^{k-1} \left(\frac{n}{2^{i-1}} - 1\right)$$

Table of Contents

Asymptotic convergence of κ_n

Combine new upper bound and the best known lower bound

Rivals & Rahmann. 2003

$$\frac{\ln \kappa_n}{\ln^2(n)} \geq \frac{1}{2\ln 2}(1 - \frac{\ln\ln n}{\ln n})^2 + \frac{0.4139}{\ln n} - \frac{1.47123 \ln\ln n}{\ln n} + o(\frac{1}{\ln^2(n)})$$

$$\Longrightarrow \frac{\ln \kappa_n}{\ln^2(n)} \geq \frac{1}{2\ln(2)} - O(\frac{\ln\ln n}{\ln n})$$

Asymptotic convergence on κ_n

$$\frac{1}{2\ln(2)} - O(\frac{\ln\ln n}{\ln n}) \leq \frac{\ln \kappa_n}{\ln^2(n)} \leq \frac{1}{2\ln(2)} + o(1)$$

$$\Longrightarrow \frac{\ln \kappa_n}{\ln^2(n)} \longrightarrow \frac{1}{2\ln 2}, \text{ when } n \to \infty$$

Table of Contents

Correlation: a binary encode of overlaps between two different strings.

Definition Correlation

For every pair of strings $(u, v) \in \Sigma^n \times \Sigma^m$, the correlation of u over v is the vector $t \in \{0,1\}^n$ such that

$$t[k] = \begin{cases} 1 & \text{if } u[i] = v[j] \text{ for all } i \in \{0,\ldots,n-1\}, j \in \{0,\ldots,m-1\} \\ & \text{with } i = j + k, \\ 0 & \text{otherwise} \end{cases}$$

for all $k \in \{0,\ldots,n-1\}$.

Examples: Correlations

Ex: The correlation of $u = aabbaa$ over $v = baabaa$ is $t = 000100$.

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

Let

- Δ_n be the set of all correlations between two strings of length n
- δ_n be the cardinality of Δ_n

Results about Δ_n and δ_n

Lemma 4 (Characterization of Δ_n)

The set of correlations of length n is of the form

$$\Delta_n = \left\{ 0^{(n-j)} s_j \mid s_j \in \Gamma_j,\ j \in [0, n] \right\},$$

where Γ_j is the set of autocorrelations of length j.

- Idea: Consider example above.

Lemma 5 (Relation between δ_n and κ_n)

Let κ_n be the number of autocorrelations of length n and δ_n the number of correlations between two strings of length n. Then

$$\delta_n = \sum_{j=0}^{n} \kappa_j.$$

- Idea: By Lemma 4.

The same convergence behavior for δ_n and κ_n

Theorem 5– Asymptotic convergence of δ_n

$$\frac{\ln \delta_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \text{as} \quad n \to \infty.$$

Proof idea:

- Notice:

$$\kappa_n \leq \delta_n \leq n \cdot \kappa_n$$

- Substitute the above inequality to the convergence result of κ_n.

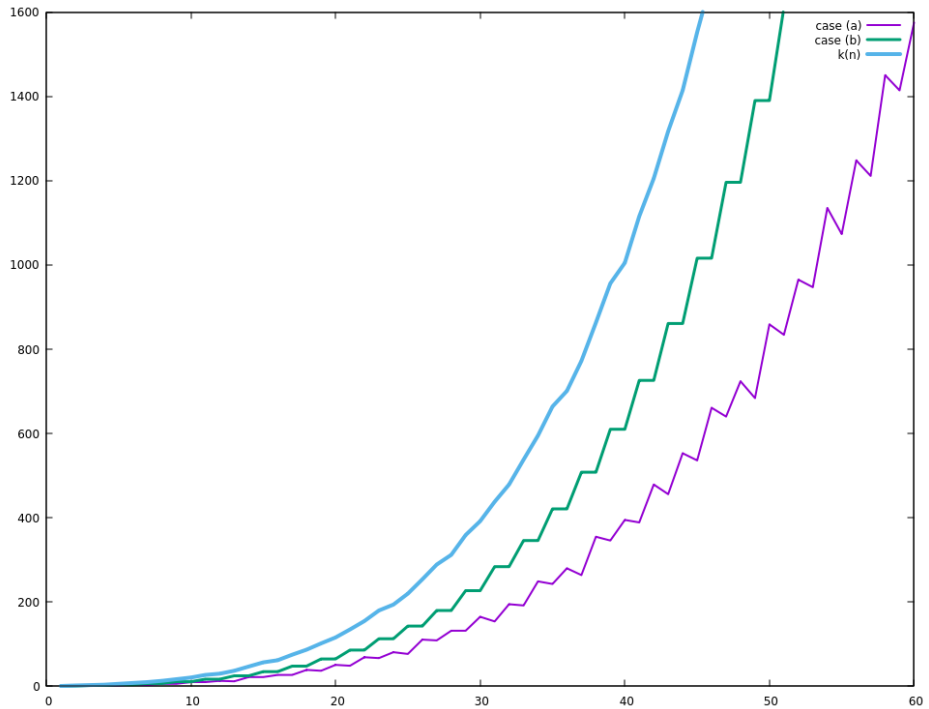
Conclusion

- Improve upper bound on the number of period sets κ_n:
- Resolve the conjecture from 1981
- Show the convergence on the number of correlations,

Scan the code to read our paper! [4]
10.48550/arXiv.2209.08926

Perspective

The Plot of κ_n:



- Case(a), basic period $p \leq \frac{n}{2}$,
- Case(b), basic period $p > \frac{n}{2}$
- Denote κ_a,κ_b as the number of period sets for case(a) and case(b) respectively. What's the exact formula for them? Can we give good bounds for them?

Funding and acknowledgements

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement N°956229 and the Netherlands Organisation for Scientific Research (NWO) through Gravitation-grant NETWORKS-024.002.003.

Thanks for your attention!

Questions?

References


Nathan J Fine and Herbert S Wilf.
Uniqueness theorems for periodic functions.
Proceedings of the American Mathematical Society, 16(1):109–114, 1965.


Leonidas J. Guibas and Andrew M. Odlyzko.
Periods in strings.
Journal of Combinatorial Theory, Series. A, 30:19–42, 1981.


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


Eric Rivals, Michelle Sweering, and Pengfei Wang.
Convergence of the number of period sets in strings.
CoRR, abs/2209.08926, 2022.
arXiv:2209.08926, doi:10.48550/arXiv.2209.08926.