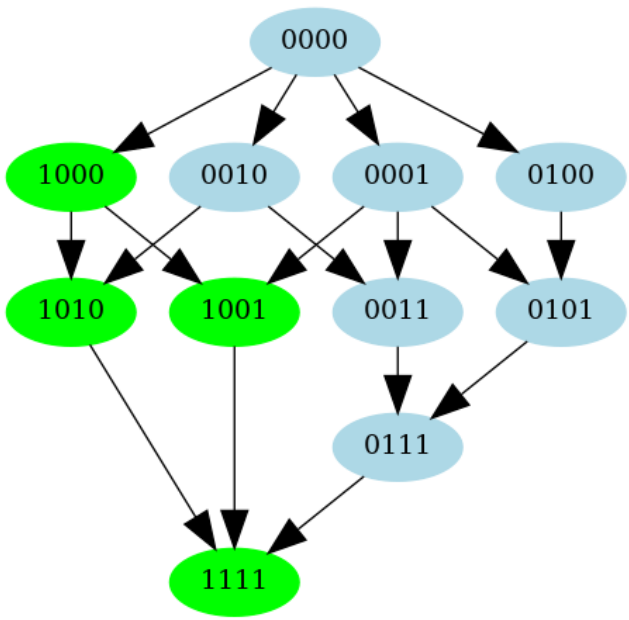
Counting overlapping pairs of strings

Eric Rivals, Pengfei Wang

LIRMM, CNRS Univ Montpellier



Periods, borders, and period set

Word $u = \texttt{abracadabra}$ of length $n = 11$

pos.	0	1	2	3	4	5	6	7	8	9	10	
u	a	b	r	a	c	a	d	a	b	r	a	border lg
	a	b	r	a	-	-	-	a	b	r	a	4
	a	-	-	-	-	-	-	-	-	-	a	1

- abracadabra has two non trivial borders: abra and a
- Its set of periods is $P(u) = \{0, 7, 10\}$
- The autocorrelation of u: $c(u, u) = 10000001001$
  it encodes periods in binary vector of length n
- Computing the period set or the autocorrelation of u takes $\Theta(n)$ time.

Suffix-prefix overlaps in a pair of words

A pair of words (u, v) of length 6 with $u = \texttt{aabbab}$ and $v = \texttt{babbba}$

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

- (u, v) has two borders `bab` and `b`
- Note: `b` is a border of `bab`
- Correlation $c(u, v) = \texttt{000101}$
  encodes starting positions in u of suffix-prefix overlaps

Many pairs of words share the same correlation (1)

Consider previous correlation $c(u, v) = 000101$
then the following pairs also share the same correlation:

1. case due to a permutation of letters $c(\texttt{bbaaba}, \texttt{abaaab}) = 000101$

1. cases with a different word structure:
   - $c(u, \texttt{babbaa}) = 000101$

- $c(\texttt{aaabab}, v) = c(\texttt{aaabab}, \texttt{babbba}) = 000101$

Combinatorial question

Mapping from pair of words to correlation

Mapping c from $\Sigma^n \times \Sigma^n$ to $\{0,1\}^n$
to an ordered pair (u, v) it associates its (unique) correlation

- c is surjective

Question: population size of a correlation

How many pairs of words of length n share the same correlation?

Note:

- It depends on both the alphabet size and string length.
- Population size of autocorrelation / period sets are known [3, 10]

Population sizes for Δ_4 and alphabet size $\sigma = 2, 3, 4$.

Correlation	Population sizes for alphabet size σ		
	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$
0000	74	3678	45132
0001	82	1866	15108
0010	30	480	3060
0011	24	216	960
0100	16	162	768
0101	8	54	192
0111	6	24	60
1000	6	48	180
1001	6	24	60
1010	2	6	12
1111	2	3	4

Related works: Vocabulary statistics

1. Question: how many k-mers are missing in random texts?
   - the average depends on autocorrelation of k-mers
   - the variance depends on correlations of pairs of k-mers

2. Applications in bioinformatics:

   comparing sequences through their k-mer spectrum [1, 9]

3. or in test of Pseudo-Random Number Generators [6, 5]

Related works: code design

1. Counting and generating the set bifix-free words of length n
   - Algorithm by Nielsen 1972 [7] and
   - analysis of the expected search time for a fixed pattern in random data [8]

2. Counting bordered and unbordered pairs of words [2]
   - Gabric extends the work of Nielsen to pairs of words
   - two open questions [2]

Remarks:

- Nielsen studies words with autocorrelation $10\ldots0$
- Gabric studies pairs with correlation $00\ldots0$ and pairs whose correlation is not $00\ldots0$

Two open questions from Gabric 2022

Question 1 How many pairs of length-n words have a largest border of fixed length i?

Question 2 What is the expected length of the longest border of a pair of words?

Gamma and Delta: alphabet independence

Definition

$$\Gamma_n := \{t \in \{0,1\}^n : \exists u \in \Sigma^n : c(u,u) = t\}$$

$$\Delta_n := \{t \in \{0,1\}^n : \exists (u,v) \in \Sigma^n \times \Sigma^n : c(u,v) = t\}$$

Lemma

Both Γ_n and Δ_n are independent of the alphabet provided the alphabet Σ has at least 2 symbols (exclude the case of an alphabet with a single symbol – $\sigma = 1$)

- Original proof for Γ_n by Guibas and Odlyzko in 1981 [3]
- Simplified proof by Halava et al. 2000 [4]
- Proof for Δ_n in [11]

Characterization of Delta – link with Gamma

Lemma

The set of correlations of length n is of the form

$$\Delta_n = \left\{ 0^{(n-j)} s, \;\; \textit{with } s \in \Gamma_j \;\; \textit{and} \;\; j \in [0, \ldots, n] \right\}.$$

Partition of Delta according to longest overlap:

$$\Delta_n = \bigcup_{j=0}^{n} \{ 0^{n-j} s \,|\, s \in \Gamma_j \} = \bigcup_{j=0}^{n} \left(0^{n-j} . \Gamma_j \right).$$

Correlation of (u, v) and autocorrelation of vu

Idea

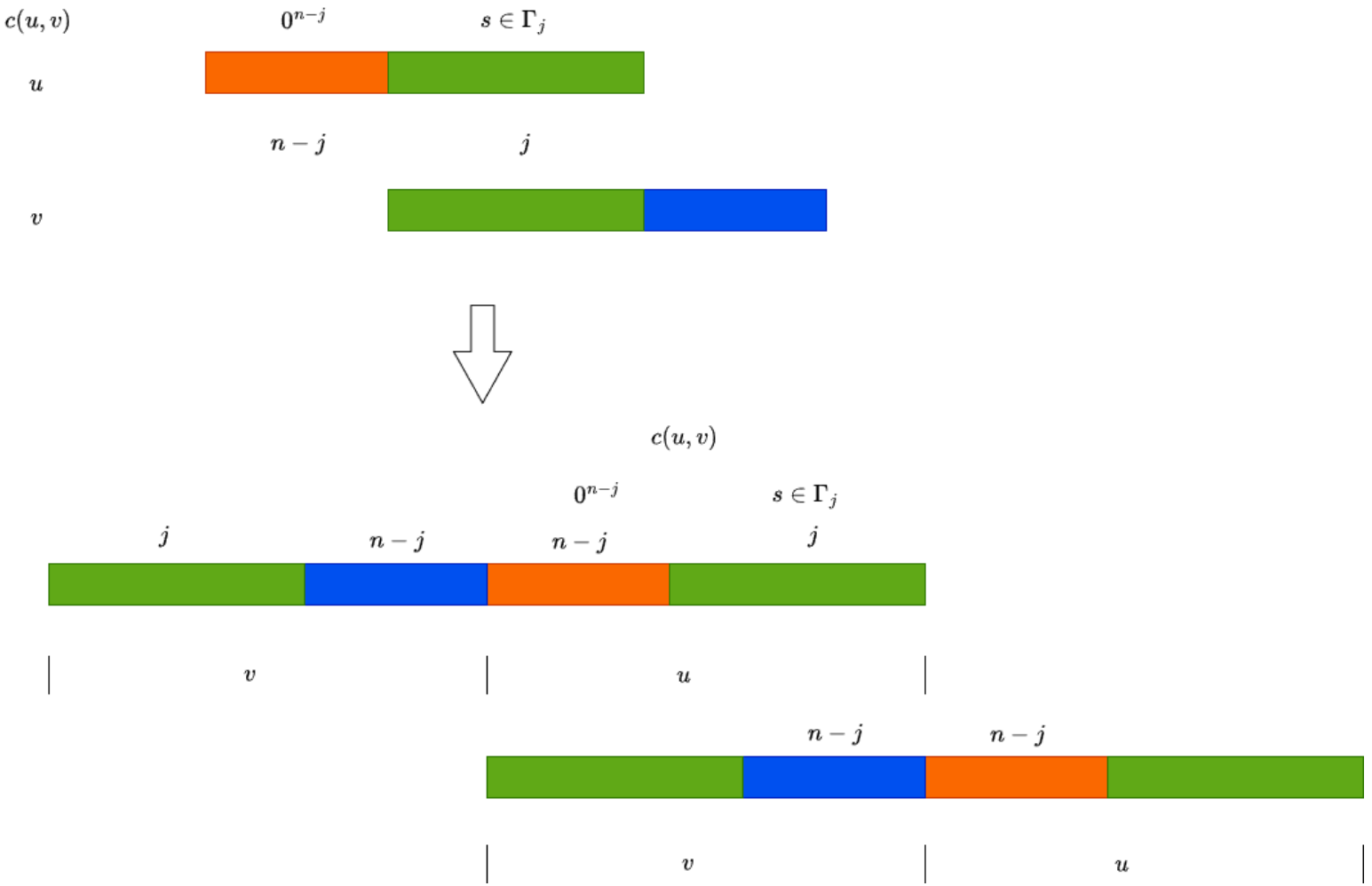
Connections between overlaps of pair (u, v) and
self-overlaps of the word made from the concatenation of v and u.

Relation between $c(u, v)$ and $c(vu, vu)$.

Theorem

Let u and v be two words of length n, and $t = c(u, v)$.
Then $c(u, v)$ is suffix of the length n of the autocorrelation of vu.

Proof idea



Population size of a correlation

Two recursive formulas for computing the population size of an autocorrelation, i.e., $pop(s)$ for $s \in \Gamma_n$ [3, 10]

Theorem

Let $\lambda, n \in \mathbb{N}$ satisfying $0 \leq \lambda < n$. Let $t \in \Delta_n$. By the characterization of Δ_n, there exists an integer j and $s \in \Gamma_j$ s.t. $t := 0^{n-j}s$ and j be the length of s. Then, the population size of t satisfies the recurrence

$$pop(t) = \sum_{\lambda=\lceil \frac{2n-j}{2} \rceil}^{n-1} pop(s_{(2n-\lambda)}) \cdot s[j+2\lambda-2n] + pop(s_{2n}).$$

where s_k denotes the binary vector $10^{k-j-1}s$.

Theorem

Let $L_{[i..k]}$ be the number of pairs of strings of length n that have a longest border within the fixed length range $[i..k]$ where $i \leq k \in \{0, \ldots, n-1\}$. Then:

$$L_{[i..k]} = \sum_{t \in (\cup_{j=i}^{k} (0^{n-j}.\Gamma_j))} pop(t)$$

Recurrence for open question #1 of [2]

Direct application of this theorem for the case where $i = k$.

$$L_j = \sum_{t \in (0^{n-j}.\Gamma_j)} pop(t)$$

Expectation of the longest border of a pair of length-n words

Define the random variable X:
the length of the longest border of a pair of strings of length n.
Then, the expectation of X is

$$E(X) = \sum_{j=0}^{n-1} j \cdot Pr(X = j) = \sum_{j=1}^{n-1} j \cdot \frac{L_j}{\sigma^{2n}} = \sum_{j=1}^{n-1} j \cdot \frac{\sum_{t \in (0^{n-j}.\Gamma_j)} pop(t)}{\sigma^{2n}}.$$

Theorem

$E(X)$ diverges when n tends to infinity.

Conclusion and perspectives

Take home messages

1. Characterization of correlations of length n
2. Formula to count the population of any correlation
3. Solution for two open questions from Gabric's paper in 2022
4. In our approach, we account for the combination of all overlaps between two strings.
5. Not shown: Bounds on the population ratio $pop(t)/\sigma^{2n}$

Open questions and future works

- Conjecture: the population ratio converges towards the limiting value of $pop(s_n)/\sigma^{2n}$
- Investigate the distribution of the length of the longest border of a pair of words

Bibliography


Stefan Burkhardt, Andreas Crauser, Paolo Ferragina, Hans-Peter Lenhof, Eric Rivals, and Martin Vingron.
q-gram based database searching using a suffix array (QUASAR).
In *Proceedings of the third annual international conference on Computational molecular biology - RECOMB '99*. ACM Press, 1999.
URL: http://dx.doi.org/10.1145/299432.299460, doi:10.1145/299432.299460.


Daniel Gabric.
Mutual borders and overlaps.
IEEE Transactions on Information Theory, 68(10):6888–6893, 2022.
doi:10.1109/TIT.2022.3167935.


Leo J. Guibas and Andrew M. Odlyzko.
Periods in strings.
J. of Combinatorial Theory series A, 30:19–42, 1981.
doi:10.1016/0097-3165(81)90038-8.


Vesa Halava, Tero Harju, and Lucian Ilie.
Periods and binary words.
J. Comb. Theory, Ser. A, 89(2):298–303, 2000.
URL: https://doi.org/10.1006/jcta.1999.3014, doi:10.1006/JCTA.1999.3014.


Paul Leopardi.
Testing the Tests: Using Random Number Generators to Improve Empirical Tests.
In Pierre L' Ecuyer and Art B. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 501–512. Springer Berlin Heidelberg, 2009.
DOI: 10.1007/978-3-642-04107-5_32.
URL: http://link.springer.com/chapter/10.1007/978-3-642-04107-5_32.


George Marsaglia and Arif Zaman.
Monkey tests for random number generators.
Computers and Mathematics with Applications, 26(9):1–10, 1993.


Peter Tolstrup Nielsen.
A note on bifix-free sequences (corresp.).
IEEE Transactions on Information Theory, 19(5):704–706, September 1973.
URL: http://dx.doi.org/10.1109/TIT.1973.1055065, doi:10.1109/tit.1973.1055065.


Peter Tolstrup Nielsen.

On the expected duration of a search for a fixed pattern in random data (corresp.).
IEEE Transactions on Information Theory, 19(5):702–704, September 1973.
URL: http://dx.doi.org/10.1109/TIT.1973.1055064, doi:10.1109/tit.1973.1055064.


Sven Rahmann and Eric Rivals.
On the distribution of the number of missing words in random texts.
Combinatorics, Probability and Computing, 12(01), Jan 2003.
URL: http://dx.doi.org/10.1017/S0963548302005473, doi:10.1017/s0963548302005473.


Eric Rivals and Sven Rahmann.
Combinatorics of periods in strings.
Journal of Combinatorial Theory, Series A, 104(1):95–113, Oct 2003.
URL: http://dx.doi.org/10.1016/S0097-3165(03)00123-7, doi:10.1016/s0097-3165(03)00123-7.


Eric Rivals, Michelle Sweering, and Pengfei Wang.
Convergence of the Number of Period Sets in Strings.
In Kousha Etessami, Uriel Feige, and Gabriele Puppis, editors, *50th International Colloquium on Automata, Languages, and Programming (ICALP 2023)*, volume 261 of *Leibniz International Proceedings in*

Informatics (LIPIcs), pages 100:1–100:14, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
URL: https://drops.dagstuhl.de/opus/volltexte/2023/18152, doi:10.4230/LIPIcs.ICALP.2023.100.