Counting overlapping pairs of strings

Eric Rivals, Pengfei Wang

▶ To cite this version:

Eric Rivals, Pengfei Wang. Counting overlapping pairs of strings. Workshop SeqBIM 2025, IRISA, Univ Rennes, Nov 2024, Rennes, France. lirmm-04831119

HAL Id: lirmm-04831119

https://hal-lirmm.ccsd.cnrs.fr/lirmm-04831119v1

Submitted on 11 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

SeqBIM

Abstract

Counting overlapping pairs of strings

Pengfei Wang, Eric Rivals

LIRMM, Université Montpellier, CNRS, Montpellier, France
Corresponding author: rivals@lirmm.fr

Abstract

A word u overlaps a word v if a suffix of u equals a prefix of v. The shared suffix-prefix is called a *border* for the ordered pair of words (u,v) (note that other authors call this a *right border*, see [1]). If (u,v) has no border it is said *unbordered*. These notions generalize to pairs of words, the well studied notions of border, bordered and unbordered words, that were originally defined for single words. Example: Consider the binary alphabet $\{a,b\}$ and the following three words denoted by u,v,w: `abaaa`, `aaabb`, and `abbbb`. The pairs (u,v) and (v,w) both have a longest border of length 3, but (u,v) has 3 distinct non empty borders `aaa`, `aa`, and `a`, while (v,w) has only one `abb`. The pairs (v,u) and (w,v) have no borders, which illustrates the asymmetry of this notion.

Overlapping and unbordered words are central in many applications: bioinformatics, pattern matching, code design, or word statistics, among others.

Other authors have proposed to encode the starting position of such overlaps in a binary vector called a *correlation* [2]. In our example, the correlation of the pair (u,v) is `00111`, while that of (v,w) is `00100`. For any word z, the correlation of (z,z) is called the *autocorrelation* of z. Clearly, multiple pairs can have the same correlation, and hence there are less correlations of length n than pairs of words of length n. Recently, Gabric [1] gave three recurrences to count bordered, mutually bordered, mutually unbordered pairs of words of length n over a k-ary alphabet [1]. In his conclusion, he raised challenging open questions: 1/ count the number of pairs having the longest border of length i (with i satisfying $0 < i < n$), and 2/ what is the expected length of the longest border of a pair of words? Here, we exhibit two solutions to compute the population size of any correlation, that is the number of pairs of words having the same correlation. For this, we exploit two recurrences to compute the population size of autocorrelations [2, 3]. With this in hand, we derive a formula for the abovementioned open question 1/ and show that the expected length of the longest border of words of length n asymptotically diverges (open question 2/). Besides this, we provide bounds for the asymptotic of the population ratio of any correlation, which extend the result known for autocorrelations [2].

An article presenting these results is available on ArXiV online [4].

References

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

[2] Leonidas J. Guibas and Andrew M. Odlyzko. Periods in strings. *Journal of Combinatorial Theory, Series. A*, 30:19–42, 1981. doi:10.1016/0097-3165(81)90038-8.

[3] Eric Rivals and Sven Rahmann. Combinatorics of Periods in Strings. In F. Orejas, P. Spirakis, and J. van Leuween, editors, *ICALP 2001, Proc. of the 28th International Colloquium on Automata, Languages and Programming, (ICALP), Crete, Greece, July 8-12, 2001*, volume 2076 of *Lecture Notes in Computer Science*, pages 615–626. Springer Verlag, 2001. doi:10.1007/3-540-48224-5_51.

[4] Eric Rivals and Pengfei Wang. Counting overlapping pairs of strings. *ArXiv*, abs/2405.09393, 2024. URL: https://doi.org/10.48550/arXiv.2405.09393.