Combinatorics of Periods in Strings
Eric Rivals

To cite this version:

HAL Id: lirmm-00193520
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00193520
Submitted on 3 Dec 2007

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.
Combinatorics of Periods in Strings

Eric Rivals
L.I.R.M.M., U.M.R. 5506
CNRS and Univ. Montpellier II
Montpellier, France
rivals@lirmm.fr
http://www.lirmm.fr/~rivals

Here, we consider a central notion of word combinatorics and string algorithmics: the periods of a string. A period is an offset (i.e., a shift) at which a word can overlap itself. A word may have several periods, which we call its set of periods, and distinct words of the same length may share the same period set. When denoted by a binary string, a period set is called the autocorrelation of a word. In the early 80’s, Guibas and Odlyzko provided the first investigation of the structure of period sets \([3, 2]\) and characterized them. Considering the set \(\Gamma_n\) of all period sets of strings of length \(n\) over a finite alphabet, they showed that \(\Gamma_n\) is independent of the alphabet (provided the cardinality of \(\Sigma\geq 2\)).

Pursuing the goal of finding an enumeration algorithm for \(\Gamma_n\), we study further the properties of \(\Gamma_n\) and exhibit the redundancy in period sets. It enables us to introduce the notion of an irreducible period set and to elucidate the structure of both \(\Gamma_n\) and the set of all irreducible period sets, denoted \(\Lambda_n\). We then propose the first efficient enumeration algorithm for \(\Gamma_n\). We also exhibit a relation between the number of binary partitions of \(n\) and the number of distinct period sets (i.e., the cardinality of \(\Gamma_n\)). It allows us to improve upon the previously known asymptotic lower bounds on the cardinality of \(\Gamma_n\) [3]. Additionally, from these results we derive a new recurrence to compute the population of a period set, as well as an algorithm to sample uniformly irreducible and classical period sets.

All above mentioned results were published in [6, 7]. Related entries of the Encyclopedia of Integer Sequences [8] are A018819 and A000123. This study has been extended to partial words [1]. The enumeration algorithm found applications for the computation of several statistics about the vocabulary of strings, like the number of missing words of length \(n\) in a text or the number of common words between two texts [4, 5].

Acknowledgements: Collaboration with S. Rahmann, now at the University of Bielefeld.
References


