An average number of suffix-prefixes

E. Furletova, M. Regnier, M. Roytberg

Definition

Consider a set of words (pattern) H of the same length m over an alphabet V. A word w is a ***suffix-prefix*** (an ***overlap***) for H iff there exist two different words h, g from H such that w is a proper prefix of h and w is a proper suffix of g. The set of all suffix-prefixes for H is called an ***overlap set*** $OV(H)$. In case of pattern consisting of one word set is an autocorrelation set [1].

Motivation and aim

We are interested in the problem to estimate an average number of suffix-prefixes of all patterns generated according to Bernoulli model. This problem appears in computational biology and combinatorics of words. In particular, we have proposed [2] an algorithm for counting probabilities of exactly p occurrences of words from a pattern H in a given text of length n. The time complexity of this algorithm mainely depends on size of overlap set. And to estimate the complexity of this algorithm we have to estimate the overlap set size $|OV(H)|$.

Hypothesis and main results

Hypothesis: Consider patterns consisting of r words of length n over an alphabet V and suppose that the patterns are generated according to the Bernoulli model. Then an average number S of suffix-prefixes of the patterns is $C\cdot r$, where C is a constant that does not depend on the word length and depend on the probability distribution.

We have already proved this assumption for patterns having uniform distribution (Bernoulli distribution, where all letters have the same probability 0.25)

In case of general (biased) Bernoulli distribution we have proven that $S = C\cdot r$ where C and a does not depend on word length n ; $a > 1$.

Experiments and results

To verify our hypothesis we have performed experiments. We have generated random sets of r words with the same length m over alphabet $V = \{A,C,G,T\}$ distributed under Bernoulli model. Let S be an average number of suffix-prefixes. The experiments show that for uniform distribution (letters have probabilities: 0.25, 0.25, 0.25, 0.25) C 1 and S r. These results are demonstrated on figure 1. Also we consider Bernoulli distribution $\{0.1,0.1,0.1,0.7\}$. For this distribution S 1.3$\cdot r$ (see Fig. 2).

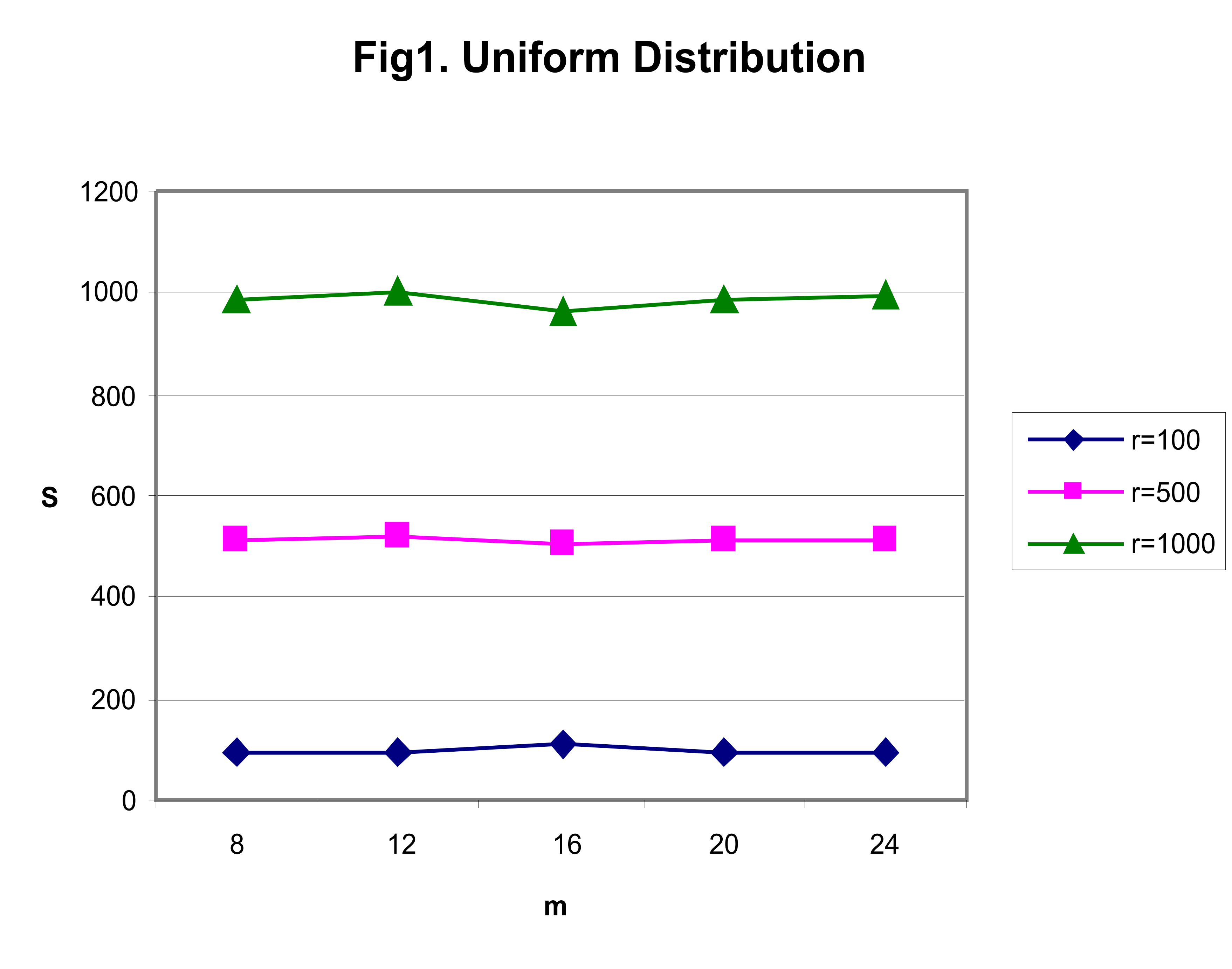

Fig1. Uniform Distribution

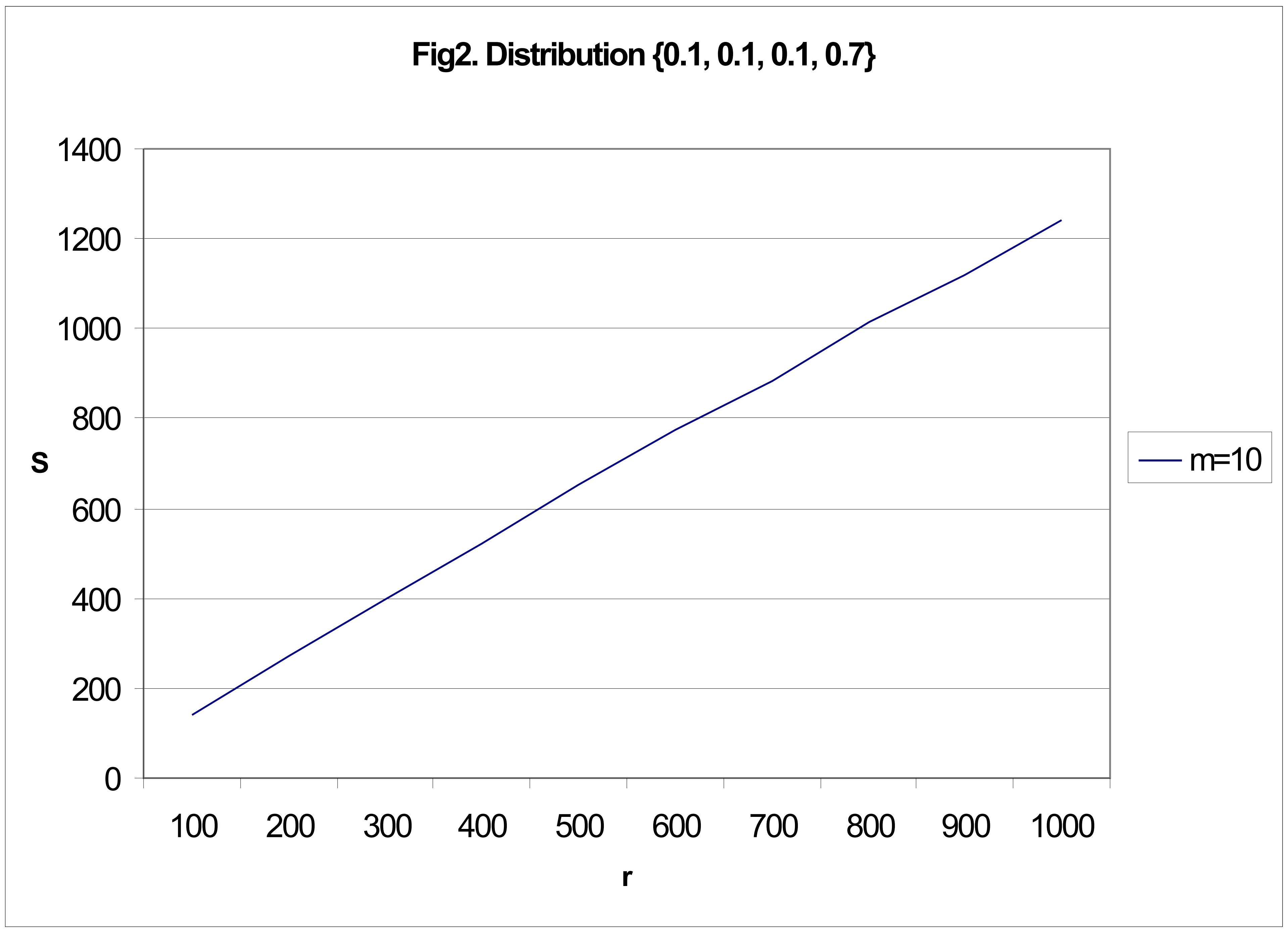

Fig2. Distribution {0.1, 0.1, 0.1, 0.7}

The work was supported by grants 08-01-92496-NCNIL-a and 01053- from RFBR (Russia), Intas grant 05-1000008-8028 and Migec-Inria associate team.

1. E. Rivals (2006). Autocorelation of Stringth. Encyclopedia of Integer Sequences.
2. M. Regnier, Z. Kirakosyan, E. Furletova and M. Roytberg. A Word Counting Graph.
(accepted to "London Algorithmics 2008: Theory and Practice").