

# On the Average Sequence Complexity\*

Svante Janson  
Dept. of Mathematics  
Uppsala University  
Uppsala, Sweden

Stefano Lonardi  
Dept. of Computer Science  
University of California  
Riverside, CA 92521

Wojciech Szpankowski  
Dept. of Computer Sciences  
Purdue University  
West Lafayette, IN 47907

In this paper, we are interested in studying a measure of complexity of a sequence called the *complexity index*. The complexity index captures the “richness of the language” used in a sequence. Formally, the *complexity index*  $c(x)$  of a string  $x$  of size  $n$  over an alphabet of cardinality  $k$  is equal to the number of distinct substrings in  $x$ . The measure is simple but quite intuitive. Sequences with low complexity index contain a large number of repeated substrings and they eventually become periodic (e.g., tandem repeats in a DNA sequence).

The idea of using the complexity index to characterize sequence statistically has a long history of applications in several fields, such as data compression, computational biology, data mining, computational linguistics, among others.

In order to identify unusually low- or high-complexity strings one needs to determine how far are the complexities of the strings under study from the average or maximum string complexity. While the maximum value for the complexity index has been studied quite extensively in the past, to the best of our knowledge there are no results concerning the average complexity. We first prove that for a sequence generated by a mixing model (which includes Markov sources) the average complexity is asymptotically equal to  $n^2/2$  which coincides with the maximum string complexity. However, for memoryless sources we establish a more precise result, which follows.

**Theorem 1** *Let  $C_{n,k}$  be the random variable associated with complexity index of a string generated by an unbiased memoryless source over a  $k$ -ary alphabet. Then the average  $l$ -subword complexity (i.e., the number of distinct words of length  $l$ ) is*

$$\mathbf{E}(C_{n,k}^l) = k^l(1 - e^{-nk^{-l}}) + O(l) + O(nlk^{-l}).$$

Furthermore, for large  $n$  the average complexity index becomes

$$\mathbf{E}(C_{n,k}) = \binom{n+1}{2} - n \log_k n + \left( \frac{1}{2} + \frac{1-\gamma}{\ln k} + \phi_k(\log_k n) \right) n + O(\sqrt{n \log n})$$

where  $\gamma \approx 0.571$  is the Euler constant and

$$\phi_k(x) = -\frac{1}{\ln k} \sum_{j \neq 0} \Gamma\left(-1 - \frac{2\pi ij}{\ln k}\right) e^{2\pi ijx}$$

is a continuous function with period 1 and  $|\phi_k(x)| < 2 \cdot 10^{-7}$ .

---

\*This work was supported in part by NSF Grants CCR-0208709, DBI-0321756 and NIH grant R01 GM068959-01.