Herzel, H.

Complexity of symbol sequences. (English) Zbl 0651.92014
Syst. Anal., Modelling Simulation 5, No. 5, 435-444 (1988).

Summary: The statistical properties of various one-dimensional strings are investigated using Shannon entropies and Hamming distances between substrings. Special attention is devoted to representative biosequences: a DNA string and two protein sequences. A rather high degree of randomness of biosequences is found whereas all considered computer languages exhibit long-range correlations.

Shannon entropies of longer "words" are significantly influenced by the finite length of any real sequence. It is shown that straight forward calculations lead to systematic underestimations of entropies. In order to compensate this effect a length correction formula is proposed.

MSC:

92Cxx Physiological, cellular and medical topics
94A17 Measures of information, entropy
68Q99 Theory of computing

Keywords:

Rous Sarcoma virus; biochemistry; reverse transcription enzyme; envelope protein; molecular biology; long-range correlations; complexity measures; mutual information; spatial structures in sequence space; one-dimensional strings; Shannon entropies; Hamming distances; DNA string; protein sequences; biosequences; length correction formula