Mr. Md. Izhar Ashraf
(RRN:1294213)
Department: Computer Applications

B.S. Abdur Rahman Crescent Institute of Science & Technology
Deemed to be University u/s 3 of the UGC Act, 1956

Network representation of symbolic sequences: A computational analysis

Abstract

Unraveling the information encoded in symbolic sequences, either natural or man-made, has always been a fundamental challenge in science. Such sequences appear at different scales, ranging from the microscopic chains of nucleotides (in DNA and RNA) and amino acids (in proteins) which constitute the very basic processes of life, to the macroscopic arrangement of growth rings in trees and layers of sedimentary deposits on the earth's surface providing a record of geological time scales. Sequences may also be temporal as in time-series, for example, that of human speech and musical sounds whose meaning depends sensitively on the precise relative arrangement of the constituent sounds. The common thread connecting these very different types of sequences is the fact that their components are arranged in a specific order that arises from the process of their construction, i.e., their generative mechanism. Knowledge of this "grammar" can potentially inform us about the properties of the possible sequences that can be generated. However, the inverse problem, i.e., inferring the generative mechanism (or at least, discovering the key set of grammatical rules) from a set of sequences remains an extremely difficult problem. In this thesis, we have developed unsupervised computational techniques (that do not use any prior knowledge of the underlying "language") that are based on network (or graph) theory in order to infer properties characterizing the underlying syntactic organization of sequences produced by the unknown generative mechanism. We have applied these techniques to several corpora of linguistic sequences and one database of as yet undeciphered inscriptions obtained from the remains of the Indus Valley civilization. Our results show evidence for syntactic structure not only in the known linguistic sequences but also in the Indus civilization inscriptions. We report a universal pattern in the structure of linguistic sequences that can be used as a tool for inferring the direction in which inscriptions have been written. We also describe a technique for segmenting inscriptions into frequently occurring blocks of sub-sequences, defined in terms of statistically significant co-occurring symbols. Apart from linguistic sequences, we have also applied the techniques to micro RNA and human-generated random number sequences.