Text Encoding for Protein Structure Representation

Jun Tan[1,⋆], **Donald Adjeroh**[1]

1. West Virginia University, Morgantown, WV 26506
⋆Contact author: jforce716@gmail.com

Keywords: Text Processing, N-Gram, Protein, Secondary structure

Given the rapidly increasing quantity of genomic and proteomic data that is now easily available even to a casual observer, the new challenge is in making sense out of the vast quantities of data. Efficient and reliable analysis of protein 3D structures is identified as a major challenge in this post genomic era. Whether the objective of the analysis is for protein classification, protein structure prediction, discovery of protein structural motifs, or assignment of a functional class to a newly discovered protein, a key aspect in the analysis is the representation used to encode the protein 3D structural information. In this work, we introduce the protein shape context, and its encoding into a protein shape string as an effective descriptor for protein 3D structures. Based on the same general idea, we develop three different ways to construct shape context and compare their performance by searching a database for proteins with similar structures. We show how the choosing of parameters affects the performance. We further investigate the effect of combine two methods together which lead to building protein structure profile using multiple strings.