Contribution ID: **28** Type: **not specified**

Modelling for Efficient Scientific Data Storage Using Simple Graphs in DNA

Wednesday 25 September 2024 10:05 (20 minutes)

Data analytics requires large data archives beyond current world storage media, causing researchers to seek alternative storage media. Scientists in fields like biology, ecology, life sciences, and medicine are using data archiving to aid their research. During the last decade, DNA (Deoxyribonucleic Acid) storage has been significantly investigated as a method for archiving data at massive scales. Digital information can be encoded at high density with synthetic DNA that is durable and long-lasting. However, expensive synthesis and sequencing processes hinder DNA data storage at a large scale and lead us to compress the data beforehand. Network science applications are eager to store graph data archives efficiently using DNA storage, even though it has been demonstrated with raw data storage. Graph-aware data archiving has a significant advantage over raw data, reducing the related data size for DNA storage in terms of nucleotides and resulting in lower database operational costs. This paper presents a theoretical model for storing scientific data efficiently in DNA using simple graphs. The Re-Pair compression algorithm is utilized to investigate individual and composite graph storage strategies, and simple graph-based datasets, particularly from the biological domain, are used for experimental analysis. Composite graphs, however, yield a higher compression ratio than aggregated standalone graphs. Noticeably, the compression ratios range from 1.18 to 1.53, saving a substantial amount of money in a DNA storage system's synthesis and sequencing processes. Consequently, data analytics could be performed cost-effectively using DNA as an emerging storage medium.

Presenter: USMANI, Asad (Goethe University Frankfurt)

Session Classification: Multiscale Models in Cell Biology I (Chair: Thomas Sokolowski)