ESIT 2018 Tutorials

New problems in coding theory and computing driven by macromolecule-based data storage

Olgica Milenkovic
University of Illinois

Abstract

Macromolecule-based data storage is an emerging technique for massive archival storage that uses DNA sequences and synthetic and naturally occurring polymers as storage media. As the storage media, as well as the recording and reading processes are highly unconventional compared to classical storage systems, new methods for addressing system reliability, random access, parallel data processing and in-memory computation are needed. We will overview the basic operational principles of systems for molecular data recording (e.g., DNA and polymer synthesis, DNA editing), molecular data retrieval (e.g., DNA high throughput and third generation sequencing platforms, mass spectrometry systems), error-correction and constrained coding (e.g., DNA-profile codes, trace and substring reconstruction codes, weakly mutually uncorrelated codes, asymmetric Lee distance codes), and DNA-based computing (e.g., string and nick displacement). In addition, we will provide an overview of relevant concepts in bioinformatics and genomic data processing needed to implement operational macromolecular storage devices.