DNA-based data storage

Contemporary global demand for storage capacity is increasing exponentially, even as traditional magnetic storage media has exhausted its potential for optimization, and requires unsustainable investments for both production and maintenance, in capital, energy and space.

One promising potential medium for archival data storage is DNA; it features high density, extreme longevity, convenient scalability and a lower maintenance footprint. Complete DNA-based storage ecosystems are in active development, raising multiple coding theoretic (as well as engineering, algorithmic, and biotechnological) challenges; correspondingly, increasing attention is recently given to the study of such systems, and they are drawing significant investments from both governments and the private sector.

The following theses are available (other topics in this domain will also be entertained):

- The torn paper channel models the effects of DNA strand breakage in storage or processing. It has been studied from both an average-case and a worst-case [1] perspective, with several distinct adversarial models. Recently, the t-break model was studied [2] as a refinement of the previously studied min-max constraint. These developments open the way to study new problems in this setting.
- Reconstruction from substring spectra is a model motivated by the process of shotgun sequencing, where short strands are drawn sufficiently many times to reconstruct a long information sequence. Adapting existing literature [3,4] to more realistic models is an open problem.
- Nanopore sequencing is a nascent technology that reads single-stranded DNA molecules by passing them through a narrow pore while passing electric current through it. More work studying its properties and designing codes capable of handling its relatively high error rate, extending existing literature [5,6], is necessary.
- Duplications are a type of mutation occurring in the process of cell replication, which may be responsible for large portions of our current genome. For data storage schemes in in vivo DNA (e.g., for watermarking research material) it is an error model that needs to be countered. We aim to extend and build upon existing literature [7,8].


Advisors

Yonatan Yehezkeally