Coding for Composite DNA

DNA Data Storage

Data storage on DNA molecules is a promising approach for archiving massive data.

In classical DNA storage systems, binary information is encoded into sequences consisting of the four DNA bases {A, C, G, T}. The encoded sequences are used to generate DNA molecules called strands using the biochemical process of DNA synthesis. The synthesized strands are stored together in a tube. To retrieve the binary information, the strand must be read via DNA sequencing and decoded back into the binary representation.

The synthesis and the sequencing procedures are error-prone, and with the natural degradation of DNA they introduce errors to the DNA strands. To ensure data reliability, the errors have to be corrected by algorithms and error-correcting codes (ECCs).

A 5min video with an overview of DNA storage: https://youtu.be/r8qWc9X4f6k?si=Yzm5sOW-a6VDnBu3

Composite DNA

Recently, to allow higher potential information capacity, [1,2] introduced the DNA composite synthesis method. In this method, the multiple copies created by the standard DNA synthesis method are utilized to create composite DNA symbols, defined by a mixture of DNA bases and their ratios in a specific position of the strands. By defining different mixtures and ratios, the alphabet can be extended to have more than 4 symbols. More formally, a composite DNA symbol in a specific position can be abstracted as a quartet of probabilities \( \{ p_A, p_C, p_G, p_T \} \), in which \( p_X \), \( 0 \leq p_X \leq 1 \), is the fraction of the base \( X \) in \( \{ A, C, G, T \} \) in the mixture and \( p_A+p_C+p_G+p_T =1 \). Thus, to identify composite symbols it is required to sequence multiple reads and then to estimate \( p_A, p_C, p_G, p_T \) in each position.

Problem description

ECCs for DNA data storage differ in many aspects from classical error correction codes. In this model, new error type gain relevance, like deletions and insertions which affect the synchronization of the sequences. Especially for composite DNA data storage, these error types received only little attention.

The most related work to this problem was recently studied by Zhang et al. in [6]. The authors initiated the study of error-correcting codes for DNA composite. They considered an error model for composite symbols, which assumes that errors occur in at most \( t \) symbols, and their magnitude is limited by \( l \). They presented several code constructions as well as bounds for this model. In this thesis, we want to analyse a different way to model the composite synthesis method and studies additional error models. We already have some results for substitution and single deletion errors. This thesis should focus on evaluating more error models in the channel model.

This should only roughly introduce the problem. No need to review all references. If you are interested, please reach out to me, and we can discuss a suitable direction for you.

References


Advisors

Frederik Walter