DNA encoding

August 2025

This article outlines a method for encoding digital data into 8-base-long DNA strands.

1 Encoding Methodology

1.1 Data Conversion: Binary to Ternary

The encoding methodology begins by converting the source data into a single, long binary stream. This stream is then converted into its ternary equivalent, which forms the foundational data sequence for DNA encoding.

1.2 Segmentation and Homopolymer Avoidance

For storage, this long ternary sequence is partitioned into smaller, manageable segments. The conversion from these ternary digits (trits) to DNA bases is specifically designed to prevent the formation of homopolymers—consecutive runs of the same base. This proactive measure significantly reduces the likelihood of synthesis and sequencing errors.

1.3 DNA Strand Structure

The information within each 8-base-long DNA strand is meticulously structured to include indexing trits, data trits, and an error correction trit. The value for this error correction trit is calculated for each segment by summing the corresponding source trits (both index and data) and taking the result modulo 3. This integrated approach aims to create an efficient and reliable method for DNA-based data storage.

1.4 Encoding Initialization

To ensure consistent encoding across all sequences, the first base of every DNA strand is processed using a fixed initial condition: the preceding base is always treated as 'G'. Any necessary padding should appear at the beginning.

1.5 Ternary-to-Base Conversion Rules

The conversion from ternary digits to DNA bases follows the rules outlined in the table below, which depend on the preceding base to prevent homopolymer formation.

Table 1: DNA Base Conversion Rules from Ternary Digits

Previous Base	**Next Base for Ternary Digit**		
	0	**1**	**2**
A	T	C	G
T	A	C	G
C	A	T	G
G	A	T	C

1.6 Decoding and Data Retrieval

The retrieval of the original information is achieved through DNA decoding, a process that directly reverses these encoding steps to reconstruct the initial binary data.

2 Conclusion

The described DNA encoding method offers a structured approach to data storage that directly addresses the challenge of homopolymer-induced errors. It achieves a great balance between efficiency and data integrity for storing with DNA strands that are 8 bases in length. This methodology not only ensures data integrity but also provides a straightforward and reversible pathway, making both the DNA encoding and DNA decoding processes highly efficient.