A Method of Encoding and Decoding Information in DNA Sequences that Includes Error Correction

Zimmerman, Anna

The purpose of this investigation was to develop a means of encoding information into DNA sequences that can then be decoded and to develop error correction methods to detect and correct possible errors in the DNA sequences. Each ASCII character value in a message is first converted to a quaternary (base 4) value. The quaternary value is then encoded by substituting the DNA base-pairs A, C, G or T for each digit. A software algorithm written in C# was developed to encode any ASCII text message into the corresponding DNA sequences. Four different error correcting algorithms (Redundant, Hamming Base 2, Hamming Base 4, and Hamming Base 2 modified) were also developed. A sample message was created and encoded into two DNA sequences, one with and one without error correction included. These two DNA sequences were sent to a laboratory for DNA synthesis and subsequent sequencing. The returned sequences were compared to the originals. The four different error correction methods were tested for comparative performance. Monte-Carlo simulations of random error content, from 0 to 10 percent in the encoded DNA sequences, were run using 10,000 samples each. The results of the investigation showed that messages can be encoded into DNA sequences. Additionally, error correction improves confidence in the decoded result. Messages encoded in DNA that was synthesized and sequenced demonstrated that the encoding and decoding algorithm work and no errors were detected in either sample. Monte-Carlo simulation of each tested error algorithm showed that the most efficient error correcting code was the modified Hamming Base 2 algorithm. The most accurate error correcting code was the Redundant algorithm.

Awards Won: