Name(s)
Nithika Karthikeyan

Project Number
J0809

Project Title
The Design of Algorithms to Encode English Text in Amino Acids Using Digital Data Compression Techniques

Abstract

Objectives/Goals
Worldwide digital data is forecast to grow to 160 zettabytes (10^18 KB) in 2025. Traditional storage solutions are not keeping up with this exponential demand, increasing the cost of data storage. Most data (75%) is archival in nature, for which storage access is not time critical.
DNA and amino acids are being considered as unconventional, high-density storage media for archival data. Their small size (sub-nanometer for DNA and nanometer for amino acids) and non-binary nature can result in 85% space savings compared to current storage solutions.
The goal of this project is to devise algorithms to further increase the space efficiency of amino acid storage media using data compression techniques and data characteristics.

Methods/Materials
Two encoding and decoding algorithms were invented. The FixedLength algorithm assigns two amino acids to a byte of data, based on size. The VariableLength algorithm uses Huffman encoding and text characteristics, to assign one or two amino acids for every byte.
Using Pascal and a Windows PC, the algorithms were implemented. The output of encoding, a sequence of amino acids (peptide), was checked for stability and structure using free simulation software. Peptide sequencing to ensure it stored the sequence correctly, was done using free Mass Spectroscopy simulation. These algorithms were compared to the baseline algorithm, Sabry. Each algorithm was run over nine English texts of differing lengths. The decoded text was compared to the original for correctness. An end-end amino acid storage system has been proposed, but not implemented, as it is outside the scope of this project.

Results
FixedLength is 35% better than baseline and VariableLength is 42% better. The encoding and decoding times of both algorithms are comparable to Sabry. Overall, both algorithms proved to be more efficient than baseline.

Conclusions/Discussion
Computations showed that VariableLength can store "The Bible" in 8% of the space needed by solid state memory. This will bring huge space savings at data centers and the electricity needed to run them, thus reducing storage cost. As slow access of archival data is acceptable, chemical processes involved would not deter the application of this storage solution.
Enhancements could be made to represent other media types. Duplicate amino acids when sequencing has to be resolved. Actual synthesis and sequencing would prove the system works.

Summary Statement
This project involves the creation of two algorithms that reduce data storage costs using unconventional storage media.

Help Received
I designed the algorithms on my own, but received some help with the understanding of amino acids from BioCurious as well as several online websites.