Readme

Algorithm Description

The proposed algorithm is designed for text storage and is particularly suitable for DNA cassette tape structures. It divides the text into several segments (or phrases) and stores each segment in a distinct zone. When the text undergoes slight modifications, only a few segments are affected, resulting in changes to the corresponding zones only. Consequently, this algorithm facilitates minimal modifications to the tape zones when the text is slightly altered.

File Function

DNACodebook.py – Generates a DNA codebook consisting of multiple 4-nt length DNA codes. The DNA codes are capable of satisfying 3-nt homopolymer and GC balance (50%) constraints.

DNAMapping.py – Maps characters, numbers, and punctuation to DNA codes and simultaneously generates a reverse mapping.

TextDNAEnc.py – Encodes the text file (before modification) into DNA sequences.

ChangeDNAEnc.py – Encodes the text file (after modification) into DNA sequences.

TextDNADec.py – Decodes DNA sequences back into a text file.

Usage

Language: python3

1. Run **DNACodebook.py** and **DNAMapping.py** to generate the mapping files.
2. Run **TextDNAEnc.py** to generate the DNA sequences.
3. Run **TextDNADec.py** to decode the DNA sequences.
4. If the original text file is modified, run **ChangeDNAEnc.py** to generate the DNA sequences for the modified text file. Note that manual input is required in this step.
5. Run **TextDNADec.py** to decode the DNA sequences.

Example

1. Suppose we have a file named **incorrect-address.txt** containing an incorrect address of SUSTech (i.e., Department of Biomedical Engineering, College of Engineering, Sudan University of Science and Technology, Nanshan Shenzhen, Guangdong Province, China), but we are unaware of the inaccuracy.

2. To encode this file, we first run **DNACodebook.py** and **DNAMapping.py** to generate the mapping files **DNAMapping.json** and **DNAMappingReverse.json**.

3. Using the mapping file, we utilize **TextDNAEnc.py** to encode **incorrect-address.txt** into DNA sequence files **incorrect-addressDNA.txt** and **incorrect-addressDNAZone.txt**. These sequences are then stored on a DNA Cassette Tape.

   Note: The generated segments are stored in a phrase array (named **A** for convenience) `["Department of Biomedical", "Engineering, College", "of Engineering, Sudan", "University of Science", "and Technology, Nanshan", "Shenzhen, Guangdong", "Province, China#####"]`, where `#` is a placeholder to satisfy the DNA length requirement.

4. Upon reading, we obtain the DNA sequence file **incorrect-addressDNA-exp.txt**. We then run **TextDNADec.py** to decode the file, retrieving the original text file containing the incorrect address of SUSTech.

5. We correct the address by modifying **incorrect-address.txt** and saving it as **correct-address.txt** (i.e., Department of Biomedical Engineering, College of Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, Guangdong Province, China).

6. To obtain the encoded sequences of **correct-address.txt**, we execute **ChangeDNAEnc.py**, generating the DNA sequence files **correct-addressDNA.txt** and **correct-addressDNAZone.txt**.

   Note: It is observed that only the 3rd and 5th phrases need to be changed. Therefore, we input `['of', 'Engineering', ',', 'Southern']` and `[2]` to indicate that the element with index 2 in **A** should be changed to store the phrase `['of', 'Engineering', ',', 'Southern']`. Similarly, we input `['and', 'Technology', '(', 'SUSTech', ')', ',']` and `[4]`. Finally, we input `###` to indicate that the modification has been completed.

7. Since **correct-addressDNA.txt** and **correct-addressDNAZone.txt** are similar to **incorrect-addressDNA.txt** and **incorrect-addressDNAZone.txt**, only a small number of sequences in some zones need to be changed.

8. Upon reading, we obtain the DNA sequence file **correct-addressDNA-exp.txt**. We then run **TextDNADec.py** to decode the file, retrieving the original text file containing the correct

address of SUSTech.

For more details, please refer to the code files.