SUMMARY

SUMMARY

The purpose of DNA sequence compression-encryption is to develop fifteen compression and two encryption techniques for completion of this work. For experimental purposes, two sets of data, named data set-I and data set-II had been used. Among ten techniques, the better compression rate is 1.6977 bits/base of data set-I in RHUFF technique and 1.999907 bits/base of data set-II and in both the cases the compression ratio is also less. This compression rate is better than so many marketable standard compression techniques. The RHUFF technique is Repeat + modified Huffman's technique.

This RHUFF technique is far better than gzip techniques with respect to compression rate and information security. The gzip technique is (Lempel-Ziv "LZ"+Huffman) without any security concept. This method is the combined method of Repeat+ modified Huffman for compression as well as security.

The single substring substitution technique, the average 43% to 45% of the data set-I and 41% to 43% of the data set-II of encryption on the actual text will be modified 95% of the text, it is better over bi & tri-combinational substitution method.

The modified Huffman's technique shows that 99% of the original text will be affected when swapping is done at a lower level, the percentage of encryption is reduced. This provides higher security if two binary values are considered as a key and encryption of the data is better over node interchange and interchanging of words instead of character. It is also observed that if interchange is done at the top level, the Lavenstein Distance is at the highest point and % of damage is also the highest.

In selective encryption, change of one byte in plain text leads to significant changes in bytes of output information. It shows that the decryption time is always less than the encryption time and independent of file size in both the data set. If consider the highest level of Hamming distance, the effect on original file is highest on the basis of top level interchanging.

The entropy and encryption & decryption throughput result is better than the others. Here all encryption techniques provided strong security with respect to security level and speed of encryption.

It also shows the randomly generated same length of artificial sequence, compression rate and ratio is highly comparative in cellular DNA sequences. The cellular DNA sequence is logical in organization, structure, systematic and non random whereas artificial data is random and unstructured. It proves that Life means order, it is non random and not topsy-turvy .

This lossless Compression technique achieves an average compression ratio and rate than that of available marketable DNA based compression algorithms and provides the best information safety with encryption and minimal decompression time. There is an optical difference in between cellular DNA and artificial DNA of equivalent length.

The experimental purpose uses cellular data that come from different origin as a results the compression rate & ratio are different every time where as compression rate & ratio remain same every time in case of artificial data.

This algorithm found compression gain, space saving percentage, disk utilization and percentage of encryption ratio percentage is best in comparison with the other. This algorithm is compression friendly because it has no or very little impact on data compression efficiency and encryption rate and ratio is also the best. This algorithms compression score is also better.

Considering algorithm evaluation parameter, it is found RHUFF to be the best compression-encryption technique, having encryption rate of 2038.99byte/see. As such RHUFF acts as a model.

This proposed compression-encryption model does encryption after compression, generating the original file, then decryption followed by decompression.

The RHUFF is the best suitable DNA sequence compression algorithm that depends on data type & size. This compression-encryption model can be used imperially to increase efficiency & security during data transmission, also reducing the overhead time of compression.