SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

ENHANCEMENT OF DNA MEMORY TECHNOLOGY USING COMPRESSION ALGORITHM

Project Guide : Ms Y. Asnath Victy Phamila, M.E
School of Computing, SRM University: 2009

Rahul Deo Vishwakarma, 10305151

SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

Abstract

The practical realization of DNA data storage is a major scientific goal. Here we enhance the potential of DNA for data storage using compression algorithm implemented on Information to be stored. Our propose scheme depends on Huffman Encoding with the adoption of BWT and MTF transforms, as compression of small text files must fulfill special requirements since they have small context. The Information is encoded in Nucleotide sequence and stored inside E.Coli genome. For data retrieval, a complete genome sequencing is made using DNA Sequencer.

SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

What is DNA Memory?

- **DNA Memory technology:** Use of Information theory and traditional Molecular Biological protocols.

- **Purpose:**

✓ To create Ultra High density storage of data.

✓ To exploit the protocols of Computational Biology for a new approach towards Data Storage for long duration.

Proposed Methodology

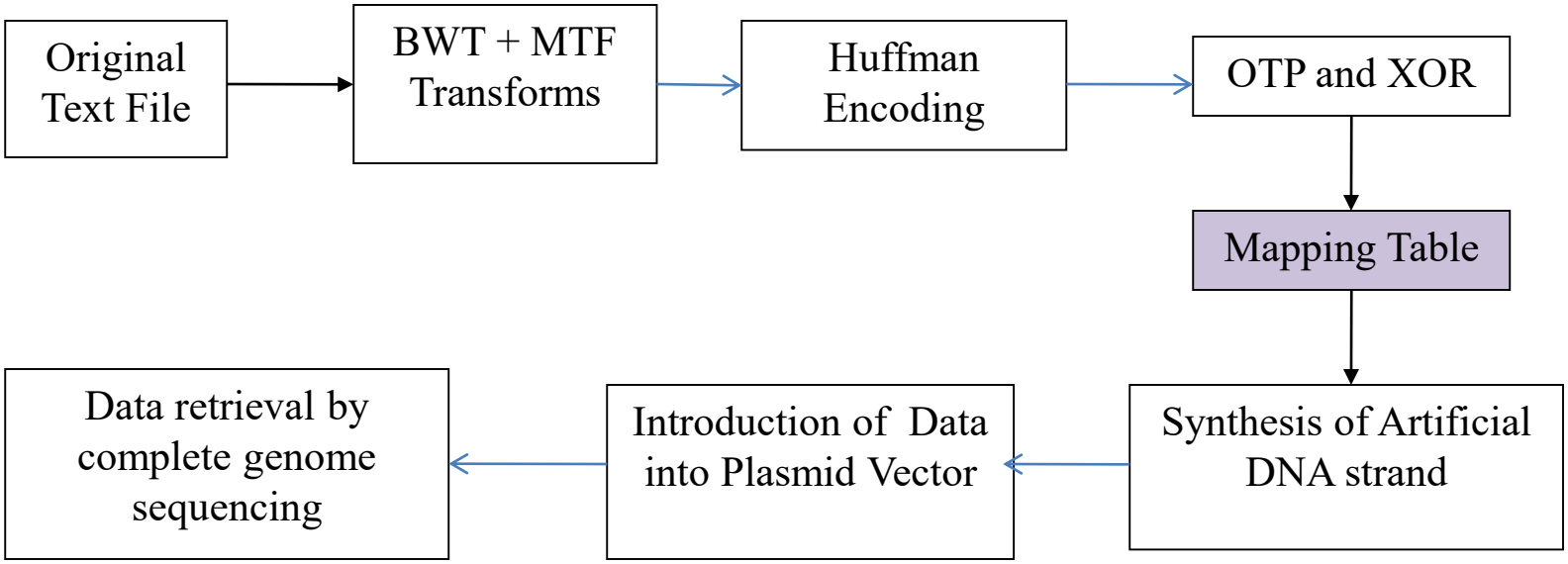


Fig.1 Project design Approach

Mapping Table

AA=0000	AG=1000
CA=0001	CG=1001
GA=0010	GG=1010
TA=0011	TG=1011
AC=0100	AT=1100
CC=0101	CT=1101
GC=0110	GT=1110
TC=0111	TT=1111

Table 1. Mapping table for Encoding and decoding information

Implementation

1.Synthesis of Information Strand

- ✓ Basic transform and Huffman Encoding is made over Information to be stored.

  OPERATION BARBAROSSA → (BWT+MTF+HUFFMAN)

  1110101111000101111001110010101101110110011100001011000100000

- ✓ For enhanced security binary strand is XOR-ed with One-Time-pad randomly generated Key.

  1010000111000010110111010000100001011011111001101100000011000

- ✓ Finally using Mapping Table, obtained binary strand is converted into Genetic base pairs.

  GGCAATGACTCTAAAGCCTGGTGCATCAAG

SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

Implementation

2. Introduction of Data into Plasmid Vector

- ✓ The chemically synthesized oligonucleotides were annealed and subcloned into the Escherichia Coli strand.
- ✓ DH5α was used as Cloning Host.
- ✓ Sequence of DNA strand was confirmed using DNA Sequencer ABI 3100.

Implementation

3. Transformation of *E.Coli*

- ✓ E.Coli competent cells were prepared and transformed using two-step culture method.
- ✓ For conformation of the introduction of Information strand in E.Coli genome, E.Coli strand was introduced to Chloroamphinicol.
- ✓ The Colonies that appeared after incubation of 37° overnight, displayed Em-sensitivity, indicating the presence of Information strand in E.Coli genome.

Implementation

4. Data Retrieval

- ✓ For Data retrieval one strand is extracted from genome of E.Coli and PCR is performed to get many copied of same stand.
- ✓ The stand is introduced to DNA Sequencer for Chromatogram sequence scanner.
- ✓ Obtained DNA Sequence is decoded using mentioned algorithms and Encryption key.

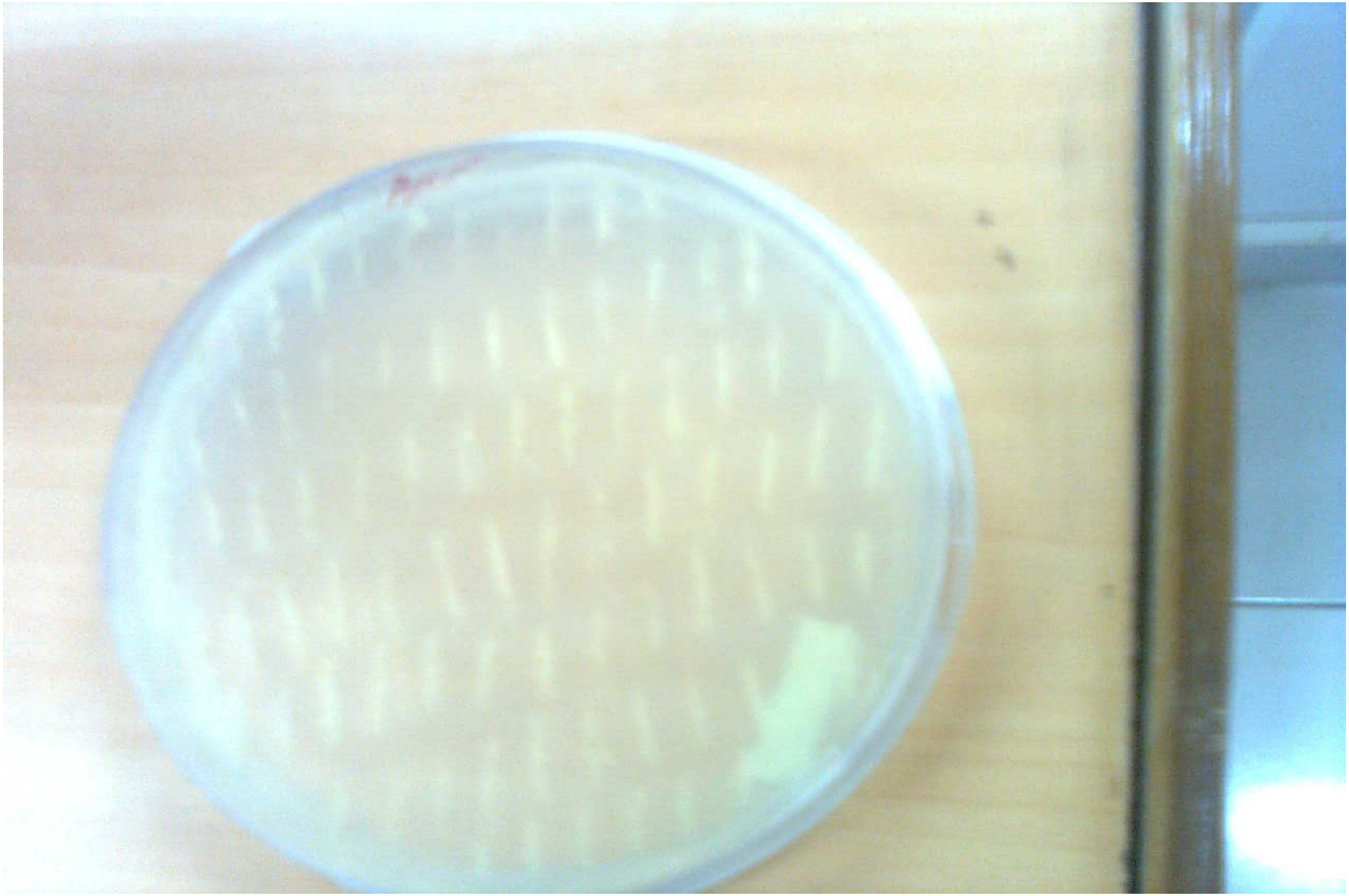


Fig.2. Cloned Information strand after PCR and before DNA Sequencing

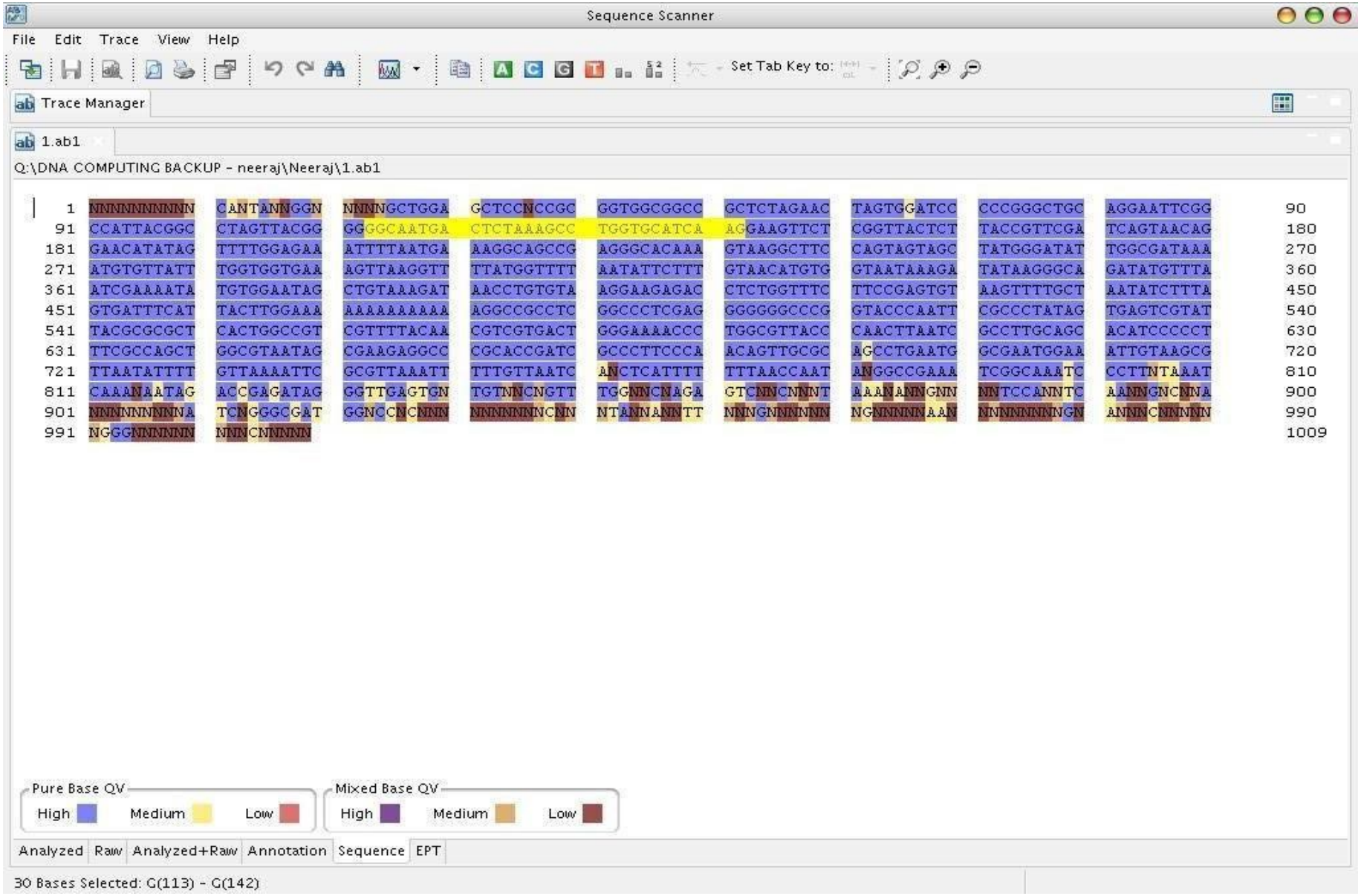


Fig.3. Base pair Sequence of DNA strand in *ABI Biosystems* Sequence Scanner

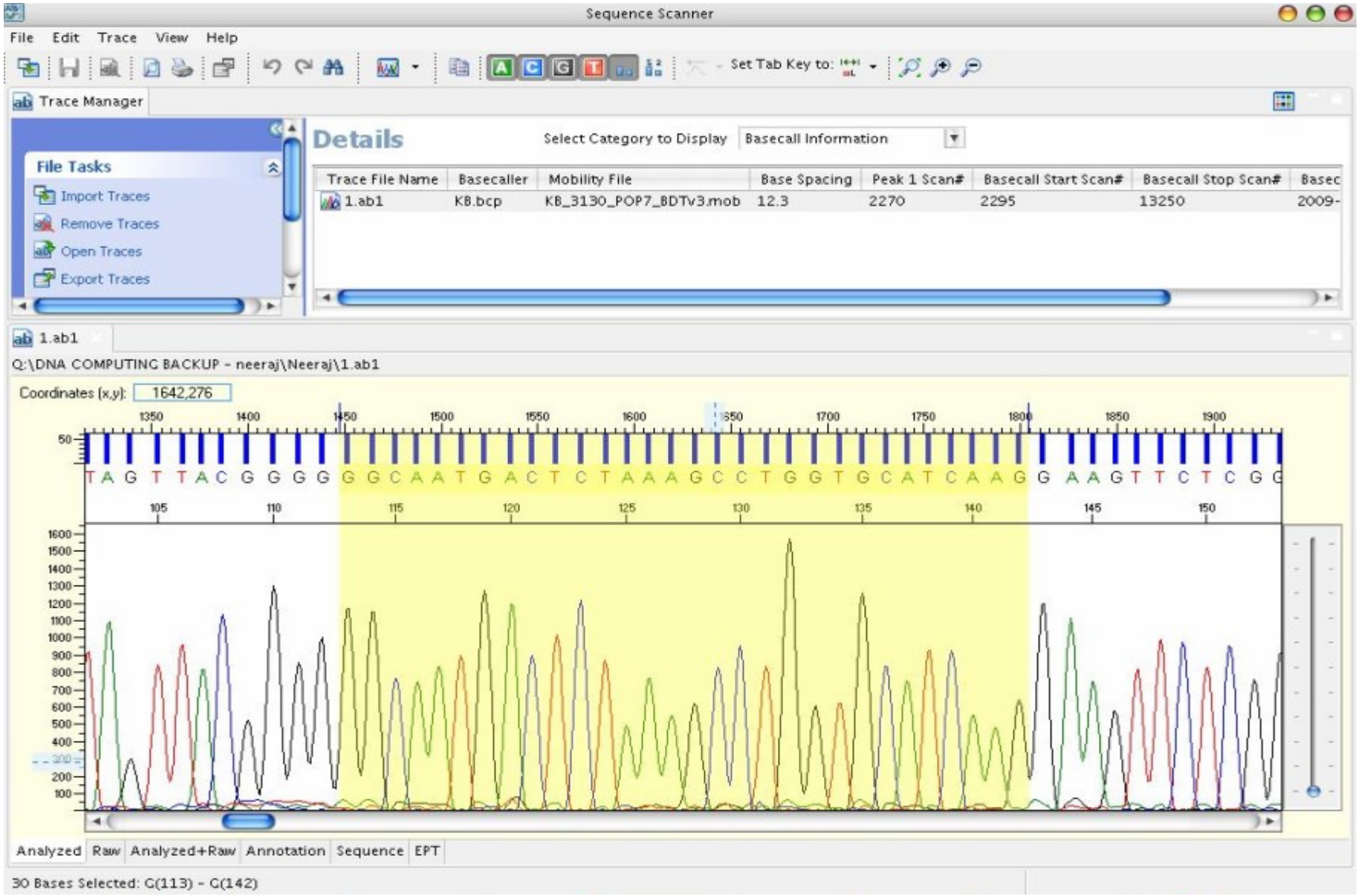


Fig.4. Analysis of DNA strand in *ABI Biosystems* Sequence Scanner

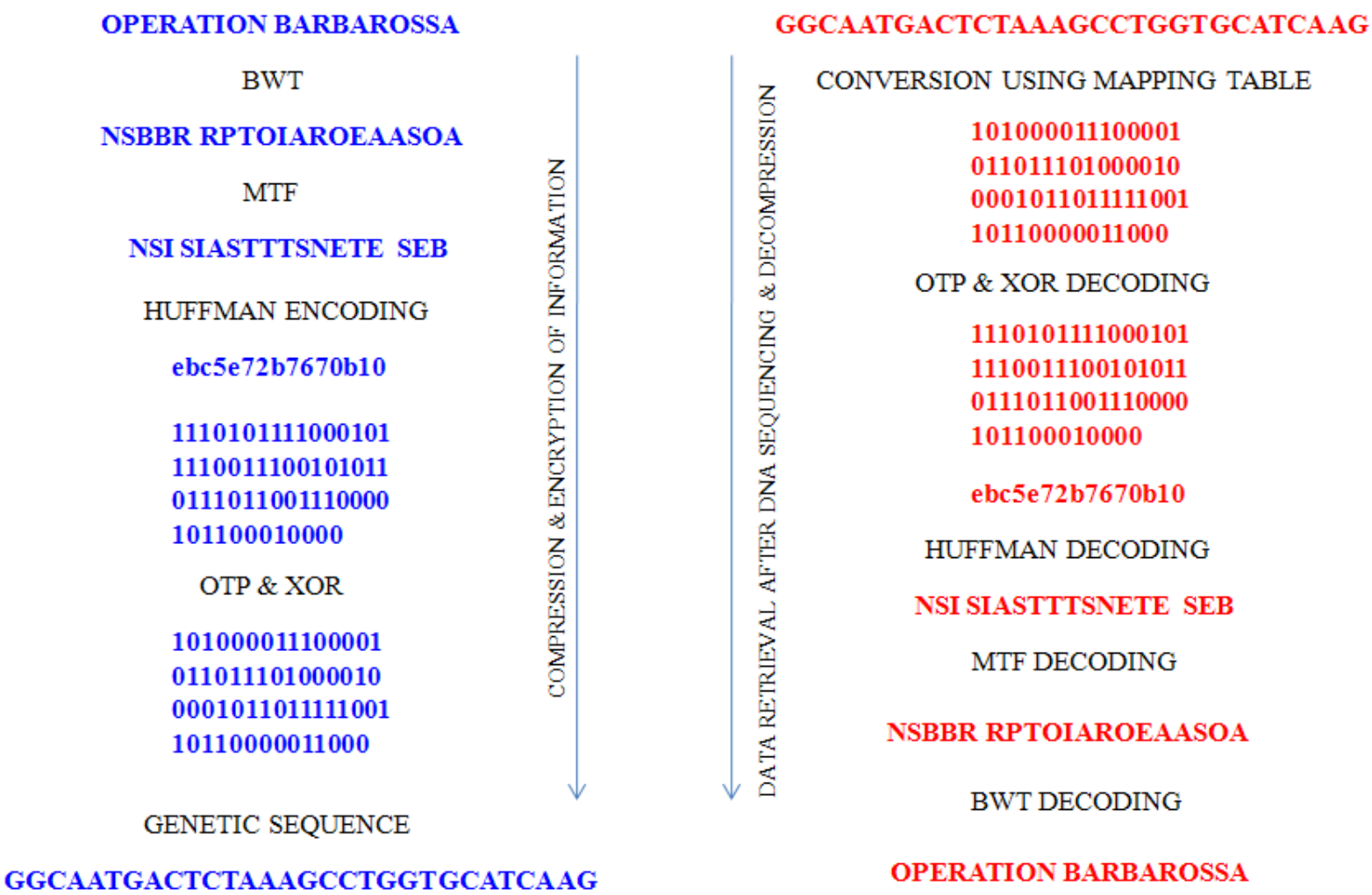


Fig.1. Compression and Decompression Algorithm

Analysis

Calculation of Data Density

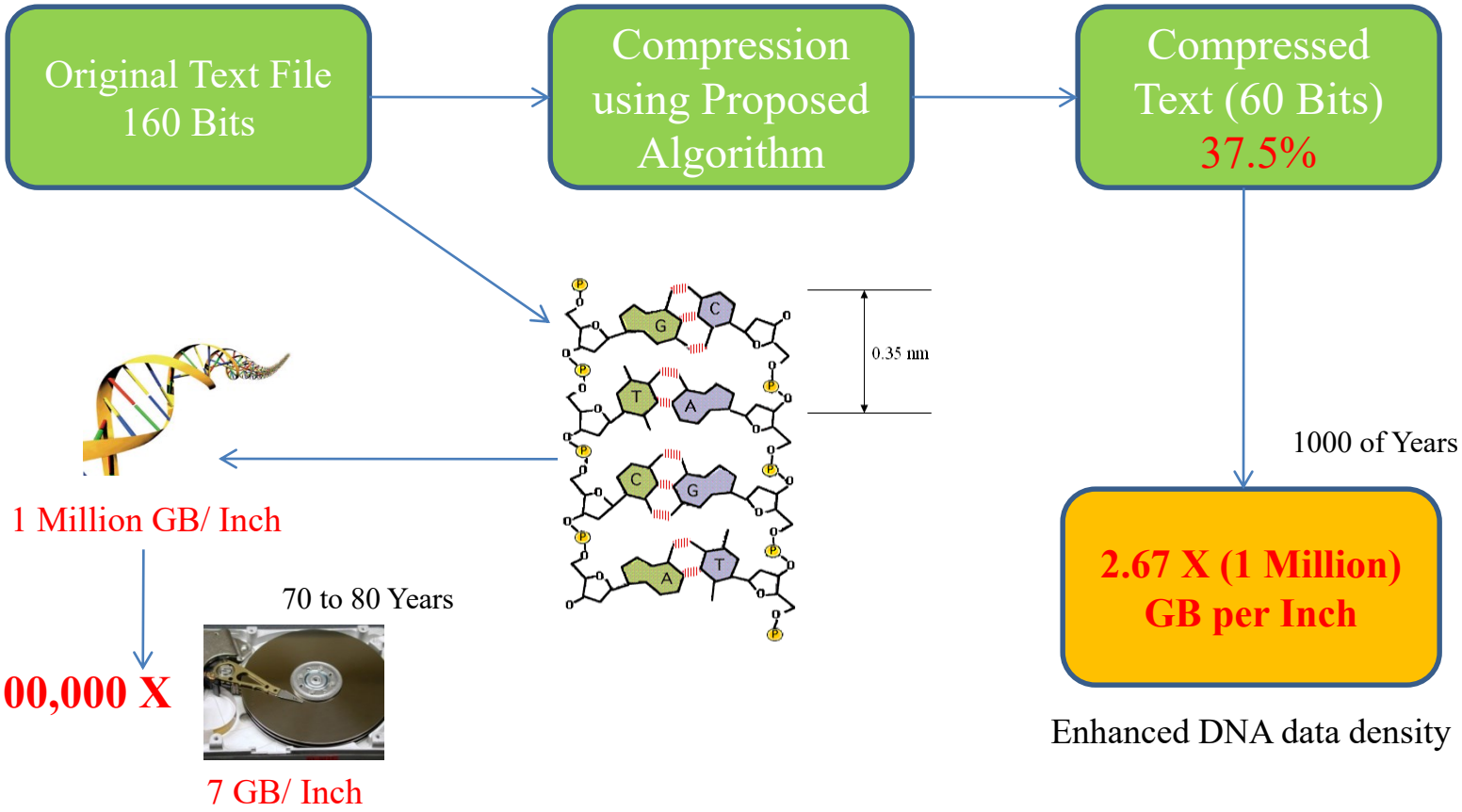


Fig.5. Analysis of enhanced DNA density after proposed Compression algorithm

Analysis

Error Correction Code scheme

Errors addressed in DNA

- Point Substitution
- Insertion & Deletion
- Inversion & Transversion

NATURAL SOLUTION

COMPLEMENTARITY

A = T

C ≡ G

Reference Information Strand

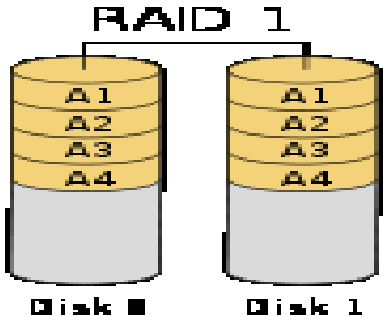
GGCAATGACTCTAAAGCCTGGTGCATCAAG

CCGTTACTGAGATTTCGGACCACGTATTTC

Complementary Strand

G

Repair Enzyme restore original DNA Sequence



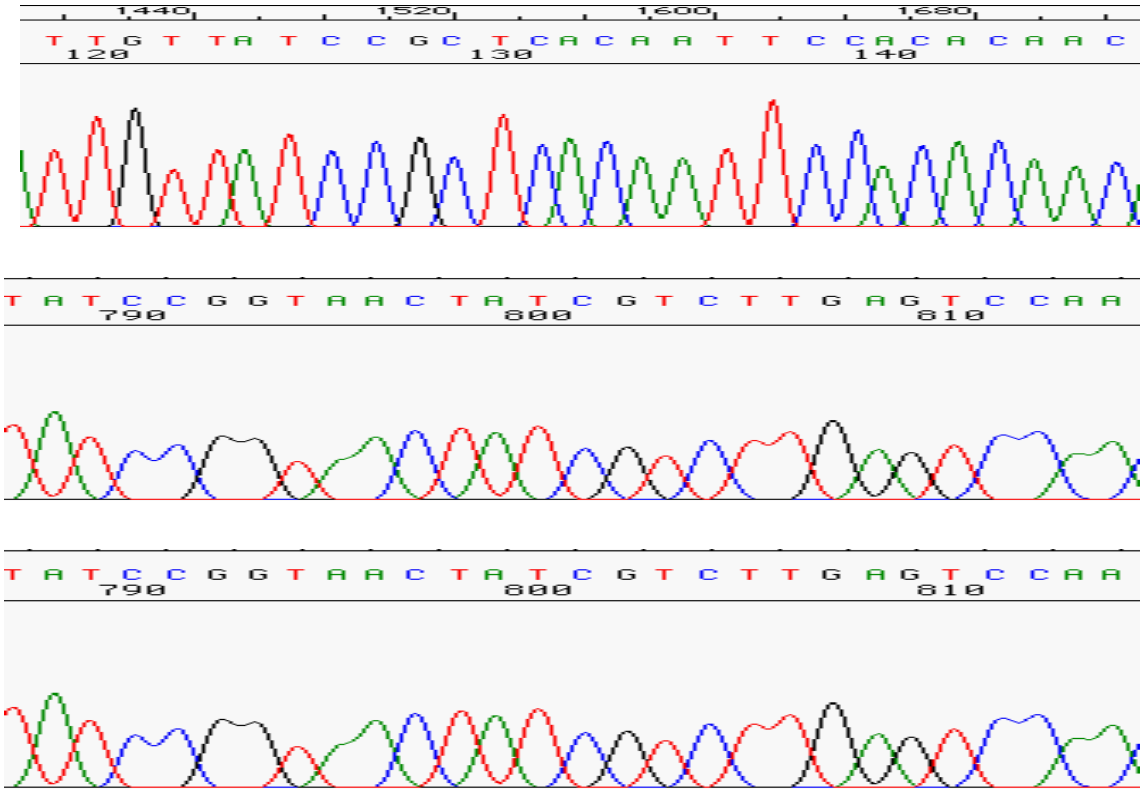
Data mirrored in two drives

RAID 1 Array

Fig.6. Analysis of DNA strand in *ABI Biosystems* Sequence Scanner

Analysis

The need of Compression Algorithm



Data from a very good sample analyzed by an ABI Biosystems DNA Analyzer

As the GEL progresses it leads to loss in resolution of the Chromatogram

As the length of the DNA strand increases, data retrieval rate decreases.

Fig.7. The need of Compression algorithm

Applications

- ✓ Durable data storage for long interval of time (1000 of years).
- ✓ Highly sophisticated DNA based Cryptosystems.
- ✓ Signatures of Living Modified organisms (LMOs).

Advantage

- ✓ BWT and MTF generates better context information for files of Small size.
- ✓ Less number of Bits represents more Information.
- ✓ Solves the problem during sequencing, if Information strand is long.
- ✓ Security for Small messages.
- ✓ Cost Minimization.

Limitation

- ✓ Extensive lab work for message encoding into DNA
- ✓ Data retrieval rate.
- ✓ Practically difficult to synthesize very long message strand.
- ✓ Financial factor

SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

Conclusion

✓ The message “OPERATION BARBAROSSA” was stored into *E.Coli* genome and retrieved successfully.

✓ Implementation of Compression algorithm increased the data density.

✓ Future work can be implementation of faster DNA Sequencing techniques and improvement of Compression algorithm.

SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

References

D.A. Huffman, "**A Method for the Construction of Minimum Redundancy Codes**", Proceedings of the I.R.E., September 1952, pp 1098-1102

Burrows M and Wheeler D (1994), **A block sorting lossless data compression algorithm, Technical Report** 124, Digital Equipment Corporation

J. L. Bentley, D. D. Sleator, R. E. Tarjan, V. K. Wei, **A Locally Adaptive Data Compression Scheme Communications** , ACM-Vol. 29, No. 4, 1986

Erskine, Ralph, "**Enigma's Security: What the Germans Really Knew**", in "Action this Day", edited by Ralph Erskine and Michael Smith, pp 370–386, 2001.

Smith, G.C.; Fiddes, C.C.Hawkins, J.P.; Cox, J.P. **Some possible codes for encrypting data in DNA**. Biotechnol let. 2003.25,1125-1130

Nozomu Yachie, Kazuhide Sekiyama, Junichi Sugahara, Yoshiaki Ohashi, and Masaru Tomita; **Alignment-Based approach for durable data storage into living organisms**, Biotechnol. Prog. 2007, 23, 501-505

Cox, J.P. **Long term data storage in DNA**. Trends Biotechnol. 2001, 19, 247-250

Vercoutere, W. et al, **Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel**. Nat. Biotechnol,19, 248-252, 2001