ISSN:2231-0711
www.ijcset.net

Image Compression using DNA sequence

Prof. Samir Kumar Bandyopadhyay
Dept. of Computer Sc. & Engg, University of Calcutta
92 A.P.C. Road, Kolkata – 700009, India
skb1@vsnl.com

Suman Chakraborty
B.P. Poddar Institute of Management and Technology
137, V.I.P. Road, Kolkata – 700052, India
sumanc75@gmail.com

Abstract -Every day an enormous amount of data is stored, processed and transferred digitally. Expenditure of these processes is very much related with amount of data. In order to minimize the processing cost if same information can be presented by reducing amount of bits. In case of image this can be achieved by image compression. In this paper an attempt has been made to present an approach of image compression using DNA sequence. Here two techniques have been presented which compresses an image by more than 99%.

1. INTRODUCTION

The term DNA sequencing refers to sequencing methods for determining the order of the nucleotide bases—adenine(a), guanine(g), cytosine(c), and thymine(t)—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematic. The advent of DNA sequencing has significantly accelerated biological research and discovery. The speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes [1-5].

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster [6-8].

A massage representing a DNA sequence, with the combination of a, c, t, g is equivalent to DNA sequence of the original data. Compressing the DNA sequence by-1) 2-bits encoding method, 2) Exact matching method, 3) Approximate matching method, 4) For the approximate matching method. These compression techniques would be produce equivalent digital form of the DNA sequence and one of procedure will produce minimum number of bits [9-10].

In this paper DNA sequence of an animal is used whose image we compress. Compression rate depends upon DNA sequence length.

2. PROPOSED METHODS FOR DNA SEQUENCE COMPRESSION

We consider three standard edit operations in our approximate matching algorithm. These are:

1) *Replace*. This operation is expressed as (*R*, *p*, *char*) which means replacing the character at position *p* by character *char*.
2) *Insert*. This operation is expressed as (*I*, *p*, *char*), meaning inserting character *char* at *p*.
3) *Delete*. This operation is written as (*D*, *p*), meaning deleting the character at position *p*.

Let *C* denote "copy," then the following are two ways to convert the string
"gaccttca" to "gaccgtca" via different edit

```
C C C C R C C C
g a  c c g t  c a
g a  c c t t  c a
```

or

```
C C C C I C D C C
g a c c g t   c a
g a c c   t t c a
```

The first involves one replacement operation. The second involves one insertion and one deletion. It can be easily seen that there are infinitely many edit sequences to transform one string to another A list of edit operations that transform a string *v* to another string *u* is called an *Edit Transcription* of the two strings [9].This will be represented by an edit operation sequence $\lambda(u,v)$ that orderly lists the edit operations. For example, the edit operation sequence of the first edit transcription in the above example is
λ(*gaccgtca*, *gaccttca*) = {(*R*,4, *g*)} ; and for the second edit transcription, λ(*gaccgtca*, *gaccttca*) = {(*I*,4, *g*),(*D*,6)}.If we know the string *u* and an edit operation sequence $\lambda(u,v)$ from *v* to *u*, then the string *u* can be

constructed correctly using λ.There are many ways to encode one string given another. Using the above example, we describe four ways to encode "gaccgtca" using string "gaccttca" supposing that the string "gaccttca" is located earlier in the sequence.

1) 2-bits encoding method. In this case,we can simply use 2 bits to encode each character; i.e., 00 for *a*, 01 for *c*, 10 for *g*, 11 for *t*. Thus "10 00 01 01 10 11 01 00" encodes "gaccgtca." It needs 16 bits in total.
2) Exact matching method.Wecan use (repeat position, repeat length) to represent an exact repeat. This way, for example,if we use 3 bits to encode an integer, 2 bits to encode a character, and use 1 bit to indicate if the next part is a pair (indicating an exact repeat) or a plain character, then the string "gaccgtca" can be encoded as {(0,4), *g*,(5,3)}, relative to "gaccttca." Thus, a 17-bit binary string "0 000 100 1 10 0 101 011" is required to encode the {(0,4), *g*,(5,3)}.
3) Approximate matching method. In this case, the string "gaccgtca" can be encoded as {(0,8),(*R*,4, *g*)}, or "0 000 111 100 100 10" in binary, with *R* encoded by 00, *I* encoded by 01, and *D* encoded by 11, and 0/1 indicating whether the next item is a doubleton or triple. A total of 15 bits is needed.
4) For the approximate matching method, if we use the edit operation sequence, then the string "gaccgtca" can be encoded as {(0,8),(*I*,4, *g*),(*D*,6)}, or "0 000 111 1 01 100 10 1 10 110," in total 21 bits.

Method I

Step1. Take DNA sequence of an animal whose image will compress
Step2. This DNA sequence assign as a DNA sequence of image
Step3. Four nucleotides in DNA sequence (A,C,G,T),2 bit is used to represent 4 nucleotides. Like 00 for A,01 for C,10 for G and 11 for T.
Step4. Represent DNA sequence in digital form D
Step5. D is the compress binary form of that image.

Method II

Step1. Take DNA sequence of an animal whose image will be compressed
Step2. This DNA sequence assign as a DNA sequence of image
Step3. Compress DNA sequence using -1) 2-bits encoding method, 2) Exact matching method, 3) Approximate matching method, 4) For the approximate matching method, one of the method will produce minimum no. of bits.
Strp4 Compress DNA sequence is the compress digital form of the image.

3. RESULTS

The above described method I and Method II are applied to the following two figures and results are shown encouraging.

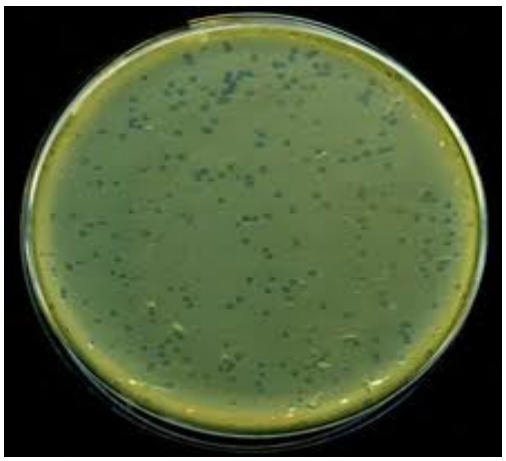

Figure-1.Lysis plaques of lambda phage on *E. coli* bacteria [11]

This is the image of Lambda phage. Lambda phage is a virus particle consisting of a head, containing double-stranded linear DNA as its genetic material, and a tail that can have tail fibers. Lambda DNA is 48,502 base pairs in length[12].

Size of the image 1366(No. of pixel in Y-axis)*768(No. of pixel in X-axis)*32(No. of bits per pixel)=33570816 bits

For Method I

Two bits represent one nucleotide. In case of Lamda phage, total no of nucleotide 48502.Total no. of bits required for DNA sequence 48502*2=97004.This is the size of the image after compression.

Compression rate= (97004/33570816)*100=0.288%

For Method II

Lambda phage has 48502 no of nucleotide. DNA sequence compressing using 1) 2-bits encoding method or 2) Exact matching method or 3) Approximate matching method or 4) For the approximate matching method. These four methods compress 8 nucleotide pair at a time.

In 48502 no. of DNA sequence has 6063(48502/8=6062.75) no. of 8-pair DNA.
After applying Approximate matching method , total no. of bits 90945(6063*15=90945).

Compression rate=(90945/33570816)*100=0.270%.

Example-2

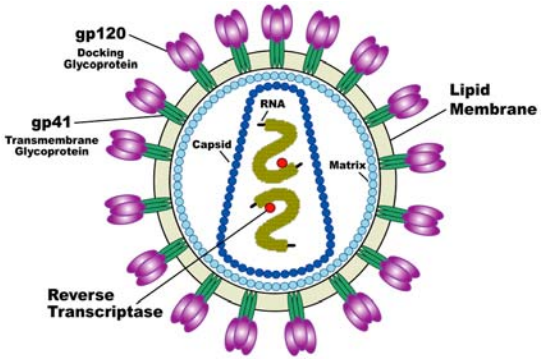

Figure-2. HIV Virus [13]

Size of the image 800(No. of pixel in Y-axis)*557(No. of pixel in X-axis)*24(No. of bits per pixel) =10694400bits.

HIV-1 is composed of two copies of single-stranded RNA enclosed by a conical cap Sid comprising the viral protein p24, typical of lentiviruses (Figure 2). The RNA component is 9749 nucleotides long [13]

For Method I

Two bits represent one nucleotide. In case of HIV Virus total no of nucleotide 9749.

Total no. of bits required for DNA sequence 9749*2=19498.This is the size of the image after compression.

Compression rate=(19498/10694400)*100=0.182%

For Method II

HIV Virus 9749 no of nucleotide. DNA sequence compressing using - 1) 2-bits encoding method or 2) Exact matching method or 3) Approximate matching method or 4) For the approximate matching method. These four methods compress 8 nucleotide pair at a time.

In 9749no. of DNA sequence has 1219 (9749/8=1218.625) no. of 8-pair DNA.

After applying Approximate matching method, total no. of bits (1219*15=18285).

Compression rate is 18285/10694400)*100=0.170%

4. CONCLUSIONS

An attempt has been made to present an approach of image compression on DNA sequence. Here two techniques has been illustrated which compresses an image by more than 99%.

REFERENCES

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell, Garland Publishing, New York & London, 1994.

[2] A. Apostolico, D. Breslauer, Z. Galil, Optimal parallel algorithms for periods, palindromes and squares, in: Proceedings of the International Colloquium on Automata, Languages, and Programming, 1992, pp. 296–307.

[3] A. Apostolico, D. Breslauer, Z. Galil, Parallel detection of all palindromes in a string, Theoretical Computer Science 141 (1995) 163–173.

[4] D. Breslauer, Z. Galil, Finding all periods and initial palindromes of a string, Algorithmica 14 (1995) 355–366.

[5] C.C. Chang, C.C. Lin, C.S. Tseng, W.L. Tai, Reversible hiding in DCT-based compressed images, Information Sciences 177 (2007) 2768–2786.

[6] C.C. Chang, T.C. Lu, Y.F. Chang, R.C.T. Lee, Reversible data hiding schemes for deoxyribonucleic acid (DNA) medium, International Journal of Innovative Computing, Information and Control 3 (2007) 1–16.

[7] C.C. Chang, W.C. Wu, Y.H. Chen, Joint coding and embedding techniques for multimedia images, Information Sciences 178 (2008) 3543–3556.

[8] C.T. Clelland, V. Risca, C. Bancroft, Hiding messages in DNA microdots, Nature 399 (1999) 533–534.

[9] M. Crochemore, W. Rytter, Jewels of Stringology, World Scientific, 2002.

[10] European Bioinformatics Institute, <http://www.ebi.ac.uk/>.

[11] http://en.wikipedia.org/wiki/Lambda_phage.

[12] http://www.neb.com/ nebecomm/ products /productn 3011.asp.

[13] http://en.wikipedia.org /wiki/ Structure_ and_ genome _of_HIV