Image Processing Of DNA Chromatogram For Removal Of Errors and Finding Maximum Length DNA Sequence

Swaroopa S.Kulkarni#1, Dr. T.B. Mohite-Patil*2

#E&TC Department, D.Y.Patil College of Engg. & Tech,Shivaji University
KasbaBawda,Kolhapur,Maharashtra,India
1swaroopa_kulkarni@yahoo.co.in

*Dr. D.Y.Patil pratishthan's D.Y.Patil college of Engg. & Tech,Shivaji University
Salokhe Nagar,Kolhapur, Maharashtra,India
2tanaji.mpatil@gmail.com

Abstract— *This paper gives a brief explanation about image processing of a chromatogram that can be used to find maximum length DNA sequence by removing errors in problematic areas of that chromatogram. Image processing facilitates the analysis and removes confusion in positions of 4 nucleobases of a DNA. Thus for getting the maximum length of a DNA sequence obtained from a sequencer and using Reverse and Compliment method, image processing is useful in reducing the time taken before analysis of DNA or other work related to DNA, compared to time required for manual work that would be involved otherwise.*

***Keywords*— DNA, Chromatogram, Image Processing, Reverse and Compliment method,**

I. INTRODUCTION

This paper explains about a chromatogram and how image processing of a chromatogram can be used to get maximum sequence of a DNA. It is useful in many fields like Bio-Chemistry, Bio-Technology where DNA sequencing is frequently implemented for analysis and research work.

II. DNA AND CHROMATOGRAM

DNA [1], deoxyribonucleic acid, is responsible for storing genetic information in cells of the body of every living individual. DNA consists of a helical ladder like structure. The

DNA consists of two long polymers of simple units called nucleotides. Each nucleotide contains a nucleobase, and the sequence of these bases is the key to store the genetic information.

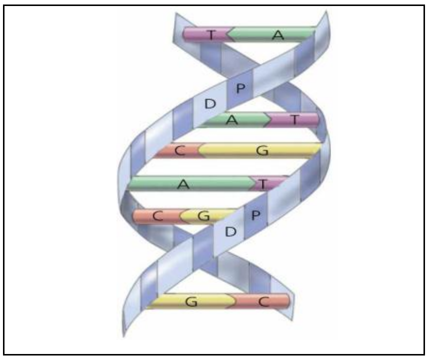


Fig.01 DNA structure

The genetic information is coded in the DNA using only four different types of nucleobases, namely: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T).

A chromatogram, sometimes also called as electropherogram is the visual representation of a **DNA** sample produced by a sequencing machine, called as digital Sequencer. The output is given in .ab1 file format, which can be opened using special applications like, Chromas and Sequencher.

Following figure shows a chromatogram. In this, four colours are used to represent four nucleobases. For example, blue for C, Black for G etc.

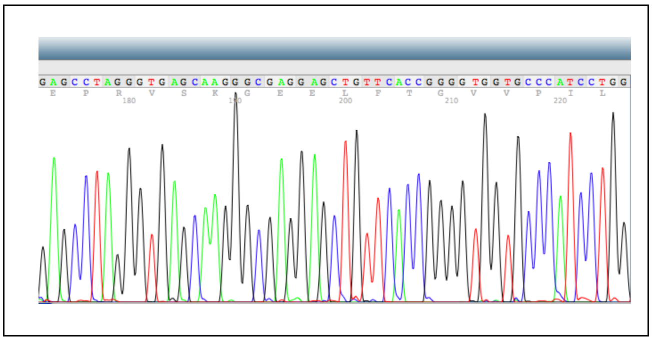


Fig.02 Chromatogram of a DNA sequence

III. PROBLEM STATEMENT

Digital sequencer gives the DNA sequence in .ab1[2] format. The ab1 file is a file type produced by Data Collection software generating sequencing data, with the extension ".ab1". This .ab1 file can be opened using applications like, Chomas [3] or Sequencher [4]. But before its analysis, it must be checked to ensure there is no mistake in the sequence. This is done by reverse and compliment method. In this method, the sequence is again crosschecked and mistakes, if any, are removed by observing both the chromatograms-one is the original forward sequence and the other is reverse and compliment of the reverse sequence. Wherever a mistake is found, taking a proper decision in this case of confusion includes manual work. This task is time consuming also. Thus instead of manual observation and decision making, image of the chromatogram is processed in MATLAB [5]. The maximum peak out of four peaks, (corresponding to four bases) must be found in the chromatogram. Then a similar procedure is used to compare the original chromatogram and the reverse and compliment form of the reverse sequence chromatogram.

IV. METHODOLOGY

A. *Reverse and compliment method:*

Normally, DNA occurs as a double strand where each A is paired with a T and vice versa, and each C is paired with a G and vice versa. The reverse complement of a DNA sequence is formed by reversing the letters, interchanging A and T and interchanging C and G. Thus the reverse complement of ACCTGAG is CTCAGGT.

Following images are showing the chromatogram of a DNA sequence (Fig 03) and the one obtained after reverse and compliment method (Fig 04)

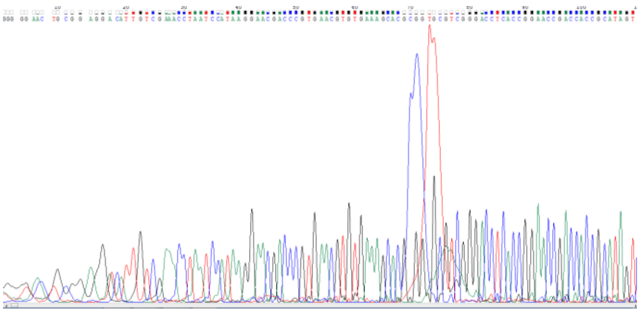


Fig 03: Original DNA sequence

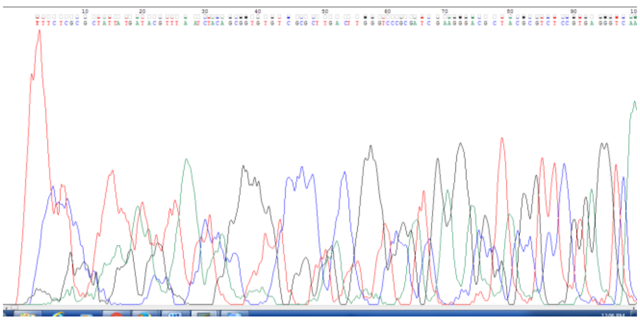


Fig 04: Reverse and compliment of original

Thus the bases in original sequence and the bases in its reverse and compliment of the reverse sequence should essentially match. But in practice, they do not sometimes match due to some errors. The mismatch indicates that there is either error in the original sequence or it is in the reverse complement sequence.

To remove these errors, generally manual observation and replacement of the wrong nucleobase with correct nucleobase is done, which increases time duration elapsed before actual analysis of that DNA sequence. This duration can be reduced if these two chromatograms are processed for every error using image processing.

1) *Fetching images:* The two images, one containing an original DNA sequence and the other containing reverse and compliment sequence are to be stored in the same folder as that of the image processing tool project folder. They are obtained by .ab1 format, opened using either a free open software application like 'Chromas' or a licensed one like 'Sequencher'.

2) *Finding the maximum peak:* Different colours indicate four different levels so firstly, the maximum amplitude is to be found within the same image and secondly it must be compared with the maximum amplitude in the reverse and compliment sequence. Whichever in the two is the greater represents the most suitable nucleobase at that position.

3) *Representing the final DNA sequence in text /word format:* The DNA sequence is to be displayed again in text or word format so as to compare it with another DNA. This is done by introducing a GUI(Graphic User Interface) for simplicity.

V. RESULTS

Following figure shows the screenshot of the GUI introduced.

Each tab enables the user to distinctly see one of the four sequences of nucleobases chromatogram and its reverse and complement chromatogram.

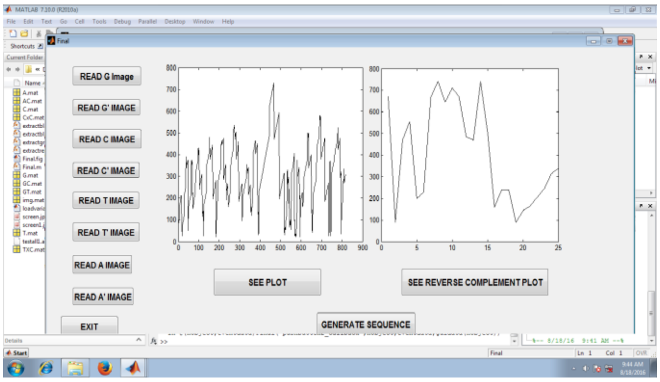
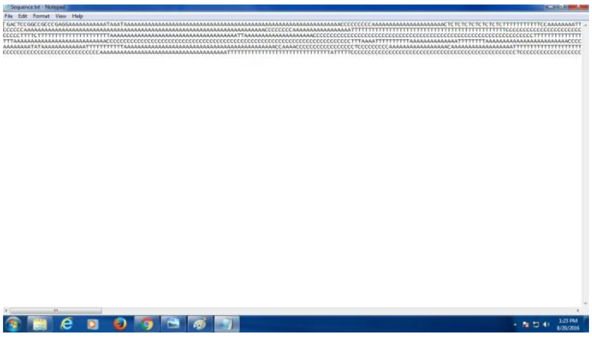


Fig. 05: A Graphic User Interface to observe each sequence chromatogram and its reverse compliment

Using this GUI user can view the chromatogram of a base in the respective colour which is used for its representation. For example, for guanine (G) graph is displayed in black colour, whereas for Cytosine (C), graph is displayed in blue colour. When the Generate Sequence tab is clicked, it adds a word file in the same folder as that of the code. In this file DNA sequence of maximum possible length is given in text format. It can be then readily used for further analysis.



[illegible]

Fig. 06: Screenshot of the text sequence obtained

VI. Conclusions

Thus using image processing the manual work included before analysis of a DNA chromatogram is reduced which also saves time. This method can be used by a non-technical person also as the image processing is completed entirely within the code and it is interfaced using only tabs in GUI. Roughly each sequence of maximum possible length can be found out in one or two minutes. It is very small compared to manual error removal time which is ten to fifteen minutes.

VII. FUTURE SCOPE

This work can be extended for other operations that are involved before comparison of the two chromatograms for example, alignment of two DNA sequences- one is forward and the other is Reverse and Complement of the reverse sequence is essential. This can be done as an extension of this work. Also, programming tools other that image processing tool of MATLAB can be tried out for above operation.

VIII. Acknowledgment

We offer our profound gratitude to the management of D.Y. Patil college of Engineering and technology for giving us an opportunity of exposure to use our theoretical knowledge with our practical experience, in a professional environment, and allowing us to use their labs.

We are also thankful to Prof. Dr. S. P. Govindwar (Former HOD, Department of Biochemistry, Shivaji University, Kolhapur) and Prof. Dr. K. D. Sonawane (HOD, Department of Micro-Biology, Shivaji University, Kolhapur) from Shivaji University, Kolhapur for their valuable guidance and support.

IX. References

[1] S. L. Wolfe (Ed.) "Molecular and cellular biology." *Wadsworth Publishing Company*, 1993.
[2] https://www.fileinfo.com/extension/ab1
[3] https://www.technelysium.com.au/wp/chromas
[4] https://www.genecodes.com
[5] www.mathworks.com/products/matlab