CHAPTER 6

CONCLUSION AND SCOPE FOR FUTURE ENHANCEMENTS

Chapter 6. Conclusion and Scope For Future Enhancements

6.1. Conclusion

Computer Engineering being the supportive Engineering field, the main objective is to support and improvise manual systems existing in various domains by securing data and optimizing processing with minimum resources. To emphasize on re-utilization of existing resources and optimize time, Distributed Computing is carried out. My research work is pertaining to applying Distributed Computing to the newly emerging domain of BioInformatics, with an urge for inter-disciplinary research work. My contribution to my field of Computer Engineering is to secure the data and optimize time involved in processing various applications pertaining to this new domain of BioInformatics with a view to tackle various issues while applying Distributed Computing. As a subsidiary to optimize time in Distributed Computing, additional work related to designing and development of algorithms using an altogether innovative approach of Signal Processing using Wavelet Transforms is applied for data reduction to the data which needs to be transmitted and processed for Distributed Computing. The security aspect of data during transmission is achieved since the DNA sequences being transmitted are transmitted in a form of digital signal. Digital signal is an encoded form of original DNA sequence.

The implementation of Web-based application using distributed approach for storage and retrieval of DNA sequence data has proved to be efficient because each component of the application, whether, it is the Web-Server, Database-Server or File-Server are all scattered on different machines each having heterogeneous platforms. These heterogeneous platforms have helped in using all the resources of the organisation, without the need of expensive special-purpose computer storing the DNA sequencing data or for performing any DNA analysis. The database that is designed has also proved efficient because the raw DNA sequencing data can be stored in an optimized manner, which in turn is effective in executing with reasonable efficiency any type of query based on combination of data using relational concept.

The signal processing with the help of wavelet transform used for analysis of DNA sequences, is a significant improvement over the conventional methods of string or regex based processing. The research work represented DNA sequences by using Integer representations of binary values, EIIP, dipole moments and used this as signals. Haar Wavelets though are primitive wavelet transforms have proved to be efficient when applied to the DNA sequencing data. The signal processing approach proved very effective because it reduces the DNA sequences in terms of the number of elements, preserving the details of the original signal. The information of original DNA sequence, in terms of position as well as content, is preserved in a signal compressed using Haar wavelet transform. Thus, using transformed data in place of original sequence for data analysis was possible, because there is no loss of data due to transformation. The search techniques applied over data transformed using wavelet transformations, have a time complexity of

O(log n). Hence, the use of Haar wavelet transforms have proved to be more efficient over other classic string-based techniques, not just in terms of processing time but also in terms of storage space requirement since it does not require complex data structures and performs transforms in place thus avoiding overheads of storage while processing. Haar Wavelets have also proved to give best results in recognizing duplicate reads obtained after DNA sequencing process. The Haar Wavelet Transforms have also proved to be efficiently effective in finding Short Tandem Repeat Regions. The study results have also found the use of detail co-efficents of Wavelet Transforms, which usually represent the fluctuations in the signal, to be very effective in finding the Short Tandem Repeat regions. The recognition of Short Tandem Repeats is possible using Haar Wavelet Transforms, without the need of supplying additional parameters like reference or pattern etc.

All the three algorithms, directly or indirectly are applied in context of reducing the data, by removing irrelevant portions from the given set of DNA data, thus, applied in context of optimizing space and time. This optimized data then can further be used in optimizing the time involved in transmission of data when Distributed Computing is applied for DNA sequence analysis.

6.2. Scope for Future Enhancement

BioInformatics is still an evolving inter-disciplinary area of research and development. Hence, there are many possibilities of improvements in this field of research studies. Moreover, due to evolving technology there are large numbers of new platforms as well as ubiquitous and pervasive computing creeping in the field of computing. Hence, this is another area, which can be probed for improving the work of Biologists.
As part of the future enhancement, the intention is to extend this work and apply distributed computing using processing environments like

- HT Condor
- Hadoop
- Matlab Distributed Computing toolbox

Wavelet Transforms can be futher used in various other analysis techniques, such as:
In multiple sequence alignment of DNA sequences
In de-novo assembling of reads to form contigs
Identifying Single Nucleotide Polymorphism (SNPs)
Base Calling of sequenced data
The enhancement to this work is, of applying wavelet transforms on large sequences using distributed tool box of MATLAB, HT Condor or Hadoop. The research also aims to identify existence or non-existence of repeat regions in newly sequenced organisms and plants, other than the ones used for study, and apply Wavelet Transforms to other related sequence analysis techniques.