Improving Bioinformatics tools and approaches of Immunoglobuline sequencing, alignment and assembly

Nazia Tasnim[1], Md. Istiak Hossain Shihab[1], Md. Moqsadur Rahman[2], and Ruhul Amin[3]

[1]Student, Department of Computer Science and Engineering, Shahjalal University of Science and Technology
[2]Lecturer, Department of Computer Science and Engineering, Shahjalal University of Science and Technology
[3]Assistant Professor, Department of Computer Science and Engineering, Shahjalal University of Science and Technology

Introduction

This thesis aims to build or improve some currently existing Bioinformatics tools, apply different algorithms on sequencing, alignment and assembly and also, add a new Machine Learning approach to it. Primarily we are working on preexisting bioinformatics tools and trying to research and Implement a new tool to provide more service.

Motivations

Bioinformatics is an interdisciplinary field of science, bioinformatics that brings together the knowledge of biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data.The domain is faced with a strong demand for immediate solutions; because the genomic data that are being uncovered encode many biological insights whose deciphering can be the basis for dramatic scientific and economic success. Multi-faceted obstacles exist here. One of many is the continuing development of high-throughput measurement techniques that lead to a constant increase in the volume of data available for analysis.Even with sophisticated methods for information reduction, data-archiving costs can be considerable. The fact that information is available is not sufficient; it also has to be made accessible and useable. Barriers to its use include a lack of standardized formats, a lack of common interfaces to data, inconsistency in identifiers for biological entities, insufficient support for data-exchange frameworks and insufficient visibility. There are too many ways to classify and organize data. There are too many standards, and within any given standard, there are different ways of implementing the codification of data.The challenge is not only to manage this data, but to integrate data that has been analyzed to different degrees of complexity using different generations of technology. Bioinformatics has many far-reaching exciting fields where it can contribute significantly. From personalized healthcare to drug discovery, forensic analysis, biodiversity management and even inferring the tree of life. So, overall we find it to be a branch of knowledge teeming with challenges that we may contribute to solve.

The Research Question

Despite our general idea regarding the domain, we haven't yet decided on a specific topic. So far, we have been exploring the various domains with our supervisors and continuously working to settle down as soon as possible.

Preliminary Literature Review

DeepAnnotator: Genome Annotation with Deep Learning[1]
Genome annotation is the process of labeling DNA sequences of an organism with its biological features, and is one of the fundamental problems in Bioinformatics. In this paper, the authors introduce a method (DeepAnnotator) that uses RNN and LSTM to demonstrate the potential of deep learning networks to annotate genome sequences, and evaluate different approaches on prokaryotic sequences.

NanoBLASTer: Fast alignment and characterization of Oxford Nanopore single molecule sequencing reads[2]
The quality of long DNA sequence reads has been, to date, lower than other technologies, causing great interest to develop new algorithms that can make use of the data.To address these challenges the authors developed a new read aligner called NanoBLASTer specifically designed for long nanopore reads that produces longer alignments with higher overall sensitivity than previous algorithms and is also significantly faster.

pRESTO: A Toolkit for Processing High-Throughput Sequencing Raw Reads of Lymphocyte Receptor Repertoires[3]
pRESTO is a handy tool for processing DNA Reads from high - throughput Lymphocyte Receptor studies. It processes raw sequence to produce error-corrected. This toolkit can be used easily to align, assemble, filter and join sequences according to their quality. So far, we have tested this on some datasets provided by SRA and now comparing its output with our algorithms.

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins[4]
The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in 1970. The algorithm essentially divides a large problem (e.g. the full sequence) into a series of smaller problems, and it uses the solutions to the smaller problems to find an optimal solution to the larger problem.

SHMPrep: A Fast and User-Friendly Program for Preprocessing Paired-end Immunoglobulin Sequence Data from Illumina Platforms[5]
Next generation sequencing platforms, predominantly the Illumina MiSeq, are now routinely used to sequence immunoglobulin (Ig) genes to measure effects such as somatic hypermutation (SHM). There is a need for software that can be used widely by experimentalists who may not have computational expertise or access to specialized computers. We have developed SHMPrep, a very fast and user-friendly program for preprocessing paired-end immunoglobulin data from the Illumina MiSeq. The program aligns paired-end reads, performs error correction and collapses sequences. Implementation of efficient string matching algorithms such as suffix automata and string alignment algorithms such as k-band global alignment, enable processing of an entire MiSeq run of 20 million reads in approximately 30 minutes using a standard desktop computer. SHMPrep can also handle barcodes (UMI) and multiple sets of primers, making it ideal for working with Ig repertoire data.

Methodology

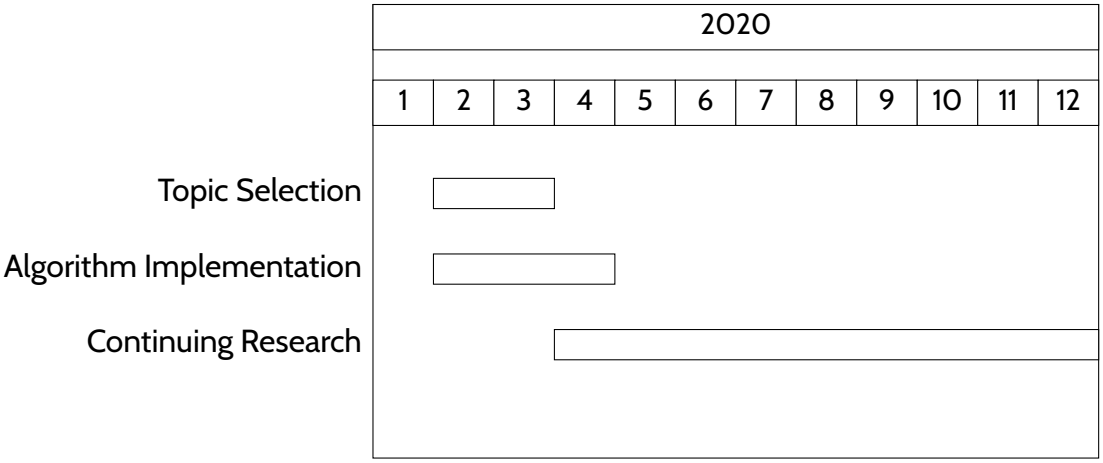
Besides reading the associated literatures of our current tasks, we are currently working on a significant amount of real-life and simulated Immunoglobulin datasets in .fastq and .fasta formats, provided by our external supervisor. We are performing different operations on these data to produce summaries

and interpret these outcomes. Currently our work is mostly algorithm based and there are chances that we would be shifting to a Machine-learning based approach soon.

Findings

Our work so far could potentially help in finding new generation Genes, function prediction, finding homology in proteins,identify, diagnose and potentially develop treatments for genetic diseases, help with predictive and preventive medicines etc.Drug-likeness could be predicted by genetic algorithm and neural network based approaches. Personalized medicine can be defined widely as a model of healthcare that is predictive, personalized, preventive and participatory. A successful and reliable drug design process could reduce the time and cost of developing useful pharmacological agents.

Timeline



References

1. Amin, R., Yurovsky, A., Tian, Y. & Skiena, S. Deepannotator: Genome annotation with deep learning. https://ruhulsbu.github.io/publication/bcb2018/ (2018).
2. Amin, R., Skiena, S. & Schatz, M. C. Nanoblaster: Fast alignment and characterization of oxford nanopore single molecule sequencing reads. https://ruhulsbu.github.io/publication/iccabs2016/ (2016).
3. JA, V. H. *et al.* presto: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. https://www.ncbi.nlm.nih.gov/pubmed/24618469 (2014).
4. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. https://pubmed.ncbi.nlm.nih.gov/5420325-a-general-method-applicable-to-the-search-for-similarities-in-the-amino-acid-sequence-of-two-protein (1970).
5. Ruhul, M. *et al.* Shmprep: A fast and user-friendly program for preprocessing paired-end immunoglobulin sequence data from illumina platforms. http://www.ams.sunysb.edu/~maccarth/software.html (2018).

6. Wooley JC, L. H. Challenge problems in bioinformatics and computational biology from other reports, catalyzing inquiry at the interface of computing and biology. https://www.ncbi.nlm.nih.gov/books/NBK25461/ (2005).

7. Fuller, J. C. *et al.* Biggest challenges in bioinformatics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3615659/ (2013).

8. Pevzner, P. A. & Jones, N. C. An introduction to bioinformatics algorithms (2013).

9. Pevzner, P. A. & Compeau, P. Bioinformatics algorithms: An active learning (2013).

10. Dahiya, B. P. Bioinformatics impacts on medicine, microbial genome and agriculture. www.phytojournal.com/archives/2017/vol6issue4/PartAB/6-4-357-863.pdf (2017).

11. Singh, H. Bioinformatics: Benefits to mankind. http://sphinxsai.com/2016/ph_vol9_no4/1/(242-248)V9N4PT.pdf (2016).

12. Quintana, Y. Bioinformatics for discovery and global collaborations. https://allergen-nce.ca/wp-content/uploads/Quintana_key-messages.pdf (2018).

67891011123