Master's Thesis Proposal

Comparison of sequence alignment on different hardware settings

Author:

Ali Alloush (22108595)

1 - Problem Introduction

Genetic sequencing is the technique used to determine the genetic composition and sequence in a DNA or RNA molecule. These sequences encode the biological information that cells use to develop and operate [1]. The sequencing of DNA/RNA is paramount in biotechnology in many different fields, such as but not limited to;

- Medical Biotechnology
- Pharmaceutical Biotechnology
- The COVID-19 pandemic [2]

While greatly influential in all these fields, in Medical Biotechnology, sequencing plays a key role in studying the human genome to identify genetic disorders and therein finding ways to prevent or cure them. Similarly, with the COVID-19 pandemic, the ability to sequence the virus in a quick manner allowed the biotech industry and pharmaceutical companies to develop the necessary vaccines and medicines needed [2].

One of the steps in genetic sequencing is sequence alignment. This is performed by comparing similarities in genomic sequences. Usually, one of these sequences is already known (such as that of a human gene or a disease, for example) and the other is 'compared' to it. A score is assigned to the matching based on the likelihood that these two genes have the same origin with respect to possible genetic mutations that may have occurred over time.

Sequence alignment can generally be grouped into two major categories;

- Global alignments, where the entirety of a sequence is aligned
- Local alignments, where regions of similarity are aligned [3]

Hybrid methods of these two methods also exist, such as;

- Burrows Wheeler transformation
- Pairwise alignment
- Multiple sequence alignment

Over the last years, there has been substantial improvement in next generation sequencing methods and machines. The limiting factor in current systems, however, is the analysis of this generated sequence data, namely in the process of sequence aligment. [5]

2 - Research question

The aim of this thesis is to compare the performance of various sequence alignment methods on various computer hardware.

The question raised is;

How does the performance of genetic sequence alignment tools differ when performed on different hardware settings?

Additional questions of observations could also be raised, such as:

1. Are any pieces of hardware inherently 'better' in terms of performance?
    a. If not, are there different sequencing methods that perform better with different hardware, i.e., would a specific computer architecture have a higher affinity for a specific sequencing method
2. Are there other overlooked parameters that may also influence the outcome, such as:
    a. Long or short reads
    b. Local or global alignment
    c. Performance per Watt

3 - Methodology

For this project, we plan on using multiple samples of both long and short reads. The motivation for using reads of different lengths is to observe any effects of long and short read processing times. The motivation for using multiple samples is to have more consistent results among a wider dataset.

For these various reads, we plan on running multiple DNA / RNA alignment algorithms spanning over local and global alignment using some but not limited to the methods mentioned in the problem introduction section of this proposal.

After outlining the exact tests to be run over the various samples, we can begin with running the tasks on various hardware. The hardware on which these sequence alignment methods are to be tested on may contain but are not limited to:

- Single core processor
- Multi core processor
- Two multi core processors connected to a mainboard
- With GPU acceleration

Initially, the tests will be run on a personal 'average' pc. Tests would be run on a single core processor, a multi-core processor, and potentially with GPU acceleration. The exact specifications of this computer will also be provided. The main purposes of this testing are to optimize the testing criteria and specifications. Also, while not spanning over all the computer architectures, we believe for it to be diverse enough to display any patterns that may be of interest, as well as bringing attention to unconsidered parameters that may also be influential in terms of performance and productivity.

The motivation for running these tests on an 'average consumer' pc is due to the more recent availability of open source alignment methods and people exhibiting an interest in conducting sequence alignment on their own personal computers.

Following this, the same primary tests on single core, multi core, and GPU acceleration will be run again on a server at "Technische Hochschule Deggendorf", with the addition of two multi core processors on a mainboard.

Following the testing, the data acquired will be analyzed. Each testing platform will have independent results in terms of runtime based on different alignment methods.

In terms of the datasets mentioned, we will be using open data sets.

Also, in terms of the alignment tools that will be used, we aim to use open-source tools, such as but not limited to Clustal omega [6], or HISAT2 [7]

4 - Goal

The goal is to output a coherent display of data and graphs demonstrating the influence of different computer hardware and architectures on varying DNA / RNA sequence alignment methods. We aim to being able to streamline this step in genetic sequencing in order to optimize the processing time, result quality, and inherently, the processing costs.

5 - Time Plan

Month	Targeted work plan
1	Literature research (approx. 2 weeks) Defining scope of alignment methods to be tested (approx. 2 weeks)
2	Script writing for testing
3	Running tests
4	Running tests Data acquisition and analysis
5	Data analysis Writing thesis
6	Writing thesis

References

[1] https://www.genome.gov/genetics-glossary/DNA-Sequencing

[2] https://www.blalbiotech.com/blog/importance-of-dna-sequencing-in-biotechnology/#:~:text=In%20Medical%20Biotechnology&text=It%20involves%20the%20use%20of,genes%20associated%20with%20the%20disease.

[3] Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol. 2011 Oct 27;6(1):25. doi: 10.1186/1748-7188-6-25. PMID: 22032267; PMCID: PMC3223492.

[4] Mount DM. (2004). *Bioinformatics: Sequence and Genome Analysis* (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 978-0-87969-608-5.

[5] Arram, J., Tsoi, K.H., Luk, W., Jiang, P. (2013). Hardware Acceleration of Genetic Sequence Alignment. In: Brisk, P., de Figueiredo Coutinho, J.G., Diniz, P.C. (eds) Reconfigurable Computing: Architectures, Tools and Applications. ARC 2013. Lecture Notes in Computer Science, vol 7806. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36812-7_2

[6] http://www.clustal.org/omega/

[7] http://daehwankimlab.github.io/hisat2/