INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 6)
Available online at: www.ijariit.com

Cloud and high-performance computing tools for sequence alignment in bioinformatics

Dindi Dhanunjai
dhanu0537@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu

Kotur Guna Pragna
gunapragna1234@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu

ABSTRACT

Pairwise Sequence alignment and Multiple Sequence Alignment (MSA) are the major challenges for the Bioinformatics application. In recent years there is good development and progress in this field by introducing good Tools or software, through which the High-Performance computing in alignment is achieved. This kind of alignment needs the best Database for searching the protein's and nucleotides. This paper talks about recent tools that evolved for Database searching and parallel computing of huge biological data. This is useful for researchers or practitioners who are working for bioinformatics applications.

Keywords— *Bioinformatics, Cloud, Clustering, Parallel and Distributed computing, Multicore, Supercomputers, Database searching, Pairwise, and Multiple Sequence Alignment.*

1. INTRODUCTION

Bioinformatics is the advance use of computers to solve biomedical and biological challenges. It is the application of use of information technology for visualization, integration, analyzing [1] and mining. It deals with the management of biological and genetic information. Bioinformatics is the Key reason for exploring drug discovery and development. Biologists and Computer scientists together called Bioinformaticians. The biologists will collect molecular data, gene expression, DNA and Protein sequences, etc. Computer scientists who are aware of Mathematics, Statistics, physics, chemistry, etc. They will develop the software, tools, and algorithms to analyze and store the data in the database [2] [3]. The job of these two people leads to form a people who are called Bioinformaticians who study the biological problems by analyzing the Biological and molecular data. Uses of the bioinformatics are DNA analysis (Genome sequencing), Protein analysis (both Structure and Sequence), Drug designing, etc. Everything [4] which is used in bioinformatics application is retrieved and stored in the database. So, the first task is to search the required protein and nucleotide sequences from the database with less time complexity. Among all these applications, Sequence alignment [5] is the major challenge that consists of pairwise and Multiple Sequence Alignment (MSA). In case of two sequences it is considered as pairwise alignment and if more then it is MSA [6]. Global alignment suits the comparison of sequences but when we consider them part of the sequences or with gaps then local alignment comes into the action. Maximum all the Multiple Sequence Alignment methods use one these two methods: Needleman and Wunsch algorithm (global alignment), Smith and Waterman algorithm (local alignment).

Multiple nucleotides [7] or amino acid sequence alignment lead to increase the sensitivity of alignment. This alignment is performed to hit any of these two tasks: Characterize protein families, identify the shared regions of homology in MSA and Determination of the consensus of all aligned sequences. Progressive global alignment, iterative methods, Alignment based on locally conserved patterns, statistical and probabilistic method (hidden Markov model) are the methods for Multiple Sequence Alignment.

```
ASIS C TGSSSNIGAG-NHVK W YQ QLPG
ATIS C TGTSSNIGS--ITVN W YQ QLPG
GRLS C SSSGFIFSS--YAMY W VR QAPG
GSLT C TVSGTSFDD--YYST W VR QPPG
VEVT C VVVDVSHED--VKFN W YVDG--
ATLV C LISDFYPGA--VTVA W KADS--
AALG C LVKDYFPEP--VTVS W NSG---
VSLT C LVKGFYPSD--IAVE W WSNG--
```

2. DATABASE SEARCHING

2.1 MMseqs2

It is a software introduced in 2017 which provides a sensitive sequence searching for computation of huge biological data sets. Since the outcome of the protein and nucleotide sequencing is increased much higher than the computational speed in past 15 years. Sensitive sequence searching is a crucial task in computation of huge biological data sets. To overcome this problem, MMseq2 is developed. It had achieved a full range of speed-sensitivity trade-off and it works better than PSI-BLAST [8]. It is nearly 350 times faster in its speed.

MMseq2 search undergoes 3 stages. They are a short word match stage ('k-mer'), vectorized ungapped alignment, and gapped Smith- Waterman alignment. By undergoing these

stages, MMseq2 had removed the performance and cost of time gap between the computation analysis and sequence alignment of the protein's sequences [2] [9]. It holds a great gain in accuracy and speed which perform the computation of huge biological data sets. It also useful for both genomic and metagenomic protein sequences.

2.2 SWAPHI

SWAPHI is a parallel way to improvise the Smith-Waterman Protein Database search on Intel Xeon Phi co-processor. It is an application or script type software with Command Line Interface. It uses Unix/Linux operating system and C++ as the Programming language. SWAPHI's algorithm [1] [3] [9] relies on Scale and vectorize method. It improves the alignment speed by exploring both coarse-grained parallelism from co-processing cores. It is fine Grained Parallelism from 512-bit wide SIMD vectors for each core. It is successful for updating up to 59 billion cells for each second by Searching against the TrEMBL/UniProtKB database.

3. PAIRWISE ALIGNMENT

3.1 Parasail

Parasail is a Single Instruction Multiple Data libraries of Smith-Waterman, Needleman-Wunsch, and semi-global pairwise sequence alignment algorithms Implementations. The other name for Parasail is Pairwise Sequence Alignment Library [5] [10]. It is package type software with Command-line Interface and there are no restrictions to use. Parasail runs on Windows, Unix/Linux, and Mac OS operating systems. C, Python is the programming language used. It uses a 375 query sequence with a speed of 135 billion cells update per second. This task is performed by two methods: Returning the alignment statistics, returning the alignment score and last locations.

3.2 UGENE

UGENE [11] is used for Creating, updating and annotating the protein and nucleic acid sequences. UGENE is an Opensource Smith-Waterman algorithm for CUDA/SSE [12] [13] it is a Fast search tool for sequences. It gets data by searching databases like NCBI, PDB, DAS servers, etc. To perform PCR primer design it is integrated with Primer3 package. Files can be viewed in 3D Structure in PDB and MMDB formats. By using GOR IV and PSIPRED algorithms, the Secondary structure of protein is predicted. For nucleic acid sequences it builds Dot plots. For labs it creates and uses shared storage.

4. MULTIPLE SEQUENCE ALIGNMENT

4.1 FASMA

FASMA will analyze and produce the standardized Visualization of the Multiple Sequence Alignments (MSA's). It is also known as Formatting and Computation of the Sequences in the multiple alignments. It contains Web User Interface. FASMA [14] allows the people to analyze and compare the sequences, with for investigating conservation in particular positions and pairwise similarity. Extremely huge protein families (10000's) undergo Progressive alignment. Input Format [15] [16] will be ClustalW, fasta, GCG MSF, PIR, etc. Then the output data will be a page having good visualized alignment, the construction of an alignment, Information of the Residue position, Frequency of the residues in particular sequence. A table with residue [17] identity percentage between all aligned sequences. The output format will be in HTML GCG MSF. Perl language is used.

4.2 PROMALS3D

PROMALS3D [7] [18] is a tool for protein Multiple Sequence Alignment constructions. PROMALS3D is enhanced with extra Evolutionary and detailed information from database searches. It automatically identifies the homologs from the sequence. Proteins are input for PROMALS3D [19] [20] and undergo construction of structure-based constraints from the 3D Structures. Its output is consensus alignment which is enriched with sequence information for given input proteins. There are different Strategies [21] at various stages of alignment which allow PROMALS3D to align 1000's of protein sequences in less time because the memory took steps and time of database search.

5. DISCUSSION

Because of the regular (more) use of the different bioinformatics tools, the scientists and bioinformaticians are having a doubt abt which tool has to be used. In the past decade some research peoples had worked on some survey papers by comparing all the attributes which will because of the success or failure of the tool. By comparing and examined the tools about their sequential analysis and had highlighted the strengths and drawbacks. Due to increase in the Biological data nowadays, the tools which we are using may have some drawbacks like delay, etc. so, by considering some parameters we compare all the tools available for database searching, Pairwise sequence alignment, Multiple Sequence Alignment are categorized and summarised below. An efficient comparison of the recent tools or software is done and came to a conclusion of specific tools.

5.1 Significance

It is used to know the Function or activity of a new gene/protein. The structure or shape of a new Protein and to determine the Location or Preferred location. Stability and Origin of the protein are obtained. We finally come to conclude the origin or phylogeny of the organelle/organism. Since, if the two sequences compared from different organisms are similar then they may have common ancestor sequence and some of them may be homologs of the other.

5.2 Accuracy

Accuracy deals with perfection in the alignment. When two/more sequences are compared we calculate the Column Score (CS). By that score comparison we get ability of the tool to align the sequences. The length of the sequence also handles a good measure for the accuracy comparison.

5.3 Availability

All the scientists and bioinformaticians working in this application field are not adequately using all the tools due to lack of awareness of some recently developed software. This comparison which brought out some efficient tools can be incorporated in their system.

5.4 Portability

Portability is an important factor because the tools which we suggest to use should work on any kind of environment. That means Software which is been selected to be used should run in any operating system. If the software is restricted to a particular operating system then that is not an efficient software to use. Tools working in Linux are also working in Windows benefits them. Mainly it should be User-friendly.

5.5 Performance

High performance is a term in this application, which talk about the overall execution time of the processor. That is the elapsed time taken from the first processor to the last moment of last processor execution. It defines the Algorithm's high computational functioning and high-speed memory access. Though it is affected by the count of the sequences and their average sequence length.

Tools	Significance	Accuracy	Availability	Portability	Performance
MMseq2	Search and cluster large sequences.	MMseqs2 matched 78% of sequences to eggNOG in only 1.5% of the CPU time that BLAST	This is widely available now in many platforms.	Sequence profiles held in 16 GB of memory of a single 2× 14-core server in any kind OS.	This tool annotated a nonredundant set of 1.1 billion of hypothetical protein sequences with Pfam domains.
SWAPHI	To accelerate the Smith-Waterman protein database search, they developed this 1st parallelized algorithm.	It searches against the TrEMBL database which contains 13,208,986,710 amino acids in 41,451,118 sequences.	It has four Xeon Phis, each of which has 60 processor cores and 7.9 GB RAM.	With Two Intel E5-2670 8-core 2.60 GHz CPUs and 64 GB RAM running the Linux OS	The performance of SWAPHI is efficient because it searching queries of lengths ranging from 144 to 5,478
PARASAIL	The first time, single instruction multiple sequence global variants, semi-global variants, and local alignments are present in a stand-alone C library.	The accuracy of the alignment will be high due to opal library is faster for single-threaded applications	Since it can be used inking of the operating system, and tool is available for free from different sources.	It can be used in Unix/Linux, Mac OS, Windows. It can be used by anyone who uses a various operating system	Intra-sequence Pairwise alignment is done very fast when compared to other tools. including the global variants, it is described and benchmarked
UGENE	By UGENE Assembly browsers, the advanced sequencing data is pictured	The accuracy of the UGENE will be more because it deals with building and combining various algorithms.	It is freely available from root sources as it can be installed on any kind of OS. so it is user-friendly	Operating systems like windows, macOS 10.7+, Linux are eligible for the installation and require a 512 Mb RAM or 2Gb for safe	Due to a suffix array-based method, the algorithm will be repeated and finally completes by Producing a Dotplot
FASMA	For huge Protein families taken in lakhs, the Progressive alignment is Performed	This serves a accurate sequence alignment due to core of the FASMA, which is a CGI script written in perl language	FASMA, which is available from many sources can be installed and use by the beginners. It is an user friendly one.	It can be used in any operating system. The language used is perl and requires a basic skill on computers	It takes the input in any format and undergo the data processing and out will be in HTML format. It completes this task in less time, which explains the high performance.
PROMALS3D	Develops the Structured based constraints, so the sequence alignment will be less complex, fast alignment is Done.	The sequences are enhanced and their Profile consistency is maintained which acquires the good accuracy	This tool can be installed on a normal computer of 2Gb RAM. It can be understandable by everyone who requires this.	It can be used in windows, Linux. Since all the users mostly use these, it can be easily incorporated by them.	The performance of the PROMALS3D will be high because of diving the query into subdivisions (profile-derived, structure-derived and user-defined) and computing.

6. CONCLUSION

The job of this paper is too aware of both scientists and bioinformaticians for using a good and well-suited Tool or Software for the required tasks. Different tools that can be used for database searching, Pairwise and Multiple Sequence alignment are been compared and on basis of their availability, accuracy, portability, Performance, and significance some tools are been selected. Since they are recent tools, then can be taken in to use. So for Database searching we suggest to use MMseq2(It is perfect for searching and clustering of big sequence data) which is similar to BLAST and PSI-BLAST but magnitude faster; SWAPHI Which is the first parallel Smith-Waterman method exploiting Intel Xeon Phi clusters to accelerate the alignment of long DNA sequences.

Coming to the Pairwise alignment, the recent tools or Software's which can be use are PARASAIL (C/C++/Python/Java Single Instruction Multiple Data dynamic programming for SSE, AVX2, and UGENE (It is an Open Source Smith-Waterman for SSE/CUDA, Array-based duplicates finder and it constructs the Dotplot). Mainly considering the Multiple Sequence Alignment, you can use FASMA (which is a Progressive Alignment for very big protein Families); PROMALS3D (It undergoes Hidden markov model, Progressive alignment, 2D and 3D structures Description.

7. REFERENCES

[1] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Y ouseff, et aI., "The eucalyptus open-source

cloud-computing system," in Proceedings of the Cloud Computing and Its Applications, 2009, pp. 124-131.

[2] Bertil Schmidt, Bioinformatics: High-Performance Parallel Computer Architectures, CRC Press, 2011.

[3] B. Rauchwerger, D. Breitgand, E. Levy, A. Galis, K. Nagin, 1. M. Llorente, et aI, "The Reservoir model and architecture for open federated cloud computing," IBM Journal of Research and Development, vol. 53, 2009, pp.535-545.

[4] Watson, P. et aI., "Cloud computing for e-Science with CARMEN," in 2nd Iberian Grid Infrastructure Conference Proceedings, 2008, pp. 3-14.

[5] J. J. Cheetham et al., Parallel CLUSTALW for PC Clusters, ICCSA '03, 2003, pp. 300–309.

[6] Guangming Tan, Shengzhong Feng, and Ninghui Sun, Parallel multiple sequences alignment in SMP cluster, HPCASIA '05, 2005, pp.426-431.

[7] Kridsadakorn Chaichoompu and Surin Kittitornkun, Multithreaded ClustalW with improved optimization for Intel multi-core processor, ISCIT '06, 2006, pp 590-594.

[8] Martin Steinegger, Johannes Soeding
DOI: https://doi.org/10.1101/079681

[9] Liu, Y., & Schmidt, B. (2014). SWAPHI: Smith-waterman protein database search on Xeon Phi coprocessors. 2014 IEEE 25th International Conference on Application-Specific Systems, Architectures Processors.doi:10.1109/asap.2014.686865

[10] Taeho Kim and Hyun Joo, ClustalXeed: a GUI-based grid computation version for high performance and terabyte size Multiple Sequence Alignment, BMC Bioinformatics, vol. 11, 2010, pp. 467-475.

[11] Okonechnikov K, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts091

[12] Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, Oler AJ, Quiñones M, Hurt D, Fursov M, Huyen Y. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. PeerJ 2014 2:e644. doi:10.7717/peerj.644

[13] Rose R, Golosova O, Sukhomlinov D, Tiunov A, Prosperi M. Flexible design of multiple metagenomics classification pipelines with UGENE. Bioinformatics, bty901, 2018/10/25. doi:10.1093/bioinformatics/bty901

[14] Y. Liu, B. Schmidt, and D.L. Maskell, MSA-CUDA: Multiple Sequence Alignment on Graphics Processing Units with CUDA, 20th IEEE International Conference on Application-specific Systems, Architectures and Processors, pp. 121-128, 2009

[15] T. Rognes, Faster Smith-Waterman Database Searches with Intersequence SIMD Parallelization, BMC Bioinformatics, vol. 12, pp. 221, 2011.

[16] Do CB, Katoh K. Protein multiple sequence alignment. In: Walker J, editor. Methods Mol Biol. 1st edn. Vol. 484. Totowa: Humana Press; 2008. pp. 379–413. [PubMed]

[17] Pei J. Multiple protein sequence alignment. Curr Opin Struct Biol. 2008; 18(3):382–386. [PubMed]

[18] Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol. 2007; 3(8):e123. [PMC free article][PubMed]

[19] Mohamed Radhouene Aniba, Olivier Poch, and Julie D. Thompson, Survey and Summary Issues In bioinformatics benchmarking: the case study of multiple sequence alignment, Nucleic Acids Res., Vol. 38, No. 21, 2010, pp. 7353–7363.

[20] J. D. Thompson, B. Linard, O. Lecompte, O. Poch, A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS One, Vol. 6, No. 3, e18093, 2011.

[21] Pei J, Kim BH, Grishin NV. Nucleic Acids Res. 2008 Apr; 36(7):2295-300. Epub 2008 Feb 20

