REVIEW ARTICLE OPEN ACCESS

Biological Sequence Alignment - A Review

Kala Karun*, Dancy Kurian**, Sheeja Y.S.***
**(Department of Computer Science & Engineering, College of Engineering Kottarakara, Kerala, India
Email: kalavipin@gmail.com)*
*** (Department of Computer Science & Engineering, College of Engineering Attingal, Kerala, India
Email: dancyk@gmail.com)*
**** (Department of Computer Science & Engineering, College of Engineering Attingal, Kerala, India
Email: yssheeja@gmail.com)*

ABSTRACT
Bioinformatics is an emerging interdisciplinary research area that deals with the computational management and analysis of biological information. Genomics is the most important domain in bioinformatics which compares genomic features like DNA sequences, genes, regulatory sequences, or other genomic structural components etc. of different organisms. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. Scientists may require weeks or months if they use their own workstations since biological big data is generated by several different bioinformatics/biological/biomedical experiments and it can be presented as structured or unstructured data. Each cell in the body contains a whole genome, yet the data packed into a few DNA molecules could fill a hard drive. Biological big data is now reaching the size of Terabytes, Petabytes and exa bytes and the different modes of representation adds complexity. It introduces many challenges such as handling of complex information; integration of heterogeneous resources; analysis on big data. Advanced methods to handle the volume of data and speed of analysis scientists may require weeks or months if they use their own workstations. Sequence alignment is a standard technique in bioinformatics for visualizing the relationships between residues in a collection of evolutionarily or structurally related proteins. In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity which may be a consequence of functional, structural, or evolutionary relationships between the sequences and are used to infer biological information. Since the sequence data bases are big databases, the existing techniques and algorithms have several computational challenges. A review of the major sequence alignment algorithms are discussed in this paper.

Keywords – Bioinformatics, Genomics, DNA molecules, Hadoop, RNA molecules

Date of Submission: 24-11-2017 Date of acceptance: 05-12-2017

I. INTRODUCTION

Sequence alignment is a technique Bioinformatics which is used for visualizing the relationship among the residues with a collection of structural or evolutionary proteins. In the amino acid sequence, there are set of proteins that need to be compared with the alignment that displays the residues of the protein in a single line with the gaps (“-“) which means it is “equivalent” residues that appears in the same column. The precise meaning of equivalence is generally contested dependent: for the phylogeneticist, equivalent residues have common evolutionary ancestry; for the structural biologist, equivalent residues correspond to analogous positions belonging to homologous folds in a set of proteins; for the molecular biologist, equivalent residues play similar functional roles in their corresponding proteins. In each case, an alignment provides a bird’s eye view of the underlying evolutionary, structural, or functional constraints characterizing a protein family in a concise, visually intuitive format [1]. In our present era, a biological data explosion has occurred and also a great acceleration in the accumulation of biological knowledge began. The reasons for the biological data explosion are the revolutionary recombinant DNA technology used for DNA sequencing and the latest evolution of Genome Sequencing Projects. So, it is easier to obtain the DNA sequence of the gene corresponding to an RNA or protein than it is to experimentally determine its function or its structure. Because of this, the size of sequence databases (e.g. Genbank maintained by NCBI, USA) is larger than the size of structure databases (e.g. PDB, maintained by RCSB, USA), to date. This provides a strong motivation for developing computational methods that can infer biological information from sequence alone. With the advent of modern computers and information technology, the biological data have not only been stored in the computer in the form of databases but also processed using computational techniques to get useful results and connections among them [2]. The omics fields of life science include the voluminous amount of data in many forms used to represent various levels of biological data that includes genomic data, epigenomic data, proteomic data and transcriptomic data of different people.

The biological data becomes too big and available in petabytes and exabytes. This data produce a lot of meaningful values. It introduces many challenges such as handling of complex information; integration of heterogeneous resources; analysis of big data. Big data analytics is one of the most booming markets. When Google search engine launched image search feature, it had indexed more than 300 million images. In every minute, so many video contents are uploaded to YouTube. Twitter handles millions of tweets per day. Facebook users update their wall in every minute. Search engines are logging 600 million queries daily. There are different data centers where people can store a vast amount of data, such as IBM Server, EMC Server, etc. On the other hand, AWS (Amazon Web Services) provide a host of services to store, process and analyze the data at scale in a cost effective manner. Big data term refers collection of large datasets that are distributed, multi-dimensional and complex that it becomes difficult to processing on hand traditional data processing applications [3]. Unstructured data includes image files, text files, audio files and video files. In general online social networks generates large amount unstructured and semi-structured data. Relational Database Management System contains structured data [4]. The three dimensions of big data are volume, velocity, and variety [5]. The following areas are identified as the most challenging in big data: Data storage, Analytics, Security, and Privacy. Data storage includes relational databases and No-SQL related processing aspects. In big data analytics machine to machine learning plays a major role [6].

II. LITERATURE SURVEY

Needleman and Wunsch [7] and Smith and Waterman [8] are the widely used sequence alignment algorithms. Sequence alignment algorithms work based on dynamic programming. These are all produces accurate alignment score. These algorithms need high computation for processing the data.

Li, [9]; Notredame, Higgins & Heringa, [10] used ClustalW and T-Coffee tools to develop heuristic algorithms. It uses progressive approximation method. These tools identify the similar sequence in very fast. ClustalW tool uses the profile to profile alignment used to represent the probability of sequence in a particular position. This gives better results compare to the previous one. Sequence alignment score improved using iteration based approach.

Holmes & Bruno [11] (2001) make use of hidden Markov model (HMM), Zhang and Wong [12] make use of generic algorithms in iteration based approach. HMM is created using already aligned sequence. It tests the sequence with respect to HMM or not. Dynamic Programming methodology produces a better result, but it needs higher computation power. Heuristics algorithms are too fast, and it needs local maxima value. Iterative based approach is relatively slow.

Liu, Schmidt, and Maskell [13] used MSAProbs a new and practical multiple protein sequence alignment algorithm designed by combining a pair-HMM and a partition function to calculate posterior probabilities. It also investigates two critical bioinformatics techniques, namely weighted probabilistic consistency transformation and weighted profile-profile alignment, to achieve high alignment accuracy. In addition, it is optimized for modern multi-core CPUs by employing a multi-threaded design in order to reduce execution time. It statistically demonstrates dramatic accuracy improvements over several top performing aligners: ClustalW 2.0.12, MAFFT 6.717, MUSCLE 3.8.31, ProbCons 1.12 [14] , and Probalign 1.3 [15].

Deng and Cheng [16] uses MSACompro a new efficient and reliable multiple protein sequence alignment programs. It incorporates predicted secondary structure, relative solvent accessibility, and residue-residue contact information into the currently most accurate posterior probability-based MSA methods. It used a multiple-threading implementation on a 32 CPU cores machine. Benchmarks clearly show improvements in accuracy over the leading tools including MSAProbs.

Zhang *et al.* [17] a recent parallel program (ParaAT) that is capable of constructing multiple protein-coding DNA alignments for a large number of homologs. It is well suited for large-scale data analysis in the high-throughput era. It assigns each homolog to one of the slave threads; enable the user to customize one of multiple sequence aligners (including ClustalW, Mafft, Muscle, T-Coffee); consolidates the results from all slave threads; then parallel back-translates multiple protein sequence alignments into the corresponding DNA alignments. Tests performed on a 24 cores machine provide good scalability and exhibits high efficiency.

Bar-Yossef, Keidar, and Schonfeld [18] was the first URL Based proposed method. Dust-Buster algorithm is used for mining DUST very effectively from a URL list. In this, the author addressed DUST detection problem as a problem of finding normalization rules able to transform given URL to another likely to have similar content. It can reduce crawling overhead by up to 26% , increases crawl efficiency and reduces indexing overhead. As the substitution rules were not able to capture many common duplicate URL transformations on the web, so the author Dasgupta presented a new formalization of URL Rewrite rules [19]. The study makes use of heuristics method in order to generalize the generated rules. There is a set of URL"s which can be classified in to classes based upon the classes the content cluster is formed then the rules are rewritten for the applied clusters. It helps in trapping replacements easily in search engine workflow. In addition, it improves the entire process of efficiency in large scale experiment. The study findings revealed that the learned rule is obtained at a maximum compression that can be expected.

Li *et al.*,[20] aimed to examine the performance of using Hadoop in a cloud computing environment. In addition, the study identifies how to establish the performance by using Hadoop in a cloud computing environment. The study findings revealed different alignment applications and tools to perform their sequence alignment in a cloud burst. A similar work is described by Karlsson et al.[21] in which they aim to use parallel platform which executes the bioinformatics tool in a cloud environment. The study has make use of Microsoft azure software in order to parallelize the task execution instead of using Hadoop.

Widera & Krasnogor [22] described the use of google app engine computing platform which can be used as a computational resource. The study has used the method for building the computer protein models which is used in the prediction of protein structure. The study findings revealed that the proposed protein model comparator is the solution in order to overcome the problem of large-scale model comparison and can be scaled for different data sizes.

Bonnel & Marteau [23], has used the recent method i.e. laplacian norm alignment which represents the performance of protein in the two phases. In the first phase a graph that denotes adjacency matrix is represented whereas in the second phase the feature values are evaluated by the laplacian operator. At last, there exist a comparison step since in this study; there are two algorithms one is based on local similarity, and the other is global.

Braberg et al.,[24] has used SALIGN, which is a flexible and unique method which permits the user to form the properties that define the structure. In this method, the dissimilarity matrix is compared first and from this dissimilarity score, some of the properties represent the proteins. Secondly, it matched the models that are molecular in the stage from the theories in which they are used to find the optimal alignment in the dynamic programming by the optimal path of the matrix.

Westfall & Young [25] has used resampling methods which provide an alternative approach for genomic inference. In this study, they have considered the only limited number of an assumption than that of the asymptotic counterparts the computational cost increases at a higher rate when there is an increase in computational analysis. The asymptotic approach has a statistics calculated for each analysis, and the resampling is based upon the multitude statistics which can be calculated regularly. The procedure for marginal score statistics and the resampling replication are parallel even though they both are calculated independently this offers the potential for massive reduction in the computation time.

Halved is also based on Hadoop. It includes a variant detection phase which is the next stage after the sequence alignment in the DNA sequencing workflow. Halve calls BWA (Burrows-Wheeler aligner) from the mappers as an external process which may cause timeouts during the Hadoop execution if the task timeout parameter is not adequately configured. Therefore, a priori knowledge about the execution time of the application is required. Note that setting the timeout parameter too high values causes problems in the detection of actual timeouts, which reduces the efficiency of the fault tolerance mechanisms of Hadoop. To overcome this issue, as it is explained in further sections, SparkBWA uses Java Native Interface (JNI) to call the BWA methods. Another approach is applying standard parallel programming paradigms to BWA. For instance, pBWA [26] uses MPI to parallelize BWA in order to carry out the alignments on a cluster. We must highlight that pBWA lacks fault tolerant mechanisms in contrast to SparkBWA. In addition, pBWA, as well as SEAL, does not support the BWA-MEM algorithm [27].

Several solutions try to take advantage of the computing power of the GPUs to improve the performance of BWA. This is the case of BarraCUDA [28], which is based on the CUDA programming model. It requires the modification of the BWT (Burrows-Wheeler Transform) alignment core of BWA to exploit the massive parallelism of GPUs. Unlike SparkBWA, which supports all the algorithms included in BWA, BarraCUDA only supports the BWA-backtrack algorithm for short reads. It shows improvements up to 2× with respect to the threaded version of BWA. It is worth to mention that due to some changes in the BWT data structure of most recent versions of BWA, BarraCUDA is only compatible with BWTs generated with BWA versions 0.5.x. Other important sequence aligners (not based on BWA) that make use of GPUs are CUSHAW [29] , SOAP3 [30] and SOAP3-dp [31]. Some researchers have focused on speeding up the alignment process using the new Intel Xeon Phi coprocessor (Intel Many Integrated Core architecture—MIC). For example, MBWA [32], which is based on BWA, implements the BWA-backtrack algorithm for the Xeon Phi coprocessor. MBWA allows using concurrently both host CPU and coprocessor in order to perform the alignment, reaching speedups of 5× with respect to BWA. Another solution for the MIC coprocessors can be found in [33] . A third aligner that takes advantage of the MIC architecture is MICA

Authors in [34] claim that it is 5× faster than threaded BWA using 6 cores. Note that, unlike SparkBWA; this tool is not based on BWA. Another researcher exploit fine-grain parallelism in FPGAs (Field Programmable Gate Arrays) to increase the performance of several short-read aligners including some based on BWT [35] [36] [37].

III. CONCLUSION

In order to address the computational challenges we can use cloud platform which is suitable to allow other researchers in order to conduct the data analyses at intermediating costs that participate the absence of access towards a large computer infrastructure [38] [39] [40]. The pay-as-you-go model of cloud computing, which removes the maintenance effort required for HPC (high performance computing) instantaneously which offers scalability, elastic which makes a genomic analysis. In this, the apache-spark is a new framework that performs Hadoop efficient in a machine learning iterative jobs. In this, the memory-computing spark is used for an interactive query of 39 GB data with the response time [41]. In the spark describes a new concept known as RDDs, which is a read-only collection which partitioned crossways with a machine set that can be constructed. Spark gets delivered by caching mechanism where the user exploits to the RDD memory across the machine and multiple maps-reduce operations like parallel operations.

REFERENCES

[1] Do, C.B. & Katoh, K. Protein Multiple Sequence Alignment. *Methods in Molecular Biology*. 484 (1), 2008. pp. 59745–59745

[2] Parthasarathy S. *Sequence Alignment Algorithms - Application to Bioinformatics Tool Development.* Bharathidasan University.

[3] Lewis, S., Csordas, A., Killcoyne, S., Hermjakob, H., Hoopmann, M.R., Moritz, R.L., Deutsch, E.W. & Boyle, J. Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. *BMC bioinformatics*. *13 (1),* 2012. pp. 324.

[4] Sahane, M., Sirsat, S. & Khan, R. Analysis of Research Data using MapReduce Word Count Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*. *4 (5),* 2015. pp. 8–11.

[5] Gu, J. & Zhang, L. Some Comments on Big Data and Data Science. *Annals of Data Science*. *1 (3-4),* 2015.pp. 283–291.

[6] Grolinger, K., Hayes M., Higashino, W. et.al Challenges for MapReduce in Big Data. *Proc. of the SERVICES - IEEE World Congress on Services*. Anchorage, AK: IEEE. 2014. pp. 182–189.

[7] Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. *48 (3),* 1970. pp. 443–453.

[8] Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *Journal of Molecular Biology*. *147 (1),* 1981.pp. 195–197.

[9] Li, K.-B. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics (Oxford, England)*. *19 (12),* 2003.pp. 1585–1586.

[10] Notredame, C., Higgins, D. & Heringa, J. T-coffee: a novel method for fast and accurate multiple sequence alignment [online]. *Journal of Molecular Biology*. 302 (1), 2000. pp. 205–217.

[11] Holmes I., Bruno W.J., Evolutionary HMMs: a Bayesian Approach to multiple alignment .Bioinformatics. Sept 17(9), 2001.pp 803-820.

[12] Zhang, C.,Wong, A.K. A genetic algorithm for multiple molecular sequence alignment. *Comput Appl Biosci*. *13 (6),*1997 pp. 565–581.

[13] Liu,Y.,Schmidt, B.Maskell, D.L.MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*. *26 (16),* 2010.pp. 1958–1964.

[14] Do, C.B., Mahabhashyam, M.S.P., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*. *15 (2),* 2005.pp. 330–340.

[15] Roshan, U. & Livesay, D.R. Probalign: Multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*. 22 (22), pp.2006.pp 2715–2721.

[16] Deng, X. & Cheng, J. MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and

residue-residue contacts. *BMC Bioinformatics*. *12 (1),* 2011.pp. 472.

[17] Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X. & Dai, L.ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications*. 419 (4),2012. pp. 779–781.

[18] Bar-Yossef, Z., Keidar, I. & Schonfeld, U. (2009) Do not crawl in the DUST. *ACM Transactions on the Web*. 3 (1), pp. 1–31.

[19] Dasgupta, A., Kumar, R. Sasturka, A. De-duping urls via rewrite rules. In: *Proc of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD "08*. New York, NY, USA: ACM. 2008. pp. 186–194.

[20] Li, X., Jiang, W., Jiang, Y. & Zou, Q. (2012) Hadoop Applications in Bioinformatics. In: *2012 7th Open Cirrus Summit. IEEE*. 2010. pp. 48–52

[21] Klambauer, G., Cano, M. & Trelles, O. (2012) Enabling Large-Scale Bioinformatics Data Analysis with Cloud Computing. In: *2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications*. *IEEE*. 2012. pp. 640–645.

[22] Widera, P. & Krasnogor, N. (2011) *Protein Models Comparator: Scalable Bioinformatics Computing on the Google App Engine Platform*.2011.

[23] Bonnel, N. & Marteau, P.-F. LNA: Fast Protein Structural Comparison Using a Laplacian Characterization of Tertiary Structure. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. *9 (5),* 2012.pp. 1451–1458.

[24] Braberg, H., Webb, B.M., Tjioe, E., Pieper, U., Sali, A. & Madhusudhan, M.S. (2012) SALIGN: a web server for alignment of multiple protein sequences and structures. *Bioinformatics (Oxford, England)*. *28 (15),* 2012.pp. 2072–2073.

[25] Westfall, P.H. & Young, S.S. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley: A Wiley-Interscience publication.1993.

[26] Peters, D., Luo, X., Qiu, K. & Liang, P. Speeding Up Large-Scale Next Generation Sequencing Data Analysis with PBWA. *J Appl Bioinform Comput Biol. 1 (1),* 2012.pp. 1–6.

[27] Abuín, J.M., Pichel, J.C., Pena, T.F. & Amigo, J. SparkBWA: Speeding Up the Alignment of High-Throughput DNA Sequencing Data Ruslan Kalendar (ed.). *Plos one*. *11 (5),* 2016.pp. 1–21.

[28] Klus, P., Lam, S., Lyberg, D., Cheung, M.et.al. BarraCUDA - a fast short read sequence aligner using graphics processing units. *BMC Research Notes. 5 (1),* 2012.pp. 27.

[29] Liu, Y., Schmidt, B. & Maskell, D.L. CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*. 28 (14), 2012.pp. 1830–1837.

[30] Liu, C.-M., Wong, T., Wu, E., Luo,et.al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics (Oxford, England)*. *28 (6),* 2012. pp. 878–879.

[31] Luo, R., Wong, T., Zhu, J., et.al. SOAP3-dp: Fast, Accurate and Sensitive GPU-Based Short Read Aligner Frederick C. C. Leung (ed.) *PLoS ONE. 8 (5),* 2013.pp. 65632.

[32] Cui, Y., Liao, X., Zhu, X., Wang, B. & Peng, S. (2014) mBWA: A Massively Parallel Sequence Reads Aligner. In: *8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*.2014.pp.113–120.

[33] You, L. & Congdon, C. (2016) *Building and Optimizing BWA ALN 0.5.10 for Intel Xeon Phi^{TM} Coprocessors*.

[34] Luo, R., Cheung, J., Wu, E., Wang, H., *et al*. MICA: A fast short-read aligner that takes full advantage of Many Integrated Core Architecture (MIC). *BMC Bioinformatics*. 16 (Suppl 7), 2015.pp. S10.

[35] Arram, J., Tsoi, K.H., Luk, W. & Jiang, P. Hardware Acceleration of Genetic Sequence Alignment. In: *9th International Symposium*. Los *Angeles, CA*. 2013.pp. 13–24.

[36] Sogabe, Y., Maruyama, T. An acceleration method of short read mapping using FPGA. In: *2013 International Conference on Field-Programmable Technology (FPT)*. *IEEE*.2013 pp. 350–353.

[37] Waidyasooriya, H.M. & Hariyama, M. Hardware-Acceleration of Short-Read Alignment Based on the Burrows-Wheeler Transform *IEEE Transactions on Parallel and Distributed Systems*. *27 (5),* 2016.pp. 1358–1372.

[38] Langmead, B., Hansen, K.D. & Leek, J.T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology*. *11 (8),* 2010.pp. R83.

[39] Schatz, M.C., Langmead, B. & Salzberg, S.L. Cloud computing and the DNA data race. *Nature biotechnology*. 28 (7), 2010.pp. 691–693.

[40] Stein, L.D. The case for cloud computing in genome informatics. *Genome Biology*. 11 (5), 2010.pp. 207.

[41] Zaharia, M., Chowdhury, M., Franklin, M.J., et.al Scott Shenker, and Ion Stoica. Spark: cluster computing with working sets. In: *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. Berkeley, CA, US: USENIX 2010.pp 10

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

Kala Karun Biological Sequence Alignment - A Review." International Journal of Engineering Research and Applications (IJERA) , vol. 7, no. 12, 2017, pp. 18-23.