International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 12 271 – 273

Identification of Protein Alignment for Elder Health Care

Sneha A. Khaire
K.C.T. Late G.N. Sapkal
College of Engineering,Nashik
sneha15khaire@gmail.com

Prof. N.R. Wankhade
Assistant professor, K.C.T. Late G.N.
Sapkal College of Engineering, Nashik
nileshrw_2000@yahoo.com

Abstract: Over many years protein sequence alignment problem has grabbed attention of biologists as it implicates, more than two biological sequences. It states all the important aspects of big data and how medical and health informatics, translational bioinformatics will benefit personalized health care both structured and unstructured, covering genomics, proteomics, metabolism. The system develop approach for biological sequence alignment to increase efficiency of analysis operation that speed up the calculation of alignment for huge real time sequences, to develop distributed scan approach in Smith-waterman algorithm for presenting fast solution and optimize the Smith Waterman(SW) alignment algorithm using Distributed approach.

Index Terms—*health informatics, bioinformatics, proteomics, smith waterman.*

*****

I. INTRODUCTION

In parallel Distributed figuring improvement the approach of consecutive arrangement is basic that helps to work out the organic affiliations or connections from an immense information sets. Such colossal work can't be settled effortlessly utilizing ordinary approach of string coordinating operation and it won't not be compelling for coordinating long succession. Another way is Smith-Waterman (SW) calculation that discovers likenesses between certain question grouping and a subject successions however consecutive approach of this calculation is are slightest productive as far as its execution, for example, time. To lessen computational handling time of succession operation this framework utilizes parallel circulated processing abilities to get precise and proficient usage. The motivation behind this framework is to quicken the utilization of organic succession arrangement utilizing circulated handling approach for finding ideal neighborhood alignment. As the current framework concentrate on the nearby arrangement so the need of the proposed framework is to chip away at worldwide arrangement. The framework create a better approach for sequential alignment using distributed approach for health care. To create dispersed sweep approach in SW calculation for exhibiting quick arrangement. To streamline the Smith Waterman arrangement calculation utilizing Distributed approach. It plots the key attributes of enormous information and how restorative and wellbeing informatics, translational bioinformatics will profit by a coordinated approach of sorting out various parts of customized data from a different scope of information sources, both organized and unstructured, covering genomics, proteomics, metabolomicWe propose to create novel approach for natural grouping arrangement to build proficiency of investigation operation that accelerate the count of arrangement for colossal constant successions, to create dispersed output approach in Smith-waterman calculation for exhibiting quick arrangement and enhance the Smith Waterman arrangement calculation utilizing Distributed approach. The Smith- Watwerman algorithm states the exact prediction of the health care benefits.

II. LITERATURE REVIEW

Over decades, the Multiple Protein Sequence Alignment problem has attracted the attention of biologists because it is one of the major techniques which is used in several areas of computational biology, such as homology searches, genomic annotation, protein structure prediction, gene regulation networks, or functional genomics. The results which is obtained by our method are compared with well-known methods published in the literature, concluding that the new approach presents remarkable accuracy when dealing with sets of sequences with a low sequence similarity, the most frequent ones in real-world. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations (substitutions) and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In the multi objective version of the MSA problem, we try to optimize simultaneously two of the most widely-used objective functions in the literature, such as the weighted sum-of-pairs function with affine gap penalties (WSP) and the number of Totally Conserved columns score. One of the main contributions of this paper is to tackle the MSA problem by using Multiobjective Optimization and Evolutionary Computation jointly in order to provide a new research line for this classical bioinformatics problem.

On the other hand, the H4MSA algorithm has been hybridized with the progressive, fast, and accurate Kalign to boost the accuracy and effectiveness of the algorithm. These methods use the Protein Data Bank (PDB) structures as templates to guide the multiple sequence alignment. Finally, some structure-based approaches have been developed for multiple sequence alignment. After comparing the accuracy and effectiveness of H4MSA against sixteen well-known aligners and six GAs published in the literature, we can conclude that H4MSA is a very promising approach for Multiple Sequence Alignment. In addition, we can highlight the main advantage of H4MSA relies on its accuracy when solving data sets with a low identity percentage which is a significant contribution to the scientific community. In addition, with this new approach, we are able to obtain a set of accurate and consistent solutions by simultaneously maximizing two well-known objective functions: the weighted sum-of-pairs function with affine gap penalties (WSP) and the number of Totally Conserved (TC) columns score. More precisely, we have observed a remarkable accuracy of H4MSA against other approaches in those sets of sequences with a low sequence similarity.H4MSA is a very promising approach for Multiple Protein Sequence Alignment. More precisely, we have observed a remarkable accuracy of H4MSA against other approaches in those sets of sequences with a low sequence similarity. These sets represent the most frequent scenarios in real-world.[1]

(1976) has a drawback in that it takes a large number of computational steps of the order of M^2N compared to the MN steps of Needle-man Wunsch Sellers' algorithm, where M and N (M^2N) are the lengths of the proteins or nucleic acids under comparison. form of gap weight has usually been used in computer systems that adopt the matching algorithm of Waterman.

Needle man-Wunsch's method has also been applied to statistical tests of relatedness between a pair of sequences[2]

The goal of this paper is to give idea the visibility it deserves by developing a linear-space version , which stores affine gap penalties. If one also sequence desires a set of operations attaining the minimum cost, then straightfor-ward implementations need 0(MN) space. Generalizing his specific treatment, the central idea is to find the 'midpoint' of an optimal conversion using a 'forward' and a 'reverse' application of the linear space cost-only variation.[3]

As the data of biological banks expands exponentially, so we develop a new hardware parallel bi- sequence and tri-sequence alignment algorithm, using the strategy of divide and extend. This paper suggests hardware architecture for the SW algorithm. However the existing parallel Smith-Waterman algorithm needs extra memory space and this limits the size of a sequence to be aligned .Nowadays, Next-generation sequencing (NGS) machines solve this problem by determining the nucleotide sequence of short DNA fragments (short reads) which lowers the cost and increases the throughput of DNA sequencing. Then, the short read mapping process is performed to determine the location in the reference genome to which each read maps best (Short-reads are aligned along the reference genome). Bioinformatics Science uses super computers to store, analyze, simulate and predict biological information outcomes by In Silicon (via computer simulation) techniques. The short reads are derived from many copies of the genome of one organism with a reference genome sequence which is already known. The Second problem is the speed which is important due to the huge data (A human genome is large as 3 billion nucleotides and requires billions of short- reads to map it) These problems are solved by software or hardware implementation; otherwise Gene banks use Hybrid approach.And the heuristic algorithms; such as BLAST and FASTA, ignore the unused data from computation and that accelerate the performance, but the alignment isn't the optimal solution. Bioinformatics researchers use two types of algorithms; Dynamic and heuristic: Dynamic algorithms such as Needleman-Wunsch and Smith Waterman algorithms, these algorithms find the optimal alignment solution. In this paper we present a new complete parallel Smith-Waterman algorithm for DNA sequences alignments using the technique of divide and extend by dividing the reference sequence into subsequences each subsequence equal to query sequence length in different processes and implement the algorithm.[4]

These alignment algorithms are tested on benchmark datasets BAliBASE 2.0 and BAliBASE 3.0. Experimental results show that MOMSA can obtain the significantly better alignments than VDGA, GAPAM on the most of test cases by statistical analyses, produce better alignments than IMSA in terms of TC scores, and also indicate that MOMSA is comparable with the leading progressive alignment approaches in terms of quality of alignments. We compare the performance of MOMSA with several alignment methods based on evolutionary algorithms, including VDGA, GAPAM, and IMSA, and also with state-of-the-art progressive alignment approaches, such as MSAprobs, Probalign, MAFFT, Procons, Clustal omega, T-Coffee, Kalign2, MUSCLE, FSA, Dialign, PRANK, and CLUSTALW. In this, firstly we compare MOMSA with the recently proposed multiple sequence alignment algorithms based on evolutionary algorithms, including

VDGA, GAPAM, IMSA, to demonstrate its superiority. In this paper, MOMSA is coded in MATLAB and implemented on the personal computer with four Intel Core i5 3.1GHz processors with 3072 MB RAM under WIN7 platform. BAliBASE(Benchmark Alignment database), which is a well- known benchmark alignments dataset for evaluating multiple sequence alignment programs, was developed by manually aligning based on those known three-dimensional structures of proteins. The one is the sum of pairs (SP) score, the fraction of aligned residue pairs in the reference alignment that are correctly found in the tested alignment, and the other one is the total column (TC) score, the fraction of aligned columns that are correctly found. There are five reference sets of unaligned sequences with different characters that contain the length and the identity among the unaligned sequences in the original BaliBASE. BAliBASE 3.0 is the latest version of the multiple sequence alignment benchmark, which has been widely used by many researchers For avoiding the limits by single objective modeling approach, this paper proposed a MOMSA algorithm that views MSA as a multi-objective optimization problem to find a set of non-dominated solutions for the decision-maker, and then obtain further a good alignment by a metric, such as WSPs. We have compared our proposed MOMSA algorithm with GAPAM,

VDGA and IMSA, which are based on evolutionary algorithms, on the widely used datasets BAliBASE 2.0 and 3.0. It experimentally shows that the proposed MOMSA achieves better alignments than VDGA and GAPAM.[5]

Conclusion

Sequence Analysis is core operation in Bioinformatics. The implementation of Smith-waterman algorithm with the Distributed scan approach using Hadoop able to align large biological sequences. Now a day's distributed computing receives lot of attention. Use of Hadoop is found helpful to accelerate analysis task in bioinformatics in which we have made a provision that the program can be executed concurrently on multiple systems. There are various algorithms available based on distributed computing but drawback is that they are able to calculate only score but not alignment for complex biological sequence. Smith waterman algorithm with distributed scan approach provides score as well as alignment for huge biological sequences

REFERENCES

[1] A Hybrid Multiobjective Memetic Metaheuristic for Multiple Sequence Alignment, Alvaro Rubio-Largo, Miguel A. Vega- Rodr iguez, David L. Gonz alez-Alvarez, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. XX, NO.XX,APRIL 2015.

[2] An Improved Algorithm for Matching Biological Sequences, J. &i'oZ,T. F. Smith and M. S. Waterman, Biol. (1982) 162, 705-708

[3] Optical alignments in linear space Eugene W.Myers 1 ' 2 and Webb Miller 2, Vol.4, n o . 1 . 1988.

[4] Hardware Acceleration of Smith-Waterman Algorithm for short read DNA Alignment Using FPGA Z.Wael Abou El-Wafa, Hesham F. A.Hamed, Asmaa G.Seliem, 2016 IEEE 40th Annual Computer Software and Applications Conference.

[5] Giovanni Causapruno, Gianvito Urgese, Marco Vacca, Mariagrazia Graziano, Member, IEEE, and Maurizio Zamboni,"Protein Alignment Systolic Array Throughput Optimization",IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS 1063-8210 © 2014.

[6] Wael Abou El-Wafa, Hesham F. A.Hamed, Asmaa G.Seliem,"Hardware Acceleration of Smith-Waterman Algorithm for short read DNA Alignment Using FPGA", 2016 IEEE 40th Annual Computer Software and Applications Conference.

[7] Heba Khaled 1,H.M. Faheem 1,2, Mahmoud fayez 1,2, Iyad Katib 3 and Naif R. Aljohani,"Performance Improvement of the Parallel Smith Waterman Algorithm Implementation Using Hybrid MPI – Openmp Mode",SAI Computing Conference 2016 July 13-15, 2016 | London, UK.

[8] Vijay Naidu and Ajit Narayanan,"Needleman-Wunsch and Smith-Waterman Algorithms for Identifying Viral Polymorphic Malware Variants",016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology.

[9] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences," J. Molecular Biology, vol. 147, no. 1, pp. 195-197, Mar. 1981.

[10] D.S. Hirschberg, "A Linear Space Algorithm for Computing Maximal Common Subsequences," Comm. ACM, vol. 18, no. 6, pp. 341-343, 1975.

[11] Huazheng Zhu, Zhongshi He, and Yuanyuan Jia, "A Novel Approach to Multiple Sequence Alignment Using Multi- objective Evolutionary Algorithm Based on Decomposition.", 2168-2194 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.