ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Survey on Multiple Sequence Alignment in Bioinformatics

Namrata Rathod
Dept. of Computer Science Eng.
University Institute of Technology, RGPV
Bhopal, (M.P) India

Anjana Deen
Dept. of Computer Science Eng.
University Institute of Technology, RGPV
Bhopal, (M.P) India

Shikha Agrawal
Dept. of Computer Science Eng.
University Institute of Technology, RGPV
Bhopal, (M.P) India

Abstract: Multiple Sequence Alignment (MSA) is a key step for performing various tasks in bioinformatics, such as identification of conserved motifs, 2D and 3D Structure prediction of proteins, etc. Various algorithms are designed to implement MSA and they are evolved too, to give biologically perfect alignment. In this paper, the survey of Multiple Sequence Alignment using Multi Objective Optimization Algorithm and Single Objective Optimization Algorithm which are evolved till now are specified**.**

Keywords: Multiple Sequence Alignment (MSA), Multiobjective Evolutionary Algorithm based on Decomposition (MOEA/D), Genetic Algorithm (GA).

I. INTRODUCTION

Sequence alignment is a method of arranging sequences so as to find area of similarity and dissimilarity among the sequences. Sequence alignment is a fundamental step used in bioinformatics. It helps in finding functional, structural or evolutionary relationship among the sequences. The alignment methods used in sequence alignment are: Global alignment, Local alignment, Multiple Sequence Alignment. Global and Local alignment are put into pairwise alignment category as they can align only two sequences at a time. The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods [1].Multiple sequence alignment is an extended part of pairwise alignment. In MSA all the sequences present in query domain are aligned at a time. Multiple sequence alignments are computationally difficult to produce and most formulations of the problem lead to NP-complete combinatorial optimization problems [2].

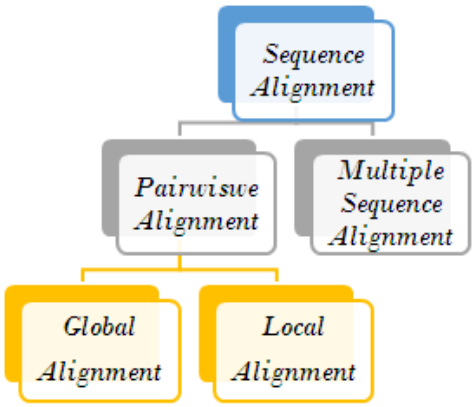

Fig 1.1:- Hierarchical representation of types of sequence alignment

Since our research is focused on MSA so in the next section we are discussing MSA in brief.

II.MSA

Multiple Sequence Alignment is an alignment that align more than two sequences at a time. MSA is a process in which length of all the sequences are made similarby inserting and deleting gapsso as to obtain aligned symbols as much as possible. MSA uses scoring function to calculate the quality of alignment.

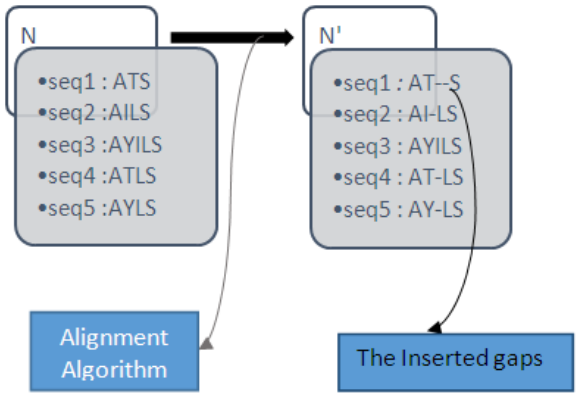

Fig 1.2:- Example of MSA

Many approaches has been made for MSA which falls in basically three categories: Exact, progressive and iterative.

(A) Exact method:

Exact method involves dynamic programming techniques which can give mathematically perfect alignment. The computational complexity of Dynamic program is very high, when more than three sequences are aligned at a time that is why progressive alignment method is proposed by Feng and Doolittle [3].

(B) Progressive and Iterative Method:
They iteratively used Needleman Wunsch pairwise alignment algorithm on multiple sequences to produce best alignment [4]. This approach is fast and simple. However the progressive methods are heuristic by nature, and follow the rule of "once a gap, always a gap," which may lead to the main disadvantage that once a mistake is made in the initial alignment, it will be impossible to be modified. In order to overcome the limitations of the progressive algorithms, mixture of iterative and progressive algorithms is recommended by researchers for instance, MUSCLE is a representative hybridAlgorithm [5].

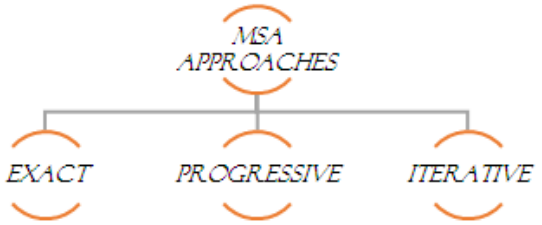

Fig 1.3:- Hierarchical representation of MSA Approaches

MSA is an optimization problem. Mostly genetic algorithm are used for MSA because this algorithm improves the Quality of sequence alignment. Genetic algorithms like SAGA [6], MSAGA[7], RBT-GA [8], GAPAM [9], and VDGA [10] are based on single objective optimization mechanism whereas MOMSA and MOEA/D are based on Multiobjective evolutionary algorithms. In this paper, the existing alignment algorithms used in MSA are categorized into single objective optimization algorithm and Multiobjective optimization algorithm. The disadvantages of single objective optimization algorithm and advantages of Multiobjective optimization algorithm are also discussed. Further the paper is structured as follow: In Section III, Literature Survey is described briefly. Finally, in Section IV, the paper is concluded and future work is mentioned.

III. LITERATURE SURVEY

A) Single Objective Optimization Algorithm for MSA
All the algorithms developed before Multiobjective evolutionary algorithm are treated as single objective optimization algorithm because MSA is consider as single objective problem and focused on only one objective function.Some single objective optimization algorithms are as follows:-

(1) Vertical Decomposition with Genetic Algorithm(VDGA):VDGA is a modified algorithm based on GA. The steps performed in VDGA are: initial population generation, child population generation by applying genetic operators, formation of next generation, vertical division and the stopping criteria. The vertical division step is use for forming first and after generating child generation. It consist of few sub steps in which the population is divided into parts and all the gaps are removed from each part and then tree based method is applied. After this each part generate an alignment. These alignments are combined together to form new alignment [10]. VDGA is an algorithm that uses an iterative mechanism to produce sequence alignment as much as possible. VDGA uses a weighted sum of pairs scoring function that measures the quality of sequence alignment. The main objective of VDGA is focused on alignment with best WSPs.

(2) Progressive Alignment Method using a Genetic Algorithm(GAPAM):GAPA is another algorithm based on GA. In GAPAM, two new mechanism are introduced to develop initial population: the first mechanism generate guide trees by using random selection and the second mechanism shuffle the sequences inside the generated trees. The steps of this method are: initial population generation, child population generation by genetic operators, next generation development by new population and the stopping condition. Weighted sum of pair method (WSPM) is a fitness measure used for multiple sequence alignment. GAPAM uses WSPM score to find the quality of alignment and hence this algorithm is also focused on single objective [9].

(3) Random Drift Particle Swarm Optimization (RPSO):Hidden Markov Models are important tools for MSA. HMMs are NP-complete problem. HMMs are very difficult to learn. Many meta-heuristic methods have been developed, including Particle Swarm Optimization (PSO) for learning HMMs. PSO is modified and a Random Drift Particle Swarm Optimization algorithm (RPSO) is designed. HMMs is based on the free electron model in metal conductors in an external electric field that employs a novel set of evolution equation that can improve the global search ability of the algorithm. RDPSO is incorporated with diversity control method for further enhancement and a new algorithm, Random Drift Particle Swarm Optimization with Diversity Guided Search (RDPSO-DGS) is proposed. This algorithm uses modified sum-of -pair (MSOP) score to evaluate alignment, if the prior knowledge of the resulting alignment is available otherwise sum-of-pair function (SOP) is used [11].

B) Multi Objective Optimization Algorithm for MSA
The goal of single objective optimization algorithms is to measure the quality of alignment by using scoring function. As the example mentioned above, WSPs is used as scoring function in VDGA. However, MSA is a kind of sequence alignment that has two objectives. The first one is to maximize the number of matching pairs among the sequences and the second one is to minimize the count of gap by adding or removing the gaps [12].Some Multiobjective optimization algorithms are as follows: -

(1) Multiobjective evolutionary algorithm based on decomposition (MOEA/D):MOEA/D is an algorithm developed by Zhang and Li in 2007. The basic idea behind this algorithm is to convert Multiobjective optimization problem into a number of scalar problem by using weighting method. The performance of MOEA/d found better than MOGLS, NSGA-II and SPEA-II. MOEA/D uses a weight vector which is preselected from N weight vectors and give same amount of effort to all aggregation functions, while MOGLS randomly generates a weight vector at each iteration, aiming at optimizing all the possible aggregation functions [13]. MOEA/D performs well when the

decomposition method and weight vectors are chosen properly [13].

(2) Multiple Sequence Alignment with Affine Gap by using Multi Objective Genetic Algorithm (MSAGMOGA):MSAGMOGA is focused on three objective: - the first one is to maximize the similarity, the second one is to minimize affine gap penalty and the third one is to maximize the support. This algorithm uses NAGA II which gives high performance.

(3) Multiple Sequence Alignment using Multiobjective Evolutionary Algorithm (MOMSA):MOMSAisbased on Decomposition.It has two objectives: - the first one is to maximize the quality and number of alignment pair into sequences and the other one is to minimize the number of gaps. This algorithm gives a new method for population generation and develop a new mutation operator [14].

TABLE: SUMMARY OF TECHNIQUE USED IN MSA

S.No	ALGORITHM	TECHNIQUE USED	RESULT
1	SAGA	An iterative stochastic alignment technique uses genetic algorithm	better scoring alignments then MSA ofLipman and ClustalW
2	T-COFFEE	A global progressive consistency-based algorithm. ClustalW and Lalign are used in it.	Higher accuracy compared to ClustalW, Prrp, Dialign2, and POA.
3	IMSA	Clonal Selection Algorithms and based on the 'weighted sum of pairs' as objective function	shows good alignments than state-of-the-art alignment algorithms
4	MAFFT	fast Fourier transform	Speedy detection of homologous segments.
5	MUSCLE	Progressive alignment uses log-expectation score as profile function, kmer counting and refinement using tree-dependent restricted partitioning.	Give more accuracy and speed than CLUSTALW and T-COFFEE
6	CLUSTALW	Neighbor-Joining method	fast and can handle sizeable sets of sequences
7	VDGA	Vertical division	VDGA with three vertical divisions was the most successful variant for most of the test cases in comparison to other divisions considered with VDGA
8	GAPAM	Guide trees with randomly selected sequences and the second is shuffling the sequences inside such trees.	GAPAM performed better for most of the test cases
9	RPSO	particle swarm optimization and meta-heuristic methods	better scores than SPSO and BW
10	MOEA/D	decomposition	lower computational complexity than MOGLS and NSGA-II [13]
11	MSAGMOGA	uses NSGA II	Better in terms of runtimethanSAGA and MSA-GA
12	MOMSA	MOEA/D framework as optimizer.	achieves better alignments than VDGA and GAPAM and better alignment than IMSA [14]

IV.CONCLUSION AND FUTURE WORK

Construction of biologically correct alignment depends on how the MSA model is constructed. Traditional method for solving MSA says that MSA was treated as single objective optimization problem such as **VDGA** which is focused on alignment with best WSPs, **GAPAM** which is focused on fitness function to find better quality of alignment, etc. this leads to two drawback: - complication in fitness functions selection and different alignments can have same fitness function value but it can find only one.

To overcome from these two problems Multiobjective optimization approaches has been developed. However, the runtime of Multi Objective Optimization Algorithms is high. Therefore,the future work is to minimize the time complexity.

REFERENCES

1. Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7.
2. Wang L; Jiang T. (1994). "On the complexity of multiple sequence alignment". J Comput Biol. **1** (4): 337–48. doi:10.1089/cmb.1994.1.337. PMID 8790475.
3. W. Lusheng and J. Tao, "On the complexity of multiple sequence alignment,"J. Comput. Biol., vol. 1, no. 4, pp. 337–348, 1994.
4. D. F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," J. Mol. Evol., vol. 25, no. 4.
5. R. C. Edgar, "MUSCLE: A multiple sequence alignment method with reduced time and space complexity," BMC Bioinformat., vol. 5, p. 113, Aug. 19, 2004.
6. C. Notredame and D. G. Higgins, "SAGA: Sequence alignment by genetic algorithm," Nucleic Acids Res., vol. 24, no. 8, pp. 1515–1524, Apr. 15, 1996.
7. J. Taheri and A. Zomaya, "RBT-GA: A novel metaheuristic for solving the multiple sequence alignment problem," BMC Genomics, vol. 10 (S1), p. S10, 2009
8. C. Gondro and B. P. Kinghorn, "A simple genetic algorithm for multiple sequence alignment," Genetic. Mol. Res., vol. 6, no. 4, pp. 964–982, 2007.
9. F. Naznin, R. Sarker, and D. Essam, "Progressive alignment method using genetic algorithm for multiple sequence alignment," IEEE Trans. Evol.Comput., vol. 16, no. 5, pp. 615–631, Oct. 2012
10. F. Naznin, R. Sarker, and D. Essam, "Vertical decomposition with genetic algorithm for multiple sequence alignment," BMC Bioinformat., vol. 12, p. 353, 2011.
11. Jun Sun, Vasile Palade, Xiaojun Wu, and Wei Fang, "Multiple Sequence Alignment with Hidden Markov Models Learned by Random Drift Particle Swarm Optimization" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 11, NO. 1, JANUARY/FEBRUARY 2014
12. Huazheng Zhu, Zhongshi He∗, and YuanyuanJia, "A Novel Approach to Multiple Sequence AlignmentUsing Multiobjective Evolutionary Algorithm Based on Decomposition" IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 20, NO. 2, MARCH 2016
13. Qingfu Zhang, Senior Member, IEEE, and Hui Li, MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 11, NO. 6, DECEMBER 2007.
14. Huazheng Zhu, Zhongshi He∗, and YuanyuanJia, A Novel Approach to Multiple Sequence Alignment Using Multiobjective Evolutionary Algorithm Based on Decomposition. IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 20, NO. 2, MARCH 2016