Inheritance of Hybrid Meta-Heuristic Approach for Multiple Sequence Alignment to optimize Genetic Aligner

Heena Arora[1], Avani Chopra[2]
[1]*Researcher, Computer Science and Engineering, D.A.V.I.T., Jalandhar, India*
[2]*Astt. Professor, Information Technology, D.A.V.I.T., Jalandhar, India*
(E-mail: aroraheena714@gmail.com)

***Abstract*—** Characteristics Sequence Alignment and Optimization of genetic aligners at each level of msa iteration is a hard optimization problem, especially where characteristics selection based on specific sequences. Bioinformatics is one of the prominent area where optimization problems occur due to mutation, uncertainty of large genetic data sets. Machine Learning and Mathematical Optimization Algorithm provides the combine solution for optimization problems related to genetic alignment. This paper proposed a native Hybrid Meta-Heuristic Optimization Genetic Alignment Algorithm (HMHOGA) to solve genomes alignment optimization problem at each phase of msa iteration and also help to recognize occurrence of heuristic characteristics. Proposed algorithm provides optimize alignment along with various evaluation factors like pair_sum(q), col_score(tc), col_dist_sum(z), col_dist_mean(r) and consensus_matrix_mean(e). All these would help to recognize and filter appropriate tasks at each level of development.

Keywords*—*MUSCLE; Bioinformatics; Multiple Sequence Alignment; Optimization; Hybrid Meta-heuristic; Machine Learning; Data Science.

I. INTRODUCTION

Optimization and Predictive Decision Making of genetic mutation in the evolution of protein has been continuously generated large amount of predictive optimization problems in field of bioinformatics. The need to understand complex information and to fill knowledge gap between them is important. The greatest challenge of today is how to develop data science & machine learning algorithms which focus to optimize these problems. Chromosome to Protein generation have various level of mutations, these type of mutations generate cancer cells, which leads towards cancer diseases. To track this mutation, we need to generate mutation tracking algorithms with the help of data science, so that it's help to track diseases mutation and to find it's medicines.

Sugawara Kohei and Fujita Hamido (2014), have further extended criteria of harmony search space to generate better results, which actually overcome the restriction generate by local optima, and this search span also avoid overmuch improvement of computing degree of negative impact by various interrupts. In this study authors try to inherit subjective attributes of worker in meta-heuristic approach to generate better results, but lack of subjective attributes this approach doesn't recommend best solution to the decision making for adaptive task selection. Nguyen Su et al. (2014), proposed heuristic framework with forward construction approach for order acceptance and scheduling. This approach focus to learn priority rules directly from optimize scheduling algorithms, which evolve based on decision making along with a set of rules for forward construction heuristic, instead of a single priority rules. This approach is very effective for large problem instances and also competitive with currently existing meta-heuristics approach.

Sugawara Kohei and Fujita Hamido (2013), have focused on the appropriate characteristics selection for worker based on workflow is often changed by occurring interruption due to assignment of external activity to the worker, which consume usable time for the current workflow by operating generated new task. To overcome this problem author, use meta-heuristic approach with subjective attributes such as experience. This approach generates optimize results, but it's not generated best results due to lack of operational parameters.

Sri R. Leena and Balaji N. (2013), have discussed about the decision making and system state estimation in multifarious environment with hybrid meta-heuristic approach by propose Hybrid Genetic and Case Based Reasoning Algorithm. This approach helps to generate fast optimize results in dynamic and diverse working environment with the help of optimal task scheduling and case specific knowledge simultaneously to utilize prediction accuracy. According to that, this approach is very suitable in Real Time heterogeneous working environment.

Multiple sequence alignment of proteins and nucleic acids has been continuously generated large amount of optimization problems in field of bioinformatics. The need to understand complex information and to fill knowledge gap between them is important. The greatest challenge of today is how to develop data science & machine learning algorithms which focus to optimize these problems. In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences shared evolutionary origins. To find this alignment, we need to generate optimistic algorithms with the help of data science, so that it's help to find the alignment of amino acids and protein. Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments.

Alam Manaar et al. (2016), have presented a parallel search technique with very less tunable parameters for solving optimization problems fast and efficient as compared to similar stochastic algorithms. This approach use rotate left and complement operator to reach distinct nodes, instead of traversing repeated nodes in the search space and flip operator use to capture variations within the search space as well. Performance of this approach is very fast and efficient and number of generations is too less than other meta-heuristics algorithms. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex.

The proposed work generates heuristic knowledge base to increase the search space to generate efficient optimize results for any genetic sequence alignment. Characteristics knowledge base focus on the issues related to features of amino acid. These characteristics inherit the protein predictive sequences, which generate better optimize results in multiple sequence alignment. Sharing quality of knowledge base between various MSA aligners to extend the scope of the study, which helps to improve other less qualitative data sets. Predictive results also analyses and track for future use to generate better results. Sharing of knowledge base helps to track the continuous mutation between genetic sequences and also provide another dimension of mutation tracking in genetic search space.

II. METHODOLOGY

Knowledge Gap occurs when two isolated sequences are co-relate to achieve the target. In Multiple Sequence Alignment, this gap is generated due to the co-relation of two different sequences based on same properties. To fill this knowledge gap, first of all we need to create characteristics based data acquisition system as per genetic sequences properties. The identification of differentially expressed genes and related molecular pathways is of great importance. Chromosome to Protein generation have various level of mutations, these type of mutations generate cancer cells, which leads towards cancer diseases. To track this mutation, we need to generate mutation tracking algorithms with the help of data science, so that it's help to track diseases mutation. To find genes which involved in cancer is difficult task. To solve this problem. we proposed a meta-heuristic function which provides optimal results. For find significant microarrays we used data analytical algorithms.

A. *Theoretical Framework (Hybrid Meta-Heuristic Approach)*

Characteristics based pairwise sequence alignment and multiple sequence alignment optimization problem occur in bioinformatics. This is a very complex NP-Hard Problem, which needs to be solved by scientific approach. We understand the biological characteristics of genetic sequences, and finally proposed a suitable solution in form of native genetic optimize aligner based on Hybrid Meta-Heuristic Approach. The existing framework has been designed for the characteristics based sequence alignment, especially for amino acid chain sequences. This framework helps to track mutation of similar characteristics proteins. This approach also helpful to predict and track cancer cells in different gene structure and it's causes, which helps to develop remedies for cancers. This framework is very suitable for local search space, especially in the case of translation from mRNA to protein sequences, but this framework is less suitable for other sequence alignment like DNA, RNA and mRNA.

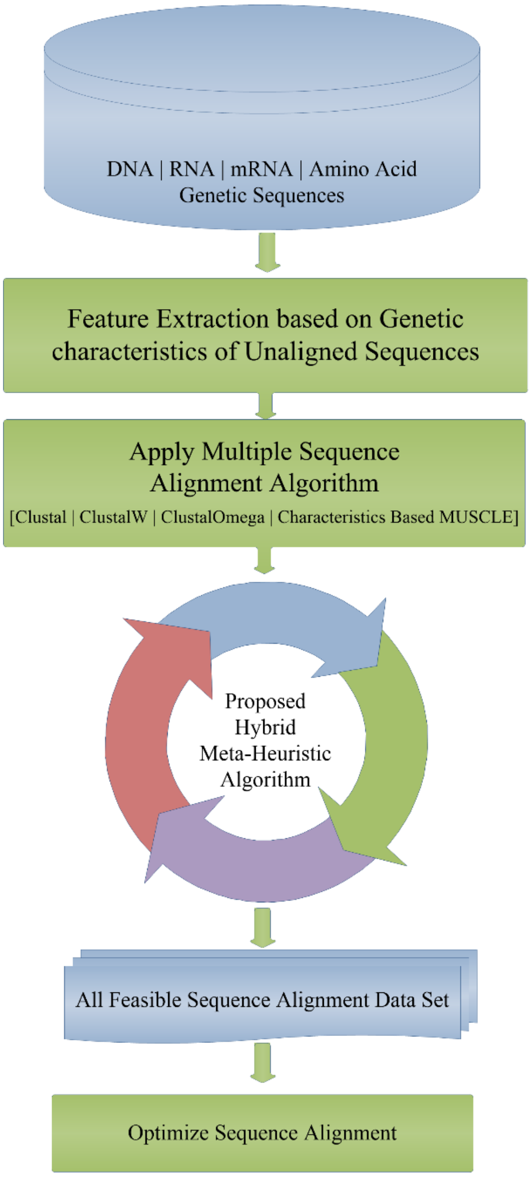

DNA | RNA | mRNA | Amino Acid
Genetic Sequences
Feature Extraction based on Genetic characteristics of Unaligned Sequences
Apply Multiple Sequence Alignment Algorithm
[Clustal | ClustalW | ClustalOmega | Characteristics Based MUSCLE]
Proposed Hybrid Meta-Heuristic Algorithm
All Feasible Sequence Alignment Data Set
Optimize Sequence Alignment

Figure 1: Block Diagram of Proposed Meta-Heuristic MUSCLE Algorithm

The existing framework also generate characteristics knowledge base, which helpful for protein translation, but due to lack of characteristics, this framework is not suitable for large search space. This approach is extended the feature of MUSCLE sequence aligner approach, which limited the search space. To overcome this problem MUSCLE sequence aligner, provide flexibility to extend the features of algorithm by further configuration. The proposed framework overcome all these drawback of existing framework, with the help of meta-heuristic approach of data science, which generate heuristic inspired knowledge base to fill the gap between different sequences alignment results and it's also help to extend the search space in optimal time frame. The proposed algorithm provides native search dimension in existing search space, which helps to improve the optimize results and also helps to track mutation between multiple sequences with effective heuristic knowledge base. Meta-heuristic approach provides artificial intelligence, which learn from past experience with the help of heuristic knowledge base and generate better results from past experiences. The proposed approach also helpful to generate effective phylogenetic trees based on mutation evolution in different genes as well as different species genes based on relative characteristics.

B. *Proposed Hybrid Meta-Heuristic MUSCLE Algorithm*

Algorithm: Multiple Sequence Alignment Hybrid Meta-heuristic MUSCLE Optimization Algorithm.

Input: A List x of n unaligned sequences.

Output: A List y* of n aligned sequences, pair_sum(q), col_score(tc), col_dist_sum(z), consensus_matrix_mean(e), and col_dist_mean(r).

```
Step 1: Start.
Step 2: Extract features from unaligned sequences.
        $\alpha \leftarrow feature\_extraction(seq, char_{\alpha})$
        $\beta \leftarrow feature\_extraction(seq, char_{\beta})$
        $\Upsilon \leftarrow feature\_extraction(seq, char_{\Upsilon})$
        $fx\_seq\_list \leftarrow push(seq, \alpha, \beta, \Upsilon)$
Step 3: Calculate consensus matrices using feature based Muscle algorithm.
    $consensus\_matrices \leftarrow fwkMuscle(fx\_seq\_list)$
Step 4: Initialize first efficient optimization msa consensus matrix.
  $efficient\_optimistic\_msa \leftarrow consensus\_matrices[0]$
Step 5: Calculate heuristic(cmatrix) in consensus_matrices until fullyAlinged == true.
        $efficient\_optimistic\_msa \leftarrow cmatrix$
Step 6: Calculate Z value of consensus Matrix.
                $Z \leftarrow \frac{\sum_{k=0}^{n} Zi}{n}$
        $dist \leftarrow mean(consensus\_matrices)$
Step 7: Calculate meta_heuristic(cmatrix) in consensus matrices and repeat step 8 until dist_cmatrix* <= dist.
        $efficient\_optimistic\_msa \leftarrow cmatrix$
Step 8: If dist_cmatrix* == one or seqTerminal == true
        set $efficient\_optimistic\_msa \leftarrow cmatrix$
Step 9: Calculate Alignment Pair Score.
                $Q \leftarrow \sum_{k=0}^{n} \frac{qi}{qri}$
Step 10: Calculate Alignment Matrix Total Column Score.
                $TC \leftarrow \frac{\sum_{k=0}^{n} Ci}{n}$
Step 11: Calculate Reliability Score based on Column Distance mean.
                $R \leftarrow \frac{\sum_{k=0}^{n} Ri}{n}$
Step 12: Calculate Efficiency Score based on Consensus Matrix Mean.
                $E \leftarrow \frac{\sum_{k=0}^{n} Ei}{n}$
Step 13: Store Optimize Efficiencey factor for future alignment.
Step 14: Stop.
```

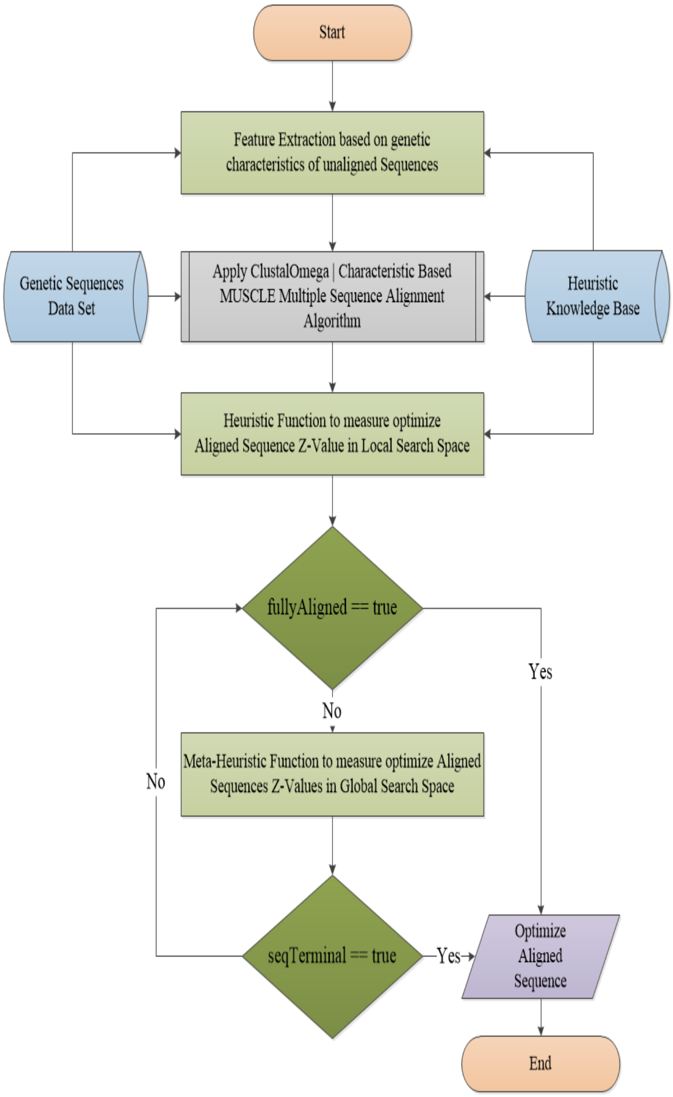

Start
Feature Extraction based on genetic characteristics of unaligned Sequences
Genetic Sequences Data Set
Apply ClustalOmega | Characteristic Based MUSCLE Multiple Sequence Alignment Algorithm
Heuristic Knowledge Base
Heuristic Function to measure optimize Aligned Sequence Z-Value in Local Search Space
fullyAligned == true
Yes
No
No
Meta-Heuristic Function to measure optimize Aligned Sequences Z-Values in Global Search Space
seqTerminal == true
Yes
Optimize Aligned Sequence
End

Figure 2: Flow Chart of Proposed Meta-Heuristic MUSCLE Algorithm

III. RESULTS AND DISCUSSION

A. *Input Data Set (X)*

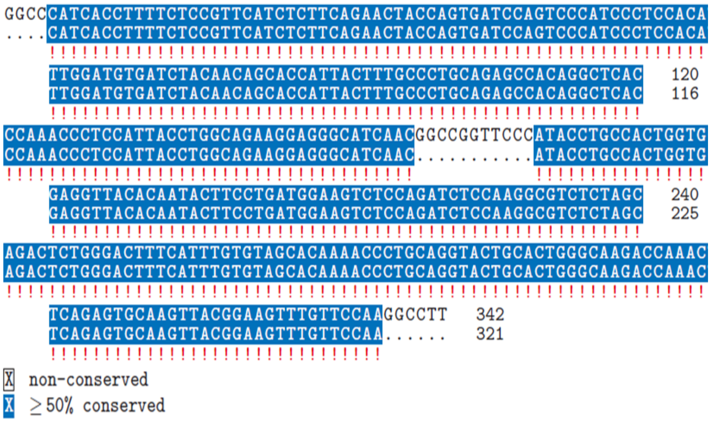
Instance of Genetic Sequences Data Set is generated using microarray. Input Data Set gather from ncbi server. This Data Set include homo sapiens mRNA and poeciliopsis prolifica tRNA sequences.

TABLE I. INPUT SEQUENCES

ID	Name
HMCN_01233.4 Homo_sapiens_1233, micro RNA Sequence	GGCCCATCACCTTTTCTCCGTTCATCTCTTCAG AACTACCAGTGATCCAGTCCCATCCCTCCACA TTGGATGTGATCTACAACAGCACCATTACTTT GCCCTGCAGAGCCACAGGCTCACCCAAACCC TCCATTACCTGGCAGAAGGAGGGCATCAACG GCCGGTTCCCATACCTGCCACTGGTGGAGGTT ACACAATACTTCCTGATGGAAGTCTCCAGATC TCCAAGGCGTCTCTAGCAGACTCTGGGACTTT CATTTGTGTAGCACAAAACCCTGCAGGTACTG CACTGGGCAAGACCAAACTCAGAGTGCAAGT TACGGAAGTTTGTTCCAAGGCCTT
GBYX01016981.1 TSA: Poeciliopsis prolifica comp29840_c0_seq1 transcribed RNA sequence	CATCACCTTTTCTCCGTTCATCTCTTCAGAACT ACCAGTGATCCAGTCCCATCCCTCCACATTGG ATGTGATCTACAACAGCACCATTACTTTGCCC TGCAGAGCCACAGGCTCACCCAAACCCTCCA TTACCTGGCAGAAGGAGGGCATCAACATACC TGCCACTGGTGGAGGTTACACAATACTTCCTG ATGGAAGTCTCCAGATCTCCAAGGCGTCTCTA GCAGACTCTGGGACTTTCATTTGTGTAGCACA AAACCCTGCAGGTACTGCACTGGGCAAGACC AAACTCAGAGTGCAAGTTACGGAAGTTTGTT CCAA

B. Output Alignment (Y*)

Optimize Sequences Alignment Results (Y*) are generated after applying proposed Hybrid Meta-Heuristic Algorithm.



```
GGCCCATCACCTTTTCTCCGTTCATCTCTTCAGAACTACCAGTGATCCAGTCCCATCCCTCCACA
....CATCACCTTTTCTCCGTTCATCTCTTCAGAACTACCAGTGATCCAGTCCCATCCCTCCACA
    !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
    TTGGATGTGATCTACAACAGCACCATTACTTTGCCCTGCAGAGCCACAGGCTCAC   120
    TTGGATGTGATCTACAACAGCACCATTACTTTGCCCTGCAGAGCCACAGGCTCAC   116
    !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
CCAAACCCTCCATTACCTGGCAGAAGGAGGGCATCAACGGCCGGTTCCCATACCTGCCACTGGTG
CCAAACCCTCCATTACCTGGCAGAAGGAGGGCATCAAC...........ATACCTGCCACTGGTG
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!           !!!!!!!!!!!!!!!!
    GAGGTTACACAATACTTCCTGATGGAAGTCTCCAGATCTCCAAGGCGTCTCTAGC   240
    GAGGTTACACAATACTTCCTGATGGAAGTCTCCAGATCTCCAAGGCGTCTCTAGC   225
    !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
AGACTCTGGGACTTTCATTTGTGTAGCACAAAACCCTGCAGGTACTGCACTGGGCAAGACCAAAC
AGACTCTGGGACTTTCATTTGTGTAGCACAAAACCCTGCAGGTACTGCACTGGGCAAGACCAAAC
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
    TCAGAGTGCAAGTTACGGAAGTTTGTTCCAAGGCCTT   342
    TCAGAGTGCAAGTTACGGAAGTTTGTTCCAA......   321
    !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

X non-conserved
X ≥50% conserved
```

C. Evaluation Factors for Best Optimization Results

TABLE II. EVALUATION FACTORS FOR BEST OPTIMIZE RESULTS

#	Evaluation Factor	Optimal Score (Probability)	Optimal Score (%)
1	$\boldsymbol{Q} \leftarrow \sum_{k=0}^{n} \frac{qi}{qri}$	0.9654	96.54%
2	$\boldsymbol{TC} \leftarrow \frac{\sum_{k=0}^{n} Ci}{n}$	0.9395	93.95%
3	$\boldsymbol{Z} \leftarrow \frac{\sum_{k=0}^{n} Zi}{n}$	0.9289	92.89%
4	$\boldsymbol{R} \leftarrow \frac{\sum_{k=0}^{n} Ri}{n}$	0.9458	94.58%
5	$\boldsymbol{E} \leftarrow \frac{\sum_{k=0}^{n} Ei}{n}$	0.9365	93.65%

IV. CONCLUSION

Multiple Sequence Alignment is very complex structure. Uncertainty in different characteristics of sequences generate various optimization problems, which are very difficult to handle by MSA algorithms. To overcome this problem, a native optimize genetic aligner is proposed, which provides optimal solutions using proposed hybrid meta-heuristic optimization algorithm. This proposed system able to work in all different sequence alignment, and it is also flexible in nature via various judgement factors those are able to modify further to get better result as per requirement. This approach generates efficient result for similar characteristics sequences.

REFERENCES

[1] Álvaro R. L., Leonardo V., Mauro C., and Miguel A. V.R. (2015) "A Hybrid Multi-Objective Memetic Metaheuristic for Multiple Sequence Alignment". IEEE Transaction on Evolutionary Computation, vol. 20, pp. 499 - 514.

[2] Álvaro R. L., Leonardo V., Mauro C., and Miguel A. V.R. (2016) "A Characteristic-Based Framework for Multiple Sequence Aligners ". IEEE Transaction on Cybernetics, vol. 48, pp. 41-51.

[3] Alam Manaar, Chatterjee Soumyajit, Banka Haider, "A novel parallel search technique for optimization" in Proceedings of the 3rd International Conference on Recent Advances in Information Technology (RAIT), (Dhanbad, India), ISBN 978-1-4799-8579-1, vol. 1, pp. 259-263, March 2016.

[4] Ánderson R. A. et al., (2016) "Performance Improvement of Genetic Algorithm for Multiple Sequence Alignment". 17th International Conference on Parallel and Distributed Computing, Applications and Technologies, Guangzhou, China, pp. 70-72.

[5] Dengfeng Y., Minghu J. J., Xu. Y., Abudoukelimu A., Renkui H. (2015) "An Algorithm of Multiple Sequence Alignment Based on Consensus Sequence Searched by Simulated Annealing and Star Alignment". International Symposium on Bioelectronics and Bioinformatics (ISBB), Beijing, China, pp. 3-6.

[6] Huazheng Z., Zhongshi H., and Yuanyuan J. (2016) "A Novel Approach to Multiple Sequence Alignment Using Multi-Objective Evolutionary Algorithm Based on Decomposition". IEEE Journal of Biomedical and Health Informatics, vol.20, pp.717-727.

[7] Julie D.T., Patrice K., Raymond R. and Olivier P. Manuel L., Cedric C., Nadia E.M., Aida O. (2005) "BALIBASE 3.0: Latest developments of the multiple sequence alignment benchmark". Proteins, vol. 61, pp. 127-136.

[8] Konstantina K., Costas P., and Dimitrios I. F. (2016) "Integration of Pathway Knowledge and Dynamic Bayesian Networks for the Prediction of Oral Cancer Recurrence". IEEE Journal of Biomedical and Health Informatics, vol.21, pp. 320-327.

[9] Liu Ruoqian, Agrawal Ankit, Liao Wei-keng, Choudhary Alok, Chen Zhengzhang, "Pruned search: A machine learning based meta-heuristic approach for constrained continuous optimization" in Proceedings of the Eighth International

Conference on Contemporary Computing (IC3), (Noida, India), ISBN 978-1-4673-7948-9, vol. 1, pp. 13-18, Aug. 2015.

[10] Manuel L., Cedric C., Nadia E.M., Aida O. (2016) "Gene Tree and Correction using Super Tree and Reconciliation". IEEE Transactions on Computation Biology and bioinformatics vol. PP, pp. 1-1.

[11] Nguyen Su, Zhang Mengjie, Johnston Mark, "A sequential genetic programming method to learn forward construction heuristics for order acceptance and scheduling" in Proceedings of the 2014 IEEE Congress on Evolutionary Computation (CEC), (Beijing, China), ISBN 978-1-4799-1488-3, vol. 1, pp. 1824-1831, July 2014.

[12] Qin A.k., Huang V.L., and Suganthan P.N. (2009) "Differential Evolution Algorithm with Strategy Adaptation for Global Numerical Optimization ". IEEE Transactions on Evolutionary Computation, vol. 13, pp. 2827–2839.

[13] Ranjani R., and Ramyachitra D. (2017) "Application of Genetic Algorithm by by Influencing the Crossover Parameters for Multiple Sequence Alignment". 4th IEEE International Conference on Electrical, Computer and Electronics (UPCON), Mathura, India, pp. 33-38.

[14] Sugawara Kohei, Fujita Hamido, "A workflow optimization by handling subjective attributes with meta-heuristic approach" in Proceedings of the 10th International Conference on Natural Computation (ICNC), (Xiamen, China), ISBN 978-1-4799-5151-2, vol. 1, pp. 497-502, Aug. 2014.

[15] Usman R. and Dennis R.L. (2005) "Probalign: Multiple sequence alignment using partition function posterior probabilities". Bioinformatics, vol. 22, pp. 1-16.

[16] Wei F.G., Ling L.H., San Y.L. and Cai D. (2015) "Artificial Bee Colony Algorithm Based on Information Learning Gene Tree and Correction using Super Tree and Re-conciliation ". IEEE Transactions on cybernetics vol. 45, pp. 2827 - 2839.

[17] Yue J. G. et al., (2016) "Genetic learning particle swarm optimization," IEEE Transaction on Cybernetics. vol. 46, pp. 2277-2290.