

Bio-Inspired Metaheuristic Framework for DNA Motif Discovery Using Hybrid Cluster Based Walrus Optimization

M. Shilpa, C. Nandini

Department of Computer Science and Engineering,
Dayananda Sagar Academy of Technological & Management, Bengaluru, Karnataka, India^{1,2}
Visvesvaraya Technological University, Belagavi, Karnataka, India^{1,2}

Abstract—Motifs are short, recurring sequence elements with biological significance within a set of nucleotide sequences. Motif discovery is the problem of finding these motifs. The problem of motif discovery has become an important problem in the field of Bioinformatics since, it finds its applications in Drug discovery, Environmental Health Research, and early Detection of Diseases by finding anomalies in gene sequences. Motif discovery is a challenging job in bioinformatics since it is NP-hard and cannot be solved within an exact time. In this study, we have proposed Hybrid Cluster based Walrus Optimization algorithm (HCWaOA) to solve the problem of motif discovery. The accuracy and efficiency of the proposed algorithm are improved using a hybrid approach. The population is initialized using Random Projection technique to generate a meaningful solution space. Then, k-means clustering is used to group similar solutions. Lastly, a population-based metaheuristic algorithm, Walrus optimization technique, is applied on each of the clusters to find the best motif. The proposed Hybrid Cluster-based Walrus Optimization algorithm (HCWaOA) is tested on both simulated and real biological datasets. The performance of HCWaOA is compared with benchmark algorithms like MEME, AlignCE and other meta-heuristics algorithms. The results of the proposed algorithm are found to be stable with a precision of 92%, a recall of 93% and an F-score of 93%. The proposed HCWaOA is tested using biological cancer-causing BARC and CTCF datasets to identify cancer causing motifs. Results show that incorporating clustering to initial solution space results in optimal solutions within a fewer iteration. The results of HCWaOA are compared with other popular motifs discovery algorithms and found to be stable.

Keywords—Motifs; walrus optimization algorithm; meta-heuristic algorithms; k-means clustering; DNA; bioinformatics

I. INTRODUCTION

The current Next Generation Sequencing (NGS) technology rapidly sequences the genomes to produce large volumes of high throughput data. The field of bioinformatics involves analyzing these high throughput data to find meaningful inferences. Analyzing these large volumes of high throughput data using traditional approaches is time consuming and prone to error. Regulation of gene expression is carried out using transcription and translation process. Before the transcription process, a specific region of gene is activated which is critical. The activation is caused by special proteins called Transcription factors which bind to specific region in gene called Transcription Factor Binding Sites (TFBS). Hence it is critical

to find TFBS to understand the gene regulation process. These TFBS are known as motifs. Motifs are short, repeated, over represented patterns present in the regulatory regions of gene sequences, which play an important role in gene regulation mechanisms. These motifs are statistically significant, meaning they appear more frequently than expected. They are conserved among the sequences and have biological significance, often related to gene regulation. Therefore, Motif Discovery is one of the important research problems in bioinformatics. The problem of Motif Discovery is challenging because the exact positions of the motif in the gene sequences are not known. Due to mutations, the pattern of motifs across gene sequences can be the same or may vary slightly. Even though a number of motif discovery algorithms have been developed over a period, it is still a complex challenge for biologists and computer researchers.

Given a set of DNA sequences, the problem of motif discovery is to find a common pattern of length l that appears in all or most sequences with a maximum of d mutations. Different Computational approaches have been employed to solve the problem of motif discovery. Enumerative and Probabilistic approaches are the two important classifications of motif discovery algorithms. Enumerative approach involves an exhaustive search to find motifs, but it is impractical and time-consuming based on the nature of motifs. The Probabilistic approach uses Position Weight Matrix (PWM) to represent motifs. But the problem with the probabilistic algorithms is that they may converge to local optima. Other important approaches used in motif discovery are Machine Learning, Metaheuristics and Deep Learning techniques. Each of these algorithms has its own advantages and disadvantages.

Metaheuristic algorithms [1] are computational optimization algorithms that explore a solution space to find acceptable solutions for complex problems. Metaheuristic algorithms have drawn inspiration from nature, some from physics, chemistry, and mathematical concepts. Nature-based metaheuristic algorithms are inspired from natural processes to solve complex problems. Some of the popular and widely used metaheuristic algorithms are Genetic Algorithms based on biological evolution, Particle Swarm Optimization inspired from behavior of flock of birds, Ant Colony Optimization from ant colonies. Other nature-based metaheuristic algorithms are Grey Wolf Optimization, Cuckoo Search Optimization, FireFly Optimization, Walrus Optimization and so on. The advantages

of using nature-based metaheuristic algorithms for DNA motif discovery are the ability to handle large volumes of DNA sequence input data, provide good global exploration and escape the local optima. The limitations of existing algorithms are, they face scalability issues i.e. with the increase in the number of sequences and length of motifs the complexity of the algorithm increases. The tuning of algorithms parameters to derive at the optimal solution is one of the common problems encountered in metaheuristics algorithms.

In the proposed study, we use the Walrus Optimization algorithm to solve the problem of motif discovery. The advantage of using WaOA is, it requires minimal parameter setting with balanced exploration and exploitation. The key contributions of the proposed study are:

- 1) First, the Random projection technique is used for initializing the population solution space. It improves the diversity of the solution space while avoiding random unwanted population space. This acts as a starting point for providing promising solutions.
- 2) A k-means clustering technique is used for clustering the solution space.
- 3) Walrus optimization algorithm is used for motif discovery.
- 4) The proposed HCWaOA algorithm is evaluated using both simulated and biological datasets.
- 5) The proposed HCWaOA algorithm is tested to identify the cancer-causing BRCA1 and CTCF motifs in promoter regions of human gene sequences.

The study is organized as follows: The literature review is presented in Section II, discussing different and current approaches to motif discovery along with metaheuristics algorithms. Section III explains the proposed methodology of HCWaOA. Section IV gives the results, with analysis and evaluation of the proposed HCWaOA. Conclusion and future research directions of HCWaOA are given in Section V.

II. LITERATURE REVIEW

Multiple approaches have been proposed to solve the problem of Motif Discovery, each having its own advantages and disadvantages. Enumerative and Probabilistic approaches are the two major categories of motif discovery algorithms [2]. The other important approaches in motif discovery are Genetic algorithms, Meta-heuristics algorithms, Machine Learning and Deep Learning Techniques. In this section, we discuss the important and latest approaches used currently.

A. Enumerative Approach

The algorithms in Enumerative approach conduct an exhaustive search in entire search space to find the patterns. This covers the entire search space, thus exponentially increasing the time required as the size of the problem space increases. Hence it is suitable for finding motifs of short length in smaller sequences. Enumeration techniques are based on simple words, suffix trees, graphs, hashing. Some of the techniques that come under this category are:

- Brute-force Search: Examines all possible motifs and selects the best one based on a scoring function.

- Consensus-based Methods: Identify the most frequent k-mers (short substrings of length k) across sequences.
- Suffix Trees: Used to efficiently store and search for repeated substrings in a dataset.
- Word-based Methods: Search for overrepresented short words across sequences.

The algorithms developed using this approach are DREME, CisFinder, Weeder, FMotif, MCES.

B. Probabilistic Approach

Probabilistic algorithms compute a position weight matrix or position-specific weight matrix for representing motifs. Some of the important algorithms are Expectation Maximization (EM) and Gibbs sampling.

- Expectation-Maximization (EM) Algorithm: A probabilistic method that iteratively refines a motif model by updating motif positions in sequences.
- Gibbs Sampling: A Markov Chain Monte Carlo method that iteratively samples motif positions to improve alignment.

MEME, STREME [3], EXTREME are based on EM, and AlignCE is based on the Gibbs-Sampling approach.

Machine Learning, Artificial Intelligence and Deep Learning Approaches.

Current trends involve the use of Machine Learning [4], Artificial Intelligence and Deep Learning Approaches [5] in motif discovery. Hidden Markov Models (HMMs) are probabilistic models. Neural Networks and Deep Learning models are used to detect motifs. Deep Learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Hybrid CNN RNN, Ensemble-based models and complex hybrid models are applied to Motif Discovery. Supervised learning, like Random Forests and Support Vector Machines (SVMs), is used in settings when labelled motif data is available, whereas Autoencoders are used for unsupervised motif discovery by reducing dimensionality and extracting patterns. DeepBind, DeepVISP [6], and Deep6mA are some of the popular deep learning approaches used in the problem of motif discovery.

C. Nature Inspired Approach

Metaheuristic algorithms are approximate methods which provide acceptable solution within reasonable computational cost. Metaheuristic algorithms have drawn its inspiration from nature. The solution to the problem is based on natural selection and survival of the fittest. Some of the popular and widely used nature inspired metaheuristic algorithms are Genetic Algorithms based on biological evolution, Particle Swarm Optimization based on behavior of flock of birds [7], Ant Colony Optimization based on ant colonies. Other nature-based metaheuristic algorithms [8] are Grey Wolf Optimization, Cuckoo Search Optimization, FireFly Optimization, Walrus Optimization [9] and so on. In [10] and [11], Chemical Reaction Optimization (CRO) and Henry Gas Solubility optimization which are based on physics and chemistry are used. Hybrid metaheuristic algorithms have given good results compared to

traditional approaches. Recent trends in motif discovery algorithms are discussed below:

The study in [12] is a physics inspired metaheuristic for DNA motif discovery proposed in 2025. Here Archimedes Optimization Algorithm (AOA) mimics the principles of buoyant force and object equilibrium in fluids, balancing exploration (global search) and exploitation (local refinement) effectively. The Information Content is used for fitness evaluation. This technique is tested using real DNA dataset and the performance is evaluated against other metaheuristics algorithm. The major drawback is scalability issue to handle huge data and risk of convergence to local optima. The algorithm needs optimal tuning of control parameters to achieve optimal solutions.

In 2023, Dang et al [13] proposed a hybrid genetic algorithm to discover DNA motifs that satisfy the 2-Optimality postulate. 2-optimality postulate means the motif should be found in minimum of two input sequences. It uses evolutionary operations selection, crossover and mutation. It is tested against a benchmark dataset and provides better accuracy. The 2-optimality postulate may sometimes not discover relevant motifs.

In 2023, Theepalakshmi, P. and Reddy, U.S., [17] have proposed a novel effective quorum seeded (ℓ , d) motif search utilizing segmentation to filter using freezing firefly approaches on ChIP-seq data. The final motif was identified using the usual firefly approach, which uses both local and global freezing techniques. The effectiveness of these techniques was assessed using both simulated and actual datasets, including the human ChIP-seq dataset, the mouse emergent stem cell dataset, with *Escherichia coli* cyclical AMP receptor protein (CRP) dataset. It attains high F-measure and low accuracy.

In [14], the authors proposes a Freezing FireFly algorithm to solve the motif discovery problem. Here local and global freezing strategy is used wherein the best possible positions of the poor solution is also preserved. This helps in the overall search process to find new possible good results. The performance of Freezing Firefly algorithm is evaluated by comparing with benchmark tools like Samselect, TraverStringRef, PMS8, qPMS9, AlignACE, FMGA, and GSGA.

In 2022, Li, Song et al. [15] proposed an improved Henry gas optimization algorithm with Levy mechanism and Brown motion. The study in [11], is Henry gas solubility optimization (MHGSO) algorithm for motif discovery. In MHGSO, the optimal solutions are obtained by evaluating the characteristics of the candidate solution space. It is based on chemistry, Henry's Law, treating the search space as gas molecules and mimics the finding of best motif to adjusting the solubility of gas molecules. The performance of MHGSO algorithm is evaluated on both synthetic and real datasets to find accurate motifs. The limitations of the algorithm is control parameters are specified which make the algorithm less flexible.

In 2024, Mohammad Hasan, et.al. [16] have proposed a Trie-PMS8. The algorithm is based on enhanced trie-tree for planted motif search problem. The problem with earlier PMS algorithm

is that the time complexity increases exponentially in worst case scenarios. The proposed trie tree uses sort row by size step to reduce the time and linked lists to reduce the space. It also uses dynamic programming techniques to avoid redundant calculations in frequent tree processing. It gives better results than earlier versions of PMS with reduced time complexity.

In 2024, Qiang Yu, et.al. [18] have proposed an exact Planted Motif Search (PMS) on large DNA sequences. The efficient and exact algorithm finds (ℓ , d) motifs using the technique of searching the branches on the pattern tree. The algorithm has good running time ratio compared to the existing PMS algorithms. The algorithm is tested on challenging problem instances of large DNA sequence datasets.

In 2024, Ledesma-Dominguez, et al. [19] have proposed a hybrid model named Deep Regulation (DeepReg) to identify transcription factor binding sites in prokaryotic and eukaryotic protein sequences. The hybrid model uses CNN, BiLSTM and attention mechanism. Feature extraction and grammar regulation is done using CNN and BiLSTM respectively. This leads to enhanced F1-score and performance in DeepReg compared to the other deep learning models. The model showed reliability and robustness for unseen experimental data. It provided low variance and eliminated overfitting problem. The model was tested on three organisms *S. cerevisiae*, *N. crassa*, and *A. nidulans* [19] and on average identified 71.8% of transcription factors.

The motif discovery problem is essential for understanding regulatory elements in DNA sequences. Various computational techniques, ranging from simple brute-force methods to advanced deep learning models, are used to identify biologically significant motifs. Despite the progress, motif discovery continues to be a challenging problem in genomics and bioinformatics research. The major issues with the existing metaheuristics algorithm are:

- Results are dependent on algorithms parameters tuning like the mutation rate, co-efficient, optimal size of population and iterations.
- Scalability issues as the number of sequences, length of sequences and motif length increases.

The aim of the research is to address the limitations of the existing metaheuristics algorithm and propose a hybrid metaheuristics algorithm to address the above issues.

III. PROPOSED METHOD

In this study, a Hybrid Cluster based Walrus Optimization algorithm (HCWaOA) is proposed for finding motifs in the promoter regions of a given set of DNA sequences. The first step of the proposed algorithm, builds an initial population solution using random projection strategy [20]. The second step involves clustering the population solutions using k-means clustering. In the third step, modified Walrus optimization algorithm is applied on the clusters to find the best solution. Here the best solution is the motif with highest fitness value found in the given set of DNA sequences. The architecture block diagram of the HCWaOA is given in Fig. 1.

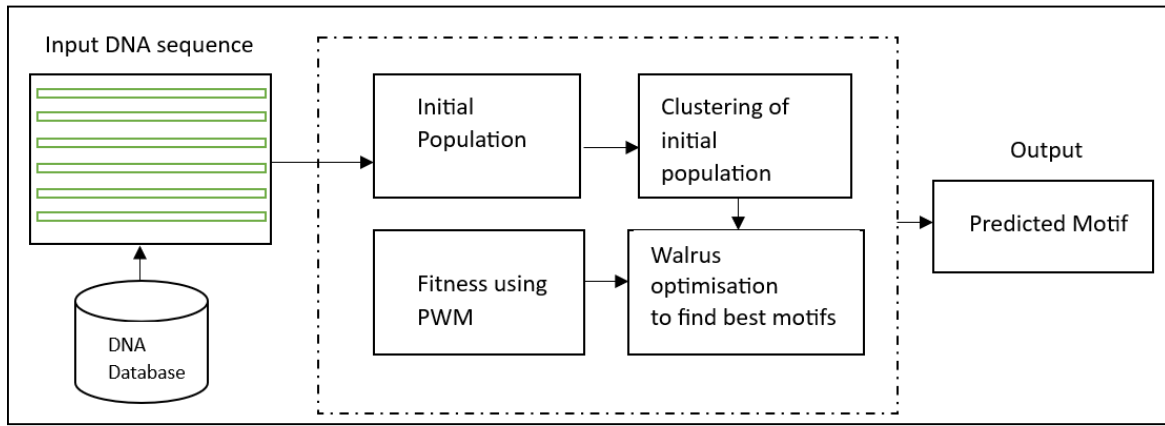


Fig. 1. Block diagram of HCWaOA.

A. Walrus Optimisation Algorithm

Walrus Optimization Algorithm (WaOA) is a new nature-based metaheuristic algorithm that is based on the behavior of walrus. It draws inspiration from Walrus behavior of feeding, migrating, escaping predators and fighting predators [21]. The Walrus Optimization Algorithm is implemented in three phases: exploration, migration, and exploitation. The working of the algorithm is given by the following steps:

1) *Initialization*: A population of walrus solutions is randomly generated within the search space, representing potential solutions to the optimization problem.

2) *Movement based on behavior*:

Phase 1: Feeding (Exploitation): When a safety signal [22] is identified, the walruses feed by refining their positions, moving closer to the optimal solution. The new position is calculated using Eq. (1):

$$X_i^{t+1} = X_i^t + r (S^t - I \cdot X_i^t) \quad (1)$$

X_i^t : Position of the i^{th} walrus at iteration t

S^t : Best fitness solution

I : control parameter

r : random vector [0,1]

Phase 2: Migration: Walruses are simulated to migrate towards areas with better potential solutions, based on the random solution. The new position for migration is calculated using Eq. (2):

$$X_i^{t+1} = X_i^t + r (X_k^t - I \cdot X_i^t) \quad (2)$$

X_k^t : Position of randomly selected walrus

Phase 3: Escape (Exploration): When a danger signal [22] is triggered, the walruses escape by making large random movements to explore new regions of the search space. The new position for exploitation is calculated using Eq. (3):

$$X_i^{t+1} = X_i^t + r \cdot \delta \quad (3)$$

$$\delta = \alpha \cdot (ub - lb)$$

α : random value [-0.1, 0.1]

ub, lb : upper and lower bounds of the search space

3) *Updating positions*: The position of each walrus (candidate solution) is updated based on the calculated danger signals and safety signals, influencing how much exploration or exploitation is performed in each iteration.

4) *Iteration and convergence*: The process of movement and position updates is repeated for a set number of iterations, with the best solutions gradually converging towards the optimal solution.

B. Random Projection Technique

The starting point of the algorithm is the generation of initial population using on random projection technique. This involves selecting random sub sequences of motif length from the given set of input DNA sequences. These sub sequences act as the initial population i.e. candidate solutions. This gives a set of candidate solutions which are closer to the optimal solution since they are already a part of the solution space. This strategy is better than generating random candidate solution which are not related to the solution space. The number of iterations and time required to reach the optimal solution is reduced and a meaningful initial candidate solution space is generated as the input to walrus algorithm.

C. Clustering

A k-means clustering is used to cluster initial candidate solutions. Clustering allows us to group highly similar motifs. This allows for easier analyzing of pattern, reducing noise and identifying biologically significant motifs. The problem with initial candidate solutions of metaheuristic algorithms is they are random and no relation exists between them. It is found that individual elements of candidate solution which are similar or closer have higher fitness then dissimilar elements.

Clustering techniques [23] like k-means help group similar motifs which may have variations due to mutations, insertions or deletions together, making the analysis more structured and meaningful. The partitioning of the population into multiple clusters allows only intra cluster operations like mutation, crossover within the cluster whereas multiple clusters allow for diversity in the population. Multiple clusters allow us identify multiple weaker motifs which are similar. Hence clustering efficiently organize motifs and improves the motif discovery

process. K-means is used in the proposed study since it is computationally efficient and can handle large-scale motif datasets.

D. Fitness Function

The fitness or the quality of the motifs is evaluated based on the information content (IC). IC is calculated using the following steps:

1) *Compute PFM (Position Frequency Matrix)*: A Position Frequency Matrix (PFM) represents how frequently each nucleotide i.e. A, C, G, T appears at each position in a set of aligned motifs.

2) *Computer PWM (Position weight matrix)*: The PWM is obtained by normalizing the PFM with background nucleotide probabilities and taking the logarithm, as in Eq. (4):

$$PWM_{b,j} = \log_2 \left(\frac{PFM_{b,j}}{P_{background}(b)} \right) \quad (4)$$

Here, $PWM_{b,j}$ is the weight for nucleotide b at position j , and $P_{background}(b)$ is the background probability of nucleotide b . This is often assumed to be 0.25 for uniform distribution.

This transformation helps identify conserved positions in the motif.

3) *Computer IC*: quantifies how well a motif is conserved. It measures how different each position is from a random sequence.

Eq. (5):

$$IC_j = \sum_b P_{b,j} \log_2 (P_{b,j} / P_{background}(b)) \quad (5)$$

Total IC is the sum across all motif positions, as in Eq. (6):

$$IC = \sum_{j=1}^L IC_j \quad (6)$$

A higher IC means the motif is more conserved and biologically significant. The calculated IC value serves as the fitness score.

E. Process of Walrus Optimisation Algorithm for Motif Discovery

The Walrus Algorithm efficiently discovers motifs by combining random exploration with local exploitation. Its adaptive approach ensures a balance between global search (exploration) and local refinement (exploitation), making it well-suited for motif discovery in DNA sequences. Walrus Optimization technique operates on each of the clusters. Thus, returning the best solution in each of the clusters. The results are evaluated to obtain global best solution based on the fitness. The global best solution is the motif with fitness score having highest IC value. Algorithm 1 presents the pseudocode of the proposed HCWAO algorithm.

Algorithm 1: Pseudocode of the proposed HCWAO algorithm

1. Set all the parameter values
 2. Initialise the population
-

3. Calculate the fitness value using the fitness function to get the best solution
 4. while ($t < \text{max_generations}$)
 5. Compute adaptive exploration and exploitation rates
 6. Update Motif Positions
 7. For each motif set in population:
 - Phase 1: Feeding (Exploitation)
 - Calculate the new position using Eq1
 - Update the position of new walrus
 - Phase 2: Migration
 - Using a random walrus, find the position of new walrus using Eq2
 - Update the position of new walrus
 - Phase 3: Escape (Exploration):
 - Calculate the new position using Eq3
 - Update the position of new walrus
 8. Compute new fitness scores using Eq4, Eq5, Eq6
 9. Update the best solution with highest fitness score
 10. Return the best motif with the corresponding fitness score
-

IV. EXPERIMENTAL RESULTS AND DISCUSSION

HCWAO algorithm is implemented in Python using Anaconda IDE under Windows OS and tested on both real-time and simulated dataset. The test results are analyzed and the performance is compared with existing state-of-art tools such as MEME, DREME and MHGSO [11] systems. The proposed algorithm is tested to identify breast and ovarian cancer causing motifs BRCA1 and CTCF in human gene sequences.

A. Performance Measure

The performance of the HCWAO is evaluated using the metrics Precision (P), Recall (R) and F-score [24]. Precision P is number of predicted motifs that are true divided by number of predicted motifs. Recall R is number of predicted motif sites that are true divided by number of true motifs. F-score is computed using the values of Precision and Recall. The best value for F-score is 1 and worst is 0. TP is true positive, TN is true negative, FP is false positive, FN is false negative.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

$$F - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

B. Test on Simulated Data

The algorithm is tested using 7 groups synthetic dataset. The details of the dataset are given in Table I. The sequences are generated randomly and motifs with mutations are added at random positions using python script.

The proposed algorithm is tested on the synthetic dataset given in Table I. The detailed results for each of the 7 dataset groups is given in Table II. The algorithm is run for 10 times and the best value of P and R is taken into consideration. Table II

gives the average values of P, R and F -score for 7 dataset groups. The results show a higher F-score attained through higher values of P and R. Thus, results show a higher F-score for dataset (8, 2) and (9,3). We can infer that as the length of the motif increases with the increase in the number of mutations. It is difficult to identify all the correct motifs since higher R value means large number of candidate instances which cause false positive. Prediction results can be sensitive to the dataset properties like the number and length of sequences generated, true and mutated motifs in the input sequence.

TABLE I. SYNTHETIC DATASET DETAILS

Dataset group	1	2	3	4	5	6	7
No of sequences (t)	40	40	40	40	40	40	40
Length of sequence (n)	400	400	400	400	400	400	400
Length of motif (l)	8	9	10	14	16	18	21
Maximum mutations	2	3	3	4	5	6	7

TABLE II. RESULTS FOR SYNTHETIC DATASET ON PROPOSED ALGORITHM: P IS PRECISION, R IS RECALL AND F IS F-SCORE

SI NO	Dataset	P	R	F - score
1	(8,2)	0.95	0.95	0.95
2	(9,3)	0.95	0.95	0.95
3	(10,3)	0.90	0.92	0.91
4	(14,4)	0.93	0.93	0.93
5	(16,5)	0.93	0.93	0.93
6	(18,6)	0.90	0.92	0.91
7	(21,7)	0.90	0.92	0.91
	Avg	0.92	0.93	0.93

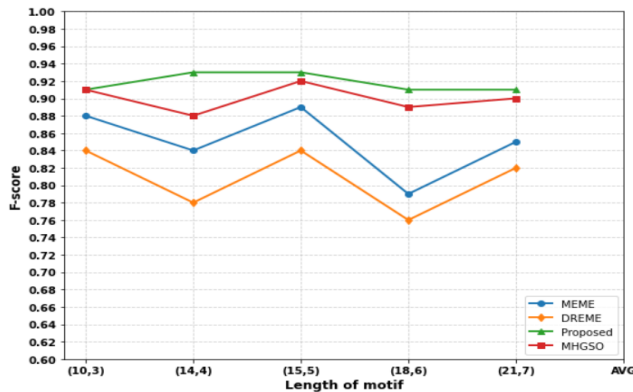


Fig. 2. Analysis of F-score results on synthetic dataset.

The synthetic datasets (10, 3), (14,4), (15,5), (18,6) and (21,7) are familiar examples in motif discovery problem. Fig. 2 shows the F-score on synthetic dataset. The results of HCWAOA are compared with the results of state-of-art motif discovery tools MEME, DREME and metaheuristic algorithm MHGSO [11]. The proposed algorithm provides an improved F-score compared to the other algorithms. The clustering approach allows for grouping of similar solution space and walrus optimization exploits the solution space to get better results. The hybrid approach of HCWAOA allows a thorough exploration of

the search space. This thorough exploration reduces the chances of missing optimal solutions and increases the overall accuracy of the optimization task.

C. Convergence of Proposed HCWAOA

The proposed HCWAOA algorithm detects the convergence based on the fitness of the motifs. The fitness, i.e. IC is calculated using Eq. (6) across fixed iterations. If the fitness value reaches the highest and remains the same for fixed number of iterations then the algorithm is terminated. This prevents the overhead of computation once convergence is reached. Fig. 3 shows the convergence curve for dataset group 1, 2 and 3 of Table I.

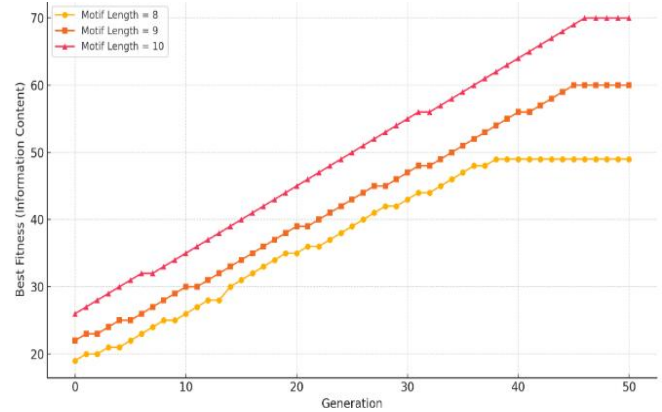


Fig. 3. Convergence curve for motifs of length 8, 9 and 10.

D. Analysis of k-means Clustering on Walrus Optimisation

The proposed Walrus optimization algorithm was tested with and without k-means clustering on the initial population for simulated dataset given in Table I. The fitness was calculated using the Eq. (6). The average fitness value was calculated across 15, 25, 35 and 50 generations. The results in Table III show an improvement in average fitness value when clustering is employed. It was also observed that new weaker motifs are identified in the clusters which was not a part of the simulated dataset. Fig. 4 infers; clustering discovers better motifs i.e. motifs with better fitness values within fewer iterations. Hence by incorporating clustering on initial population, the proposed HCWAOA allows for discovering of motifs with higher fitness values within given iterations. This overcomes the complexity issue of metaheuristics algorithms.

TABLE III. PERFORMANCE COMPARISON OF HCWAOA WITH CLUSTERING AND WITHOUT CLUSTERING

SI NO	Dataset	Without clustering	With Clustering
		Average Fitness	Average Fitness
1	(8,2)	51.14	53.18
2	(9,3)	53.46	57.95
3	(10,3)	58.78	62.29
4	(14,4)	82.9	98.29
5	(16,5)	92.6	101.74
6	(18,6)	100.49	126.85
7	(21,7)	110.39	146.17

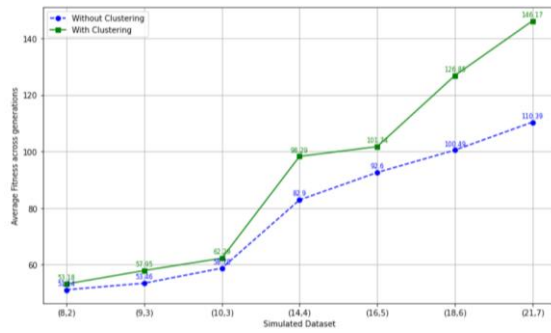


Fig. 4. Performance comparison of HCWAOA with and without clustering.

E. Test on Real-Time Data

The proposed algorithm is tested using a CRP benchmark dataset. The benchmark dataset is a real-time dataset consisting of 18 sequences of *Escherichia coli* of 105 bp-long. The proposed HCWAOA algorithm finds the motifs and their starting positions for the CRP *E. coli* input sequence. Table IV shows the results of the proposed algorithm in comparison with MEME and AlignCE. Table IV gives the actual starting position of the known motif, the predicted motif start position from different algorithms and the deviations between actual and predicted results. The proposed algorithm gives most of the positions correctly, and the results are better than existing algorithms MEME and AlignCE.

F. Test on Breast Cancer Gene Input Dataset


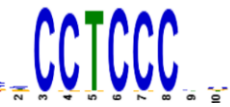


The proposed algorithm is tested to check its ability to discover real biological data in human DNA sequence. The input gene sequences are collected from well curated JASPAR database and government authorized website NCBI. These are breast cancer gene input sequences containing BRCA1 [25] and CTCF motifs. These BRCA1 and CTCF motifs are associated with breast and ovarian cancer. The identification of BRCA1

and CTCF motifs in the input data sequence infers the detection of cancer. Table V gives the details of input sequences along with their accession numbers as well as true motifs and predicted motifs. The proposed algorithm was run on the cancer dataset, and the experimental results show that the predicted motifs match with known cancerous motifs.

TABLE IV. COMPARISON RESULTS OF HCWAOA, MEME, ALIGNCE FOR CRP BENCHMARK DATASET

Sequence No	Starting position of the known motif	MEME	AlignCE	Proposed algorithm (HCWAOA)
1	17,61	61	63(2)	61
2	17,55	55	57(2)	55
3	76	76	78(2)	76
4	63	63	65(2)	63
5	50	13 (-37)	52(2)	50
6	7,60	7	9(2)	7
7	42	42	26(-16)	24(-14)
8	39	39	41(2)	39
9	9,80	9	11(2)	9
10	14	14	16(2)	14
11	61	35	63(2)	61
12	41	34	43(2)	51(10)
13	48	48	50(2)	48
14	71	71	73(2)	71
15	17	75 (58)	19(2)	17
16	53	6	55(2)	53
17	1,84	27 (26)	68(16)	5(4)
18	78	76 (-2)	80(2)	78

TABLE V. LIST OF CANCER CAUSING BRAC1 AND CTCF GENE INPUT DATASET WITH PUBLISHED AND PREDICTED MOTIFS

Name	Seq No	Accession no.	Published Motif using MEME & TOMTOM	Predicted Motif from proposed HCWAOA algorithm
BRCA1	15 sequences	AF507075.1 AY093484.1 AF507076.1 AF507077.1 AF507078.1 AY093484.1 AY093486.1 AY093487.1 AY093488.1 AY093489.1 AY093492.1 AY093493.1 AY093490.1 AY093491.1 AF284812.1	 Name : SP5(C2H2 zinc finger factors) Matrix ID: MA1965.1	
CTCF	30 sequences	JASPAR	 Name : CTCF Matrix ID: MA0139.1	

The predicted motifs identified by the proposed algorithm is validated by checking the similarity with annotated motifs. Motif similarity analysis was evaluated using TomTom from the MEME tool-kit. It was also found that the discovered motif had clear resemblance with known motif in JASPAR database. The predicted motifs logo was created using the tool <https://weblogo.berkeley.edu/logo.cgi>.

V. CONCLUSION

The study proposes, hybrid cluster-based motif discovery approach using a new metaheuristic walrus optimization algorithm. Using the proposed technique, adjusting the algorithm parameters is considerably reduced compared to the other metaheuristic algorithms and this in turn efficiently discovers the motifs in gene sequences. The performance of the proposed HCWAO algorithm is evaluated using both synthetic and real dataset. The results on synthetic and real dataset are better compared to existing well-known traditional methods. Experimental results have shown that the proposed algorithm can discover motif of both short and long length sequences with mutations. By employing clustering, better motifs can be identified within fewer iterations thereby reducing the complexity of the algorithm. Here the real time application of the proposed algorithm is verified using cancer gene sequences. The proposed HCWAO algorithm is successfully applied for the detection of CTCF and BRCA1 binding sites in homosapiens, whose deletion or inactivation has been detected in various cancers.

Although, the proposed algorithm gives good results, it can be extended to handle motifs of unknown length, thus providing more flexibility in discovering motifs of any length. In future a better clustering technique can applied which utilizes the properties of motifs. Even though multimodal metaheuristics algorithm solves some of the issues with motif discovery, deep learning techniques that are robust and scales well to handle large gene dataset can be explored in future.

REFERENCES

- [1] Almufti, Saman & Ali, Rasan & Fuente, Jayson. (2023). Overview of Metaheuristic Algorithms. *Polaris Global Journal of Scholarly Research and Trends*. 2, 10-32. 10.58429/pgjst.v2n2a144.
- [2] Shilpa, M., and C. Nandini. "A Survey on Motif Discovery Algorithms for analysis of Gene Sequences of Interest." In 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), vol. 1, pp. 275-278. IEEE, 2021.
- [3] Timothy L Bailey, STREME: accurate and versatile sequence motif discovery, *Bioinformatics*, Volume 37, Issue 18, September 2021, Pages 2834–2840, <https://doi.org/10.1093/bioinformatics/btab203>
- [4] A. Yang, W. Zhang and J. Wang, "Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA", *Frontiers in Bioengineering and Biotechnology*, Vol. 8, No. 2, September, pp. 1– 13, 2020, DOI: 10.3389/fbioe.2020.01032.
- [5] Quang D,XieX. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;44:e107–7
- [6] Xu, Haodong & Jia, Peilin & Zhao, Zhongming. (2021). DeepVISP: Deep Learning for Virus Site Integration Prediction and Motif Discovery. *Advanced Science*. 8. 2004958. 10.1002/advs.202004958.
- [7] Uyyala Srinivasulu Reddy, Michael Arock, and A.V. Reddy. 2020. Discovering of gapped motifs using particle swarm optimisation. *Int. J. Comput. Intell. Bioinformatics Syst. Biol.* 2, 1 (2020), 1–21. <https://doi.org/10.1504/ijcibsb.2020.106858>
- [8] Vasuki, A. (2020). Nature-Inspired Optimization Algorithms. 10.1201/9780429289071.
- [9] Trojovský, P., Dehghani, M. A new bio-inspired metaheuristic algorithm for solving optimization problems based on walrus behavior. *Sci Rep* 13, 8775 (2023). <https://doi.org/10.1038/s41598-023-35863-5>
- [10] Saha, Sumit & Islam, Md & Hasan, Mredul. (2021). DNA motif discovery using chemical reaction optimization. *Evolutionary Intelligence*. 14. 10.1007/s12065-020-00444-2.
- [11] Fatma A. Hashim, Essam H. Houssein, Kashif Hussain, Mai S. Mabrouk, and Walid Al-Atabany. 2020. A modified Henry gas solubility optimization for solving motif discovery problem. *Neural Comput. Appl.* 32, 14 (Jul 2020), 10759–10771. <https://doi.org/10.1007/s00521-019-04611-0>
- [12] Hashim, F.A. *et al.* (2025). Archimedes Optimization Algorithm for DNA Motif Discovery. In: Dulhare, U.N., Houssein, E.H. (eds) *Deep Learning and Computer Vision: Models and Biomedical Applications. Algorithms for Intelligent Systems*. Springer, Singapore. https://doi.org/10.1007/978-981-96-1285-7_1
- [13] Dang, D.T., Nguyen, N.T. & Hwang, D. Hybrid genetic algorithms for the determination of DNA motifs to satisfy postulate 2-Optimality. *Appl Intell* 53, 8644–8653 (2023). <https://doi.org/10.1007/s10489-022-03491-7>
- [14] Theepalakshmi, P & Reddy, U. Srinivasulu. (2022). Freezing firefly algorithm for efficient planted (l, d) motif search. *Medical & biological engineering & computing*. 60. 10.1007/s11517-021-02468-x.
- [15] Li, Song et al. "An improved Henry gas solubility optimization algorithm based on Lévy flight and Brown motion." *Applied Intelligence* 52 (2022): 12584 - 12608.
- [16] Mohammad Hasan, Abu Saleh Musa Miah, Md. Humaun Kabir, Mahmudul Alam, Trie-PMS8: A trie-tree based robust solution for planted motif search problem, *International Journal of Cognitive Computing in Engineering*, Volume 5, 2024, Pages 332-342, ISSN 2666-3074, <https://doi.org/10.1016/j.ijcce.2024.07.004>
- [17] P. Theepalakshmi, U. Srinivasulu Reddy, A new efficient quorum planted (l, d) motif search on ChIP-seq dataset using segmentation to filtration and freezing firefly algorithms, *Soft Computing*, 10.1007/s00500-023-09236-z, 28, 4, (3049-3070), (2023).
- [18] Yu, Qiang, et al. "An efficient exact algorithm for planted motif search on large DNA sequence datasets." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2024).
- [19] Ledesma-Dominguez, L., Carbajal-Degante, E., Moreno-Hagelsieb, G. et al. DeepReg: a deep learning hybrid model for predicting transcription factors in eukaryotic and prokaryotic genomes. *Sci Rep* 14, 9155 (2024). <https://doi.org/10.1038/s41598-024-59487-5>
- [20] Ge, Hongwei & Yu, Jinghong & Sun, Liang & Wang, Zhen & Yao, Yao. (2019). Discovery of DNA Motif Utilising an Integrated Strategy Based on Random Projection and Particle Swarm Optimization. *Mathematical Problems in Engineering*. 2019. 1-12. 10.1155/2019/3854646.
- [21] Muxuan Han, Zunfeng Du, Kum Fai Yuen, Haitao Zhu, Yancang Li, Qiuyu Yuan, Walrus optimizer: A novel nature-inspired metaheuristic algorithm, *Expert Systems with Applications*, Volume 239, 2024, 122413, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2023.122413>
- [22] P. Trojovský and M. Dehghani, "Walrus Optimization Algorithm (WAO), MATLAB Central File Exchange," 2023. <https://doi.org/10.21203/rs.3.rs-2174098/v1>
- [23] F. B. Ashraf, A. Matin, M. S. R. Shafi and M. U. Islam, "An Improved K-means Clustering Algorithm for Multi-dimensional Multi-cluster data Using Meta-heuristics," 2021 24th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2021, pp. 1-6, doi: 10.1109/ICCIT54785.2021.9689836
- [24] Mabrouk, Mai & Abdelhalim, mohamed b & Elewa, Ebtehal. (2018). A developed system based on nature-inspired algorithms for DNA motif finding process. *Neural Computing and Applications*. 30. 10.1007/s00521-018-3398-0.
- [25] Bhargavi, Peyakunta & Lakshmi, Kanchi & Jyothi, Singaraju. (2020). Gene Sequence Analysis of Breast Cancer Using Genetic Algorithm. 10.1007/978-981-15-0135-7_16.