www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

AN OPTIMIZATION APPROACH FOR MULTIPLE MOTIFS DISCOVERY IN DNA SEQUENCE USING LVCGRO

Prince Joseph[1]

[1]Lecturer in Computer Engineering, Govt Polytechnic College Pala, India.

DOI: https://www.doi.org/10.58257/IJPREMS32117

ABSTRACT

Extracting regulatory motifs from the DNA sequence seems to increase with intense enthusiasm. The existing motif discovery models failed to cover multiple motifs that occurred in DNA sequence, did not run the risk of getting stuck in a local optimum, and were time-consuming in handling dozens of sequences. Therefore, in this work, an approach based on multi-objective optimization is used for motif discovery. Initially, the input DNA sequences are preprocessed to remove the inconsistencies and divide the lengthy sequence into shorter lengths. After that, the gram-tree representation is used to find all motifs effectively in the target sequence. The resultant output in the form of the alphanumeric DNA sequence is converted into the numerical form using the OH-ECE, and the MNBIRCH is used to categorize the motifs DNA sequence. Finally, to filter out the optimal motifs present in the DNA sequence, LVCGRO is used. Experimental results show that the proposed method is successful in discovering the multiple motifs from the DNA sequence with a high accuracy of 98.561%.

Key Words: gram-tree construction, DNA sequence, Linearly Varying Constricted Gold Rush Optimization (LVCGRO), Multi-dimensional Nested Balanced Iterative Reducing And Clustering Using Hierarchies (MNBIRCH), One Hot Entropy-weight Correlated Encoding (OH-ECE), kmers Percent identify mean.

1. INTRODUCTION

In the line of many biological challenges, Deoxyribose Nucleic Acid (DNA) motif identification is a preliminary and crucial step for studying the gene function. In general, all cellular activities of living organisms are determined by the proper gene expression under the modulation of transcription regulatory elements and corresponding transcription factors (Yousif et al., 2023)(Tapan, 2023). The transcription regulatory elements are referred to as the transcription factor binding sites or DNA motif, which helps in learning the gene expression regulation mechanisms (Li, 2023). They are fairly short nucleic acid sequence patterns with specific structures(Wang et al., 2022) bounded by a specific group of proteins known as transcription factors, which occur repeatedly within the gene. Thus, the gene bounded by the same transcription factors expresses the same biological functions (Alam et al., 2021). As a result, the DNA motif discovery problem can be defined as a process of searching for the transcription factor binding site location information from a set of genes with similar biological functions (He et al., 2021). A large number of algorithms, which are divided into word-based and profile-based categories, have been developed for finding DNA motifs (Donohue et al., 2022)(Hammelman et al., 2021). However, these algorithms are not guaranteed to find globally optimal solutions, forcing users to repeatedly try the discovery process, which slows down its processing performance (Zhang et al., 2022). Several deep learning-based approaches have been introduced to distinguish sequences that contain a motif (Rahman et al., 2021).

1.1 Problem Statement

The existing research on DNA motif identification has certain limitations such as,

- Extracting multiple motifs from the discovered motifs is a complex task that is rarely done.
- Existing techniques are time-consuming and easily trapped in a local optimum solution in handling dozens of sequences.
- Human intervention for discovering parameters is required in prevailing techniques.
- The lack of scalability is another problem that exists in prevailing techniques.

Under these motivations, the proposed model contributes the following as,

- To discover multiple motifs, a novel objective function is introduced.
- To prevent the local optima solution, the improved optimization with a linearly time-varying constriction factor is used.
- To reduce human interventions, motif discovery using the MNBIRCH algorithm is carried out.
- To improve the scalability, the gram-tree construction and numeric conversion are performed.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

The remaining paper is organized to have an overview of the literature in section 2, describe the proposed model in section 3, evaluate the proposed model in section 4, and conclude the paper in section 5.

2. RELATED WORK

(Ge et al., 2021) presented an improved version of the Bayesian Markov Model of motif, which was termed BaMMmotif2. The BaMMmotif2 had no signs of overtraining in cross-cell and cross-platform tests, with similar improvements on the next-best tool. However, the understanding was missing to know which Markov chain was converging quickly or slowly.

(Choi et al., 2022) recommended a minimal motif for motif sequence recognition by mitochondrial transcription factors. The interactions were established with two guanine bases separated by random nucleotides. The results demonstrated that the specified random nucleotides were essential for binding. It only considered specific DNA binding and did not focus onthe common binding process.

(Maity et al., 2020) presented an NMR structure of a sequence named AT26, consisting of irregularly spaced G2 tracts and two isolated single guanines. The structure was a four-layered G4 featuring two bi-layered blocks.The finding of the study specifies the sequences containing irregularly spaced multiple short G-tracts. It doesn't prove the results for similar sequences.

(Li et al., 2023) presented a Weighted Two-Stage Alignment tool (TESA). The framework implemented that the algorithms were tested on an experimental dataset. The TESA effectively measured the sequence. In experimental evaluation, the TESA had higher accurate results. However, the alignment of sequence segments demanded a large computational expense.

(Fakharzadeh et al., 2022) represented the molecular Dynamics (MD) simulations to explore all possible DNA. Characterization of stability and structural features were identified in DNA. The experimental results showed that strong left-handedness was presented. The study failed to examine the unusual structure formed by the genes, which shows the inefficiency of research.

3. PROPOSED MOTIF DISCOVERY MODEL

The proposed model explores the ability of the optimization model to find motifs in a biological sequence of DNA. The block diagram of the proposed model is shown in Figure 1,

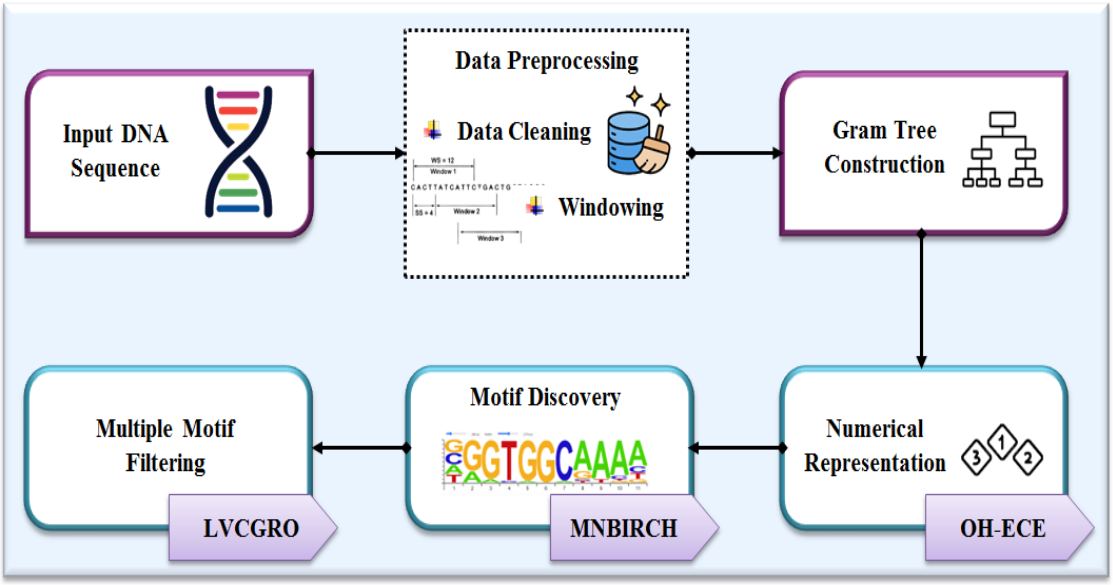


Figure 1: Block diagram of the proposed methodology

3.1 Data Collection &Preprocessing

Input to the proposed model is of (M)DNA sequence $\{D_m\}_{m=1,2,\ldots,M}$ collected from publically available sources and preprocessed for high data quality.

Data Cleaning: Removing some inconsistencies, such a sun necessary symbols, and characters present in the DNA sequence signifies the process of data cleaning.

$$D_{DC(m)} = \Re_{rem}\{D_m, \aleph\} \quad (1)$$

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

Where, $D_{DC(m)}$ denotes the cleaned data, and $\Re_{rem}$ is the function to remove hidden characters or symbols $(\aleph)$ from the input DNA sequence (D_m).

Windowing: The sliding window technique decomposes $D_{DC(m)}$ into multiple windows $(D_{w_i(m)})$ of shorter length (d). If $\{D_{w_i(m)} = D_{DC(1,2,\ldots m+d-1)}\}$ is a window defined in length (d), the window series obtained is given as,

$$D_{w_i(m)} = \langle D_{w_1(m)}, D_{w_2(m)}, \ldots, D_{w_{i+M-1}(m)} \rangle \quad (2)$$

Where, $D_{w_i(m)}$ contains all the short segments of $D_{DC(m)}$ from $w_{(i)}$ to $w_{(i+M-1)}$.

3.2 Gram-tree Construction

To analyze the DNA sequence in more detail, the subset of DNA samples obtained is $(D_{w_i(m)})$ divided into n-gram representations. When using these n-gram representations, considerable redundancy occurs; this may increase memory consumption. Hence, to reduce the memory constraint problem, the suffix tree model is then used to map these n-gram representations into a gram tree.

Given a subset $(D_{w_i(m)})$ of (d) sequence, there are $(d^{(n)})$ possible unique n-grams. Thus, the n-gram of $(D_{w_i(m)})$ is,

$$D_{ng(k)\in w_i} = \{j, d\lfloor j, j+l-1 \rfloor\} \quad (3)$$

Where, the n-grams $(D_{ng(k)\in w_i})$ for each $(D_{w_i(m)})$ are a pair of a subsequence of length (d) starting at the $j-th$ element and the positional integer (l). Then, a suffix tree (F) for $(D_{ng(k)\in w_i})$ with (R) leaves is constructed as,

$$F = \{LN, IN, g, SL\} \quad (4)$$

Where, LN is the leaf node with respect to the suffix, IN are the internal nodes in the tree in which each node except the root consists of two children, g is any leaf under the internal node, and SL are the suffix links used to form a tree.

3.3 Numerical Representation

Here, the alphanumeric DNA sequences $(D_{ng(k)\in w_i} \in F)$ in a gram-tree structure are converted into a numerical form for further processing. For this, the OH-ECE technique is utilized. The conventional one-hot encoding technique is selected for its efficiency in terms of memory and computational cost. But, it increases dimensionality so training becomes slower and more complex. To prevent this problem, entropy weighted Correlation coefficient is introduced that reduces the dimensionality of the data.

The OH-ECE receives the input data $\lfloor D_{ng(k)\in w_i} \rfloor$ and encodes the categorical data into numerical data $(D_{num(k,w_i\in F)})$. It can be expressed as,

$$D_{num(k,w_i\in F)} = \Im_{ohe}\{D_{ng(k)\in w_i}, W, q\} \quad (5)$$

Where, $\Im_{ohe}$ is the encoding function to map characters (k) to the integers (q), and W is the entropy-weighted correlation coefficient.

$$W = \frac{\sum_k \theta_k \Phi(D_{ng(k)}, D_{ng(k+1)})}{\sqrt{\sum_k \theta_k (\Phi(D_{ng(k)})^2 + \Phi(D_{ng(k+1)})^2)}} \quad (6)$$

Where, θ_k denotes the entropy weights, and Φ is the membership degree.

Hence, the encoding based on the weighted correlation coefficients selects the most desirable variables, which reduces the dimensionality and makes the motif discovery process easier.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

3.4 Motif Discovery

The numeric data $\left(D_{num(k,w_i \in F)}\right)$ is used for categorizing the motif sequence using the MNBIRCH algorithm. This algorithm is selected for its capability to cluster multidimensional data with a single scan. However, the clustering results for the non spherical dataset are not good. Hence, a Nested Pair of clusters and a Multi-Dimensional Distance Scaling method are used for effective clustering.

MNBIRCH first takes the input data points $\left(D_{num(k,w_i \in F)}\right)$ and the desired number of clusters (c). In the first phase, a clustering feature tree was built using the nested pair of clusters defined as,

$$\partial_{cf} = \begin{cases} \gamma_{pos}(G_c) = \left(G_{c(low)}\right) \\ \gamma_{bou}(G_c) = \left(G_{c(up)} - G_{c(low)}\right) \\ \gamma_{neg}(G_c) = \left(U - G_{c(up)}\right) \end{cases} \quad (7)$$

Where, ∂_{cf} is the cluster feature, $\gamma_{pos}(G_c), \gamma_{bou}(G_c)$, and $\gamma_{neg}(G_c)$ are the positive, boundary, and negative regions to form clusters, $G_{c(up)}$ and $G_{c(low)}$ are the upper and lower bounds of the clusters, and $U \in \left\{G_{c(up)}, G_{c(low)}\right\}$. Thus, the three-way clusters can discover clusters of different sizes and shapes, where the relationship between objects and a cluster is described by a pair of nested sets.

Then, the clusters are summarized by the computation of cluster centroid (η), cluster radius (cr), and cluster diameter (cd).

$$\eta = \frac{\sum_{k=1toK} D_{num(k,w_i \in F)}}{\chi} \quad (8)$$

$$cr = \sqrt{\frac{\sum_{k} \left(D_{num(k,w_i \in F)} - \eta\right)}{\chi}} \quad (9)$$

$$cd = \sqrt{\frac{\sum_{k} \left(D_{num(k)} - D_{num(k+1)}\right)}{\chi(\chi - 1)}} \quad (10)$$

Then, the clustering feature tree is constructed in which objects are dynamically inserted into the leaf node, and the diameter of the sub-cluster is stored at the leaf node. When the diameter exceeds the threshold, the leaf node is splitted. Here, the Multi-Dimensional Distance Scaling method $\left(msd\left(D_{num(k,w_i \in F)}\right)\right)$ is used to find the nearest leaf node as,

$$msd\left(D_{num(k,w_i \in F)}\right) = \left(\frac{\hbar_x \left| D_{num(k,w_i \in F)}, sm \right|}{C}\right)^{\frac{1}{C}} \times \frac{\max(\hbar_x)}{x} \quad (11)$$

Where, $\hbar_x$ is the x neighbourhood of $D_{num(k,w_i \in F)}$, and sm is the similarity matrix obtained using Euclidean distance.

Once the tree has been constructed, the partitioning method is used to obtain the final clusters.

$$\chi_{DM} = \left\{\chi_1, \chi_2, ..., \chi_N\right\} \quad (12)$$

Where, χ_{DM} is the total number of clusters, and χ_N is the N^{th} cluster.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

3.5 Multiple Motif Filtering

Finally, multiple motifs are filtered from the discovered motifs using the LVCGRO method. The method is chosen as it is based on human thinking power and decision-making processes. However, in the migration phase, the gold seekers assume the location of gold mines, which will not be always efficient in attaining the global optimum and reducing the convergence rate. Therefore, the Linearly Varying Constriction factor is introduced in the proposed work to select the gold mine position instead of assumption.

The initial locations of gold prospects (i.e., the resultant clusters to be filtered) are initialized as,

$$\chi_{DM} = \{\chi_{mn} \mid m = 1,2,..,M, n = 1,2,...,N\} \quad (13)$$

Where, χ_{DM} is the matrix that stores the location of prospectors, χ_{mn} is the location of m^{th} prospect at n^{th} dimension, M is the number of gold prospects, and N is the dimension size.

Then, the objective function based on k-mer percent identity mean $\left(fit(\chi_{mn})\right)$ is used to evaluate the location of prospects, which is given as,

$$fit(\chi_{mn}) = Z_i + |H| \log it(\psi_{(o)}) + \sum_{k-mer=1}^{u} k - mer + \sum_{i=1}^{i_{\max}} \log\left(\frac{\psi_c(k-mer)}{\Gamma(|H| + k - mer)}\right) \quad (14)$$

$$k - mer = \frac{ob(\chi_{mn}) - ex(\chi_{mn})}{ob(\chi_{mn}) + ex(\chi_{mn})} \quad (15)$$

Where, Z_i is the other terms to k-mer, H is the number of the sequence, $\psi_{(o)}$ is the log odds of k-mer, ψ_c is the prior count of k-mer, Γ is the gamma function, and $ob(\chi_{mn})$ and $ex(\chi_{mn})$ are the observed and expected occurrences of a given k-mer.

The migration of prospectors is moving towards the position of gold at the moment of discovering the gold mines. Let (χ_m) and $(\chi_{m(cf)})$ are the location of the gold prospector and the position selected by the constriction factor. The new location of the gold prospect $(\chi_{m(new)})$ is updated as,

$$\chi_{m(new)}^{i+1} = \chi_m^i + V_1.B_1 \quad (16)$$

$$B_1 = V_2.\chi_{m(cf)}^i - \chi_m^i \quad (17)$$

$$\chi_{m(cf)}^i = \frac{\cos\left(\frac{\pi}{i_{\max}}\right) \times i + 2.5}{4} \quad (18)$$

Where, V_1 and V_2 are the vector coefficients, i is the number of iterations, B_1 is the distance between (χ_m) and $(\chi_{m(cf)})$, and $i_{\max}$ is the maximum iterations. The constriction factor enlarges the searching range of the gold seekers to enrich solution diversity for preventing early convergence to local minima.

In mining, the location of each gold prospector is regarded as an approximate location of a gold mine.

$$\chi_{m(new)}^{i+1} = \chi_{m(ran)}^i + V_3.B_2 \quad (19)$$

$$B_2 = \left(\chi_m^i - \chi_{m(ran)}^i\right) \quad (20)$$

Where, $\chi_{m(ran)}^i$ is the randomly selected individual, B_2 is the distance between (χ_m) and $\chi_{m(ran)}$, and V_3 is used to increase the exploration capability of mining.

When the gold prospecting is performed through teamwork, three-person collaboration is performed between the prospectors.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

$$\chi_{m(new)}^{i+1} = \chi_m^i + s_1.\left(\chi_{m(ran1)}^i - \chi_{m(ran2)}^i\right) \quad (21)$$

Where, s_1 is the random vector, and $\chi_{m(ran1)}^i$ and $\chi_{m(ran2)}^i$ are the two randomly selected gold prospectors.

Using these methods, each individual searches for the gold mine, and the used objective function is a critical parameter in decision-making.

$$\chi_{m(opt)}^{i+1} = \chi_{m(new)}^{i+1}, \qquad fit\left(\chi_{m(new)}^{i+1}\right) > fit\left(\chi_m^{i+1}\right) \quad (22)$$

Where, $\chi_{m(opt)}^{i+1}$ denotes the multiple motifs identified using the LVCGRO algorithm.

Algorithm 1: Motif filtering using LVCGRO

Input: Clustered groups χ_{DM}

Output: Filtered motif $\chi_{m(opt)}$

Begin

Initialize population χ_{DM} , parameters V_1, V_2 , number of iteration i

While $\left(i < i_{\max}\right)$

For each $\left(\chi_{mn}\right)$ **do**

Evaluate $fit\left(\chi_m\right)$

Select $\chi_{m(best)}$

Calculate $\chi_{m(new)}$ using migration, mining, and collaboration

End for

Compute fitness of $\chi_{m(new)}$

Update $\chi_{m(best)}$

If $\left(fit\left(\chi_{m(new)}^{i+1}\right) > fit\left(\chi_m^{i+1}\right)\right)$

Move to $\chi_{m(new)}$

Else

Keep χ_m

End if

$i = i + 1$

End While

Return $\chi_{m(opt)}$

End

In this phase, each individual is defined as the list of clustered DNA sequences along with its motif positions. The best solution obtained during searching is chosen as the global optimum i.e., the multiple motif positions. The best solution found so far is considered the final solution of the algorithm.

4. RESULTS AND DISCUSSION

This section presents the evaluation of the proposed model implemented in the working platform of PYTHON.

4.1 Dataset Description - Combining chromatin immune precipitation (ChIP) assays with sequencing data is used for the experiments. The dataset contains sequence data obtained from the conversion of base calls data. From this dataset, 80% of data is used for training and 20% for testing.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

4.2 Performance Analysis- Here, the comparison results of the proposed LVCGRO and MNBIRCH methods are given.

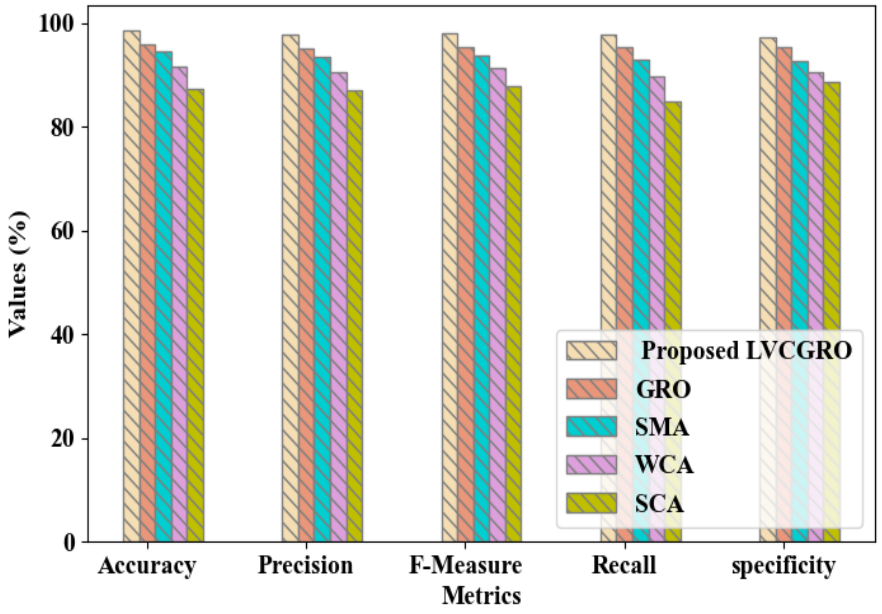


Figure 2: Performance Comparison of proposed LVCGRO

Figure 2 shows the performance comparison between the proposed LVCGRO and existing GRO, Slime Mold Algorithm (SMO), Water Cycle Algorithm (WCA), and Sine Cosine Algorithm (SCA) with respect to accuracy, precision, recall, F-score, and specificity. It was observed that the accuracy (98.561%) attained by the proposed method is higher than the existing methods. Hence, the modified objective function and the constriction factor used in the model showed its ability to find a large number of motifs from the DNA sequence.

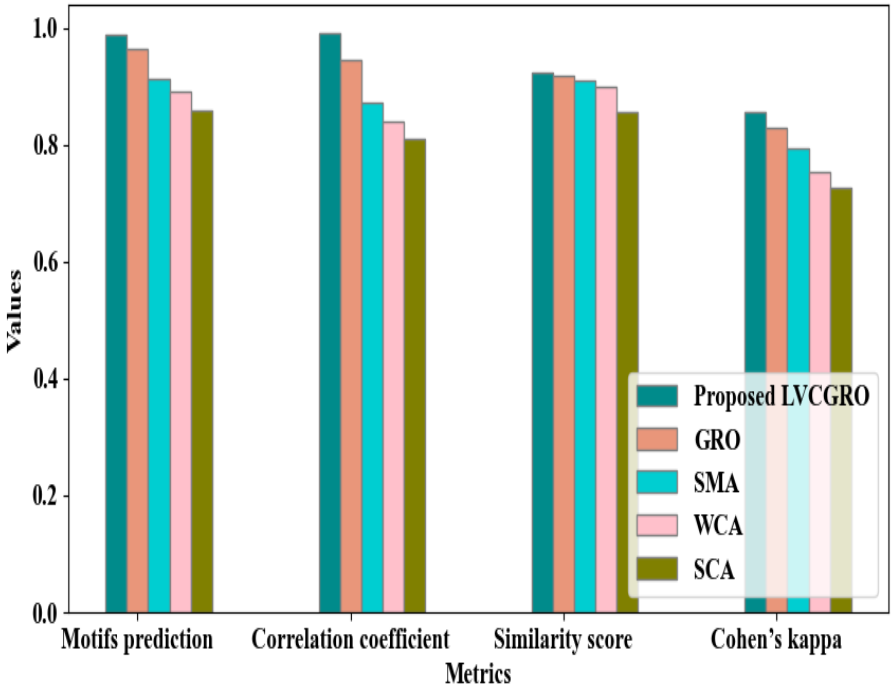


Figure 3: Performance analysis of motif filtering

Figure 3 depicts the performance of the proposed and existing motif discovery methods. It was found that the proposed LVCGRO method showed a higher prediction percentage (0.98784) and correlation coefficient (0.9899). In terms of similarity score and Cohen's kappa, the proposed method has a relatively better performance than the existing methods. This is because the objective function considers a sequence of multiple characters for filtering performance and the constriction factor makes the method escape from the local optima solutions.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

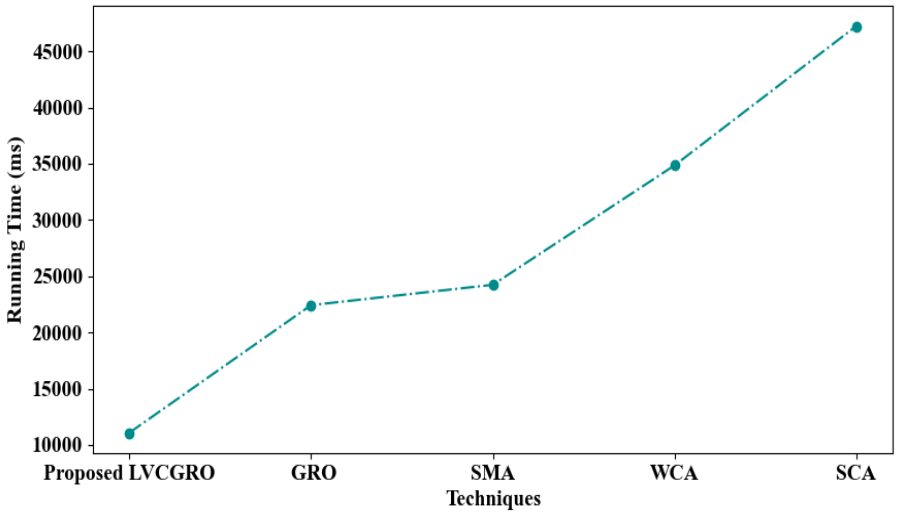
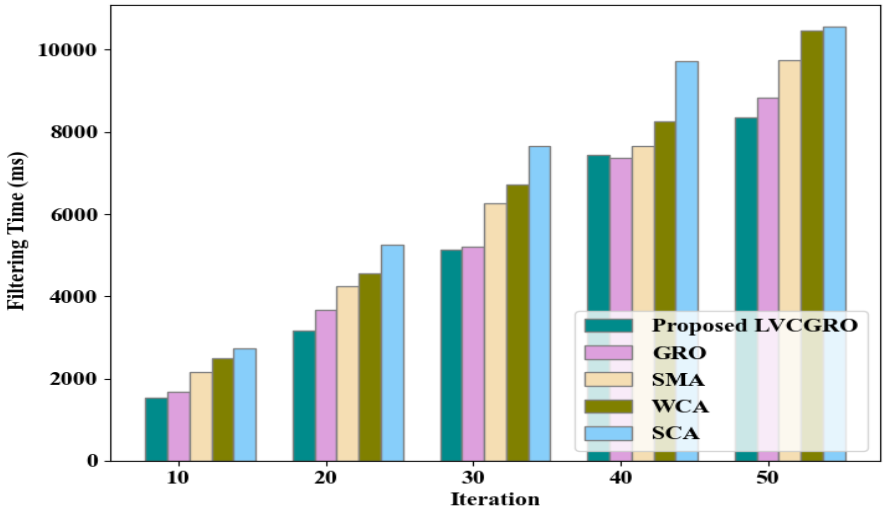
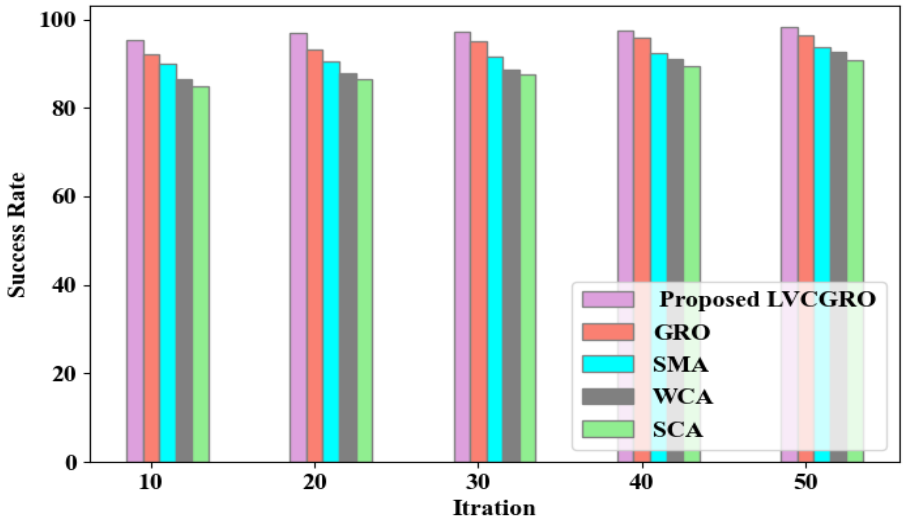


Figure 4: Running Time Analysis

Figure 4 shows the running time of the proposed LVCGRO and existing motif filtering techniques. It can be seen that the running time of the proposed method is minimum by 11350 ms than the existing methods. Therefore, the motif analysis on the shorter sequences helped the model to run faster than the existing methods.



(a)



(b)

Figure 5: Analysis of (a) filtering time and (b) success rate

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 10, October 2023, pp : 105-115

e-ISSN : 2583-1062

Impact Factor : 5.725

The filtering time and success rate of the proposed LVCGRO and existing methods are plotted in Figure 5 under various iterations. At the 50th iteration, the filtering time of the proposed method is 8354, which is less than the existing GRO, and the success rate is 98.197,which is higher than the existing methods.

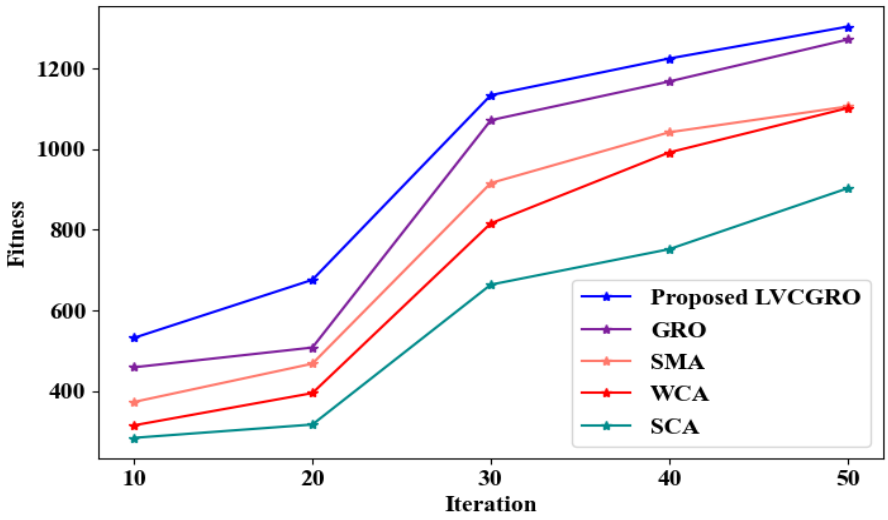


Figure 6: Fitness Vs Iterations

Figure 6 shows the graph of fitness comparison for different iterations between proposed and existing motif filtering methods. At the 50th iteration, the fitness attained by the proposed model (1304) is much higher, than the existing methods. Hence, the proposed model converges faster than the existing methods.

Table 1: Performance analysis of motif discovery

Methods/Metrics	Clustering Time (ms)	Silhouette Score
Proposed MNBIRCH	1526	0.94
BIRCH	1685	0.92
DBSCAN	2154	0.91
FCM	2488	0.88
KMA	2735	0.86

Table 1 analyzes the performance of proposed MNBIRCH and existing BIRCH, DBSCAN, Fuzzy C-Means (FCM), and K-Means Algorithm (KMA). As per the clustering time,the proposed MNBIRCH is 159 ms faster than the existing methods. It can also be found that the proposed model gives a higher silhouette score than the existing methods. The three-way clustering processing and distance scaling method-based clustering are reasonable so that the performance of the proposed MNBIRCH is better than the existing methods.

4.3 Comparative Analysis

The comparative analysis of the proposed model with the closely related approaches is presented in this sub-section.

Table 2: Comparative Analysis

Author Name	Technique Used	Dataset	Evaluation Criteria			
			Precision	Recall	Accuracy	Running Time
Proposed	LVCGRO	ChIP-seqdata	97.859	97.852	98.561	11075 ms
(Saha et al., 2021)	Chemical Reaction Optimization	E. coli transcription factor binding sites dataset	-	-	-	2334 ms
(Hashim et al., 2020)	modified Henry gas solubility optimization	Synthetic and real datasets.	96.215	90.037	-	-

(Ashraf&Shafi, 2020)	evolutionary approach	real dataset motifs	-	-	98.04	-
(Masood et al., 2021)	Emerging Substring based Motif Detection	Huge DNA datasets	-	-	83.04	150 s
(Poccia et al., 2021)	salient multi-variate motif algorithm	MoCap, BirdSong dataset	87	62	-	6.67s

Table 2 compares the proposed model with existing motif discovery approaches. When compared with the existing methods, the proposed model achieved better performances in terms of accuracy and running time. The existing optimization approaches were the most successful. However, the methods failed to identify multiple motifs present in the DNA sequence and suffered from the local optima solutions, which slowed down the processing performance. But, the proposed model is found to be efficient with the sliding windowing, objective function to discover multiple motifs, and gram tree representation-based analysis.

5. CONCLUSION

This paper proposed a new technique based on LVCGRO to find multiple motifs from the DNA sequences. Data analysis on gram tree representations, motif discovery using MNBIRCH, and multiple motifs filtering based on LVCGRO are the main phases of the proposed model. For the performance evaluation, the results of the proposed LVCGRO and MNBIRCH methods are compared with the existing methods. From the results, it can be concluded that the analysis of shorter length sequences and gram representations aid the proposed model to augment its accuracy to 2.51% than the existing methods. However, the negligence of motif patterns and features is the only limitation of the proposed model.

Recommendations for future work: In the future, the work can be extended to analyze disease susceptibility through the motif discovery process.

6. REFERENCES

[1] Alam, S. M. S., Kowser, I., Islam, M. A. J., Zaman, S. S., Kabir, T. T., & Ashraf, F. B. (2021).An Efficient Metaheuristic Approach for Finding Motifs from DNA Sequences.In 2021 5th International Conference on Electrical Information and Communication Technology (EICT), 1-5.

[2] Ashraf, F. B., &Shafi, M. S. R. (2020).Mfea: An evolutionary approach for motif finding in dna sequences. Informatics in Medicine Unlocked, 21, pp.1-9.

[3] Choi, W. S., & Garcia-Diaz, M. (2022).A minimal motif for sequence recognition by mitochondrial transcription factor A (TFAM). Nucleic Acids Research, 50(1), 322-332.

[4] Donohue, L. K., Guo, M. G., Zhao, Y., Jung, N., Bussat, R. T., Kim, D. S., ...&Khavari, P. A. (2022). A cis-regulatory lexicon of DNA motif combinations mediating cell-type-specific gene regulation. Cell genomics, 2(11), 1-12.

[5] Fakharzadeh, A., Zhang, J., Roland, C., &Sagui, C. (2022). Novel eGZ-motif formed by regularly extruded guanine bases ina left-handed Z-DNA helix as a major motif behind CGG trinucleotide repeats. Nucleic Acids Research, 50(9), 4860-4876.

[6] Ge, W., Meier, M., Roth, C., &Söding, J. (2021). Bayesian Markov models improve the prediction of binding motifs beyond first order. NAR Genomics and Bioinformatics, 3(2), 1-12.

[7] Hammelman, J., Patel, T., Closser, M., Wichterle, H., & Gifford, D. (2021). Ranking Reprogramming Factors for Directed Differentiation. 19(7), 812–822.

[8] Hashim, F. A., Houssein, E. H., Hussain, K. Mabrouk, M. S., & Al-Atabany, W. (2020).A modified Henry gas solubility optimization for solving motif discovery problem. Neural Computing and Applications, 32, 10759-10771.

[9] He, Y., Shen, Z., Zhang, Q., Wang, S., & Huang, D. S. (2021).A survey on deep learning in DNA/RNA motif mining. Briefings in Bioinformatics, 22(4), 1-10.

[10] Li, X. (2023). DNA motif discovery using evolutionary computation technique, theis, La Trobe University Melbourne, pp. 1-190.

[11] Li, Y., Wang, Y., Wang, C., Fennel, A., Ma, A., Jiang, J., ...& Liu, B. (2023). A Weighted Two-stage Sequence Alignment Framework to Identify DNA Motifs from ChIP-exo Data. bioRxiv, 1-12.

[12] Maity, A., Winnerdy, F. R., Chang, W. D., Chen, G., &Phan, A. T. (2020).Intra-locked G-quadruplex structures formed by irregular DNA G-rich motifs. Nucleic Acids Research, 48(6), 3315-3327.

[13] Masood, M. M. D., Arunarani, A. R., Manjula, D., &Sugumaran, V. (2021).An efficient algorithm for identifying motif from huge DNA datasets. Journal of Ambient Intelligence and Humanized Computing, 12, 485-495.

[14] Poccia, S. R., Candan, K. S., &Sapino, M.L. (2021). SMM: Leveraging metadata for contextually salient multi-variatemotif discovery. Applied Sciences, 11(22), pp.2-24.

[15] Rahman, C. R., Amin, R., Shatabda, S., &Toaha, M. S. I. (2021). A convolution based computational approach towards DNA N6-methyladenine site identification and motif extraction in rice genome. Scientific Reports, 11(1), 1-13.

[16] Saha, S. K., Islam, M. R., &Hasan, M. (2021). DNA motif discovery using chemical reaction optimization. Evolutionary Intelligence, 14, 1707-1726.

[17] Tapan, M. S. Z. (2023). DNA motif discovery using fuzzy self-organizing maps, thesis, La Trobe University Melbourne, pp. 1-233. https://figshare.com/articles/thesis/DNA_motif_discovery_using_fuzzy_self-organizing_maps/21857637/1

[18] Wang, W., Wang, J., Si, S., Huang, Z., & Xiao, J. (2022). RL-MD: A Novel Reinforcement Learning Approach for DNA Motif Discovery. In 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), 1-7. https://doi.org/10.48550/arXiv.2209.15181

[19] Yousif, A. B., Abbas, T., & Al-Khafaji, H. K. (2023). Motif Discovery in DNA Sequences Using Scaled Conjugate Gradient Neural Networks. Journal of Education for Pure Science-University of Thi-Qar, 13(1), 56-76.

[20] Zhang, Y., Liu, Y., Wang, Z., Wang, M., Xiong, S., Huang, G., & Gong, M. (2022). Uncovering the relationship between tissue-specific TF-DNA binding and chromatin features through a transformer-based model. Genes, 13(11), 1-18.