An Efficient System for Finding Functional Motifs in Genomic DNA Sequences by Using Nature-Inspired Algorithms

Ebtehal S. Elewa[1(✉)], Mohamed B. Abdelhalim[1], and Mai S. Mabrouk[2]

[1] College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport Cairo, Cairo, Egypt
ebtehale@hotmail.com
[2] Biomedical Engineering Department, Misr University for Science and Technology, 6 October, Giza, Egypt

Abstract. Motifs are short patterns in Deoxyribonucleic Acid (DNA) that indicate the presence of certain biological characteristics. Motifs finding is the process of successfully finding meaningful motifs in large DNA sequences. Nature-inspired algorithms have been recently gaining much popularity in solving complex and large real-world optimization problems similar to the motif finding problem. This work aims on investigating the application of nature-inspired algorithms in motif finding problem. The investigation methodology is divided into three main approaches; the first is to apply well-known nature-inspired algorithms in solving the problem, then the enhancement of an algorithm is investigated, and finally the hybridization between two algorithms is investigated. Experiments are performed on synthetic as well as real data sets. The results show that the combination provides the best results, however, individual and modified algorithms provide also good results compared to some state-of-the-art tools.

Keywords: Cuckoo search · Gravitational search algorithm · Particle swarm optimization · Motif finding · DNA · Nature-inspired algorithms

1 Introduction

The basic unit of living cells is the Deoxyribonucleic Acid (DNA). DNA sequences consist of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) [1], and contains the genetic information required for the proper functioning of all living organisms [2] Motifs are short, recurring patterns in DNA that are presumed to have a biological function. Motif finding is the process of finding these short and meaningful patterns. The purpose of motif finding is to discover repeated patterns in DNA sequences in order to understand the structure of certain cells, find the triggers of some diseases and create new treatments. So far, nature-inspired algorithms have been successfully able to solve many complex problems in almost all areas [3–6]. The success of these algorithms is due to their ability to mimic nature in solving dynamic and complex problems in a reasonable amount of time and optimal cost.

A.E. Hassanien et al. (eds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016*, Advances in Intelligent Systems and Computing 533, DOI 10.1007/978-3-319-48308-5_21

This work introduces enhanced use of nature-inspired algorithms to solve the motif finding problem. We followed the methodology of applying individual algorithms, then we modified one of them while hybridizing the other two to explore the ability to enhance individual algorithms. The first individual algorithm used is the cuckoo search algorithm, and then the modified adaptive cuckoo search that uses new strategies to improve the original algorithm is implemented. The gravitational search algorithm and the famous Particle Swarm Algorithm (PSO) are the other algorithms used and finally the hybrid of them is implemented.

The rest of the article is organized as follows: Sect. 2 provides a brief explanation of the motif finding problem and related work. Section 3 describes the implemented algorithms. Section 4 contains the problem formulation and the overall process. Section 5 shows the experimental results and discussion. Section 6 concludes the article and introduces possible directions for future work.

2 Problem Definition and Related Work

2.1 The Motif Finding Problem

Given a set of DNA Sequences defined by the alphabets {A, C, G, T}, motifs are defined as short repeated segments in the DNA. The motif finding problem has been studied since the early years of bioinformatics, a number of methods, algorithms and tools have been developed in the recent years to solve this problem However, recent show that their prediction accuracy is still low. Also most of the algorithms suffer with local optima [7].

2.2 Current Motif Finding Algorithms and Related Work

Most motif finding algorithms have their own strengths and weaknesses and fall into two major groups based on the combinatorial approach used: The first group is Word-based methods. The advantage of the word-based method is that it exhaustively searches the whole search space and therefore guarantees finding a global optimum. However, this also means that they are only suitable for short motifs. The main drawback of word-based algorithms is that they are usually computationally expensive. Some popular tools that use word-based methods are MITRA [8], and Weeder [9]. The second group is probabilistic sequence methods. The most popular methods MEME [10], AlignACE [11]. However, these algorithms are not guaranteed to find globally optimal solutions. In literature, evolutionary algorithms have been also used to find DNA motifs, however most of them depend on Genetic Algorithm (GA) [12, 13] and Particle swarm optimization (PSO) [14, 15]. So far, surveys have been made to evaluate motif finding tool [16]. The results show that combining more than one approach in finding motifs provides better results than using a single approach, and that no specific tool provides the best performance for all datasets.

3 Methods and Materials

The proposed system is comprised of the following three fundamental building phases: (1) The first phase 'Read Input' includes reading the file that contains the DNA Sequence, (2) The second phase' Execute optimization algorithm' is to execute the optimization algorithm, and (3) The last step 'Calculate evaluation criteria' is to compare the predicted motif and compare it with the actual motif and calculating the evaluation criteria. These three phases of the introduced approach is described in Fig. 1.

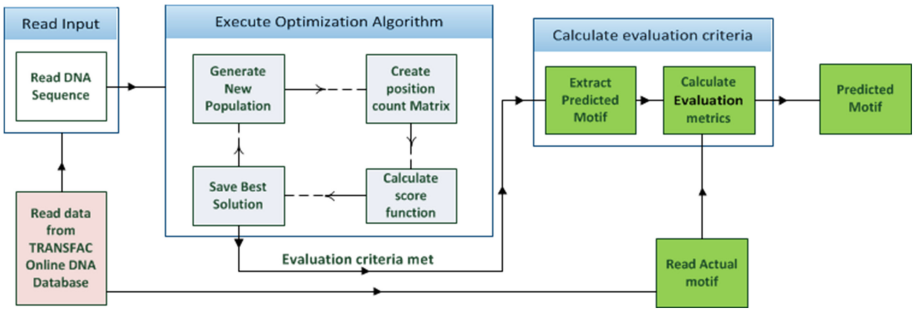


Fig. 1. Overall system block diagram

3.1 Cuckoo Search Algorithm (CS)

Cuckoo search was introduced by Yang and Deb in 2009 [17], it is inspired by the cuckoo birds and their parasitism reproduction behavior in nature.

Cuckoo search has both global search and exhaustive local search capabilities, which explains why it is so efficient in finding almost optimal solutions. These capabilities are due to the idea of Lévy flights used by the algorithm. The Lévy flight process, is a random walk that is characterized by a series of instantaneous jumps chosen from a probability density function in which the step lengths are distributed according to a heavy-tailed probability distribution [25].

3.2 Modified Adaptive Cuckoo Search (MACS)

Despite the success of the original cuckoo search algorithm, many variants and hybridizations have been suggested by many researches to improve its efficiency. One important variation is the modified adaptive cuckoo search suggested by Zhang [18]. It introduces new strategies such as grouping, parallel, information sharing and adaptive strategies.

3.3 Hybridizing GSA and PSO

GSA has been used to solve many optimization problems, however the main drawback of GSA is that it can easily fall into local optima and it also suffers from slow

exploitation. These drawbacks have been addressed by researches by using hybrid algorithms to overcome the slow exploitation of GSA [19, 20]. GSA and PSO have been used alone to solve the motif finding problem in [14, 15, 21, 25]. However, combining the features of the two algorithms has never been used to find motifs. PSO is widely used in hybrid algorithms due to its ability to find global optimal, it's speed of convergence and simplicity.

In this work, PSO and GSA are combined to produce a hybrid GSA-PSO algorithm to find motifs. The hybrid algorithm combines between the social behavior of PSO in updating the particles locations and velocities according to gbest and the local search capabilities of GSA The new hybrid GSA-PSO algorithm is illustrated in the following steps:

1. Generate the initial population.
2. Evaluate the fitness function of each object.
3. Calculate the gravitational constant and update the local best.
4. Calculate inertial mass (M) for each object.
5. Calculate forces and acceleration.
6. Update the velocity and position using the following classical PSO Eqs. (1), (2) and (3):

$$V_i(t+1) = Rand \times V_i(t) + C_1 \times ac_i(t) + C_2 \times (gbest - X_i(t)) \quad (1)$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (2)$$

$$ac_i(t) = f_i(t)/M_i(t) \quad (3)$$

   Where $f_i(t)$ is the total forces that act on the agent object i and $M_i(t)$ is the mass of the object *i*.
7. Repeat steps from 2 to 6 until stop criteria is reached.

4 Problem Formulation

4.1 Motifs Representation

An individual is represented by a vector of integers (X); each integer in the vector represents the starting positions of the potential motif in each DNA sequence. Accordingly, the length of the vector equals to the number of input sequences, so each vector (X) contains (d) items and (d) is the number of DNA sequences. Hence a candidate vector can be represented in the following form given in Eq. (4):

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots\ldots\ldots, x_{id}) \quad (4)$$

4.2 Objective Function

In this work, the score function described by Reddy and Arock [14] is used. Equation (5) describes the objective function where p(s) is the profile matrix corresponding to starting positions and $M_{p(s)}(J)$ is the largest count in column j of P(s).

$$Score = \sum_{j=1}^{l} M_{p(s)}(J) \tag{5}$$

5 Experimental Results

The proposed algorithms have been tested on synthetic as well as real data and the average results are calculated. MATLAB Bioinformatics Toolbox is used to implement the algorithms. Synthetic data used in testing the algorithms was implemented by the model provided by Pevsner and Size [22]. The goal of any motif finding algorithm is to find and discover these motifs without previously knowing their locations in the DNA sequence regardless of the type of DNA sequence. TRANSFAC database (available at http://www.gene-regulation.com/pub/databases.html#transfac) is used in this work to evaluate the algorithms on real datasets. The reason for that choice is that the real sequence data from the biological database TRANSFAC are a part of the freely accessible benchmark data ensemble constructed by Tompa et al. [23], also this real data belongs to different kinds of species mainly human, mouse, rat, yeast, and plants to ensure that the algorithms work on different types of species.

In order to analyze the performance of these algorithms; recall, precision and F-score metrics are used. Recall and precision are defined in Eqs. (6) and (7):

$$Recall = n_c/n_t \tag{6}$$

$$Precision = n_c/n_p \tag{7}$$

Where n_c is the number of motifs that are correctly predicted, n_p is the total number of predicted motifs and n_t is the total number of actual true motifs? F-score value that combines both recall and precession terms is defined in Eq. (8).

$$F - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{8}$$

6 Results and Discussions

For synthetic data, different motifs are used by using different (l, d) pairs to test the implemented algorithms where l is the length of the motif and d in the maximum number of mutations. The famous challenging instances of (13,4), (15,5), (17,6) [22] are used in this work. The algorithms have also been tested against the real data

described in the previous section and compared with well-known algorithms that have been used in finding motifs, such as AlignACE [11] that uses Gibbs sampling to find sequence elements conserved in a set of DNA sequences, Multiple EM for Motif Elicitation or MEME [10] which is a tool for discovering motifs in a group of related DNA or protein sequences. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern and finally compared with GALF [24] which is a tool based on Genetic Algorithm with Local Filtering. Figure 2 shows the average precision of motifs instances of (13,4), (15,5), (17,6), Fig. 3 shows the average recall and finally Fig. 4 provides the average F-score on the synthetic data with different motif lengths. The results on synthetic data show that GSA-PSO provides the best recall and precision values and accordingly provide the best F-score values. In addition, GSA-PSO is able to provide the best results for short and long motifs as well. The results also show that MACS is the second best in precision, recall and F-score values.

	CS	MACS	PSO	GSA	GSA-PSO	AlignACE	MEME	GALF
(13,4)	0.91	0.93	0.87	0.89	0.95	0.91	0.85	0.9
(15,5)	0.93	0.94	0.89	0.9	0.96	0.93	0.88	0.92
(17,6)	0.94	0.97	0.89	0.92	0.99	0.94	0.89	0.95

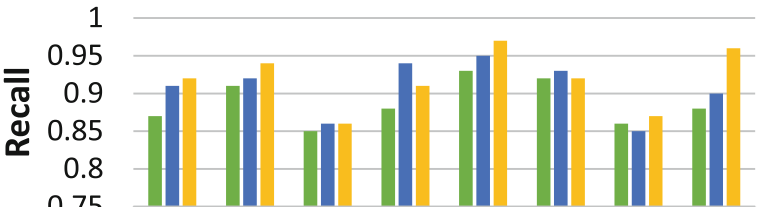
Fig. 2. Precision of instances (13,4), (15,5), (17,6) (Color figure online)

The F-score values show that GSA-PSO correctly identifies motifs and provides better results than GALF which is based on Genetic Algorithm. The enhancement in precession recall and F-score values is up to 0.05.

Table 1 shows the average precision, recall and F-score respectively on real DNA datasets. The results are the average of 20 runs on all the species.

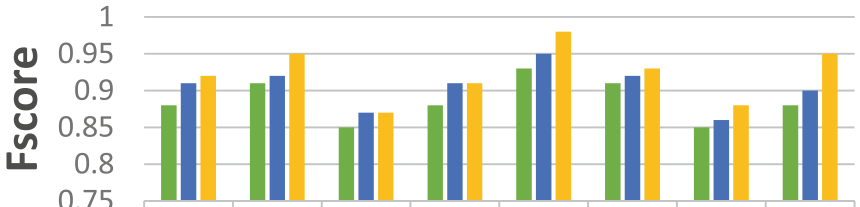
7 Discussion

The results show that implemented algorithms perform better than well-known algorithms on the benchmark data set. Except the basic PSO algorithm that performs less than GALF but almost the same as MEME & AlginACE, that have the highest precision.



	CS	MACS	PSO	GSA	GSA-PSO	AlignACE	MEME	GALF
(13,4)	0.87	0.91	0.85	0.88	0.93	0.92	0.86	0.88
(15,5)	0.91	0.92	0.86	0.94	0.95	0.93	0.85	0.9
(17,6)	0.92	0.94	0.86	0.91	0.97	0.92	0.87	0.96

Fig. 3. Recall of instances (13,4), (15,5), (17,6) (Color figure online)



	CS	MACS	PSO	GSA	GSA-PSO	AlignACE	MEME	GALF
(13,4)	0.88	0.91	0.85	0.88	0.93	0.91	0.85	0.88
(15,5)	0.91	0.92	0.87	0.91	0.95	0.92	0.86	0.9
(17,6)	0.92	0.95	0.87	0.91	0.98	0.93	0.88	0.95

Fig. 4. F-score of instances (13,4), (15,5), (17,6) (Color figure online)

Table 1. Average precision, recall and f-score values on real datasets

Algorithm	Precision	Recall	f-score
CS	0.59	0.64	0.61
MACS	0.73	0.81	0.76
PSO	0.53	0.57	0.55
GSA	0.81	0.76	0.78
GSA-PSO	**0.85**	**0.81**	**0.83**
AlignACE (10)	0.64	0.57	0.59
MEME (11)	0.67	0.53	0.58
GALF (24)	0.80	0.74	0.76

MACS, GSA and PSO-GSA algorithms have higher recall values than MEME & AlginACE. This means that these algorithms have less false positives. The results show that standard particle swarm algorithm, cuckoo search and gravitational search algorithms are able to successfully identify motifs. Modified adaptive cuckoo search is also able to give values better than standard cuckoo search, due to the adaptive step that modifies the lévy step size according to the search stage.

Regarding Synthetic data, the F-score values show that GSA-PSO correctly identifies motifs and provides better results than GALF which is based on Genetic Algorithm. The enhancement in precession recall and F-score values is up to 0.05.

PSO-GSA, MACS and GSA have the first, second and third highest F-score value respectively. This means that they have picked up the significant portion of the real motifs.

Hybrid GSA-PSO has higher precision, recall and F-score values than PSO and GSA, this is due to its ability to balance between exploration and exploitation the average enhancement in F-score of GSA-PSO is up to 0.24.

The results of the suggested algorithms vary according to the parameters used in each algorithm. The parameters were tuned to achieve the best result in this specific problem. The average results show that precision and recall of MACS are higher than GALF. Also GSA-PSO provides up to 3 % enhancement compared to results obtained from GALF

8 Conclusions

In this work, nature-inspired algorithms are used to the challenging motif finding problem. The results show that standard particle swarm algorithm, cuckoo search and gravitational search algorithms are able to successfully identify motifs. Modified adaptive cuckoo search is also able to give values better than standard cuckoo search, due to the adaptive step that modifies the lévy step size according to the search stage.

Hybrid GSA-PSO has higher precision, recall and F-score values than PSO and GSA, this is due to its ability to balance between exploration and exploitation. The average enhancement in F-score of GSA-PSO is up to 0.24. The average results show that precision and recall of MACS are higher than GALF. Also GSA-PSO provides up to 3 % enhancement compared to results obtained from GALF. The results obtained show that these nature-inspired algorithms are generally able to correctly identify meaningful motifs instances in synthetic as well as real datasets and the quality of the results is better than the other tools considered for comparison. In addition, hybrid PSO-GSA provides the best results. Moreover, the enhanced CS and GSA algorithms provide good results that are better than the other considered tools.

References

1. D'haeseleer, P.: What are DNA sequence motifs? Nat. Biotechnol. **24**(4), 423–425 (2006)
2. Marbrouk, M., Hamdy, M., Mamdouh, M., Aboelfotoh, M., Kadah, Y.M.: BIOINFTool: bioinformatics and sequence data analysis in molecular biology using Mat Lab. In: Proceedings of Cairo International Biomedical Engineering Conference, 01–09 October 2006
3. Zelinka, I.: A survey on evolutionary algorithms dynamics and its complexity–Mutual relations, past, present and future. Swarm Evol. Comput. **25**, 2–14 (2015)
4. Smolinski, T.G., Milanova, M.M., Hassanien, A.E.: Applications of Computational Intelligence in Biology: Current Trends and Open Problems. Studies in Computational Intelligence Springer, Heidelberg (2008)
5. Smolinski, T.G., Milanova, M.M., Hassanien, A.E.: Applications of Computational Intelligence in Bioinformatics and Biomedicine: Current Trends and Open Problems. Springer, Heidelberg (2008)
6. Abdelhalim, M.B., Habib, S.E.D.: Particle swarm optimization for HW/SW partitioning. In: Lazinica, A. (ed.) Particle Swarm Optimization, pp. 49–76. Tech Education and Publishing, New York (2009)
7. Wei, W., Xiao-Dan, Yu.: Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. Genom. Proteomics Bioinf. **5**(2), 131–142 (2007)
8. Eskin, E., Pevzner, P.A.: Finding composite regulatory patterns in DNA sequences. Bioinformatics **18**(suppl 1), S354–S363 (2002)
9. Pavesi, G., et al.: Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulared genes. Nucl. Acids Res. **32**, 199–203 (2004)
10. Bailley, T., Elkan, C.: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach. Learn. **21**(1–2), 51–80 (1995)
11. Roth, P., et al.: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. **16**(10), 939–945 (1998)
12. Vijayvargiya, S., Shukla, P.: Identification of transcription factor binding sites using genetic algorithm. Int. J. Res. Rev. Comput. Sci. **2**(2), 100–107 (2011)
13. Basha Gutierrez, J., Frith, M., Nakai, K.: A genetic algorithm for motif finding based on statistical significance. In: Ortuño, F., Rojas, I. (eds.) IWBBIO 2015, Part I. LNCS, vol. 9043, pp. 438–449. Springer, Heidelberg (2015)
14. Reddy, U., et al.: A particle swarm solution for planted(l, d)-Motif problem. In: IEEE Symposium in Bioinformatics and Computational Biology (CIBCB), pp. 222–229 (2013)
15. Lei, C., Ruan, J.: A particle swarm optimization-based algorithm for finding gapped motifs. BioData Min. **3**(1), 3–9 (2010)
16. Das, M.K., Dai, H.K.: A survey of DNA motif finding algorithms. BMC Bioinf. **8**(7), 1 (2007)
17. Yang, X., Deb, S.: Cuckoo search via Lévy flights. In: World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), pp. 210–214. IEEE Publications (2009)
18. Zhang, Y., Wang, L., Wu, Q.: Modifed adaptive cuckoo search (MACS) algorithm and formal description for global optimisation. Int. J. Comput. Appl. Technol. **44**(2), 73–79 (2012)
19. Sinaie, S.: Solving shortest path problem using gravitational search algorithm and neural networks (Doctoral dissertation, Universiti Teknologi Malaysia) (2010)
20. Zhang, Yu., Wu, L., Zhang, Y., Wang, J.: Immune gravitation inspired optimization algorithm. In: Huang, D.-S., Gan, Y., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2011. LNCS, vol. 6838, pp. 178–185. Springer, Heidelberg (2011)

21. González-Álvarez, D.L., Vega-Rodríguez, M.A., Gómez-Pulido, J.A., Sánchez-Pérez, J.M.: Applying a multiobjective gravitational search algorithm (MO-GSA) to discover motifs. In: International Work-Conference on Artificial Neural Networks, pp. 372–379 (2011)
22. Pevzner, P.A., Sze, S.H.: Combinatorial approaches to finding subtle signals in DNA sequences. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, California USA, pp. 269–278 (2000)
23. Tompa, M., et al.: Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. **23**(1), 137–144 (2005)
24. Chan, T.M. et al.: TFBS identification by position and consensus-led genetic algorithm with local filtering. In: GECCO 2007: Proceedings of the 2007 Conference on Genetic and Evolutionary Computation, pp. 377–384. ACM, London, England (2007)
25. Hassanien, A.E., Alamry, E.: Swarm Intelligence: Principles, Advances, and Applications. CRC – Taylor & Francis Group (2015). ISBN 9781498741064 - CAT# K26721