Identifying the definite base of COI for extraction of DNA sequences using LPBS

Sharifah Lailee Syed Abdullah [a] , Hazaruddin Harun [ac] , Naimah Mohd Hussin [a] , Jiwa Noris Hamid [a] and Roziana Mohamed Hanaphi [b]

[a] Faculty of Computer Science and Mathematical, UiTM Arau, Perlis, Malaysia
[b] Faculty of Applied Science, UiTM Arau, Perlis, Malaysia
[c] School of Computing, UUM CAS, UUM Sintok, Kedah, Malaysia

***Abstract*—This paper presents the use of the 'Linear-PSO with Binary Search' (LPBS) algorithm for discovering motifs, especially species-specific motifs. In this study, two samples from different fragments of 'mitochondrial cytochrome C oxidase subunit I' (COI/COX1) were collected from the Genbank online database. DNA sequences for the first sample are a mix of different fragments of COI and the second sample is from the same fragment of COI. The genome of COI was used as a reference set and other DNA sequences were used as a comparison set. All the collected DNA sequences are from the same species. The results show that the LPBS algorithm is able to discover motifs of greater length when using DNA sequences from the same fragment of COI. The experiment also found that 139 can be used as a starting base for COI DNA sequences extraction to discover species-specific motifs.**

Keywords—Particle Swarm Optimization; Linear-PSO; motif discovery; species specific

I. INTRODUCTION

The 'Linear-PSO with Binary Search' (LPBS) algorithm was first introduced for motif discovery by Syed Abdullah and Harun [1]. This algorithm is an extension of the Linear-PSO algorithm [2] based on the modified PSO used by Chang et al. [3] and Zhou et al. [4] for motif discovery. LPBS does not generate any random number; instead the target was determined sequentially, thereby ensuring that all possible motifs were tested. Results from previous research showed that the LPBS algorithm was able to discover motifs with higher validity and efficiency [1]. The LPBS algorithm was the first PSO based algorithm that was developed to discover motifs which can be used for identification of specific species.

In order to discover the right motif, a suitable fragment of DNA sequence needs to be identified. This fragment must contain the possible target motif to be investigated. Many researches focus on 'mitochondrial DNA' (mtDNA) as a suitable fragment for a species identification motif [5, 6]. Mitochondrial DNA is another structure of DNA that exists in each cell. Each cell contains 100 to 1000 copies of mtDNA and therefore the possibility of extracting a DNA sequence is great [7]. A few genes and fragments have been identified including the cytochrome b gene, the 16S rRNA gene, the 12S rRNA gene, the mtDNA control region, and the COI gene.

Hebert et al. [5] proposed the used of 'mitochondrial cytochrome C oxidase subunit I' (COI) for species identification. COI is one of the genes existing in mtDNA. In their research, they successfully developed a COI species profile for lepidopteran species.

According to Hebert et al. [5], COI was chosen as a target gene because COI appears to have a greater range of phylogenetic signals than any other mitochondrial gene [6]. Phylogenetic signals are used in research on relations among species.

The remaining sections are organized as follows: Section II introduces the basic concept of LPBS, Section III discusses the method used and the results of this study, and finally the conclusions are discussed in Section IV.

II. LINEAR-PSO WITH BINARY SEARCH

The Particle Swarm Optimization (PSO) algorithm is an evolutionary optimization algorithm introduced by Kennedy and Eberhart [8]. Development of this algorithm was motivated by animal social behaviors such as those found in schools of fish or flocks of birds and the way these animals find food sources and avoid predators. The original PSO algorithm starts with the random initialization of a population of individual particles in the search space. Each particle goes through the fitness calculation. A new position is calculated for each particle based on the current position and velocity values. The velocity value for each particle uses random number generation.

In motif discovery, PSO was first modified and used by Chang [3]. Later the algorithm was extended by integrating hybrid algorithms [9, 4], adding a dissimilarity graph [10], and applying the stochastic local search concept [11].

However, previous studies only focus on general motif discovery in individual DNA sequences, which permits the use of randomization of the selected population. In this study, sequential selecting and linear searching are not a choice but are compulsory, because comprehensive selecting and searching must be used to identify the right motif to represent a species. All possible motifs will be tested in order to get a higher fitness value. Therefore, Syed Abdullah et al. [2] proposed a Linear-PSO algorithm using linear selection for population initialization and next-position updating.

978-1-4673-1310-0/12/$31.00 ©2012 IEEE

However, referring to the previous research, the existing Linear-PSO algorithm required too much time and resources compared to the original PSO [2]. Therefore, there is a need for speed improvement that will allow faster motif detection.

Linear-PSO is an improved version of the PSO algorithm for discovering motifs of DNA sequences [2]. Linear-PSO uses a reference set to replace the random numbers used in PSO. The first DNA sequence was selected as a reference set, and other DNA sequences will be used to compare the similarity of motifs among different animals from the same species. The next stage in identifying motifs is to extract the target motif from the reference set. A new target motif was extracted sequentially from the reference set.

LPBS is an improved version of Linear-PSO developed by the researcher and the results showed that it is better than Linear-PSO [1]. Two improvements were made to the Linear-PSO algorithm. First, the preprocessed data needed to be sorted before applying this algorithm. Second, for similarity searching, a binary search is used instead of a linear search.

The flows of the LPBS algorithm are as follows (see Figure 1): Step 1 refers to the initialization of the population by selecting the target motif from the reference set. The first DNA sequence becomes a reference set for a possible target motif. Step 2 refers to searching for similar motifs using the binary search. Step 3 refers to the calculation of the fitness value for each individual particle; the parameter of the particle's highest fitness value (pBest) will store the highest fitness value for that particle. Step 4 refers to the updating of the global highest fitness value (gBest). Step 5 refers to the updating of the new position of each particle by referring to a new target motif from the reference set. Step 6 refers to the termination condition where the process flow will be terminated if the condition is met; otherwise the steps are repeated from Step 2.

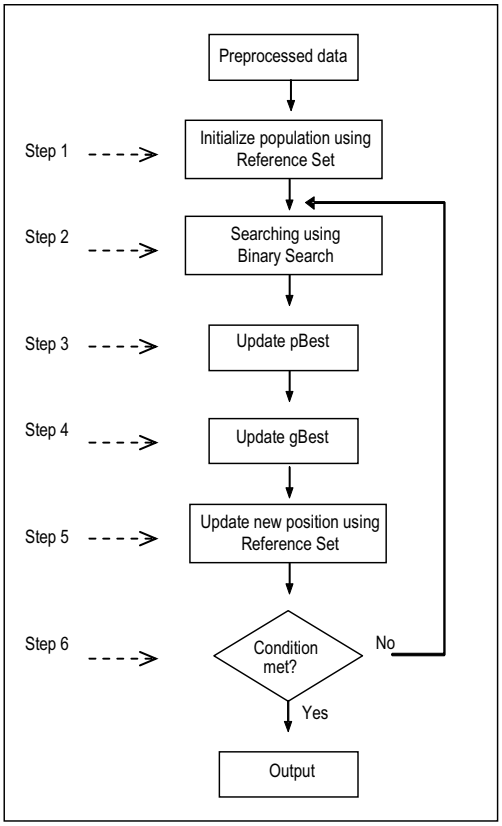

Figure 1. Flow of Linear-PSO with Binary Search

III. EXPERIMENTAL METHOD AND RESULTS

In this paper, two experiments were conducted with different samples of COI data.

A. Experiment 1

A genome is a complete set of DNA sequences of a species. The use of a genome as a reference set will allow all the possible target motifs to be extracted. Therefore, after going through the process of motif identification, the extracted motifs will represent the selected species.

In this experiment, two sets of sample data were collected from the Genbank database. The first set consists of 15 DNA sequences of *Sus scrofa* species but from different fragments of COI and is used as a comparison set. The second set is a complete genome of COI for *Sus scrofa* species and is used as a reference set.

Table I shows the detailed information on the genome of *Sus scrofa* COI used in this experiment. The accession number is an identification number used by the Genbank database for easier access.

TABLE I. GENOME OF *SUS SCROFA* COI (REFERENCE SET)

Accession Number	CO1 DNA Sequence	Length
NC_000845	ATGTTCGTAAATCGTTGACTATA…	1545bp

The DNA sequences of the comparison set can be categorized in three different groups: the first category (C1) has a length of 808 bp to 848 bp, the second category (C2) has a length of 613 bp to 798 bp, and the last category (C3) has a length of 256 bp to 575 bp. Table II shows the categories of DNA sequences used as a comparison set. The accession number is given instead of the DNA sequence for easier access and reference in the future.

The results show that the length of discovered motifs depended on the length of the DNA sequences. The longer the samples, the more motifs of greater length were discovered. This is shown in Table III.

Table III shows two discovered motifs with lengths of 10 bp, 9 bp, and 8 bp. Motifs with a length of 10 bp were the longest motifs discovered by this experiment. These longest motifs were only managed to discover using the sample in Category 1 and Category 2. The sample in Category 3 only enabled motif TTCTTTCC, with a length of 8 bp, to be discovered.

TABLE II. LIST OF DNA SEQUENCES (COMPARISON SET)

Category	Accession Number	Length (bp)
Category 1 (C1)	BW969362	808
	BP167907	835
	CJ013243	842
	BP162499	848
	BP165083	848
Category 2 (C2)	CJ016463	613
	EW340915	625
	EW379512	640
	BP165435	793
	BW979127	798
Category 3 (C3)	EW021716	256
	DT326666	359
	BP162131	380
	EW379414	527
	BP165895	575

However, motifs discovered by this experiment are not satisfactory because the length is too short due to the use of input data from different fragments of COI. Shorter motifs will permit duplication in other species' DNA sequences.

The above results were obtained because of the unavailability of good input data from the database during the experiment. This experiment only used a group of *Sus scrofa* DNA sequences under UniGene ID: 4642852. DNA sequences from this group from different resources were mixed and extracted from different parts of COI fragments.

Therefore, Experiment 2 was conducted in order to discover more reliable and longer motifs using COI DNA sequences.

B. Experiment 2

The DNA sequences used in Experiment 2 were extracted from the same fragment of COI starting base 53 to base 705. The lengths of all DNA sequences are 653 bp, which is nearly half of the genome length. The complete genome of COI was selected as a reference set and 14 other DNA sequences were selected as a comparison set. Table IV shows the details of a comparison set.

TABLE III. RESULT OF THE DISCOVERED MOTIF

Number	Motif Discovered	Length (bp)	C1	C2	C3
1	TTCCACTATG	10	✓	✓	
2	CATTCTTTCC	10	✓		
3	CCTTTCCAC	9	✓	✓	
4	TTCCACTAT	9	✓	✓	
5	TTCTTTCC	8	✓	✓	✓
6	AGCAGGTG	8	✓	✓	

TABLE IV. COMPARISON SET

Number	Accession Number	Length (bp)
1	JN714161.1	653
2	JN714162.1	653
3	JN714163.1	653
4	JN714166.1	653
5	JN714167.1	653
6	JN714168.1	653
7	JN714169.1	653
8	JN714170.1	653
9	JN714171.1	653
10	JN714172.1	653
11	JN714173.1	653
12	JN714174.1	653
13	JN714178.1	653
14	JN714187.1	653

TABLE V. MOTIF DISCOVERED

Number	Motif Discovered	Length (bp)
1	CTACTTGGCGATGATCAAATCTA…	77
2	CTACTTGGCGATGATCAAATCTA…	76
3	TACTTGGCGATGATCAAATCTAT…	76
4	CTACTTGGCGATGATCAAATCTA…	75
5	TACTTGGCGATGATCAAATCTAT…	75
6	ACTTGGCGATGATCAAATCTATA…	75

The results show that the greatest motif length discovered was 77 bp. The longer the motif, the better, because it will be

more difficult to discover the same motif in other species' DNA sequences.

Referring to Table V, all motifs discovered were actually located on the same fragment of DNA sequence and started almost from the same base position. Referring to the complete genome of *Sus scrofa* mitochondrial COI; motifs 1, 2, and 4 started at base 139, while motifs 3 and 5 started at base 140, and lastly motif 6 started at base 141.

Therefore, based on the above results, for species motif discovery it is reasonable that all extraction of COI DNA sequences should be started at base 139. However, further analysis with more data needs to be done before any final conclusions can be drawn.

IV. CONCLUSION

The Linear-PSO with Binary Search algorithm is able to discover a possible motif that can be used for identification of specific species. The results show that more reliable motifs can be discovered when the right data are used. Based on the results, base 139 is one of the options that can be used as a starting fragment of COI for further study in order to obtain the right motif for the right species. The above experiments are the only process of motif discovery where the patterns of DNA sequences are discovered. The next step, which is not included in this paper, is the process of motif identification, where all discovered motifs will be compared with the COI genomes of other species in order to identify the right motif. It should be possible to use the identified motifs for identification of specific species.

REFERENCES

[1] S. L. Syed Abdullah and H. Harun, "Motif discovery using linear-pso with binary search," 2nd World Conference on Information Technology. Turkey, 2011.

[2] S. L. Syed Abdullah, H. Harun and M. N. Taib, "A modified algorithm for species specific motif discovery," International Conference on Science and Social Research, Kuala Lumpur, 2010.

[3] B. C. H. Chang, A. Ratnaweera, and S. K. Halgamuge, "Particle swarm optimization for protein motif discovery," Genetic Programming and Evolvable Machines, vol. 5, pp. 203-214, 2004.

[4] W. Zhou, H. Zhu, G. Liu, Y. Huang, Y. Wang, D. Han, and C. Zhou, A Novel Computational Based Method for Discovery of Sequence Motifs from Co expressed Genes, International Journal of Information Technology, vol. 11, 2005.

[5] P. D. N. Hebert, A. Cywinska, S. L. Ball and J. R. deWaard, "Biological identification through DNA barcodes." Proc. R. Soc. Lond. B 270, pp. 313-322. 2003.

[6] O. Folmer, M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek, "DNA primers for amplication of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates," Molecular Marine Biology and Biotechnology, 3(5), pp. 294-299, 1994.

[7] B. Verge, Y. Alonso, J. Valero, C. Miralles, E. Vilella, and L. Martorell, Mitochondrial DNA (mtDNA) and Schizophrenia. European Psychiatry, 26, pp. 45-56, 2010.

[8] J. Kennedy and R. Eberhart, Particle Swarm Optimization, IEEE International Conference on Neural Networks, Perth, Australia, 1995.

[9] C.T. Hardin and E. C. Rouchka, DNA Motif Detection Using Particle Swarm Optimization and Expectation-Maximization, IEEE Symposium on Swarm Intelligence, 2005.

[10] C. Lei and J. Ruan, A Particle Swarm Optimization Algorithm for Finding DNA Sequence, IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, 2008.

[11] R. Akbari and K. Ziarati, An Efficient PSO Algorithm for Motif Discovery in DNA, IEEE International Conference of Emerging Trends in Computing, Tamil Nadu, India, 2009.