Jurnal
Teknologi

Full Paper

DNA MOTIF IDENTIFICATION USING LPBS

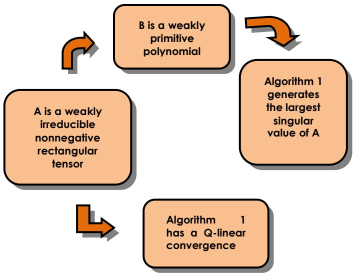
Hazaruddin Harun[a]*, Sharifah Lailee Syed Abdullah[b], Hamirul Aini Hambali[a]

[a]School of Computing, UUM College of Art and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah Darulaman, Malaysia
[b]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 02600 Arau, Perlis, Malaysia

Article history
Received
12 August 2015
Received in revised form
14 December 2015
Accepted
11 January 2016

*Corresponding author
hazaruddin@uum.edu.my

Graphical abstract


B is a weakly primitive polynomial
Algorithm 1 generates the largest singular value of A
Algorithm 1 has a Q-linear convergence
A is a weakly irreducible nonnegative rectangular tensor

Abstract

In recent years, several deoxyribonucleic acid (DNA)-based approaches have been developed for species identification including DNA sequencing. The search for motif or patterns in DNA sequences is important in many fields especially in biology. In this paper, a new particle swarm optimization (PSO) approach for discovering species-specific motifs was proposed. The new method named as Linear-PSO with Binary Search (LPBS) is developed to discover motifs of specific species through DNA sequences. This enhanced method integrates Linear-PSO and binary search technique to minimize the execution time and to increase the correctness in identifying the motif. In this study, two fragments samples of 'mitochondrial cytochrome C oxidase subunit I' (COI or COX1) were collected from the Genbank online database. DNA sequences for the first sample are fragments of COI for one species and the second samples are a complete COI from a different species. The genome of COI was used as a reference set and other DNA sequences were used as a comparison set. The results show that the LPBS algorithm is able to discover motifs of a species when using DNA sequences from the same fragment of COI.

Keywords: Generalized hyperbolic distribution, portfolio optimization

1.0 INTRODUCTION

Over the last few decades, the field of DNA sequencing has emerged as an important technique for species identification. DNA or Deoxyribonucleic acid is a molecule which made up of smaller units called nucleotides. The nucleotides sequence is the most basic level of genome which contains genetic information that make each species unique. In general, the length of DNA sequence ranges from a few hundred to several billions of nucleotides for different species. Therefore, sequence motif discovery algorithm has become important in computational genetics to discover patterns in nucleotide sequences with the aim to identify the specific species. The order of nucleotides in DNA sequence is made up of four letters, A, C, G and T which represent adenine, cytosine, guanine and thymine, respectively. These letters are used in a fragment of DNA to determine genetic code that belongs to specific species. The example of fragment of DNA sequence for different species is shown in Figure 1.

Homo Sapiens (human) = AGGAGGGGTCCAGCCCT
Sus Scrofa (pig) = ACTTTTTCCTCAGCCGG
MusMusculus (mouse) = TCCTGCTATAGGGCCAG

Figure 1 Example of DNA Sequence for Different Species

However, a motif discovery is one of the most challenging tasks in computational genetic due to two main reasons. The first reason is an inaccuracy in detecting motif for dataset which contains noise [1, 2]. The second challenge is the lack of knowledge about the structure of investigated patterns.

In order to find the accurate motif, an identification of suitable fragment of DNA sequence is required. In that case, each of the investigated fragments should contain potential target motif. Previously, many researchers used 'mitochondrial DNA' (mtDNA) as a suitable fragment for motif identification of specific species [3, 4]. Mitochondrial DNA is another structure of DNA which provides valuable information for investigating DNA in certain conditions. It is found in mitochondria which contains 100 to 1000 copies of mtDNA genome in each mitochondrion. The high number of mtDNA allows a better extraction of DNA sequence [5]. The mtDNA contains few genes such as the cytochrome b (cytb) gene, the large (16S) and the small (12S) ribosomal RNA (rRNA) gene and the mitochondrial cytochrome C oxidase subunit I (COI or COX1) gene.

Previous researches have shown that COI gene has been used in many phylogenetic studies which determine relations among species. The COI was selected because it had been widely used to detect species in biology and has been accepted as a practical and universal species-level barcode for animals [3, 6]. Hebert *et al*. [3] have proposed the used of COI gene and successfully developed COI species profile for identifying lepidopteran species. In this study, COI was chosen as a target gene because COI provides two important advantages. The first is the universal primers for COI gene are very robust and reliable. The second is COI gene has a greater range of phylogenetic signals than any other mitochondrial gene [6]. Moreover, it shows a high commonness of base substitutions, and thus leads to more accurate result.

However, there is still a lack of procedure standardization in species identification using DNA and only a few studies exist that examined the effectiveness of COI as a DNA barcode. Therefore, this study has proposed an innovative technique that is able to identify motif of DNA sequence with higher validity.

2.0 LINEAR-PSO WITH BINARY SEARCH

The Linear-PSO with Binary Search (LPBS) algorithm is an integration of Linear-PSO algorithm and Binary Search technique. The integration of both techniques is required to increase the validity in discovering motif of DNA sequence for specific species. The Linear-PSO algorithm was the improvement of original Particle Swarm Optimization (PSO).

Particle Swarm Optimization (PSO) is an established population-based optimization technique which has been shown to be effective in solving different type of motif finding problems [7]. The PSO algorithm was introduced by Kennedy and Eberhart [8] and this algorithm was inspired by social behavior of animal as those found in schools of fish or flocks of birds. It was developed to simulate and represent the movement of these animals in finding food sources and avoid predators. The original PSO algorithm starts with the random initialization of a population of individual particles. These particles are moved around in the search space according to the fitness calculation. The movement and the new position of each particle were determined based on the current position and velocity values which used random number generation.

There are few studies which used PSO to discover motif of DNA sequence. Chang [9] was first modified and used PSO in his research. In this research, PSO was integrated with symbolic data for discovering motifs in protein sequences. Later the algorithm was extended by integrating hybrid algorithms [10, 11], adding a dissimilarity graph [2], and applying the stochastic local search concept [12].

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*. [13] proposed a Linear-PSO algorithm using linear selection for population initialization and next-position updating. However, the algorithm required too much time and resources compared to the original PSO [13].

Therefore, Syed Abdullah *et al*. have improved the speed of algorithm by integrating Binary Search technique in this algorithm. This new integrated algorithm namely LPBS allows more efficient and faster motif detection. The results of LPBS showed that this algorithm is superior to Linear-PSO [14]. Two improvements were made to the Linear-PSO algorithm. First, the preprocessed data was sorted first 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 2): 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.

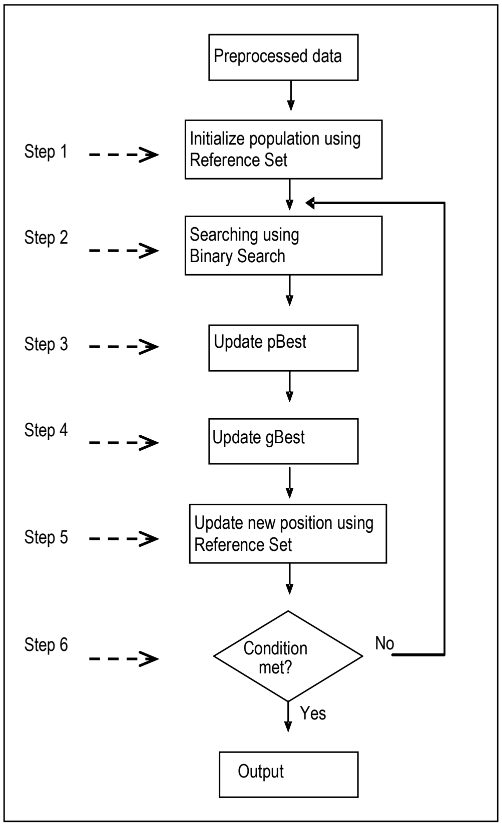

Preprocessed data
Step 1
Initialize population using Reference Set
Step 2
Searching using Binary Search
Step 3
Update pBest
Step 4
Update gBest
Step 5
Update new position using Reference Set
Step 6
Condition met?
No
Yes
Output

Figure 2 Flow of Linear-PSO with Binary Search

3.0 EXPERIMENTAL METHOD AND RESULTS

Evaluation of the algorithm is a process where the Linear-PSO with Binary Search algorithm was used for solving Motif Discovery and Motif Identification. Motif discovery is a process of discovering a meaningful pattern in a DNA sequence that is commonly shared by molecules of the same species, whereas motif identification is a process to identify the correct motif that can represent the species.

For species specific motif discovery, the genome of *Mitochondrial Cytochrome C Oxidase Subunit I* (COX1) was used as a Reference Set of the algorithm. Other DNA sequences (fragment of COX1) were used as a Comparison Set.

A species was selected based on the availability of data; namely the genome of *Mitochondrial Cytochrome C Oxidase Subunit I* (COX1) and DNA sequence fragments of COX1. All Target Motifs were extracted from the genome and compared with other DNA sequences of several selected species. There were only two species that were selected for the experiments, which were *BosTaurus* (cow) and *GallusGallus* (chicken).

In this paper, each experiment was conducted with different samples of COI data.

Table 1 shows the details of the COX1 genome used as a Reference Set.

Table 1 Genome of COX1

Accession No	Species	Length
NC_006853	*BosTaurus*	1545bp
ATGTTCATTAACCGCTGA...		

Because of the unavailability of data, only 10 DNA sequences of the COX1 fragments were selected from the Genbank database and used as a Comparison Set. Table 2 shows the details of DNA sequences from *BosTaurus* species that were used as a Comparison Set. The lengths of DNA sequences were between 658bp and 715bp.

Table 2 Selected Bos Taurus DNA sequences

No	Accession Number	Length
1	FJ958332.1	708
2	FJ958333.1	708
3	FJ958334.1	708
4	FJ958335.1	710
5	FJ958336.1	710
6	GU130589.1	715
7	GU130590.1	715
8	HQ860420.1	658
9	JF700140.1	658
10	JF700141.1	658

For this experiment, Genome and fragment of COX1 from *Bos Taurus* species were used as an input to the Linear-PSO with Binary Search algorithm.

3.1 Motif Discovery

The first column in Table 3 contains the sequence number for discovered motifs, while the second column shows the list of motifs discovered by the algorithm. The third column contains the lengths of each motif.

Table 3 Discovered Motif – Bos Taurus

Discovered Motif	Length
AGTTGTAACCGCACACGCAT...GCAGG	294bp
GTTGTAACCGCACACGCATT...GCAGG	293bp
AGTTGTAACCGCACACGCAT...AGCAG	293bp
TTGTAACCGCACACGCATTT...GCAGG	292bp
AGTTGTAACCGCACACGCAT...TAGCA	292bp
GTTGTAACCGCACACGCATT...AGCAG	292bp

As shown in the table, the longest motif discovered was AGTTGTAACCGCACACGCAT...GCAGG, with the length of 294bp. The longer the motif is better because it would reduce the possibility of similarity with other species.

3.2 Motif Identification

The first column in Table 4 represents the sequence number for discovered motifs. The second column shows the list of motifs discovered by the algorithm, while the third column exhibits the lengths of each motif. Meanwhile, columns four to eleven are representative of other species, where there is an indicator to highlight the similarity of the motif with other DNA sequences originating from other species. An indicator of ('✓') will show the similarity and an indicator of ('X') will show the dissimilarity.

As shown in Table 4, there are no similarities between all the motifs that were discovered and COX1 genome from other species. Therefore, all discovered motifs have a potential use as a motif that can represent the *BosTaurus* species.

Table 4 Motif Similarity

Motif Discovered	Len	S1	S2	S3	S4	S5	S6	S7	S8
AGTTGT…AGG	294bp	X	X	X	X	X	X	X	X
GTTGTA…AGG	293bp	X	X	X	X	X	X	X	X
AGTTGT…CAG	293bp	X	X	X	X	X	X	X	X
TTGTAA…AGG	292bp	X	X	X	X	X	X	X	X
AGTTGT…GCA	292bp	X	X	X	X	X	X	X	X
GTTGTA…CAG	292bp	X	X	X	X	X	X	X	X

* S1: *SusScrofa* (pig), S2: *HomoSapiens* (human), S3: *Ovis Arise* (sheep), S4: *CanisLupus* (dog), S5: *XenopusLaevis* (frog), S6: *MusMusculus* (rat), S7: *BosGrunniens* (yak) & S8: *GallusGallus* (chicken)

* X - Not Similar, ✓ - Similar

3.3 Motif Alignment

Motif alignment is used to compare all the discovered motifs. The result from Table 5 shows that all the discovered motifs have been extracted from the same fragment of COX1, which is located between bases 168 to base 461 in the genome. However based on other comparison study, this fragment location of COX1 is only applicable for *BosTaurus* species motifs.

Table 5 Motif Alignment for Bos Taurus

Motif Alignment	Length
AGTTGTAACCGCACACGC…ACTTAGCAGG	294bp
bGTTGTAACCGCACACGC…ACTTAGCAGG	293bp
AGTTGTAACCGCACACGC…ACTTAGCAG**b**	293bp
bbTTGTAACCGCACACGC…ACTTAGCAGG	292bp
AGTTGTAACCGCACACGC…ACTTAGCA**bb**	292bp
bGTTGTAACCGCACACGC…ACTTAGCAG**b**	292bp

* '**b**' - represents missing base.

However, the discovered motifs are only possible to compare with the COX1 fragment that only have a length between 658bp and 715bp, which is only almost half of the COX1 length (1545bp). Therefore, in order to discover more potential motifs, experiments using the whole COX1 fragment should be done when the data becomes available.

3.4 Experiment 2

Table 6 shows the details of the COX1 genome used as a Reference Set.

Table 6 Genome of COX1

Accession No	Species	COX1	Length
AP003580	*Gallus Gallus*	GTGACCTTCATCAACCGA…	1551 bp

Since there was an unavailability of data, only nine DNA sequences of the COX1 fragment were selected from the database and used as a Comparison Set. Table 7 shows the details of DNA sequences from the *Gallus Gallus* species. The lengths of DNA sequences were between 537bp and 699bp.

Table 7 Selected Gallus Gallus DNA sequences

No	Accession Number	Length
1	JF498860.1	669
2	JF498861.1	694
3	JF498862.1	694
4	JF700165.1	658
5	JF700166.1	658
6	JN793565.1	604
7	JN793568.1	642
8	GQ922621.1	699
9	HM102301.1	537

For this experiment, the genome and fragment of COX1 from the *Gallus Gallus* species were used as input for the Linear-PSO with Binary Search algorithm.

3.5 Motif Discovery

The first column in Table 8 shows the sequence numbers for discovered motifs. Meanwhile, the second column contains the list of motifs discovered by the algorithm and the third column gives the lengths of each motif.

As shown in Table 8, the longest motif discovered was CTTATAATCGGTGCCCCAGA…ACATA with the length of 261bp. The longer the motif, the better it is for use because it would reduce the possibility of similarity with other species.

Table 8 Discovered Motif – Gallus Gallus

Discovered Motif	Length
CTTATAATCGGTGCCCCAGA...ACATA	261bp
CTTATAATCGGTGCCCCAGA...AACAT	260bp
TTATAATCGGTGCCCCAGAC...ACATA	260bp
CTTATAATCGGTGCCCCAGA...CAACA	259bp
TTATAATCGGTGCCCCAGAC...AACAT	259bp
TATAATCGGTGCCCCAGACA...ACATA	259bp

3.6 Motif Identification

The first column in Table 9 shows the sequence numbers for discovered motifs. The second column has the list of motifs discovered by the algorithm, and the third column contains the lengths of each motif. Meanwhile, columns four to eleven represent other species, where there are indicators to show the similarity of the motif with other species DNA sequences. An indicator of ('✓') will show the similarity and an indicator of ('X') will show the dissimilarity.

Table 9 Motif Similarity

Motif Discovered	Len	S1	S2	S3	S4	S5	S6	S7	S8
CTTATA...ATA	261bp	X	X	X	X	X	X	X	X
CTTATA...CAT	260bp	X	X	X	X	X	X	X	X
TTATAA...ATA	260bp	X	X	X	X	X	X	X	X
CTTATA...ACA	259bp	X	X	X	X	X	X	X	X
TTATAA...CAT	259bp	X	X	X	X	X	X	X	X
TATAAT...ATA	259bp	X	X	X	X	X	X	X	X

* S1: *Sus Scrofa* (pig), S2: *Bos Taurus* (cow), S3: *Homo Sapiens* (human), S4: *Ovis Arise* (sheep), S5: *Canis Lupus* (dog), S6: *Xenopus Laevis* (frog), S7: *Mus Musculus* (rat) & S8: *Bos Grunniens* (yak)

* X - Not Similar, ✓ - Similar

As shown in Table 9, there are no similarities between all the motifs that were previously discovered (as shown in Table 6.11), and the COX1 genome from other species. Therefore, all discovered motifs have the potential to become a motif that is representative of the *Gallus Gallus* species.

3.7 Motif Alignment

Motif alignment is used to compare all the discovered motifs. The results from Table 10 show that all the discovered motifs were extracted from the same COX1 fragment, which is located between base 281 and base 541 in the genome. However when compared to other studied species, this fragment location of COX1 is only applicable for the potential identifying motif for the *Gallus Gallus* species.

Table 10 Motif Alignment for Gallus Gallus

Motif Alignment	Length
CTTATAATCGGTGCCCCA...CATCAACATA	261bp
bTTATAATCGGTGCCCCA...CATCAACATA	260bp
CTTATAATCGGTGCCCCA...CATCAACAT**b**	260bp
bbTATAATCGGTGCCCCA...CATCAACATA	259bp
CTTATAATCGGTGCCCCA...CATCAACA**bb**	259bp
bTTATAATCGGTGCCCCA...CATCAACAT**b**	259bp

* '**b**' - represents missing base.

However, the discovered motifs were only for the COX1 fragment with the length between 537bp to 699bp, which is less than half of the COX1 length (1551bp). Therefore, to discover more potential motifs, future experiments using the whole COX1 fragment should be performed when the data becomes available.

The results showed that the Linear-PSO with Binary Search algorithm had successfully discovered motifs for two different species. Although the findings showed that all the discovered motifs can be used as a motif potential identifier for the *BosTaurus* and *Gallus Gallus* species, it is better to run more tests with other species when the data becomes available in the database. The rest of the COX1 fragments also need to be tested to discover more possible motifs.

4.0 CONCLUSION

Although the results showed that all the discovered motifs can be used as a potential motif for identifying the *BosTaurus and Gallus Gallus* species, it is better to have more tests with other species when the data becomes available for use in the database. The rest of the COX1 fragments also need to be tested over time to discover more possible motifs.

References

[1] Karabulut, M. & Ibrikci, T. 2012. A Bayesian Scoring Scheme Based Particle Swarm Optimization Algorithm to Identify Transcription Factor Binding Sites. *Applied Soft Computing*, 12: 2846-2855.

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

[3] Hebert, P. D. N., Cywinska, A., Ball, S. L. and deWaard, J. R. 2003. Biological Identification through DNA Barcodes. *Proc. R. Soc. Lond.* B. 270: 313-322.

[4] Folmer O., Black M., Hoeh W., Lutz R., and Vrijenhoek R. 1994. DNA Primers for Amplication of Mitochondrial Cytochrome C Oxidase Subunit I from Diverse Metazoan Invertebrates. *Molecular Marine Biology and Biotechnology*. 3(5): 294-299.

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

[6] Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. 2003. Barcoding Animal Life: Cytochrome C Oxidase Subunit 1 Divergences Among Closely Related Species. *Proc Biol Sci. 270*(Supp1_1): 96-99.

[7] Arock, M., Reddy, S. and Reddy, A. V. 2010. A Parallel Combinatorial Algorithm for Subtle Motifs. *Int. J. Bioinformatics Research and Application*. 6(3): 260-269,

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

[9] B. C. H. Chang, A. Ratnaweera, and S. K. Halgamuge, 2004. Particle Swarm Optimization for Protein Motif Discovery. *Genetic Programming and Evolvable Machines*. 5: 203-214.

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

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

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

[13] S. L. Syed Abdullah, H. Harun and M. N. Taib. 2010. A Modified Algorithm for Species Specific Motif Discovery. International Conference on Science and Social Research, Kuala Lumpur.

[14] S. L. Syed Abdullah and H. Harun. 2011. Motif Discovery using Linear-Pso with Binary Search. 2nd *World Conference on Information Technology*. Turkey.