SPECIES MOTIF EXTRACTION USING LPBS

Sharifah Lailee Syed Abdullah[1] and Hazaruddin Harun[2]

[1]*Universiti Teknologi MARA, Malaysia, shlailee@perlis.uitm.edu.my*
[2]*Universiti Utara Malaysia, Malaysia, hazaruddin@uum.edu.my*

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, fragments of mitochondrial cytochrome C oxidase subunit I (COI/COX1) and genome of COI were collected from the Genbank online database. For the first experiment, 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. For the second experiment, all the discovered motifs were used as a reference set and the genome of COI from other species were used as a comparison set. The results show that the LPBS algorithm is able to identify correct motifs for species identification.

Keywords: Particle Swarm Optimization, Linear-PSO, motif discovery, species specific

INTRODUCTION

The Linear-PSO with Binary Search (LPBS) algorithm was first introduced for motif discovery by Syed-Abdullah and Harun (2011). This algorithm is an extension of the Linear-PSO algorithm (Syed-Abdullah, Harun, & Taib, 2010) based on the modified 'Particle Swarm Optimization' algorithm (PSO) used by Chang et al. (2004) and Zhou et al. (2005) for motif discovery. LPBS does not generate any random number; instead the target motif was determined sequentially, thereby ensuring that all motifs were tested. Results from previous research showed that the LPBS algorithm was able to discover motifs with higher validity and efficiency (Syed-Abdullah, & Harun, 2011). The LPBS algorithm was the first PSO-based algorithm that was developed to discover motifs which can be used for identification of specific species.

For the purpose of discovering 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 (Hebert et al., 2003; Folmer et al, 1994). Mitochondrial DNA is another DNA structure 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 (Verge et al, 2010). 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. (2003) proposed the used of mitochondrial cytochrome C oxidase subunit I (COI) for species identification. COI is one of the genes existing in mtDNA.

According to Hebert et al. (2003), COI was chosen as a target gene because it appears to have a greater range of phylogenetic signals compared to any other mitochondrial gene (Folmer et al., 1994). Phylogenetic signals are used in research on relations among species.

LINEAR-PSO WITH BINARY SEARCH

The Particle Swarm Optimization (PSO) algorithm is an evolutionary optimization algorithm introduced by Kennedy and Eberhart (1995). 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 improved and used by Chang (2004). Later the algorithm was extended by integrating hybrid algorithms (Hardin & Rouchka, 2005; Zhou et al., 2005), adding a dissimilarity graph (Lei & Ruan, 2008), and applying the stochastic local search concept (Akbari & Ziarati, 2009).

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, comprehensive selecting and searching must be used to identify the correct motif to represent a species. All possible motifs will be tested in order to get a higher fitness value. Therefore, Syed-Abdullah et al. (2010) proposed a Linear-PSO algorithm using linear selection for population initialization and next-position updating.

However, referring to the previous research, the existing Linear-PSO algorithm required more time and resources compared to the original PSO (Syed-Abdullah et al., 2012). Therefore, there is a need for time improvement to allow faster motif detection.

LPBS is an improved version of Linear-PSO developed by us and the results showed that it is better than Linear-PSO (Syed-Abdullah, & Harun, 2011). Two improvements were made to the Linear-PSO algorithm. First, the preprocessed data were sorted before the processing of motif discovery was started. Second, for motif similarity searching, a binary search was used instead of a linear search.

The steps of the LPBS algorithm, as shown in Figure 1, are as follows: Step 1 refers to the initialization of the population by selecting the target motif from the reference set. One of the DNA sequence is selected as a reference set for a possible target motif. Other DNA sequences are selected as a comparison set. 'Target motif' is a new term that replaces the used of 'particle' in the original PSO. Step 2 refers to a search for similar motifs using the binary search. Step 3 refers to the calculation of the fitness value for each individual target motif; the parameter of the highest fitness value (pBest) will store the highest fitness value for that target motif. Step 4 refers to the updating of the global highest fitness value (gBest). Step 5 refers to the updating of the new target motif 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.

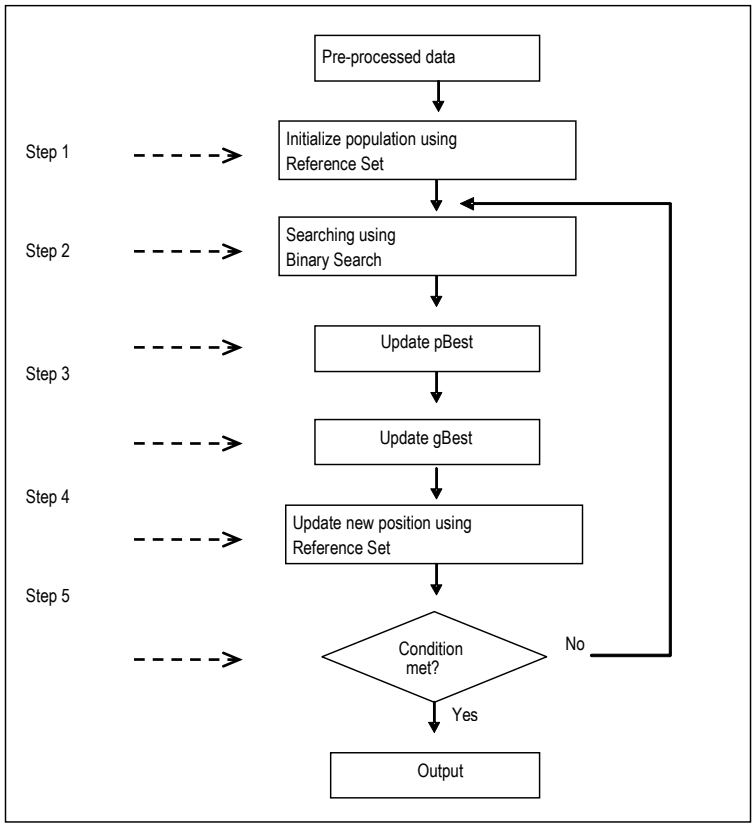

Figure 2. Flow of Linear-PSO with Binary Search (Syed-Abdullah et al., 2012)

EXPERIMENTAL METHOD AND RESULTS

In this paper, two experiments were conducted; the first experiment concerned motif discovery and the second, motif identification.

Experiment 1 (Motif Discovery)

Motif Discovery is a process used to discover motifs from DNA sequences. In this experiment, two datasets were collected from the Genbank database. The first dataset consists of 10 fragments of DNA sequences of cow (*Bos taurus*) species. All the fragments extracted from the same fragment of COI with lengths between 658 to 715 bp are used as a comparison set. The second dataset is a complete genome of COI for cow species and is used as a reference set.

Table 1 shows the information of the COI genome used in this experiment. The accession number is an identification number used by the Genbank database for easier access.

Table 2. Genome of cow (Bos Taurus) COI (Reference Set)

Accession Number	CO1 DNA Sequence	Length
NC_006853	ATGTTCATTAACCGCTGA…….	1545bp

Table 2 shows the 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.

Table 3. List of DNA Sequences (Comparison Set)

Category	Accession Number	Length (bp)
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

Table 3 shows the six longest motifs discovered. The first column is the sequence number for the motif discovered. The second column is the list of motifs discovered by the algorithm and the third column is the length of each motif.

As shown in the table, the longest motif discovered was 'AGTTGTAACCGCACACGCAT…GCAGG' with a length of 294 bp. The longer the motif, the better, because it reduces the possibility of motif similarity with other species.

Table 4. List of Discovered Motifs

No	Motif Discovered	Length
1	AGTTGTAACCGCACACGCAT…GCAGG	294bp
2	GTTGTAACCGCACACGCATT…GCAGG	293bp
3	AGTTGTAACCGCACACGCAT…AGCAG	293bp
4	TTGTAACCGCACACGCATTT…GCAGG	292bp
5	AGTTGTAACCGCACACGCAT…TAGCA	292bp
6	GTTGTAACCGCACACGCATT…AGCAG	292bp

The next experiment was conducted in order to identify correct motifs that can represent the standard motif for species identification. To represent the standard motif, the selected motif must not exist in the DNA sequences of other species.

Experiment 2 (Motif Identification)

Motif Identification is a process used to identify correct motifs that can represent a specific species. All motifs that were discovered in experiment 1 were selected as a reference set. The genome of COI from other species were collected from the database and used as a comparison set. Table 4 shows the information of all collected genome. Genome collection was based on the availability of genome of COI data in the database, and data of only eight species were collected for this experiment: pig, human, sheep, dog, frog, rat, yak and chicken.

Table 5. Genome of COI from Other Species

No.	Accession Number	Species	Length bp
1	NC_000845	Pig (Sus Scrofa)	1545
2	NC_012920	Human (Homo Sapiens)	1542
3	NC_001941	Sheep (Ovis Arise)	1545
4	NC_002008	Dog (Canis Lupus)	1545
5	NC_001573	Frog (Xenopus Laevis)	1555
6	NC_005089	Rat (Mus Musculus)	1545
7	NC_006380	Yak (Bos Grunniens)	1545
8	AP003580	Chicken (Gallus Gallus)	1551

Table 5 shows the similarity between the discovered motifs and genome of COI from other species. The first column is the sequence number for the discovered motif. The second column is the list of motifs discovered by the algorithm in experiment 1 and the third column is the length of each motif. Columns 4 to 11 represent other species where there is an indicator showing the similarity of the motif with other species' DNA sequences. A tick, '✓', indicates similarity and a cross, 'X', dissimilarity.

As shown in Table 5, there are no similarities between any of the motifs that have been discovered and the genome of COI from other species. Therefore, all of the discovered motifs have the potential to be motifs that can represent cow species.

Table 6. Motifs Similarity

No	Motif Discovered	Length bp	S1	S2	S3	S4	S5	S6	S7	S8
1	AGTTGT……AGG	294	X	X	X	X	X	X	X	X
2	GTTGTA…….AGG	293	X	X	X	X	X	X	X	X
3	AGTTGT…….CAG	293	X	X	X	X	X	X	X	X
4	TTGTAA…….AGG	292	X	X	X	X	X	X	X	X
5	AGTTGT…….GCA	292	X	X	X	X	X	X	X	X
6	GTTGTA…….CAG	292	X	X	X	X	X	X	X	X

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

* X - Not Similar, ✓ - Similar

CONCLUSION

The Linear-PSO with Binary Search algorithm is able to discover possible motif that can be used for the identification of specific species. Although the findings show that all of the discovered motifs can be used as a motif for cow species, it would be better to carry out more testing with other species when data are available. The accuracy of the identified motifs depends on the number of comparison with the other species. The greater the number of comparisons, the better the result, because the result shows the degree of similarity of existing motifs with other species. Without genomes of COI from many species, it is difficult to finally conclude that a motif that can be used as a motif for the selected species. This is because, the motif can only be compared with a few species and could be duplicated in other species that have not been compared due to data limitation. Therefore, more genomes of COI from other species are needed to actually finalize the motif for the selected species.

ACKNOWLEDGMENTS

This research was supported in part by FRGS Grant from Ministry of Higher Education, Malaysia (MOHE).

REFERENCES

Akbari, R., & Ziarati, K. (2009). An Efficient PSO Algorithm for Motif Discovery in DNA. Paper presented at IEEE International Conference of Emerging Trends in Computing, Tamil Nadu, India.

Chang, B. C. H., Ratnaweera, A., & Halgamuge, S. K. (2004). Particle Swarm Optimization for Protein Motif Discovery. Genetic Programming and Evolvable Machines, vol. 5, pp. 203-214.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & 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), pp. 294-299.

Hardin, C. T., & Rouchka, E. C. (2005). DNA Motif Detection Using Particle Swarm Optimization and Expectation-Maximization. Paper presented at IEEE Symposium on Swarm Intelligence.

Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. (2003). Biological Identification through DNA Barcodes. Proc. R. Soc. Lond. B 270, pp. 313-322.

Kennedy, J., & Eberhart, R. (1995). Particle Swarm Optimization. Paper presented at IEEE International Conference on Neural Networks, Perth, Australia.

Lei, C., & Ruan, J. (2008). A Particle Swarm Optimization Algorithm for Finding DNA Sequence. Paper presented at IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia.

Syed-Abdullah, S. L., Harun, H., & Taib, M. N. (2010). A Modified Algorithm for Species Specific Motif Discovery. Paper presented at International Conference on Science and Social Research, Kuala Lumpur.

Syed-Abdullah, S. L., & Harun, H. (2011). Motif Discovery using Linear-PSO with Binary Search. Paper presented at 2nd World Conference on Information Technology, Turkey.

Syed-Abdullah, S. L., Harun, H., Mohd Hussin, N., & Abd-Khalid, N. E. (2012). Comparative Study of Random-PSO and Linear-PSO Algorithms. Paper presented at International Conference on Computer & Information. Kuala Lumpur.

Syed-Abdullah, S. L., Harun, H., Mohd-Hussin, N., Hamid, J. N., & Mohamed-Hanaphi, R. (2012). Identifying the Definite Base of COI for Extraction of DNA Sequences using LPBS. Paper presented at The 2012 IEEE Symposium on Humanities, Science and Engineering Research. Kuala Lumpur, 2012.

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

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