A Modified Algorithm for Species Specific Motif Discovery

Sharifah Lailee Syed Abdullah, Hazaruddin Harun
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Arau, Perlis, Malaysia
shlailee@perlis.uitm.edu.my, udin_71@streamyx.com

Mohd Nasir Taib
Faculty of Electrical Engineering
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia
dr.nasir@ieee.org

***Abstract*-Motif discovery can be used to categorize unknown DNA sequences into their corresponding families. For this study, PSO was modified for discovering motif. The modified Linear-PSO is chosen even though it is a slower because linear search is not a choice but a necessary criteria for identifying motif of pig (Sus Scrofa). Pig motif identification is a critical for halal authentication. The modified Linear-PSO algorithm used linear number for population initializing and next position updating. For each cycle, only a particle called 'target motif' was selected and compared with other DNA sequences for fitness calculation. Motif discovered can be used as a standard motif for species identification. Experimental results show that the modified algorithm is able to identify motifs as expected. This study showed that a slower algorithm is still needed and has value based on how critical the problem is.**

Keywords-Motif Discovery; Species Specific; PSO

I. INTRODUCTION

Sequence patterns (motifs) are useful and important in molecular biology because of their capabilities to describe and classify features in DNA, RNA and protein sequences. The first pattern in DNA was discovered in 1970 by Hamilton Smith after the discovery of the Hind II restriction enzyme [1] and now numerous algorithms have been developed to discover pattern such as MEME [2], AlignACE [3], GibbsSampler [4], BioProspector [5]. Some of these algorithms specialized on DNA patterns only while others focus on both protein and DNA.

Motif discovery is a process of discovering a meaningful pattern for DNA, RNA or protein sequence that is commonly shared by two or more human and animal molecules. Motif discovery is a search problem and therefore search methods such as genetic algorithms, particle swarm optimization and local searches can be applied to discover existing motifs [6]. This process is important in the study of gene function [7] because motif discovery can be used to categorize unknown sequences into their corresponding families.

In this paper, we present a modified motif search algorithm based on a population-based stochastic optimization technique called Particle Swarm Optimization (PSO). The algorithm use linear number to discover the right motif for representing a specific species. Currently there is no computational based research on species specific motif discovery using DNA sequence.

The remaining sections are organized as follows; Section II, data used in this experiment, Section III, introduction of Linear-PSO, Section IV, method used for this study and finally the result are discussed in Section V.

II. EXPERIMENT DATA

All data used in this research is collected from TRANSCompel database. TRANSCompel 7.0 Public contains 322 composite regulatory elements on genes and transcription factors that correspond to entries in the TRANSFAC database. TRANSFAC database catalogs Transcription Factor Binding Site (TFBS) and their known motif. It represents the largest repository and most commonly used database for derived TFBS [8]. TRANSFAC and TRANSCompel are based on experimental evidence which is extracted from peer-reviewed papers [9].

For this study, DNA sequences from the same species and motif were searched. From the searched results, four DNA sequences from Human (Homo Sapiens) with the same motif (GAAATTCC) were selected (Table 1). Human (Homo Sapiens) DNA sequence was used instead of pig (Sus Scrofa) because Sus Scrofa DNA was not available in TRANSCompel database. In our experiment, each selected DNA sequences were searched for common motifs that exist in all sequences.

TABLE 1: DETAILS OF SELECTED DNA SEQUENCES

Access. No	Species	Base	Motif
C00327	Homo Sapiens	506	….GAAATTCC..
C00155	Homo Sapiens	1479	….GAAATTCC..
C00057	Homo Sapiens	1835	….GAAATTCC
C00189	Homo Sapiens	6684	GAAATTCC….

III. LINEAR-PSO

Particle Swarm Optimization (PSO) algorithm is an evolutionary optimization technique introduced by Kennedy and Eberhart [10]. Development of this technique was motivated by the animal social behaviors such as school of fish or flock of birds and the way these animals find food sources and avoid predators.

PSO algorithm starts with the random initialization of a population of individual particle in the search space. Each particle goes through the fitness calculation. New position for each particle is calculated based on current position and velocity values. The velocity value for each particle used random number generation.

In motif discovery, particle swarm optimization was first used by B Chang [11]. Later the algorithm was extended by integrating hybrid algorithm [7, 12]; adding a dissimilarity graph [13]; and applying the stochastic local search concept [6].

However previous researches only focus on motif discovery in individual DNA sequence that permit randomization of selected population. In this study, linear search is not a choice but a necessary criteria because identifying motif for pig (Sus Scrofa) is critical for halal authentication.

Therefore, we proposed a modified Linear-PSO algorithm by using linear number for population initializing and next position updating. For each cycle, only one particle called 'target motif' was selected and compared with other DNA sequences for fitness calculation.

IV. MOTIF DISCVOVERY USING LINEAR-PSO

Our main objective was to search a species specific motif existed in each DNA sequence from the same species. All possible motif combination was tested and compared to each other.

In order to test all possible sets of motif in a sequence, linear search must be used to ensure that there were no missing combinations of possible motif.

Linear-PSO started with the selection of one DNA sequence as a Target Set. Target motif was extracted from the Target Set and compared with other DNA sequences.

A. Particle Representation

Particle was represented by target motif. Target motif is the possible combination of bases from the 'Target Set'. In this experiment, first DNA sequence was selected as 'Target Set'. First target motif started from the first base and continued with the second base until completion. The possible number of target motif can be calculated as follow:

$$t = (n - tm) + 1 \quad (1)$$

Where *n* is the length of the first DNA sequence and *tm* is the length of target motif. For this experiment, length of target motif is fixed to eight bases. The details target motif representation is shown in Fig. 1.

B. Fitness Calculation

For motif discovery, similarity and complexity are required to evaluate individual motif [14]. Similarity between two short sequences is very important but complexity also needs to be addressed because the need to avoid low complexity motif, for example, sequence 'TTTTT' has low complexity compared to sequence 'CTTTC'. The main objective of the fitness function is to maximize the similarity of the sequence motif found while avoiding low complexity solution.

Selected target motif was compared with the entire possible combination motif extracted from other DNA sequences. Fitness function used in this experiment is adapted from Chang [11] where exact match for each base is given 1 point and exact match for each motif is given additional points (total point multiply by 2). This will differentiate the fitness value and allow researcher to easily identify exact match. The comparison process was repeated until the entire possible target motifs from the first DNA sequence were tested. Fig. 2 shows the process of comparison.

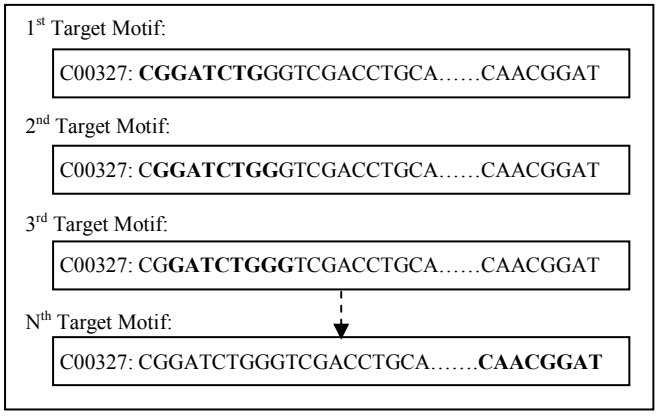

Figure 1: Target Motif representation

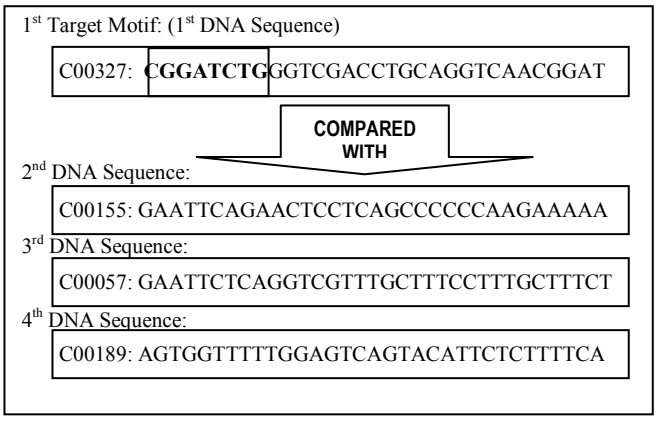

Figure 2: Comparison between Target Motif with other DNA sequences

C. Velocity

Calculation of velocity was removed from this algorithm due to the used of linear number. New particle (target motif) was selected accordingly to the next linear number. This process is repeated until the last combination of DNA bases is compared.

V. RESULTS AND DISCUSSION

The proposed Linear-PSO was executed five times using only four DNA sequences as inputs. As shown I Fig. 2, the length of target motif was set to eight bases without mutation (8, 0). Based on the first DNA sequence, there were 499 ((506 – 8) + 1) target motifs generated. The entire target motif was compared with all possible combinations of bases from 2nd until 4th DNA sequences. Table 2 shows the details number of possible combination for each sequence. Although Linear-PSO is slower, this algorithm need to be executed only once for every species with unknown motif. The discovered motif becomes a standard motif for identifying specific species such as human and pig.

Table 3 and 4 show the result of motif with the mutation parameter equal to zero and one indicating no or minor mutation. Table 3 is the result for exact motif matching because the mutation parameter was set to zero (without mutation). Only target motifs with fitness value of 16 (8 points multiply by 2) in each sequence were selected. Target motif GAAATTCC was the only target motif that exists in each DNA sequence. Table 4 shows the output when mutation is allowed. Number of motif existed in each DNA sequence increases when frequency of mutation increases.

TABLE 2: POSSIBLE MOTIF

	Number of Base	Number of Possible Combination Available
2nd DNA Seq	1479	1472
3rd DNA Seq	1835	1828
4th DNA Seq	6684	6677

TABLE 3: RESULT FOR (8,0) PROBLEM

Target Motif	Ranking	2nd Seq C00155		3rd Seq C00057		4th Seq C00189		Total Fitness Value
		Frequency Exist	Highest Fitness	Frequency Exist	Highest Fitness	Frequency Exist	Highest Fitness	
GAAATTCC	1	2	16	1	16	1	16	48
GGAAATTC	2	2	16	1	16	0	0	32
CTTTTTCT	3	1	16	0	0	2	16	32
TCTTTTTC	4	0	0	1	16	1	16	32

TABLE 4: RESULT FOR (8,1) PROBLEM

Target Motif	Ranking	2nd Seq C00155		3rd Seq C00057		4th Seq C00189		Total Fitness Value
		Frequency Exist	Highest Fitness	Frequency Exist	Highest Fitness	Frequency Exist	Highest Fitness	
GAAATTCC	1	3	16	1	16	1	16	48
GGAAATTC	3	2	16	3	16	1	7	39
CTTTTTCT	4	1	16	2	7	3	16	39
TCTTTTTC	2	1	7	2	16	5	16	39

Table 4 shows that there are 3 motifs which are similar to target motif 'GAAATTCC' in 2nd DNA Sequence compared to Table 3 which is similar to only 2 motifs. This result show that frequency of similar motif increases when mutation allowed increases.

The speed of Linear-PSO is based on the number of target motif and length of the motif. Higher number of target motif is better because more possible combination of target motif can be tested thus leading to higher fitness value. Linear-PSO has higher content validity because all possible target motifs are covered during motif search.

VI. CONCLUSION

Experimental results show that the modified algorithm (Linear-PSO) consistently detect the same possible motif, whereas Random-PSO detects different possible motif for the same set of DNA sequences. This emphasizes that Linear-PSO is important because it allows all possible set of motif in a sequence to be tested thus increasing the accuracy of the discovered motif. Although the speed is slow, the validity of the result is high.

REFERENCES

[1] P. P. Pevzner and S.-H. Sze, "Combinatorial Approaches to Finding Subtle Signals in DNA Sequences," presented at International Conference on Intelligent Systems for Molecular Biology, 2000.

[2] T. L. Bailey and C. Elkan, "Fitting a Mixture Model By Expectation Maximization to Discover Motifs in Biopolymers," presented at International Conference on Intelligent Systems for Molecular Biology, 1994.

[3] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA Regulatory Motifs Within Unaligned Noncoding Sequences Clustered by Whole-genome mRNA Quantitation," *Nature Biotechnology*, vol. 16, pp. 39 - 45, 1998.

[4] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwarld, and J. C. Wootton, "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment," *Science*, vol. 262, pp. 208 - 214, 1993.

[5] X. Liu, D. L. Brutlag, and S. J. Liu, "BioProspector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-expressed Genes," presented at Pacific Symposium on Biocomputing, 2001.

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

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

[8] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, A. M. Craven, H. B. Harlow, E. W. Su, J. E. Onyia, and C. Su, "A Statistical Analysis of the TRANSFAC Database," *Biosystems*, vol. 81, pp. 37 - 54, 2005.

[9] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, "TRANSFAC and its Module TRANSCompel: Transcriptional Gene Regulation in Eukaryotes," *Nucleic Acids Research*, vol. 34, pp. 108 - 110, 2006.

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

[11] B. C. H. Chang, A. Ratnaweera, and S. K. Halgamuge, "Particle Swarm Optimisation for Protein Motif Discovery," *Genetic Programming and Evolvable Machines*, vol. 5, pp. 203 - 214, 2004.

[12] 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 Coexpressed Genes," *International Journal of Information Technology*, vol. 11, 2005.

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

[14] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, H. B. Harlow, J. E. Onyia, and C. Su, "Discovery of Sequence Motif Related to Coexpression of Genes Using Evolutionary Computation," *Nucleic Acids Research*, vol. 32, pp. 3826 - 3835, 2004.