Comparative Analysis of Similarity Check Mechanism for Motif Extraction

A. Makolo & A.O. Osofisan
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
amakolo@ui.edu.ng, aosofisan@ui.edu.ng

E. Adebiyi
Department of Computer and Information Sciences
Covenant University
Ota, Nigeria
ezekiel.adebiyi@covenantuniversity.edu

ABSTRACT
In this work, a comparative analysis of the similarity check mechanism used in the most effective algorithm for mining simple motifs GEMS (Gene Enrichment Motif Searching) and that used in a popular multi-objective genetic algorithm, MOGAMOD (Multi-Objective Genetic Algorithm for Motif Discovery) was done. In our previous work, we had reported the implementation of GEMS on suffix tree –Suffix Tree Gene Enrichment Motif Searching (STGEMS) and shown the linear asymptotic runtime achieved. Here, we attempt to empirically proof the high sensitivity of the resulting algorithm, STGEMS in mining motifs from challenging sequences like we have in *Plasmodium falciparum*. The results obtained validates the high sensitivity of the similarity check mechanism employed in GEMS and also shows that a careful deployment of this mechanism in the multi-objective genetic algorithm, improved the sensitivity level of the resulting algorithm. The end results gave us room to exhaustively mine structured motifs.

Keywords: Motifs, GEMS, MOGAMOD, STGEMS, Suffix tree.

1. INTRODUCTION

Sequence motif discovery is one of the fundamental concerns in Bioinformatics and it has important applications in locating regulatory sites and drug target identification. The extraction of structured motifs (i.e. several words with well defined gaps) is particularly interesting because of its application to the detection of binding sites.

African Journal of Computing & ICT Reference Format:

A. Makolo, A.O. Osifisan & E. Adebiyi (2012) - Comparative Analysis of Similarity Check Mechanism for Motif Extraction. Afr J. of Comp & ICTs. Vol 5, No.1 pp 53-58

- ISSN 2006-1781

This binding sites respect distance constraints. In this paper we consider the extraction of structured motifs from the deadly organism, the malaria parasite, *P. falciparum*. The highly repetitive and specific alphabet (in this case AT) bias sequences of *P. falciparum* makes the task of extracting structured motifs a very challenging one. The GEMS algorithm by [8] has been shown to successfully mine simple motifs in the *P. falciparum* kind of sequences. This success is attributed mainly to the unique approach employed in GEMS similarity check mechanism. It involves the use of hypergeometric-based scoring function to compute the p-value of the candidate motifs, ranking them according to this value and computing the position weight matrix as a neighborhood in the sequence space.

Long distance metric can also be used to which considered positional information to merge non-

unique motif candidates. Position-Specific Scoring Matrices (PSSMs) and their derivatives (i.e. position frequency matrix, position weight matrix) have become the standard representation of a transcription factor's DNA-binding preference. For example, experimentally derived DNA-binding preferences for a growing number of transcription factors are stored as frequency matrices in databases such as JASPAR [6] and TRANSFAC [5]. In addition, most de novo motif-finding software tools report statistically over-represented degenerate sequence features in the form of frequency matrices or consensus sequences. [7]. PSSMs have been used in this work as a representative model for the extracted candidate motifs.

Similarity check is a measure of the degree of closeness of two strings. This is useful in computational biology where two slightly different patterns can represent the same motif due to the presence of a number of mismatches allowed. For instance motifs AAAATGC, AACATGC, AAATTGC are similar motifs with one mismatch.

MOGAMOD algorithm by [3] used a well-known high-performance multi-objective genetic algorithm called NSGA II.[1] to find a large number of tradeoff motifs with respect to conflicting objectives of similarity, motif length and support maximization. MOGAMOD's similarity check involves performing a measure of similarity among all motif instances defining a candidate motif. This is achieved by first generating a position weight matrix from the motif patterns found in every sequence. Then, the dominance value of the dominant nucleotide is computed from the position weight matrix which forms the basis of the similarity function.

In [4], we introduced the STGEMS algorithm which implemented the GEMS algorithm on the suffix tree and also demonstrated the improved run time of STGEMS over GEMS algorithm. This present work is an extension of that work; here we present the result of the comparative analysis of the similarity check mechanism used in GEMS and MOGAMOD after implementing them in C programming language on Linux platform.

The structure of this paper is as follows. In section 2, we discuss the technical details of GEMS and MOGAMOD's similarity check. In section 3, we show the experimental experience of running the algorithm on the data of interest and a discussion of the result and we conclude the paper in section 4.

2.0 GEMS AND MOGAMOD SIMILARITY CHECK MECHANISM

In [8] MOGAMOD algorithm was introduced using multi-objective genetic algorithm and it was used to discover optimal motifs in sequential data. Multi-objective optimization involves having a solution which is a family of pareto-optimal set or nondominated solutions . He converted the optimal motif discovery problem into three conflicting optimization problem which is to maximize similarity, motif length and support for candidate motifs, thus obtaining a large number of optimal motifs by a single run of the algorithm. The implementation of MOGAMOD which was based on a well known high performance multi-objective genetic algorithm called NSGA II [1], which is a global multi-objective optimization problem solver can be applied to any field of optimization with conflicting objectives. NSGA II is unique in that unlike other optimization solvers which convert multiple objectives into a single one by using some subjective preference information, NSGA II is capable of finding a well distributed set of trade-off optimal solutions for two or more conflicting objectives of design. MOGAMOD was compared with three well-known motif discovery methods AlignACE, MEME and Weeder using yeast data from TRANSFAC. The result showed that MOGAMOD outperformed them in terms of accuracy and runtime.

The similarity check used in MOGAMOD measures similarity among all motif instances defining an individual solution. To compute the similarity, it first generates a position weight matrix from the motif patterns found in every sequence. Then, the dominance value (dv) of the dominant nucleotide in each column is found using the formula:

$$dv(i) = \max_b\{f(b,i)\} \text{ , } i = 1,\ldots., l$$

Where *f(b, i)* is the score of the nucleotide *b* on column *i* in the position weight matrix,
dv(i) is the dominance value of the dominant nucleotide on column *i*,
and *l* is motif length.

The similarity objective function of motif *M* is the average of the dominance values of all columns in the position weight matrix.

i. e $Similarity(M) = \sum_{i=1}^{i} \mathrm{dv(i)}/l$

The likelihood of the candidate motif been discovered as a real motif depends on the value of the similarity score. In other words, the closer the value of the similarity *M* is to one, the greater the probability that the candidate motif *M* will be discovered as an optimal motif.

Figure 1 below depicts the steps involved in the similarity check of MOGAMOD

The motif discovery tool by [8] Gene Enrichment Motif Searching (GEMS) used hyper geometric-based scoring function (to calculate p-values for position weight matrices (PWMs)) and a position weight matrix optimization routine to identify with a high degree of accuracy simple motifs in the nucleotide-biased and repeat sequence rich genome of *P. falciparum* The PWMs were built from seeds with the most enriched candidates i.e. those with lowest p-values, while identifying all sequences with one mismatch from the seed words, then a *p*-value enrichment score is computed using a hypergeometric formula below.

$$P(X,x,Y,y) = \sum_{i=y}^{min(x,Y)} \frac{\binom{x}{i}\binom{X-x}{Y-i}}{\binom{X}{Y}} \quad \text{.....(1)}$$

Where X is the total set of genes, i.e. positive and negative set, *x* a subset of the gene of interest,

Y is the total promoter sequence that matches the genes, *y* is the subset of the promoters which fall within the cluster of interest. The smaller the p-value scores for a motif, the higher the likelihood of it being an optimal motif.

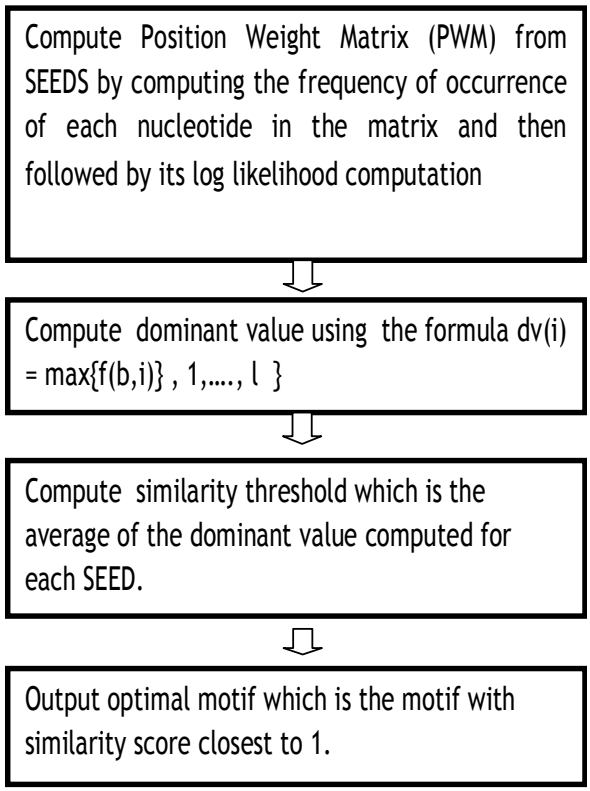

Fig 1. MOGAMOD's Similarity Check Implementation

The success of GEMS in extracting significant motifs for *P. falciparum* was based on using this hypergeometric scoring function i.e. using geometric mean to calculate similarity function and setting a threshold based on the p-value of the position weight matrix. The threshold setting was achieved by utilizing an exhaustive parameter optimization routine similar to the probability minimization protocol used in the OPI clustering algorithm of [10]. This threshold helps to determine how similar any given sequence in a promoter region must be to the PWM to be considered an actual motif. In addition, GEMS merged non-unique motif candidates using a distance metric. This approach is better for *P. falciparum* which has highly repetitive sequences present.

The approach canceled out these repetitions where others motif discovery tools especially those using background modeling approach

would identify the same repetitive sequence as potential motif because they vary significantly from the background estimation.

The details of the algorithm are encapsulated in figure 2.

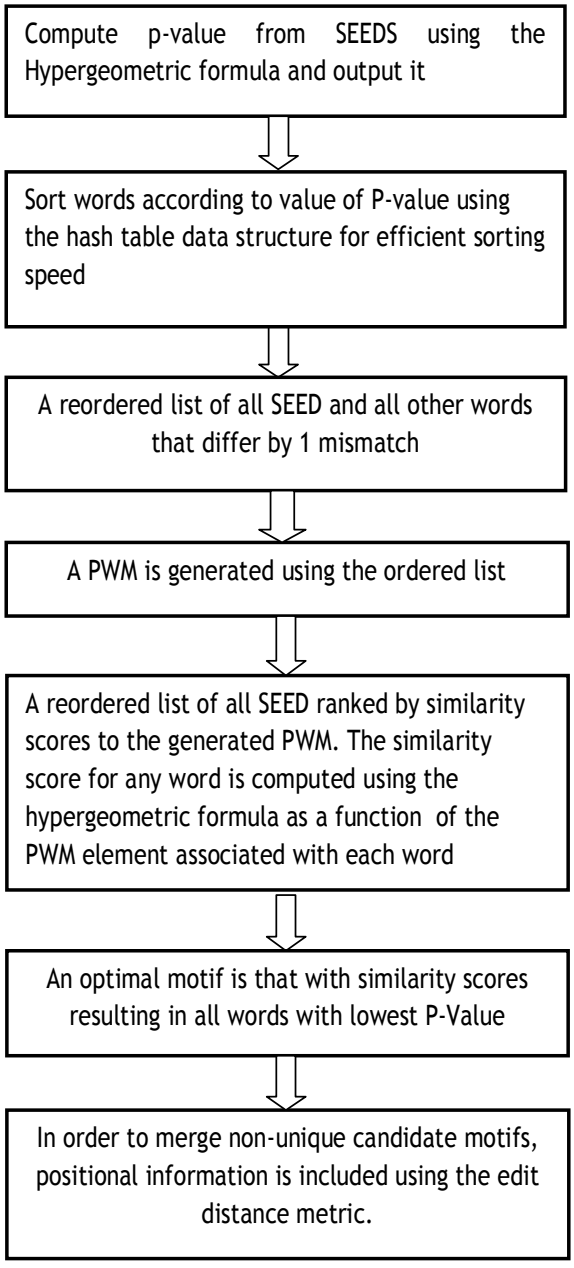

Figure 2: Similarity Check Implementation of GEMS

3.0 EXPERIMENTAL EXPERIENCE AND DISCUSSION OF RESULTS

In this section, we show the effectiveness (in mining binding sites (motifs) of our approach. To this effect, two sets of sample genes in *P.falciparum*, which have been experimentally proven to co-regulate via structured motifs were used for testing. The implementation and testing of algorithm was done in C.

In our implementation of gene enrichment searching, we used the hash table to store the extracted SEED while sorting it according to the p-values computed. The hash table data structure is a desirable choice because of the speed advantage in sorting large data sizes.

The first experiment used the set of genes in the work of [2] which experimentally extracted regulatory elements for P.*falciparum*. i.e 100 base pairs upstream of gene start codons as shown in table 1. Table 2 shows the result obtained running the data set on the two implementations, i.e. the first used GEMS's similarity check while the second used the similarity check used in MOGAMOD. The second experiment used the set of genes used by [9] which identified transcription factors in the mosquito-invasive stage of malaria parasite shown in table 3. The resulting output using the two implementations are depicted in table 4.

From [2], experimentally, the set of genes in table 1 co-regulate using the following motif: N(C/G/A)TGCA-4to5-(A/G/C)GTGC(A/G). 'N' indicates any of the four nucleotides A/C/G/T can occur at this position, while four to five gaps are between the two boxes. In [9], it was also experimentally shown that TAGCTA-100to1500-TAGCCA and TAGCTA-100to1500-TGGCTA are the structured motifs used in their co-regulation.

Table1. Set of genes from Plasmodium falciparium Intraerythrocytic stage

Accession Number	Description
PFF0645c	*Plasmodium falciparum* 3D7 , integral membrane protein, putative
PFI0265c	*Plasmodium falciparum* 3D7, high molecular weight rhoptry protein
PFE0075c	*Plasmodium falciparum* 3D7, rhoptry-associated protein 3
PFE0080c	*Plasmodium falciparum* 3D7, rhoptry-associated protein 2
PFC0120w	*Plasmodium falciparum* 3D7 , cytoadherence linked asexual protein 3.1
MAL7P1.208	*Plasmodium falciparum* 3D7, rhoptry-associated membrane antigen
PFI1730w	*Plasmodium falciparum* 3D7, cytoadherence linked asexual protein 9
PF14_0102	*Plasmodium falciparum* 3D7, rhoptry-associated protein 1
PFD0295c	*Plasmodium falciparum* 3D7 , apical sushi protein
MAL7P1.119	*Plasmodium falciparum* 3D7 , rhoptry-associated leucine zipper-like protein 1
PFI1445w	*Plasmodium falciparum* 3D7, high molecular weight rhoptry protein 2

Table 2. Output from running the two algorithms on the DNA sequences of the above genes

	IDENTIFIED BY	
Consensus	**GEM's Similarity Check**	**MOGAMOD's similarity Check**
GGTGCG	NO	NO
CGTGCG	NO	NO
CTGCA	YES	NO
GTGCA	YES	NO
ATGCA	YES	NO
AGTGCG	YES	NO

Table 3. Set of genes from the Mosquito invasive stage of malaria parasite.

Accession Number	Description
PF08_0136b	*Plasmodium falciparum* 3D7 , von Willebrand factor A-domain related protein
PFC0905c	*Plasmodium falciparum* 3D7, oocyst capsule protein
PFL0550w	*Plasmodium falciparum* 3D7, HSP20-like chaperone
PFC0640w	*Plasmodium falciparum* 3D7,CSP and TRAP-related protein
PFD0425w	*Plasmodium falciparum* 3D7 , sporozoite invasion-associated protein 1, putative
PF08_0030	*Plasmodium falciparum* 3D7, conserved Plasmodium protein, unknown function
PFL2135c	*Plasmodium falciparum* 3D7, conserved Plasmodium protein, unknown function
MAL13P1.203	*Plasmodium falciparum* 3D7 , secreted ookinete protein, putative
PF10_0027	Plasmodium falciparum 3D7,conserved Plasmodium protein, unknown function
PFL2510w	*Plasmodium falciparum* 3D7, chitinase
PF13_0355	*Plasmodium falciparum* 3D7, secreted ookinete protein
PFD0435c	*Plasmodium falciparum* 3D7, conserved Plasmodium protein
PFE0360c	*Plasmodium falciparum* 3D7, conserved Plasmodium protein
PF14_0040	*Plasmodium falciparum* 3D7, secreted ookinete adhesive protein
PFF0975c	*Plasmodium falciparum* 3D7, conserved Plasmodium protein
PF10_0302	*Plasmodium falciparum* 3D7, 28 kDa ookinete surface protein
PF10_0303	*Plasmodium falciparum* 3D7, 25 kDa ookinete surface antigen precursor
PFC0420w	*Plasmodium falciparum* 3D7, calcium dependent protein kinase 3
PFI1145w	*Plasmodium falciparum* 3D7, perforin like protein 3

Table 4. Output from running the two algorithms on the DNA sequences of the above genes

	IDENTIFIED BY	
Consensus	**GEM's Similarity Check**	**MOGAMOD's similarity Check**
TAGCTA	NO	NO
TGGCTA	NO	NO
TAGCCA	NO	NO

From table 2 above, we observed that the experimentally extracted motif was also mined by GEMS's similarity check but not by MOGAMOD's. However, in table 4, none of the experimentally extracted motifs was found by GEMS's similarity or by MOGAMOD's. This inability to mine the motifs in table 4 makes it obvious that a number of fine tunings, which are not necessarily algorithmic, are needed to effectively mine the desired structured motifs in the set of table 3.

A novel algorithm, STGEMS which incorporated the similarity mechanism used in GEMS has been shown to be more effective when compare to MOGAMOD's similarity check. This is because of the hypergeometric scoring function used in GEMS's similarity check mechanism which made it successful in discovering motifs from the challenging highly repetitive elements and base bias sequence of the malaria parasite, *Plasmodium falciparum*.

4.0 CONCLUSION AND FUTURE LEADS

A comparative analysis of the similarity mechanism of two popular motif discovery algorithms was achieved in this work. The result shows the superiority of the similarity check of GEMS over that of MOGAMOD especially when discovering motifs from organisms with some peculiarities in their genomic sequences such as the malaria parasite, *Plasmodium falciparum*. A possible future work would be to formalize the fine tunings required to effectively extract biologically motivated motifs as indicated in the observation made in session 3 above.

References

[1] Deb, K et al., "A fast and elitist multi-objective genetic algorithm II." IEEE Transactions on Evolutionary Computation, 6, 182–197, 2002

[2] Flueck C, Bartfai, R, Niederwieser, I, Witmer, K, Alako, B, Moes, S, Bozdech, Z, Jenoe,P, Stunnenberg, H, Voss T., "A major Role for the Witmer, K, Alako, B, Moes, S, Bozdech, Z, Jenoe,P, Stunnenberg, H, Voss T., "A major Role for the Plasmodium *falciparum* ApiAP2 Protein PfSIP2 in Chromosome End Biology", PLoS Pathog 6(2): e1000784, 2010.

[3] Kaya, M., "MOGAMOD: Multi-Objective Genetic Algorithm for Motif Discovery", Expert Systems with Applications, 36 (2): 1039-1947, 2009.

[4] Makolo, A, Adebiyi E and Osofisan A., "STGEMS: Mining Structured Motifs with Gene Enrichment Motif Searching on Suffix tree", Journal of Computer Science and its Applications 18(1) : 79-91, 2011.

[5] Matys V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. et al.,"TRANSFAC: transcriptional regulation, from patterns to profiles", Nucleic Acids Res., 32,D81–D96, 2003.

[6] Sandelin, A., Alkema,W., Engstrom ,P., Wasserman,W.W. and Lenhard,B. "JASPAR: an open-access database for eukaryotic transcription factor binding profiles". Nucleic Acids Res., 32,D91–D94, 2003

[7] Tompa. M., "An Exact Method for Finding Short Motifs in Sequences with Application to the Ribosome Binding Site Problem", 7th Intl. Conf. Intelligent Systems for Molecular Biology, Heidelberg, Germany, Aug 1999 10-12, 2003.

[8] Young J, Johnson, J, Benner, C, Yan, F, Chen, K, Roch, K, Zhou, Y, Winzeler, E., "In silico discovery of transcription regulatory elements in Plasmodium falciparum", BMC Genomics ,9:70, 2008.

[9] Yuda, M, Iwanaga, S, Shigenubu, S, Mair, G, Janse, C, Waters, A, Kato, T, Kaneko, I., "Identification of a transcription factor in the mosquito-invasive stage of malaria parasite. Molecular Microbilogy, 71, 1402-1414, 2009.

[10] Zhou Y, Young JA, Santrosyan A, Chen K, Yan SF, Winzeler EA. "In silico gene function prediction using ontology-based pattern identification". *Bioinformatics*, 21(7):1237-1245,2005