Algorithm based on the Ant Colony Optimization for the DNA motif fusion

JUNIOR ALTAMIRANDA, JOSE AGUILAR, RAFAEL TORRES
CEMISID, Department of Computer Science
University of Los Andes
Faculty of Engineering, Campus La Hechicera, Mérida
VENEZUELA
altamira@ula.ve, aguilar@ula.ve

CHRISTIAN DELAMARCHE
Structure et Dynamique des Macromolécules
University of Rennes I
Campus de Beaulieu, Nb 13, Rennes
FRANCE
christian.delamarche@univ-rennes1.fr

Abstract: - Motifs (patterns, signatures, domains) are useful to determine nucleotides/amino-acids that are likely involved in structures, functions, regulations and evolutions, or to infer homology between genes/proteins. The main objective of this paper is the fusion of motifs. Our task is to analyze a set of possible motifs and to detect if similarity exists between them, to construct a general motif. The motifs fusion method is based on the algorithm of combinatorial optimization called Artificial Ants System. This method uses the nucleotides of the first motif to construct the graph where the ants will walk. Then, the graph is crossed by the ants according to the path of the second motif, using a transition function that promoves to flow the path between similar nucleotides. The ants when walking leave pheromone in the nodes, in a way that at the end several have a lot of or little pheromone. Finally the graph is crossed again to construct the resultant motif composed by the nodes with much pheromone.

Key-Words: - Ant Colony Optimization, Motif, DNA, Bioinformatics

1 Introduction

This paper defines and develops a computational method for the fusion of DNA motifs. We propose an algorithm based on ACO [1], [2], with some modifications. This algorithm can efficiently find the union between two motifs and allows the generation of a new motif.

Currently, there are several methods of patterns discovery (using Regular Expressions, Hidden Markov Model (HMM), Automata, and PSSM Matrix). The regular expressions are the most commonly used by biologists, as well as the graphical method of LOGOS, since visually are simpler to understand and interpret for them [3], [4]. To discover of DNA motif historically has been used the Pratt method [5], which is based on the algorithm Knuth-Morris-Pratt [6], but there are other tools, between the most well-known we have [7], [8], [9], [10], [11], [12]: TEIRESIAS, MEME. The discovery of common motifs between sequences that are distant in evolutionary level (non-homologous or non-related sequences) is a very complex problem. In addition, there are tools that allow comparing DNA motifs and SLM (Short Linear Motifs) defined as regular expressions, such as CompariMotif [13], FunClust [14], and Bio.motif [15]. However, these tools do not allow fusing them into a common expression.

Specifically, our task is to analyze a set of DNA motifs stored in a database, detect if there are similarities between them, and construct general patterns. The patterns found can be explained by the existence of segments that have been preserved during the natural evolution of proteins, and suggest that the obtained regions play a functional role in their mechanisms and structure.

On general, finding the common motif to a set of motifs is a problem. Most of the algorithms of motifs search use heuristic techniques to obtain near optimal solutions with a relatively low computational cost [3]. For example, some works based on bio-inspired algorithms are: in [16] presents an approach based on the Ants Colony Optimization method combined with a Max-Min strategy for DNA sequences. In [17] implements a

ISBN: 978-1-61804-106-7

method based on the Ant Colony Optimization algorithm and the expectation maximization (EM) to find DNA motifs (specifically, for collections of TFBSs) in a set of bio-sequences.

2 Theoretical Framework

2.1 Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) is a type of metaheuristic whose philosophy is inspired by the behavior of real ants searching foods [1], [2]. The main aspect of ACO is the transition probability of an ant walking in the graph. This one is defined by (see Eq. 1):

$$P_r^k = \begin{cases} \dfrac{[\tau_r]^\theta [\eta_r]^\beta}{\sum_{u \in N_k(i)} [\tau_u]^\theta [\eta_u]^\beta} & \text{If } r \in N_k(i) \\ 0 & \text{in other case} \end{cases} \quad (1)$$

Where τ_r is the quantity of pheromone track in the node r of the graph, η_r is the visibility of the node r (frequently is $1/d_{ir}$, where d_{ir} is the distance that exists between the current node i and the node r), $N_k(i)$ is the neighborhood of the ant k when it is in the node i, θ and β are two parameters that consider the relative importance between the pheromone tracks and the visibility.

Additionally, ACO uses a reinforcement learning mechanism to update the pheromone on the graph. The pheromone update can be carried out once all the ants have completed its solutions, (see Eq. 2):

$$\tau_r(t+1) = \alpha \tau_r(t) + \Delta \tau_r \quad (2)$$

Where: $\tau_r(t)$ is the intensity of the trace deposited on node r at time t; α is a coefficient such that $(1 - \alpha)$ represents the evaporation rate of the pheromone between the time t and t +1. $\Delta\tau_r$ is the pheromone quantity let in a node r for the k^{th} ant between the interval t and t+1 (see Eq. 3).

$$\Delta \tau_r = \sum_{k=1}^{m} \Delta \tau_r^k \quad (3)$$

ACO has demonstrated its effectiveness in the resolution of different combinatory optimization problems considered difficult [2], [18].

2.2 Motif

A Motif is a region or portion of a protein sequence that has a specific structure and is functionally significant. Protein families are often characterized by one or more such motifs. Detection of motifs in proteins is an important problem since the motifs carry out and regulate various functions, and the presence of specific motifs may help to classify a protein [19].

A motif, in the context of biological sequence analysis, is a consensus pattern of DNA bases or amino acids which accurately captures a conserved feature common to a group of DNA or protein sequences. DNA motifs are sometimes termed signals: examples are regulatory sequences, scaffold attachment sites, and messenger RNA splice sites. Examples of protein motifs, which are also known as fingerprints, include enzyme active sites, structural domains, and cellular localization tags. Motif discovery is the act of identifying and characterizing motifs, and underlies a number of important biomedical activities. For example: the identification of regulatory signals has applications for gene finding in sequenced genomes, understanding of regulatory networks, and the design of drugs for regulating specific genes; and protein motifs are routinely used to identify the function of newly-sequenced genes and to understand the basis of a protein's cellular function [20].

3 Problem Formulation

The problem consists to construct a common motif for the motifs that have a high degree of similarity. We developed a method to fusion of similar motifs. We used the Ant Colony Optimization to construct a graph with the nucleotides of the first motif. Then, the graph is crossed by the ants according to the path of the second motif. Finally the graph is crossed again to construct the resultant regular expression. In each execution of our algorithm, two motifs are fused. In general, the macro-algorithm for the fusion process is:

1. Create the route graph.
2. Walk of the ants on the route graph.
3. Choose the best nodes
4. Construct the resultant motif

3.1 Create the route graph

Because the problem of motif fusion emerges from the study of the primary structure of DNA molecules, which is a linear structure consisting of nucleotides, there are two basic conditions for the design of the graph where will walk the ants:

The first stems from an analysis in the construction of motifs, which shows that is essential for this task the position of different nucleotides along of the chains that can be viewed as one-dimensional arrays. In the second, we establish that the product of the motif fusion must generate a new

motif that contains the nucleotides chains that belong to motifs fused.

For the previous reasons, our graph will be represented in the plane, and each node will have arcs at the right and left sides, in this way the ants can only move them in horizontal direction. The nodes must store the pheromone level deposited by the ants that visit them and the nucleotide that represent (see Fig. 1). This information will be constituted by the type of nucleotide that represents (A, C, G, T) or an identifier for special nodes (see Table I).

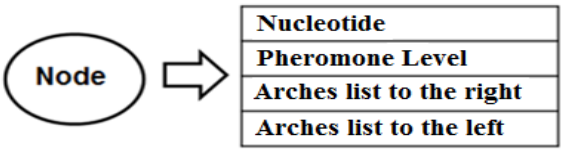

Fig. 1. Data strusture of a node

TABLE I. IDENTIFIER FOR SPECIAL NODES

Information	Special Identifier
Gap	X
Start	Start
End	End

For the graph construction we transform the first motif in a stack data structure (for example, see the motif TGAGCA in Fig. 3).

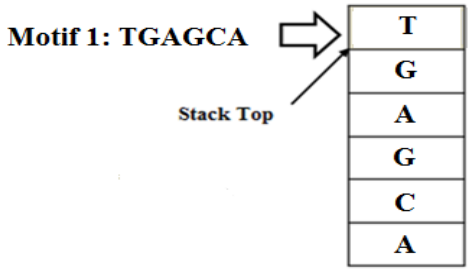

Fig 2. Transformation of a motif 1 in a stack

Additionally, two nodes are defined that serve as guide for the construction of the graph, to indicate the start and the end of the route (Fig. 3). Then, we proceed to extract the elements that are at the top of the stack iteratively, and built the nodes in the graph (nucleotides) which are in the same position in the chain. Also adds a node gap, which will serve as an auxiliary route for cases in which the ants must not continue for any of the available nodes. In this way, we avoid that an ant stops itself. Finally, when the stack is empty we stop the construction of the graph.

In our approach, we build the route graph using the first motif to fuse (motif1) (see Fig 3).

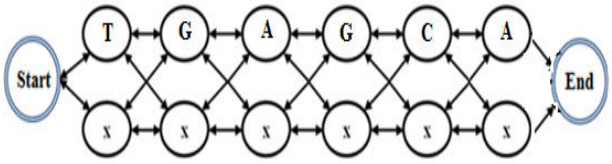

Fig.3 Route graph of the motif 1

3.2 Walk of the ants on the route graph

The artificial ant's colony, as in natural ant colonies, evolves by the actions performed by its members. This way, the route graph is walked by the N-ants that constitute the colony. So, it is necessary to define the number of individuals of the colony, before they begin to walk on the route graph. In our case, each ant has a route map defined by the second motif to fuse. We define an ant type data structure composed of 6 elements, whose characteristics are described in Table II. It contains the needed information for the ants to walk on the route graph.

TABLE II. ELEMENTS OF THE ANT DATA STRUCTURE

Element	Characteristics
Start node	Address of the node where start the ant to walk the route graph
Route map	Stack that contains the motif that must follow the ant, and serves to know that nodes should be visited by the ant in the route graph or not.
Pheromone increase coefficient	Real number (0,1), it's used to establish the pheromone concentration that deposits the ants in each visited node of the route graph.
Equalities similarity index	Integer number [0.10], it determines the pheromone level deposited by the ant, when the node found in the graph is identical to the expected to the route map.
Differences similarity index	Integer number [0.10], it determines the pheromone level deposited by the ant, when the nucleotide found in the route graph is not equal to the route map
Gaps similarity index	Integer number [0.10], it serves to mark the selected node, if node type is a Gap.

We use the second motif (motif2) to construct the route map of the ants, transforming the motif in a stack data structure.

At the start, the ant is placed in the initial node of the route graph, and with the route map it observes the contiguous nodes at the right side (see Fig. 4).

ISBN: 978-1-61804-106-7

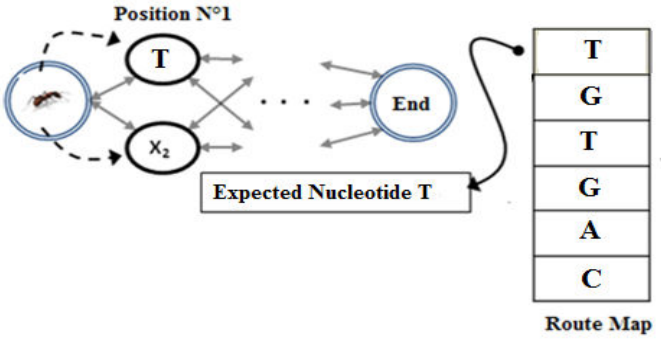

Fig 4. Ant n the initial node of the route graph.

The ant executes the function of transition to each one of the nodes that can visit in the next position. This function consists of two phases; the first phase calculates the probability to visit each contiguous node ($P_n^k(r)$) (see Eq. 4) (based on its pheromone level 'τ_r' and the index of similarity `$\varphi_r$' of each node (**'r'** indicates one of the neighbouring nodes in the position `k', and `n' is the number of neighbouring nodes at the right side for that position `k'):

$$P_n^k(r) = \begin{cases} \dfrac{\tau_r * \varphi_r}{\sum_{i=1}^{n} \tau_r * \varphi_r} & \text{si } n > 1 \\ 1 & \text{si } n = 1 \end{cases} \quad (4)$$

The second phase decides the node to visit using the simulation of Monte Carlo. When the ant moves to a node, it deposits pheromone that increases the pheromone concentration in the node. The quantity of deposited pheromone depend on the similarity index with respect to the nucleotide waited according to the route map (see Eq. 5)

$$\tau_r^k = \tau_r^k + \sigma * \varphi_r^k \quad (5)$$

The similarity index is defined as follows: if the nucleotide of the route graph is equal to the nucleotide of the route map of the ant, then we use the equalities similarity index; otherwise if the visited node contain gap, then the gaps index is used; otherwise, is used the differences similarity Index. In our example, the final route of an ant is observed in the Fig. 5.

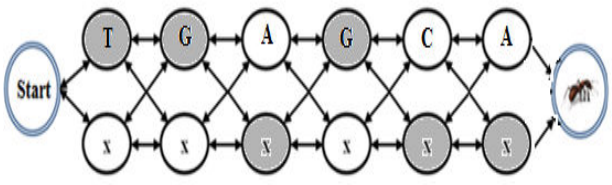

Fig. 5. The final route of the ant

For a colony, the previous process is repeated for each ant. Additionally, the same process is executed recursively until the number of colony cycles desired. At the end of a cycle, there is an evaporate pheromone traces, decrementing the pheromone levels of all nodes in the graph (see Eq. 6), where "ρ" is the pheromone evaporation coefficient.

$$\tau_r^k = (1-\rho) * \tau_r^k \quad (6)$$

3.3 Choose the best nodes

Once the colony has completed its work, we delete the arcs that lead to those nodes with a pheromone level below the pheromone threshold that the user has defined (for our example, we fix the pheromone threshold to 1.0), which help to preselect to the nucleotides that contribute to the best solutions. Fig. 6 shows the selected nodes because they exceeded the threshold (in blue).

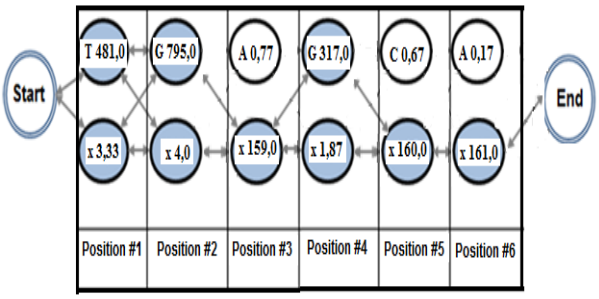

Fig. 6. Route graph with the pheromone levels of each node

3.4 Construct the resultant motif

Finally, the selected route graph is filtered to delete irrelevant information and to define the resulting sequence. To make this task, we have to analyze the marked nodes of the graph and insert the nucleotides selected in a list of chains that will contain the value corresponding to each motif position. To achieve this goal the following criteria are used:

1. If in the position one node (nucleotide or gap) which has passed the pheromone threshold exists, it will be inserted in the list.
2. If more than one node in the same position (a nucleotide and a gap) that has passed the pheromone threshold exist, the following conditions are applied:
   a) If the level of pheromone of the nucleotide node is superior we insert the nucleotide in the list.
   b) In other case we insert the gap in the list.

ISBN: 978-1-61804-106-7

We take the list that contains the nucleotide corresponding to each position in order to construct the resultant motif (see Fig. 7).

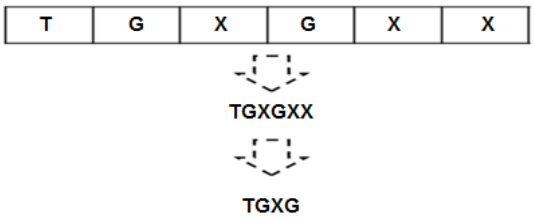

Fig. 7. Resultant fusion motif

4 Experiments of Fusion of Motifs

To run the system, it is necessary to adjust a set of parameters. Because the number of adjustable parameters in the developed system is extensive, some values for the tests were fixed with a default value (see Table III). The only parameters that we have varied are the parameters that determine the collective behaviour: the cycle's number and the ant's number. By this way, the solution depends fundamentally on the behaviour of the colony.

TABLE III. PARAMETER LIST

System Parameters	Value
Pheromone increase coefficient	0,1
Similarity index for the nucleotides that are the same	10
Similarity index for the nucleotides that are different	1
Similarity index for gaps	3
Pheromone initial level on the graph nodes	1,0
Pheromone evaporation coefficient	0,05

To a qualitative comparison of our method with previous work, we carry out experiments on real datasets previously constructed in [21] (see the results in Table IV). In this case we compare S1 with S2, the resultant motif with S3, and so on.

TABLE IV. MOTIFS FUSION

Sequences	[8]	Our approach
S1:ATCATCCGTGTAGCTCAAAA **S2**:ATCATCCGTGTAGCTCAAAA	ATCATCCGTGTAGCTCAAAA	ATCATCCGTGTAGCTCAAAA
S3:AGATCCGTAACGAAGTTTAC	ATCCGT	AxxATCCGTxxxGxxxxxxA
S4:CCCCATCCGTAATTACCTAT	ATCCGT	xxxxATCCGTxxxxxxxxxx

The subsequence ATCCGT is the consensus sequence. This study suggests that the results provided by our system are similar to the results that are found in [21], with the additional advantage that our system does not require the use of post – processing.

We carry out a second qualitative comparison with real datasets of the Escherichia coli sequences (they have two highly conserved parts, called the -35 and -10 regions) [22]. The fusion of a set of these sequences is shown in Table V. In [22] is not presented the consensus motif of each fusion. In our case, we fuse the motif resulting of the two previous rows with the following until the end.

TABLE V. RESULTS OF THE FUSION OF SEQUENCES OF ESCHERICHIA COLI

Sequence	Fusion Sequences
Bgl R mut : A A C T G T G A G C A T G G T C A T A T T T Bgl R mut : A A C T G T G A G C A T G G T C A T A T T T	RS : A A C T G T G A G C A T G G T C A T A T T T
RS : A A C T G T G A G C A T G G T C A T A T T T Deo P2 site 1 : A A T T G T G A T G T G T A T C G A A G T G	RS : A(2)-x-T-G-T-G-A-x(6)-T-C-x(2)-A-x-T-x
RS : A(2)-x-T-G-T-G-A-x(6)-T-C-x(2)-A-x-T-x Lac site 1: T A A T G T G A G T T A G C T C A C T C A T	RS : x-A-x-T-G-T-G-A-x(6)-T-C-x(6)
RS : x-A-x-T-G-T-G-A-x(6)-T-C-x(6) Lac site 2: A A T T G T G A G C G G A T A A C A A T T T	RS: x-A-x-T-G-T-G-A-x(14)
RS : x-A-x-T-G-T-G-A-x(14) Mal k: T T C T G T G A A C T A A A C C G A G G T C	RS: x(3)-T-G-T-G-A-x(14)
RS: x(3)-T-G-T-G-A-x(14) Mal T: A A T T G T G A C A C A G T G C A A A T T C	RS: x(3)-T-G-T-G-A-x(14)
RS: x(3)-T-G-T-G-A-x(14) Tna A: G A T T G T G A T T C G A T T C A C A T T T	RS: x(3)-T-G-T-G-A-x(14)
RS: x(3)-T-G-T-G-A-x(14) Uxu AB: T G T T G T G A T G T G G T T A A C C C A A	RS: x(3)-T-G-T-G-A-x(14)
RS: x(3)-T-G-T-G-A-x(14) pBR P4: C G G T G T G A A A T A C C G C A C A G A T	RS: x(3)-T-G-T-G-A-x(14)
RS: x(3)-T-G-T-G-A-x(14) Cat site 2: A C C T G T G A C G G A A G A T C A C T T C	RS: x(3)-T-G-T-G-A-x(14)
RS: x(3)-T-G-T-G-A-x(14) Tdc: A T T T G T G A G T G G T C G C A C A T A T	RS: x(3)-T-G-T-G-A-x(14)

According to [22] the consensus sequences are TTGACA and TATAAT. In our case, the consensus sequence is TGTGA. In contrast with [29], we obtained a consensus sequence for all sequences in the Table VI. Our system features well-conserved positions, in [22] it is unclear what positions are absolutely conserved, and the consensus sequences

ISBN: 978-1-61804-106-7

presented do not found within the sequences shown in Table V.

5 Conclusion

We propose a motif for the construction of the route graph and other motif defines the route map that the ants use to walk. In addition, the ants execute the transition function to each one of the nodes that it can visit in the next position using the similarity index between the nodes of the route map and of the route graph. This approach has very good results for ADN sequences, we are going to test this approach in proteins (which are composed of sequences of more than four nucleotides), this is not possible for the approaches [16, 17].

.

References:

[1] Dorigo M., Birattari M., Stutzle T. "Ant Colony Optimization". *Computational Intelligence Magazine*, IEEE. 28-39, 2006

[2] Aguilar J., Rivas F. (Ed). *Introducción a la Computación Inteligente,* MERITEC, Venezuela, 2001

[3] *Sandve G., Drablos F. A survey of motif discovery methods in an integrated framework, Biology Direct,* 1(11), 2006.

[4] Habib N., Kaplan T., MArgalit H., Friedman N. A novel Bayesian DNA Mofit Comparison Method for clustering and retrieval, Plos *Comput. Biol,* 4(2), pp. 1-17, 2008

[5] Pratt Pattern Matching. *Available in:* http://www.ebi.ac.uk/Tools/pratt/.

[6] Gusfield D., *Computer Science and Computational Biology*, Press University of Cambridge, 1999.

[7] Teiresias. *Available in :* http://cbcsrv.watson.ibm.com/Tspd.html.

[8] Meme. *Available in:* http://meme.sdsc.edu/meme/doc/examples/meme_example_output_files/meme.html.

[9] Bailey TL., Boden M., Buske FA., Frith M., Grant CE., Clementi L., Ren J., Li WW., Noble Ws."MEME Suite: tools for motif discovery and searching", *Nucleic Acids Research.,* 37, pp. W202-W208, 2009.

[10] Dogruel M., Down T., Hubbard T. "NestedMICA as an ab initio protein motif discovery tool,. *BMC Bioinformatics*, 9(19), pp 1-12. 2008.

[11] Corne D., Meade A., Sibly R."Evolving core promoter signal motifs", *Proc. 2001 Congress on Evolutionary Computation,* pp. 1162-1169. 2001

[12] Fogel G., Weekes D., Varga G., Dow E., Harlow H., Onyia J., Su C.,*"Discovery of sequence motifs related to coexpression of genes using evolutionary computation," Nucleic Acids Research 32:*3826-3835. 2004

[13] Edwards R., Davey N., Shields D. "CompariMotif: quick and easy comparisons of sequence motifs", *Bioinformatics,*24(19), pp. 1307-1309, 2008.

[14] FunClust. *Available in:* http://pdbfun.uniroma2.it/funclust/.

[15] Bio.Motif. *Available in :* http://www.bio-cloud.info/Biopython/en/ch13.html#motif-objects.

[16] Bouamama S., Boukerram A. and Al-Badarneh A., Motif Finding using Ant Colony Optimization, Dorigo M. et al (Eds.). *ANTS 2010,* Springer-Verlang Berlin Heidelberg, LNCS, 6234, 464, 2010.

[17] Chen-Hong Y., Yu-Tang L., and Li-Yeh C. DNA Motif Discovery Based on Ant Colony Optimization and Expectation Maximization. *IMECS 2011, International MultiConference of Engineers and Computer Scientists,* 1, 169, 2011.

[18] Dorigo M., Birattari M., Stutzle T. "Ant Colony Optimization". *Computational Intelligence Magazine IEEE*. Vol. 1(4). pp. 28-39. 2006.

[19] Ferreira P., Azevedo P., Evaluating deterministic motif significance measures in protein databases*"Algorithms for Molecular Biology,* 16(2), 2007

[20] Lones M., Tyrrell A., "The evolutionary computation approach to motif discovery," *GECCO 2005 Proceedings,* pp. 1-11. 2005

[21] Wei Z., Jensen T. "GAME: detecting cis-regulatory elements using a genetic algorith". *Bioinformatics,* Vol. 22, pp. 1577-84. 2006

[22] Stormo G., Hartzell G. "Identifing protein-binding site from unaligned DNA fragments". *Proc. Natl. Acad. Sci,* Vol. 86(4), pp. 1183-1187. 1989