Pairwise Alignment using ABC Optimization

Ankit Choubey
PG-Scholar
IET-Devi Ahilya University
Indore - 452017, India

G. L. Prajapati, PhD
Dept. of Computer Engg.
IET-Devi Ahilya University
Indore - 452017, India

ABSTRACT
In Artificial Bee Colony (ABC) optimization, we find better solution by employing neighbourhood search strategy using the current solutions. Researchers have tested ABC in many practical optimization problems. In this paper, we propose an application of ABC for the pairwise DNA sequence alignment in order to observe its performance in bioinformatics computation. We compare our results with the pairwise alignment algorithm FASTA. The results are encouraging. We also demonstrate ABC on Graph Coloring problem using different traversing strategies.

Keywords
Graph Coloring problem, Artificial bee colony optimization, DNA pairwise sequence alignment.

1. INTRODUCTION
Artificial Bee Colony algorithm (ABC) was initially published by Karaboga in 2005 as a technical report for numerical optimization problems. ABC is based on foraging behavior of honey bees [1]. It has been applied to solve many practical optimization applications.

2. THE GROWTH OF ABC ALGORITHM
After the invention of ABC by Karaboga in 2005 [1]. The first journal article describing ABC and evaluating its performance was presented by Karaboga and Basturk [2], in which the performance of ABC was compared to GA, PSO and particle swarm inspired evolutionary algorithm.

3. HOW ABC ALGORITHM WORKS?
ABC consists of employed and unemployed foragers and food sources. The ABC consists of three groups of artificial bees: employed forgers, onlookers and scouts. Employed bees and onlookers are equal and comprise half of population size.

In the basic ABC [2], there are 3 kinds of bees: employed, onlooker and scout bees.

3.1 Phases of ABC
It generally consists of four phases.

1) Initialization of ABC.

Determine the population size. Half of population size are employed bees and half are onlooker's bees.

Generate the random initial candidate solutions for employed bees using the equation in [2]. Determine the parameter max iteration, limit.

2) Employed bees phase

For all employed bees

Generate new candidate solution using the equation given in [1]. Also, calculate the fitness value of the new solution using the equation given in [1].

If fitness of new candidate solution is better than the existing solution replace the older solution.
Calculate the probability for each individual.

3) Onlooker bee phase.

For all onlooker bees
Select an employed bees using roulette wheel.

Produce new candidate solution.

Compute fitness of individual.

If fitness of new candidate solution is better than the existing solution replace the older solution.

4) Scout bee phase

If any food source exhausted then replace it by randomly generated solution by scout memorize the best solution.
Until (stopping criteria is not met).

4. APPLICATIONS OF ABC
4.1 Graph Coloring Problem.
Graph Coloring Problem (GCP) is the assignment of colors to the nodes of a graph such that no two adjacent vertices can have the same color. In this paper we have used two new traversing techniques DFS and BFS. By using BFS and DFS in ABC algorithm it takes less number of iteration to find chromatic number. However using DFS or BFS results are same.

4.1.1 ABC Algorithm for Graph Coloring problem
In this section we describe the ABC for Graph Coloring Problem.

4.1.1.1 Initialization
Randomly generate NB/2 food source in search space using DFS or BFS and generate a corresponding random number to nodes in a graph X_i. Obtain the sequences by sorting the X_i. Calculate the fitness for NB/2 sequences. Find the best solution and stored in global best GB.

4.1.1.2 Employed Bees
Obtain the new sequences using the policy described in next section. Calculate the fitness of new sequences. Share the information with onlooker bees.

4.1.1.3 Onlooker bees
Obtain the new sequences by using information shared by employed bees. Calculate the fitness of new sequences obtained. Update the GB by new result obtained.

4.1.1.4 Scout bees

If one bee dissatisfied from a food source then generate new solution using the traversing strategy.

4.1.2 Policy used for neighborhood discovery.

The most important is to choose the node which is going to change its order. Only one node will change its order. Here, we randomly choose one dimension from current solution denoted by j. And acc. to the equation we update the solution.

$$NS_i = OS_j + |OS_j - OSF_j| X\ rand[-1,1]$$

Where NS_i is new solution, OS_j is old solution, OSF_j is old solution fitness and i & j is the dimension chosen. Fitness=$1/c_i$ where C_i is the color used.

4.1.3 Fitness assignment and search space

The search in this problem for a graph having n nodes is constructed by n! Sequences and it is NP-Complete problem. For Fitness assignment we have used sequence based coloring. Our fitness assignment colors the node in the priority of a produced sequence rather than searching for a largest degree of nodes.

4.1.4 Experimental results.

We have used Erdos Renyi random graph generator to generate random graphs. Table 1 shows the comparison of ABC with some relevant algorithms. Edge density is also varied from sparse to dense for all graphs and results are taken on 50 graphs on edge density 0.10, 0.20, 0.30, 0.50, and 0.75 and then average is taken to compare the results with the relevant algorithms FF [3], LDO [4], SDO [5].

Table 1: Comparison with relevant algorithms

|V| Number of nodes	**Number of colors used**			
	First Fit	**Largest degree ordering (LDO)**	**Saturation Degree Ordering (SDO)**	**Artificial Bee Colony with DFS&BFS**
20	5.7	5.02	4.76	4.36
50	11.32	10.42	10.11	9.82
100	18.32	17.17	15.64	15.60

It is mentioned that all experiment are conducted on a Notebook PC with CPU 1.2 GHz using the Matlab R2008b.

4.2 DNA pairwise sequence alignment.

A DNA Sequence is arrangement of l characters randomly. For DNA Sequence it consist of 4 different nucleotides A (Adenine), C (Cytosine), G (Guanine), and T (Thymine). A sequence alignment is a way of arranging the letters of DNA, RNA, or protein in order to identify similarity that may be functional, structural, or evolutionary relationships between the sequences. There are two types of Alignment

1) Local Alignment. 2) Global Alignment

In this paper we proposed an algorithm for pairwise alignment [7]. It is the method used to find the best matching piece wise (local) or global alignment of two sequences.

We have to align two sequence in such way that they give maximum match score. To align two sequence we have to insert gaps in the sequences. However gaps are 20% of the sequences length. We have used score matrix as for match value (1), mismatch (0) and for gaps penalty (0).

4.2.1 Proposed DNA sequence alignment using ABC optimization.

In this section we describe our proposed algorithm for DNA-pairwise sequence alignment. Figure 1 shows the pseudo code for ABC-DNA.

No.	Main body of ABC-DNA sequence alignment
1.	Input: - Two sequences to be aligned.
2.	Output:-Aligned Sequences.
3.	//variable GBS stores global best alignment score
4.	//GB is a cell array holds the current sequences and //their alignment score.
5.	Initialization: - Randomly find NB/2 sequence by inserting gaps between the two sequences.
6.	Calculate fitness for each NB/2 Sequence and store
7.	global best score in GBS.
8.	While(iter < maxiteration)
9.	Begin
10.	Employed Bees()
11.	Calculate the fitness of new Sequences obtained.
12.	Onlooker bees()
13.	Update GB by new result obtain by the onlooker
14.	bees.
15.	End while
16.	If solution is not going to improve after maxiteration
17.	Repeat the algorithm by generating randomly new
18.	NB/2 sequences //Scout bees()
19.	End if
20.	Return GBS
21.	End.
22.	//sub functions
23.	Employed Bees()
24.	Begin
25.	For all NB/2 Sequences
26.	Select an existing gap randomly, and shift this gap
27.	randomly anywhere in the sequence.
28.	Update cell array GB
29.	End

Figure 1: Algorithm for ABC-DNA

An explained earlier ABC algorithm there are three kinds of bees which are employed bees, onlooker bees, scout bees. Here the sequence is food source .number of initial solution is denoted by NB.

By starting algorithm an initialization is done in which we randomly choose NB/2 sequence by inserting gaps in the sequences. When initialization stage is completed employed bees phase starts in which we select an existing gap randomly and shift this gap randomly anywhere in the sequence, after doing this employed bees phase shares its information with the onlooker bees. Onlooker bees choose sequences from employed bees and does the same work as employed bees, and update the result in GB. If solution does not improve after max iteration then repeat the algorithm. Here iteration is the current loop. Maxiteration is the maximum number of iteration. Formally algorithm is described in Figure 1.Here Fitness is the maximum match score.

Comparison has been done with FASTA [6] algorithm and results have been taken. ABC-DNA have been tested on fixed

length DNA sequence of 5, 10, 20 length. Result has been taken by running algorithm on different sequences than average has been taken. Table 2 and Table 3 show comparison between FASTA and ABC-DNA. Results show that match score of the two algorithms are same but number of gaps in FASTA algorithm is more than our proposed ABC-DNA algorithm. ABC-DNA also generate different alignment for sequence with less number of gaps.

Table 2: Difference in alignment between FASTA and ABC-DNA for pairwise alignment.

Length of a Sequence	FASTA	ABC-DNA
5	g-c--gcg gactt---	gcgc-g ga-ctt
10	g-c--gcgtgcg-c gacttg--tg-ga-	g-cgcgtgcgc gacttgtg-ga
20	g-c--gcgtgcgcggaa---gga---gcc gacttg--tg-g---aacct--actt-cc	gcgcgtgcgcggaagg-a--gcc g-ac-t-tgtggaacctacttcc

Table 3: Comparison between FASTA and ABC-DNA for pairwise alignment.

Length of a Sequence	FASTA		ABC-DNA	
	Score	Gap	Score	Gap
5	2.20	2.40	2.20	1.20
10	5.40	4.40	5.40	2.40
20	8.40	6.20	8.60	4.40

4.2.2 Parameters of ABC-DNA

Important parameters of ABC are population size, number of iteration and limit. Here Number of Bees (NB) is 40, number of iteration is 200, and limit is 10.

We have conducted all the experiments on a Notebook PC with CPU 1.2 GHz using the Matlab R2008b

5. CONCLUSIONS

ABC algorithm optimally colors the vertices in a graph as compared with other greedy algorithms LDO, SDO, FF. It is also found that ABC-DNA algorithm aligns the sequences slightly better than the FASTA algorithm. ABC-DNA results better in local alignment of sequences. However, as a future work the performance of ABC-DNA algorithm can be tested in global alignments for DNA and protein sequences.

6. REFERENCES

[1] D. Karaboga. An idea based on honey bee swarm for numerical optimization. *Techn. Rep. TR06, Erciyes Univ. Press, Erciyes,* 2005.

[2] Dervis Karaboga · Bahriye Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm" J Glob Optim (2007), pp 459-471.

[3] A. H. Gebremedhin, "Parallel graph coloring." PhD Thesis, University of Bergen, Norway, 1999.

[4] D. De Werra, "Heuristics for Graph Coloring Computational Graph Theory." Comput Suppl, Springer, Vienna 7:19 11-208, 1990.

[5] D. Brelaz, "New methods to color the vertices of a graph." *Commun ACM*, 22 (4):251-256, 1979. doi:10.1145/359094.359101

[6] Lipman, DJ; Pearson, WR (1985). "Rapid and sensitive protein similarity searches". Science **227** (4693):1435–41. Doi:10.1126/science.2983426. PMID 2983426.

[7] Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2nd Ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. *ISBN 0-87969-608-7*.