Journal of Mechanics Engineering and Automation 12 (2022) 57-63
doi: 10.17265/2159-5275/2022.02.004

Optimization of a Classical Algorithm for the Alignment of Genomic Sequences with Artificial Bee Colony

Raul Magdaleno Peñaloza, Andrea Magadan Salazar and Gerardo Reyes Salgado

Department of Computer Science, Centro de Investigacion y Desarrollo Tecnologico, Cuernavaca Morelos 62490, Mexico

Abstract: This article shows genomic alignment methods using the classic "Needleman" and "Smith-Waterman" algorithms, the latter they were optimized by the ABC (artificial bee colony) algorithm. In the genomic alignment, a goal state is not presented, the experiments that are carried out show alternative alignments by ABC were proposed. Different types of alignments could exist within the classical algorithm, based on a horizontal, vertical, diagonal and inverse search mechanism on a match value table. Our ABC-Smith Waterman algorithm was generated from the genomic sequences written in rows and columns for the search for similarities that will provide values that ABC uses to process and provide more results of alignments that can be used by scientists for their experiments and research.

Key words: Algorithm, genomic alignment, ABC, Needleman, Smith-Waterman.

1. Introduction

Sequence alignment is a basic tool that allows the extraction of functional, structural and evolutionary information contained in biological sequences. These similarities may indicate functional or evolutionary relationships [1].

The main algorithm used in the genomic areas, was propose by Saul B. Needleman and Christian D. Wunch [2] in 1970, T. F. Smith and M. S. Waterman [3] used this method and optimized in 1981. In 2017 M. A. Lopez and J. V. Medina [4] from the Universidad del Valle, Santiago de Cali, Colombia, proposed an optimization of these algorithms using a parallel architecture.

The objective of this paper is to optimize the alignment methods, which facilitates experimentation in the alignment of two sequences and generates new ones, better with different results that can be used by specialists in the genomic area.

The article is organized as follows way: Section 2 presents genomic sequences. Section 3 talks about sequences alignment. Section 4 describes Needleman-Wunch algorithm. Section 5 presents Smith-Waterman algorithm. Section 6 presents ABC (artificial bee colony) algorithm. Section 7 describes the optimization ABC Smith-Waterman. Finally, Section 8 shows the results with ABC optimization.

2. Genomic Sequences

DNA (deoxyribonucleic acid) is a finite chain built from an alphabet N = {A, C, G, T} of nucleotides and the GENOME is a set of all the DNA sequences associated with an organism [1].

According to Needleman and Wunch [2], nucleic acids are the biomolecules that carry genetic information. They are biopolymers, of high molecular weight, formed by other structural subunits or monomers, called nucleotides. From the chemical point of view, nucleic acids are macromolecules formed by linear polymers of nucleotides, linked by phosphate ester bonds, with no apparent periodicity.

DNA sequences contain the genetic information in all living things. The more similar two sequences are, the more similar the functions of the proteins encoded by them will tend to be. Genes having same ancestors reduce the chances that the sequences may be homologous.

Corresponding author: Raul Magdaleno Peñaloza, mechatronic engineer, research field: genomic alignment and artificial bee colony.

DNA undergoes mutations over the years and through their descendants, more time that passes since the last common ancestor, the more different the sequences will be.

3. Sequences Alignment

Sequence alignment is a basic tool that allows the extraction of functional, structural and evolutionary information contained in biological sequences.

The main goals of comparing two or more genomic sequences are:

- Determine and quantify the degree of similarity between them.
- Determine if there is some kind of relationship between them or if the resemblance is simply the result of chance.
- Detect the presence of conserved structural and functional motifs.
- Build phylogenetic trees that reflect their evolutionary relationships.

It is proposed that in order to find the degree of similarity in two sequences, the first thing to do is to look for similar characters, which consists of writing one sequence, on top of the other so that the number of symbols that coincide in the same position be maximum [5]. These matches are displayed in Fig. 1.

If necessary, gaps can be introduced in the sequences and considered as the insertion of a residue in one of the sequences as the disappearance of a residue in another. In an alignment, when there is no coincidence and in order not to move the whole sequence, it is left in the closest place as shown in Fig. 2.

Sales et al. [6] show that in each position of the alignment there will be two identical characters (MATCH), different (No MATCH) or one character aligned with a gap. What cannot be are two gaps aligned.

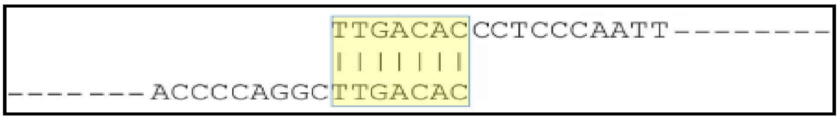
Two sequences can be aligned in many ways. To determine which is the best alignment, a scoring system is used that gives each pair of characters a different value depending on whether they are the same, different or whether there is a gap. The score of an alignment is calculated by adding the score of each of the positions helping to determine if the sequences are really related or if their similarity is due to chance.

The alignment that gets the highest score is called the optimal alignment, shown in Fig. 3.

4. Needleman-Wunch Algorithm

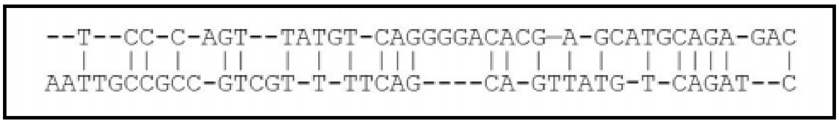
Needleman and Wunch [2] introduced an approach in 1970 to calculate the optimal global alignment of two sequences.

The algorithm [2] is a way to massively reduce the number of possibilities to consider finding new alignment.



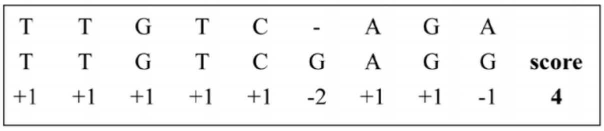
```
                TTGACACCCTCCCAATT--------
                |||||||
-------ACCCCAGGCTTGACAC
```

Fig. 1 Example of a sequence [7].



```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

Fig. 2 Match NoMatch Gap/Indel Alignment [7].



T	T	G	T	C	-	A	G	A	
T	T	G	T	C	G	A	G	G	**score**
+1	+1	+1	+1	+1	-2	+1	+1	-1	**4**

Fig. 3 Score alignment [7].

According to Coll [8] and Backofen [9], under the assumption that both input sequences come from the same origin, a global alignment tries to identify the parts that coincide and the changes necessary to transfer a sequence to the other.

The dynamic programming approach tabulates the optimal sub-solutions in a 2D matrix [2].

Needleman-Wunch consists of the following three steps:

(1) Start the score matrix

(2) Calculate the score and fill in the back matrix

(3) Deduce the alignment of the posterior matrix

To determine which is the best alignment, a scoring system is used that gives each pair of characters a different value depending on whether there is a MATCH, a No MATCH or a GAP. Taking the example of two words: SEND with AND, we get:

SEND

AND score: +1

A-ND score: +3 ← This was the best score found.

AN-D score: -3

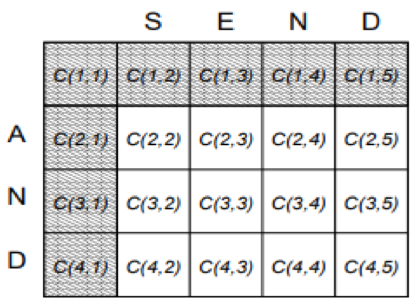
AND- score: -8

Now, in order to create a Needleman-Wunch matrix [2], the following procedure is used, which is:

- Align the first sequence horizontally at the top;
- Align the second sequence vertically, glued to the left.

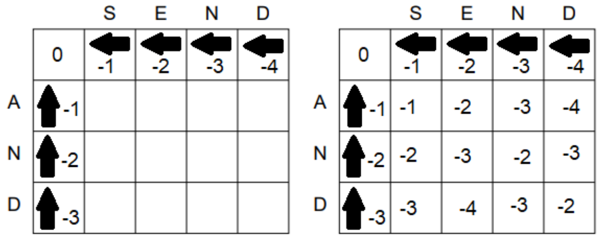
In each cell, a value is added or subtracted depending on whether the characters generate a MATCH, a No Match or a GAP [9].

Figs. 4 and 5 show how the matrices should be filled with their respective score, which can be followed by a path in the matrix, starting from the lower right corner to the upper left corner and following the best score, as shown in Fig. 6.



	S	E	N	D
C(1,1)	C(1,2)	C(1,3)	C(1,4)	C(1,5)
A C(2,1)	C(2,2)	C(2,3)	C(2,4)	C(2,5)
N C(3,1)	C(3,2)	C(3,3)	C(3,4)	C(3,5)
D C(4,1)	C(4,2)	C(4,3)	C(4,4)	C(4,5)

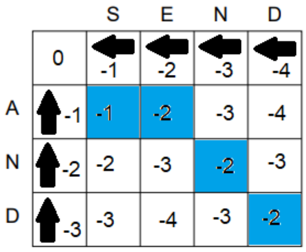
Fig. 4 Needleman matrix [10].



		S	E	N	D
	0	← -1	← -2	← -3	← -4
A	↑ -1				
N	↑ -2				
D	↑ -3				

		S	E	N	D
	0	← -1	← -2	← -3	← -4
A	↑ -1	-1	-2	-3	-4
N	↑ -2	-2	-3	-2	-3
D	↑ -3	-3	-4	-3	-2

Fig. 5 Score alignment [10].



		S	E	N	D
	0	← -1	← -2	← -3	← -4
A	↑ -1	-1	-2	-3	-4
N	↑ -2	-2	-3	-2	-3
D	↑ -3	-3	-4	-3	-2

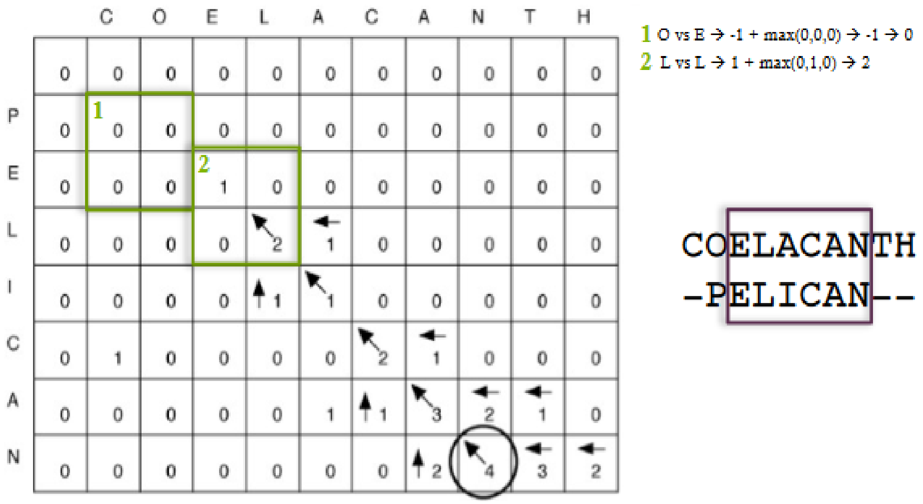
Fig. 6 Score tracking in matrix [10].

5. Smith-Waterman algorithm

The dynamic programming approach of Temple F. Smith and Michael S. Waterman [3] calculates the optimal local alignments of two sequences that are best conserved. Manavski and Valle [11] argue that this algorithm is designed to find the optimal local alignment between two sequences. Based on Needleman's matrix alignment computation, the number of rows and columns is given by the sequence database.

Sales et al. [11] propose the example where he aligns the words "COELACANTH" with "PELICAN". The matrix is built the same as Needleman [11]. Its difference lies in expanding an additional data "0". This lower bound on the similarity score excludes the alignments that eventually are not similar.

Since the row and column are 0 instead of negative numbers (-1, -2, -3…), the MATCH & No MATCH values remain the same. Now, when counting, any number that, if negative, will be set to 0, here, the score of the highest obtained is followed until reaching 0. This method can be seen in Fig. 7.



		C	O	E	L	A	C	A	N	T	H
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0
L	0	0	0	0	↖2	←1	0	0	0	0	0
I	0	0	0	0	↑1	↖1	0	0	0	0	0
C	0	1	0	0	0	0	↖2	←1	0	0	0
A	0	0	0	0	0	1	↑1	↖3	←2	←1	0
N	0	0	0	0	0	0	0	↑2	↖4	←3	←2

1 O vs E → -1 + max(0,0,0) → -1 → 0

2 L vs L → 1 + max(0,1,0) → 2

```
COELACANTH
-PELICAN--
```

Fig. 7 Matrix Smith-Waterman [1].

6. ABC algorithm

The ABC is one of the most recent algorithms in the domain of collective intelligence proposed by Dervis Karaboga in 2005 [12], in this work, a new optimization algorithm based on the intelligent behavior of honey bee swarm has been described. The new swarm algorithm is very simple and very flexible when compared to the existing swarm-based algorithms. The algorithm can be used for solving unimodal and multi-modal numerical optimization problems [13]

The objective of these bees is to discover the food sources with the greatest nectar [12].

The behavior of bees was modeled as an optimization heuristic [12] based on the biological model that consists of according to Ref. [14] the following steps:

(1) Food sources: although the value depends on many factors, it is summarized in a numerical value that indicates their potential.

(2) Collector bees employed: these bees exploit a food source; they are also in charge of communicating their location and profitability to the observer bees.

(3) Unemployed collector bees: They are continually looking out for a food source to exploit. There are two types of unemployed foragers: scouts, searching the environment surrounding the nest for new food sources

The model also defines a main mode of behavior that is necessary for self-organization, known as collective intelligence, that the main function is recruiting food collectors for rich food sources results in positive feedback and abandonment of poor sources by food collectors, causing negative feedback [15]. Therefore, ABC's exploration capacity is restricted [16].

7. ABC Smith-Waterman

In the Smith-Waterman algorithm it is one of the most efficient and easiest to understand methods to find genomic alignments.

The steps used for algorithm optimization are as follows:

(1) Generate the table of records.

(2) Generate the first alignment by the Smith-Waterman method.

(3) Generate the second alignment in reverse.

(4) Schedule bees for every possibility.

(5) Send bees for both line-ups.

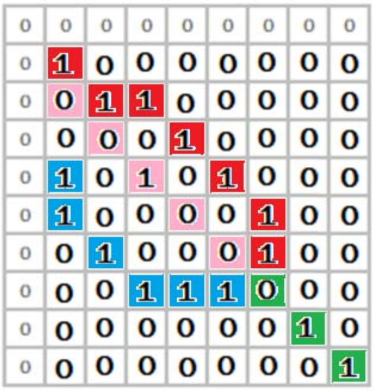
(6) Register possibilities.

(7) Save possible alignments.

The Smith-Waterman algorithm is an algorithm that is based on similarity data record tables, which seeks the best alignment possibility. However, the algorithm can find possibilities where the alignment forks in different ways. There is no goal state in genomic alignment, any possibility of generating an alignment is considered valid.

When running the method, it was found that, by aligning with a section of the sequence, the result is seen more directly, however, by aligning the entire sequence; it generates a completely different result. This experimentation raised the question: Can a different alignment possibility be found in an established method?

Fig. 8 shows how they can find a possibility, in which, choosing any of the marked paths, it could be used as an option for genomic alignment.



0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	1	0	1	0	1	0	0	0
0	1	0	0	0	0	1	0	0
0	0	1	0	0	0	1	0	0
0	0	0	1	1	1	0	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1

Fig. 8 Matrix possibilities.

8. Results with ABC Optimization

The genomic sequences were obtained from the public database NCBI (National Center for Biotechnology Information) by reason of the INSDC (International Neuclotide Sequence Databank Collaboration) which is an international collaboration between the three largest genomic databases in Europe and Asia [17, 18].

The genomes of the animals that were used for this experiment were:

- Ceratotherium simum: Used for the columns
- Bubo bubo: Used for rows.

In Figs. 9-13, the paths that the bees were found to generate different alignments are graphically shown in different data graphs.

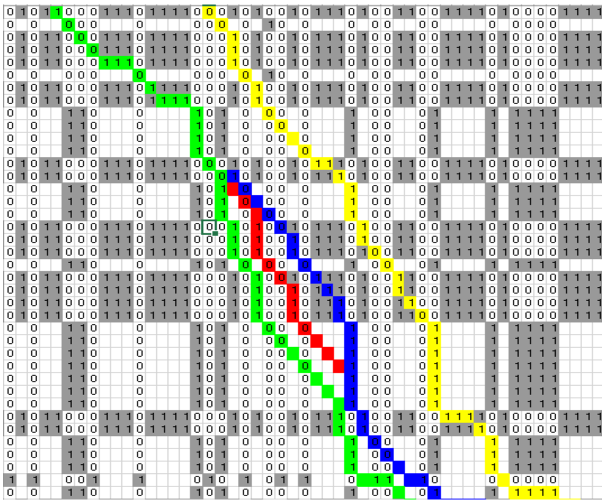


Fig. 9 First result with ABC optimization.

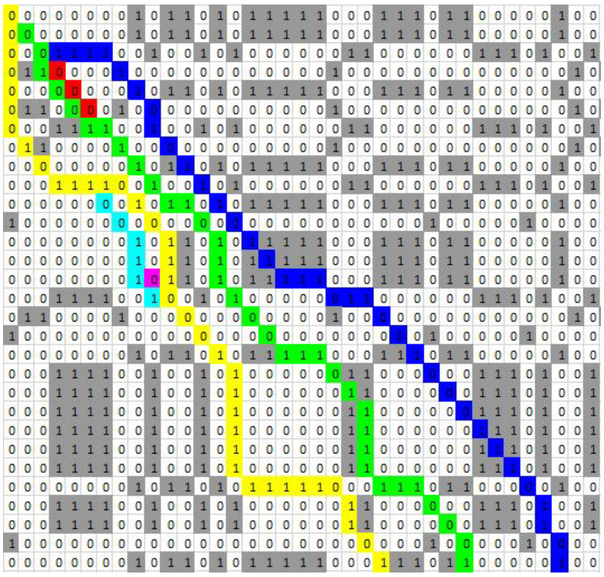


Fig. 10 Second result with ABC optimization.

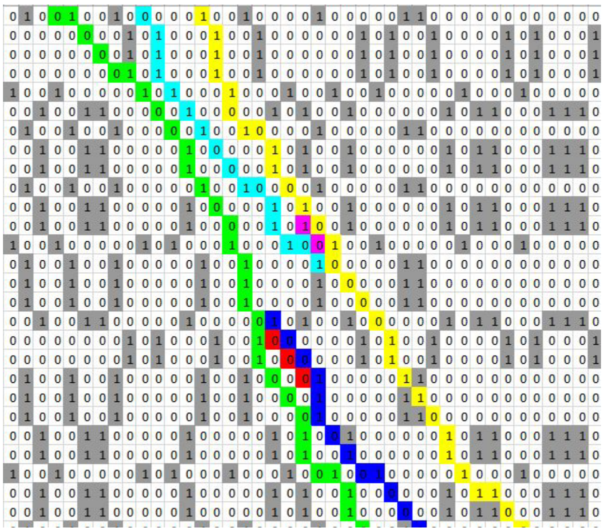


Fig. 11 Third result with ABC optimization.

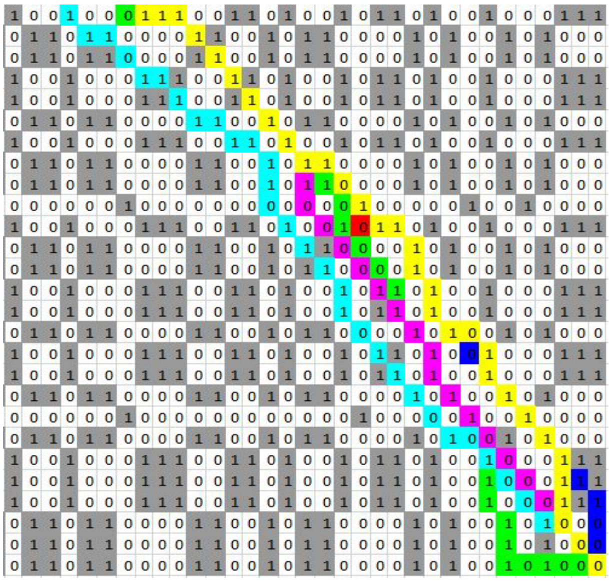


Fig. 12 Fourth result with ABC optimization.

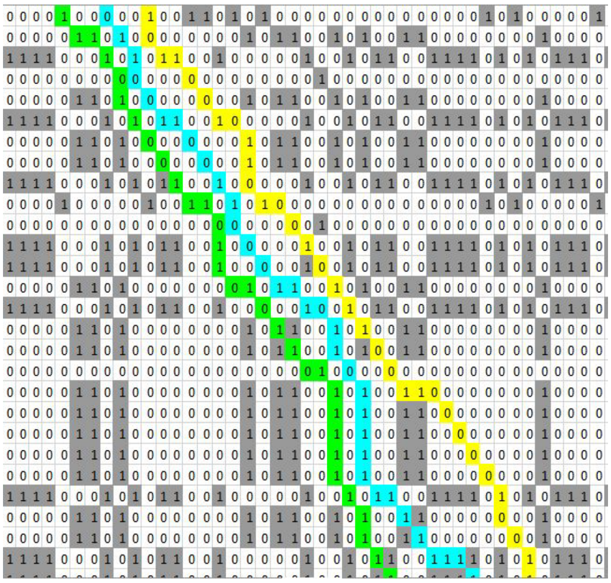
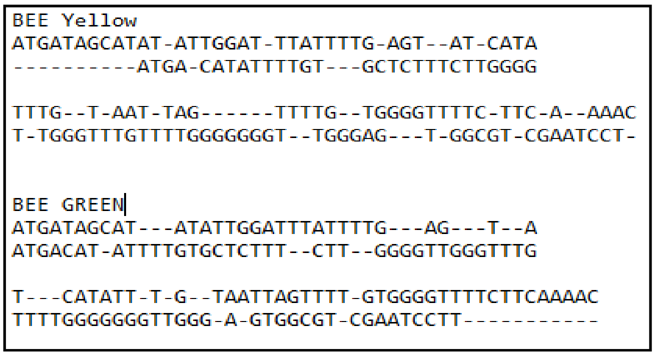


Fig. 13 Fifth result with ABC optimization.



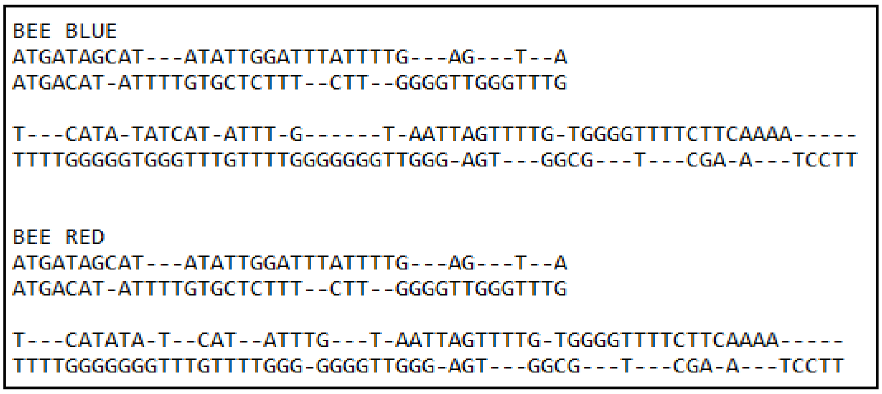
```
BEE Yellow
ATGATAGCATAT-ATTGGAT-TTATTTTG-AGT--AT-CATA
----------ATGA-CATATTTTGT---GCTCTTTCTTGGGG

TTTG--T-AAT-TAG------TTTTG--TGGGGTTTTC-TTC-A--AAAC
T-TGGGTTTGTTTTGGGGGGGT--TGGGAG---T-GGCGT-CGAATCCT-


BEE GREEN
ATGATAGCAT---ATATTGGATTTATTTTG---AG---T--A
ATGACAT-ATTTTGTGCTCTTT--CTT--GGGGTTGGGTTTG

T---CATATT-T-G--TAATTAGTTTT-GTGGGGTTTTCTTCAAAAC
TTTTGGGGGGGTTGGG-A-GTGGCGT-CGAATCCTT-----------
```

Fig. 14 Alignment result with ABC optimization.



```
BEE BLUE
ATGATAGCAT---ATATTGGATTTATTTTG---AG---T--A
ATGACAT-ATTTTGTGCTCTTT--CTT--GGGGTTGGGTTTG

T---CATA-TATCAT-ATTT-G------T-AATTAGTTTTG-TGGGGTTTTCTTCAAAA-----
TTTTGGGGGTGGGTTTGTTTTGGGGGGGTTGGG-AGT---GGCG---T---CGA-A---TCCTT


BEE RED
ATGATAGCAT---ATATTGGATTTATTTTG---AG---T--A
ATGACAT-ATTTTGTGCTCTTT--CTT--GGGGTTGGGTTTG

T---CATATA-T--CAT--ATTTG---T-AATTAGTTTTG-TGGGGTTTTCTTCAAAA-----
TTTTGGGGGGGTTTGTTTTGGG-GGGGTTGGG-AGT---GGCG---T---CGA-A---TCCTT
```

Fig. 15 Second alignment result with ABC optimization.

The yellow color shows the first alignment made with the classic Smith-Waterman algorithm. The green color shows an alignment made in the reverse way. In the section where the colors are born, those are the possibilities found by the bees in case of taking a different path by the alignment of the classical method in reverse.

Figs. 14 and 15 show an example of the different alignment results generated with the aforementioned algorithms.

9. Conclusions

The method ABC for the optimization of the algorithm Smith Waterman showed good results at the time of implementation and generation of the sequences. Actually, the ABC-Smith Waterman algorithm gives six different results as maximum alignment possibilities versus traditional algorithms alignment only give one result.

The algorithm does not take a lot of time in the process, even if the sequences are long, the program takes less than ten minutes.

The genomic sequences did not generate any problem or change in the work plans, because its manipulation is as any text file, giving a versatility when generating the algorithms and the generation of alignment results.

Expectations were exceeded when the algorithm gave the first results, extending the possibilities of alignments and multiple future works with new different possibilities that can be searched by generating new artificial employed bees with a new specification to search food sources that could give a new kind of alignment.

References

[1] Santamaria, R. 2013. *Alineamiento de pares de secuencias*. (in Spanish)

[2] Needleman, S. B., and Wunch, C. D. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *J Mol Biol*. 48 (3): 443-53.

[3] Smith, T. F., and Waterman, M. S. 1981. *Identification of Common Molecular Subsequences*. London: Academic Press Inc.

[4] López, M. A., and Medina, J. V. 2017. "Implementación hardware del algoritmo de Needleman-Wunsch modificado usando una arquitectura paralela." *Revista Ingeniería Biomédica* 12 (23): 53-62. (in Spanish)

[5] Coll, V. B. 2008. "Estructura y Propiedades de Los Ácidos Nucléicos." M.Sc. thesis, Ingeniería Biomédica (UV-UPV). (in Spanish)

[6] Sales, J. C., Blanca, J., and Ziarsolo, P. 2019. Alineamiento de secuencias (s. f.-b). bioinf.comav.upv.es. (in Spanish)

[7] Juan Manuel González Mañas. 2020. COMPARACIÓN DE SECUENCIAS. 7 diciembre 2020, de Universidad del País Vasco (in Spanish)

[8] Backofen, R. 2011. *Sequence Alignment Needleman-Wunsch*. Obtenido de Uni Freiburg Bioinformatics.

[9] Backofen, R. 2018. *Teaching-Smith-Waterman*. Obtenido de Uni Freiburg Bioinformatics.

[10] Likic, V. (2016). "The Needleman-Wunsch Algorithm for Sequence Alignment." Molecular Science and Biotechnology Institute The University of Melbourne, 46.

[11] Manavski, S. A., and Valle, G. 2008. "CUDA Compatible GPU Cards as Efficient Hardware Accelerators for Smith-Waterman Sequence Alignment." *BMC Bioinformatics* 9: S10.

[12] Salto, C. 2017. *Optimización mediante el algoritmo de colonia de abejas artificial*, edited by General Pico and La Pampa. Argentina: Universidad Nacional de La Pampa. (in Spanish)

[13] Karaboga, D. 2005. *An Idea Based on Honey Bee Swarm for Numerical Optimization*. Technical Report-TR06, Engineering Faculty Computer Engineering Department, Erciyes University.

[14] Kumar, A., Kumar, D., and Jarial, S. K. 2016. "A Comparative Analysis of Selection Schemes in the Artificial Bee Colony Algorithm." *Información Tecnológica* 20 (1): 55-66.

[15] Karaboga, D. 2010. "Artificial Bee Colony Algorithm." Accessed on 29 October, 2021. Scholarpedia.Org.

[16] Tsai, P. W., Pan, J. S., Liao, B. Y., and Chu, S. C. 2009. "Enhanced Artificial Bee Colony Optimization." *International Journal of Innovative Computing, Information and Control* 5 (12): 12.

[17] National Center for Biotechnology Information Ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/.

[18] International Nucleotide Sequence Database Collaboration INSDCInsdc.org. http://www.insdc.org.