Universidad Nacional de San Agustín

Artificial Intelligence

Multiple Sequence Alignment using Particle Swarm Optimization

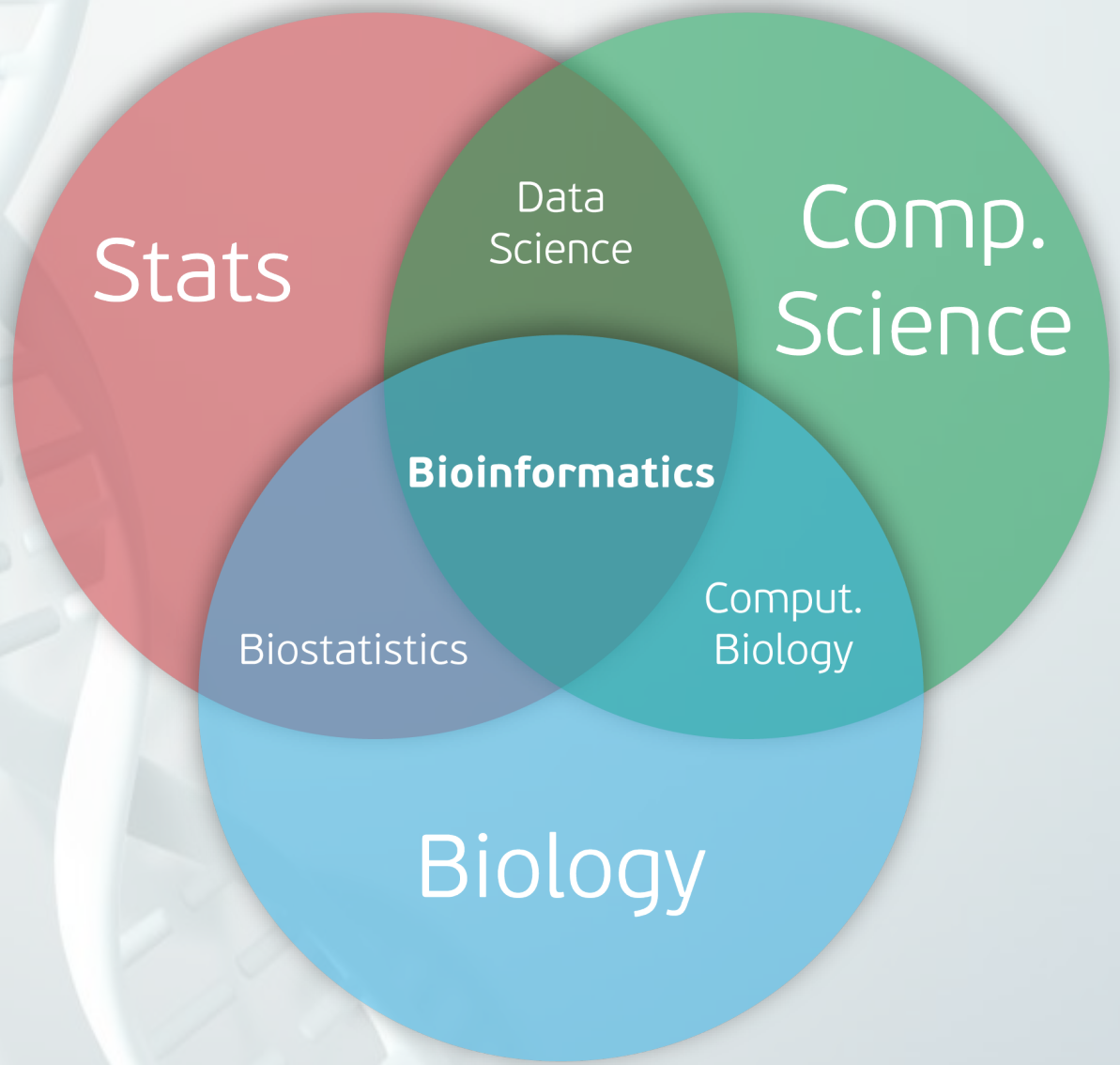
MSc. Vicente Machaca Arceda

2021

Content

Overview

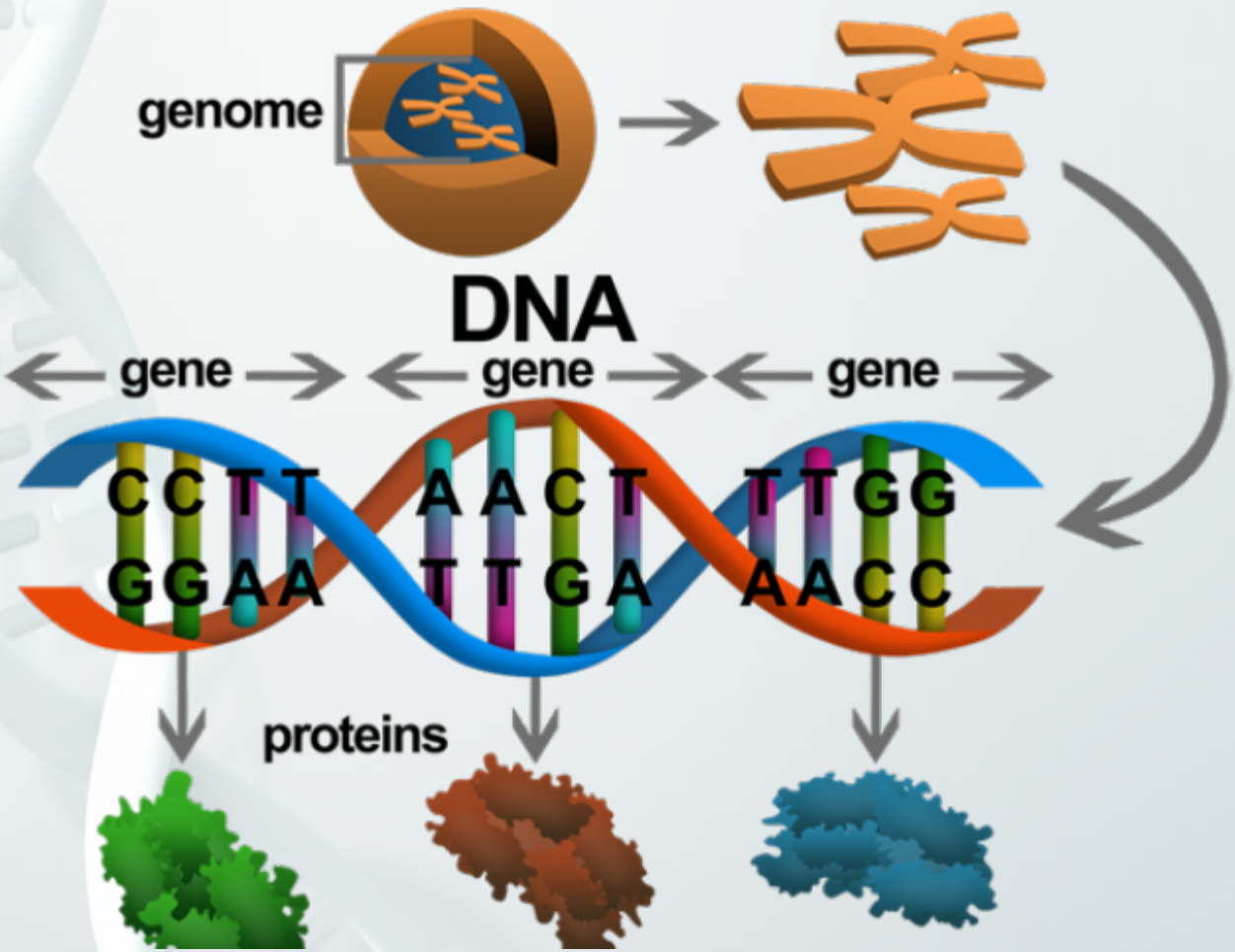
Bioinformatics



Bioinformatics
What is Bioinformatics?

According to Luscombe et al.: **Bioinformatics** involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins [1].

Bioinformatics




```
>J01859.1 Escherichia coli 16S ribosomal RNA, complete sequence
AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGT
AACAGGAAGAAGCTTGCTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATG
GAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCG
GGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACG
ATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGG
CAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTT
CGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCG
CAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAAC
TGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGT
AGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCG
TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCC
TTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACT
CAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCT
TACCTGGTCTTGACATCCACGGAAGTTTTCAGAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTGC
TGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCCT
TTGTTGCCAGCGGTCCGGCCGGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGA
CGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCGCATACAAAGAGAAGCGA
CCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATG
AAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCG
CCCGTCACACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACTT
TGTGATTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTGGATCACCTCCTT
```

Figure: 16S ribosomal DNA of Escherichia coli with FASTA Format.

Overview

Problem
Sequence alignment


```
No alignment

 CGATGCTAGCGTATCGTAGTCTATCGTAC
              |     ||
 ACGATGCTAGCGTTTCGTATCATCGTA

Alignned

 -CGATGCTAGCGTATCGTAGTCTATCGTAC
  |||||||||||| ||||||||||||||
 ACGATGCTAGCGTTTCGTA-TC-ATCGTA-
```

Figure: No alignment versus alignment.

In the alignment process there could be substitutions, changes of residues and gaps. Gaps could cause by insertions or deletions.

Problem
Sequence alignment


```
No gaps (10 matches)

 a:  ATATTGCTACGTATATCAT
         ||||||||||
 b: ATATATGCTACGTATCAT

With one gap (14 matches)

 a: ATAT-TGCTACGTATATCAT
    |||| ||||||||||
 b: ATATATGCTACGTATCAT

With two gaps (16 matches)
 a: ATAT-TGCTACGTATATCAT
    |||| |||||||  ||||||
 b: ATATATGCTACG--TATCAT
```

Figure: Alignment and gaps.

Algorithms should take into account the possibility of introducing gaps. **Several alignments can be constructed** between two sequences.

Problem

Multiple Sequence alignment

Figure: Example of Multiple Sequence Alignment (MSA) in amino acid sequences.

Overview

Objective

Propose Particle Swarn Optimization (PSO) to solve the Multiple Sequence Alignment (MSA) [2].

Overview

Proposal
Particle definition

Leader:	W	G	K	V	-	-	N	V	D
	W	D	K	V	-	-	N	-	-
	S	-	K	V	G	G	N	-	-

Particle:	W	G	K	-	-	V	N	V	D
	-	W	D	K	-	-	-	V	N
	S	-	K	V	G	G	-	N	-

	5	6		
Leader:	5	6	8	9
	2	8	9	

	4	5		
Particle:	1	5	6	7
	2	7	9	

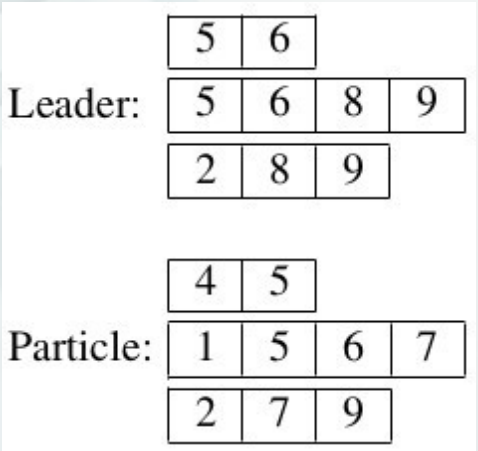
Figure: Example of particle representation.

Overview

Proposal
Movements

$$distance = \frac{matchingGaps}{totalGaps} \quad (1)$$

$$crossPoint = rand(1, distance * length) \quad (2)$$



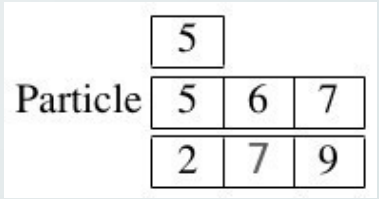


Figure: Example of particle movement.

Overview

Proposal
Objective function

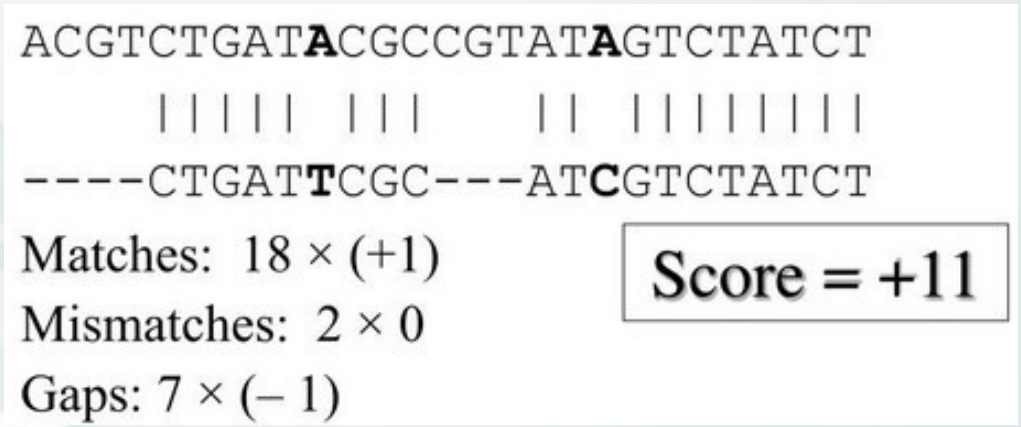


Figure: Example of score in sequence alignment.

Overview

Proposal
Mutations

The mutation operator inserts a gap in a random position in a random sequence inside a particle.

Overview

Table: Dataset used in the experiments.

Dataset	min. length	max. length	num. bases
S6	8	17801	153
S7	457	457	8
S8	7	10	5

Experiments
Datasets S8

1	A	T	G	C	A	A	G		
2	T	A	A	G	T	C	A	A	G	T
3	A	T	G	C	A	A	C	T	
4	T	A	A	G	T	C	A	T	A
5	A	T	G	G	A	T	T	C	

Figure: Sequences of S8 dataset.

Experiments
Params

Table: Params used in the experiments.

Param	Value
Iterations	30
Num. of particles	25
Mutation probability	0.2
Gaps	30%
Num. of experiments test	10

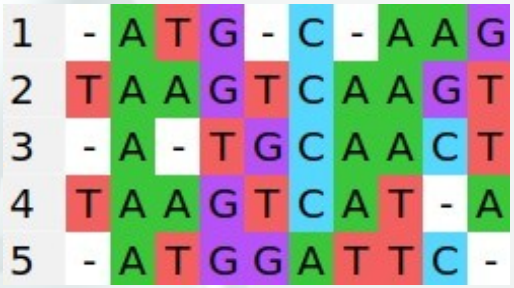
Overview

Results
PSO vs CLUSTALW

Table: Score comparison of PSO and CLUSTALW.

Dataset	PSO-mutation	PSO	CLUSTALW
S6	12678	10012	18045
S7	11105	9054	12564
S8	32	28	49

Results
PSO vs CLUSTALW



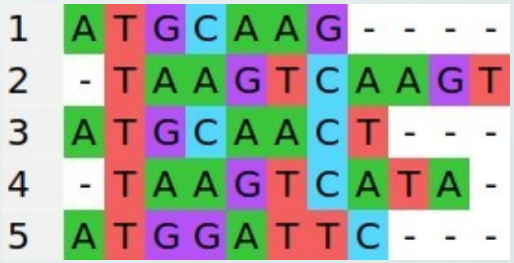


Figure: Left: S8 alignment with PSO-mutation. Right: S8 alignment with CLUSTALW.

Results
PSO

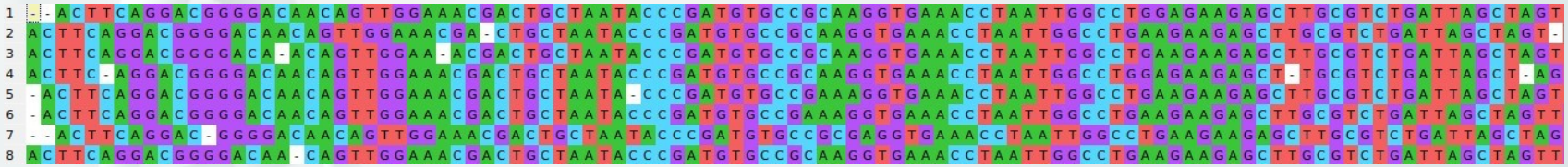


Figure: Extract of S7 alignment using PSO-mutation

Conclusions

This thesis proposed the use of PSO to solve the MSA problem. The author used a set of datasets from NCBI.

The author proposed a mutation operator to avoid local solutions. This operator just inserts a gap.

The score of PSO was accepatble and very similar to CLUSTALW. Currently, there is more research on this topic.

References I

[1] N. M. Luscombe, D. Greenbaum, and M. Gerstein, “What is bioinformatics? a proposed definition and overview of the field,” *Methods of information in medicine*, vol. 40, no. 04, pp. 346–358, 2001.

[2] F. B. R. Zablocki *et al.*, “Multiple sequence alignment using particle swarm optimization,” Ph.D. dissertation, University of Pretoria, 2009.