Particle Swarm Optimizer Variants for Multiple Sequence Alignment

Nicholas Aksamit

Brock University

na16dg@brocku.ca
Supervisor: Prof. Ombuki-Berman

COSC 4F90

December 20, 2021

Overview

Objective

- To provide a comparative evaluation of four Particle Swarm Optimization (PSO) variants applied to the Multiple Sequence Alignment (MSA) problem.

- Investigate the change of parameters for each variant.
- Perform comparisons between algorithms.
- Compare PSO versions with a state-of-the-art method.

Multiple Sequence Alignment (MSA)

A	G	G	T	C	$\rightarrow$	A	G	G	T	C	-
G	A	T	C	A		-	G	A	T	C	A

Figure: Sample Alignment, with $\alpha = \{\text{A,G,C,T}\}$

MSA is the process of aligning three or more sequences, following some alphabet α. Spaces are added, or "-", and are referred to as indels (INsertion/DELetion).

Notice the insertion, deletion, and mutation that all occur in the two sequences of the above figure.

Particle Swarm Optimization (PSO) [12]

Created by James Kennedy and Russell Eberhart in 1995 as a stochastic optimization method.

Features

1. A set of particles traversing the search space.
2. Particle positions represent solutions.
3. Particles move using velocity equation.

Velocity Equation

$$
\begin{aligned}
v_i(t+1) = & \\
\omega \times v_i(t) + & \quad (1) \\
r_1 \times c_1 \times (y_i(t) - x_i(t)) + & \quad (2) \\
r_2 \times c_2 \times (\hat{y}_i(t) - x_i(t)) & \quad (3)
\end{aligned}
$$

Position Update

$$x_i(t+1) = x_i(t) + v_i(t+1)$$

Binary PSO (BPSO) [13]

Differences from PSO

1. Binary position vector.
2. Position update equation.
3. Interpretation of the velocity function.

Position Update

$$x_{ij}(t+1) = \begin{cases} 1, & \text{if } \mathrm{rand}[0,1] < S(v_{ij}(t+1)) \\ 0, & \text{otherwise} \end{cases}$$

Angular Modulated PSO (AMPSO) [14]

Differences from PSO

1. 4-dimensional position and velocity vectors.
2. Uses a generation function to make binary solution matrix.

Generation Function

$$g(x) = \sin(2\pi(x-a) \times b \times cos(A)) + d$$
$$\text{where } A = 2\pi \times c(x-a)$$

- a, b, c, d come from position vector
- x values are evenly spaced within an interval

$$\text{bit} = \begin{cases} 1, & g(x) > 1 \\ 0, & \text{otherwise} \end{cases}$$

Multi-guided PSO (MGPSO) [15]

Differences from PSO

1. Multiple swarms; one for each objective function.
2. Addition of the archive.
3. Velocity update equation.
4. λ coefficient.

Velocity Equation

$$
\begin{aligned}
v_i(t+1) = & \\
\omega * v_i(t) + & \quad (1)\\
r_1 * c_1 * (y_i(t) - x_i(t)) + & \quad (2)\\
\lambda_i * r_2 * c_2 * (\hat{y}_i(t) - x_i(t)) + & \quad (3)\\
(1 - \lambda_i) * c_3 * r_3 * (\hat{a}_i(t) - x_i(t)) & \quad (4)
\end{aligned}
$$

- Archive filled with pareto non-dominated solutions from either sub-swarm.
- λ controls trade-off from social and archive component.
- Archive component the **only** form of inter-swarm knowledge.

Problem Representation

A	G	G	T	C	-	$\rightarrow$	0	0	0	0	0	1
-	G	A	T	C	A		1	0	0	0	0	0

Figure: Bit Matrix Representation

Objective Functions

- Aligned Characters (max f_1)
- Number of inserted indels (min f_2)

Single-objective

$$f(x) = w_1 f_1(x) + ... + w_n f_n(x)$$

Weighted Aggregation

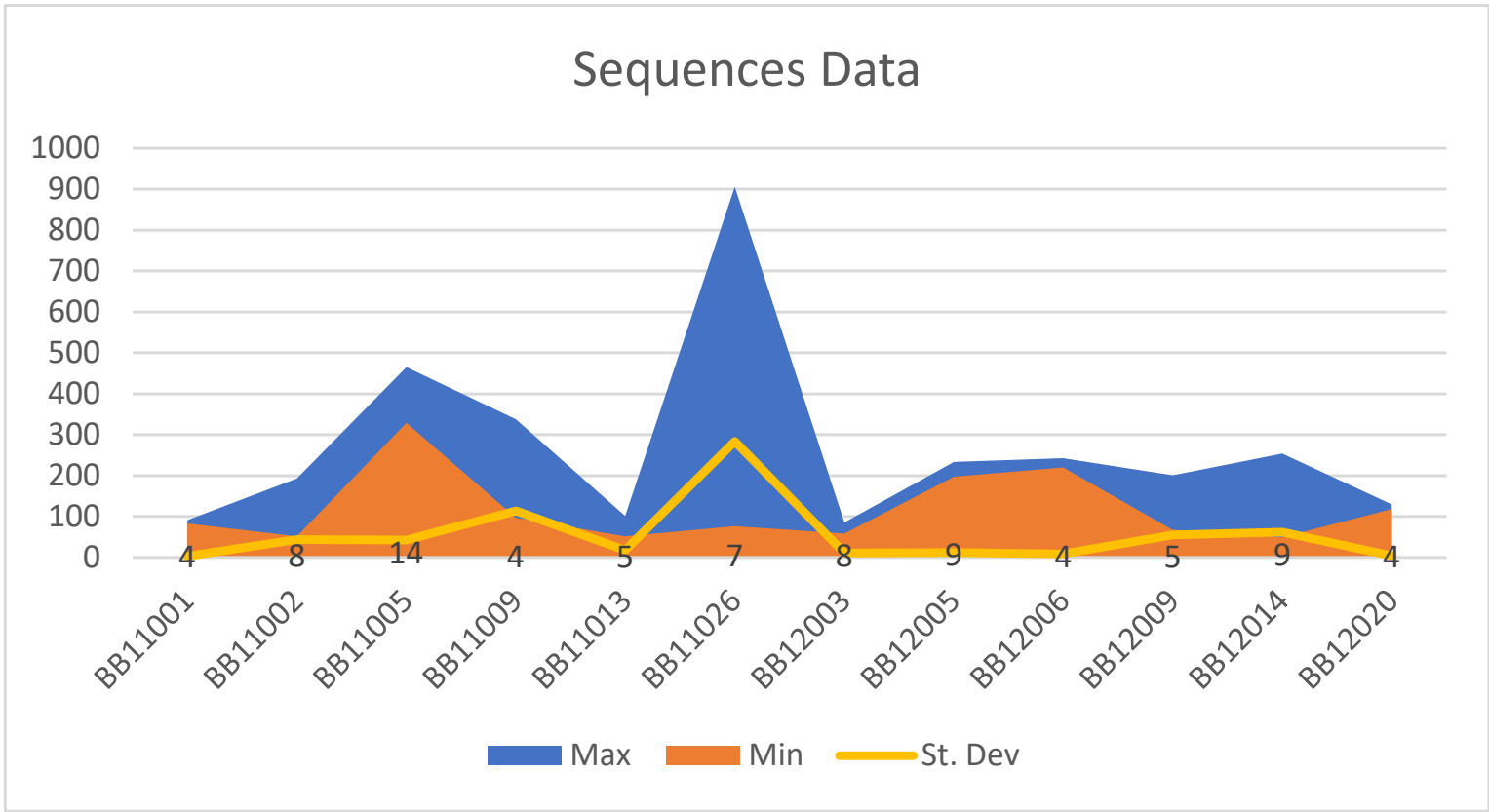
Feasibility Constraints

- *lt*: infeasible if # of 0's $<$ sequence length
- *lt-gt*: infeasible if # of 0's $\neq$ sequence length

Using *lt*, if 0 value exceeds characters of sequence, replace remaining 0's with trailing indels.

Experimental Sequences

All sequences were obtained from BaliBase [16].



Experimental Format

For AMPSO and BPSO, the velocity coefficient, objective function weights, and swarm size are changed.

For AM-MGPSO and B-MGPSO, only the velocity coefficient and swarm size are changed.

Introspective study is first completed, followed by comparison between PSO variants (on their superior parameters).

Then, additional comparisons with ProbCons are made.

Introspection: BPSO and B-MGPSO

Both are infeasible immediately after initialization.

BPSO

Most alignments when $\{w_1 : 1, w_2 : 0\}$; least inserted indels when $\{w_1 : 0, w_2 : 1\}$.

Largest swarm size of 50 performed best on both criteriums.

B-MGPSO

Most alignments and best archive quality when $\{n_1 : 40, n_2 : 20\}$.

Most trailing indels when $\{n_1 : 20, n_2 : 40\}$.

Introspection: AMPSO

Weights

- $\{w_1 : 1, w_2 : 0\}$ best for alignments.
- $\{w_1 : 0, w_2 : 1\}$ best for trailing indels.

Swarm Size

- $n = 50$ is best.

Superior Coefficients

- $\omega : 0.709815031402303 4$
  $c_1 : 1.6788775458244407$
  $c_2 : 0.899446463381824$
- $\omega : 0.7861878289341226$
  $c_1 : 1.2489605895810598$
  $c_2 : 1.211314077689869$

Introspection: AM-MGPSO

Swarm Size

- $\{n_1 : 20, n_2 : 40\}$ best for trailing indels
- $\{n_1 : 40, n_2 : 20\}$ best for alignments and archive quality

Superior Coefficients

- ω : 0.7499142729146882
  c_1 : 0.6484731483137582
  c_2 : 1.8292436916244121
  c_3 : 0.4718894923887376
- ω : 0.600264926750806 1
  c_1 : 1.39119402217411
  c_2 : 1.7645977300331588
  c_3 : 0.390304220116424 2
- ω : 0.268990284735547
  c_1 : 1.5805667363521902
  c_2 : 1.2012673626684758
  c_3 : 1.1353123776616685

Inter-algorithm Comparison - 1

BPSO and AMPSO

- For both alignments and inserted indels, AMPSO better on all sequence sets.
- AMPSO $>$ BPSO.

BPSO and B-MGPSO

- Alignments: BPSO better on $7/12$ sequence sets.
- Inserted Indels: BPSO better on $8/12$ sequence sets.
- BPSO $>$ B-MGPSO.

BPSO and AM-MGPSO

- For both alignments and inserted indels, AM-MGPSO better on all sequence sets.
- AM-MGPSO $>$ BPSO.

Inter-algorithm Comparison - 2

AM-MGPSO and B-MGPSO

- For alignments, inserted indels, and archive quality, AM-MGPSO better on all sequence sets.
- AM-MGPSO $>$ B-MGPSO.

AMPSO and B-MGPSO

- For both alignments and inserted indels, AMPSO better on all sequence sets.
- AMPSO $>$ B-MGPSO.

AMPSO and AM-MGPSO

- Alignments: AMPSO better on 5/12 sequence sets, AM-MGPSO better on 1.
- Inserted Indels: no algorithm significantly better.
- AMPSO $>$ AM-MGPSO.

Inter-algorithm Comparison - 3

Overall Rankings:

B-MGPSO $<$ BPSO $<$ AM-MGPSO $<$ AMPSO

ProbCons Comparison - AM-MGPSO

	Inserted Indels			Alignments		
Sequence Set	Mean	Std. Dev.	ProbCons	Mean	Std. Dev.	ProbCons
1	**12.57**	3.91	39	112.2	4.48	**146**
2	**23.17**	7.14	1608	243.07	3.88	**300**
3	**36.97**	11.28	6346	3002.63	11.03	**3399**
4	**19.03**	11.22	1460	**164.53**	3.93	162
5	**8.3**	2.94	277	**115.53**	2.33	112
6	**12.8**	4.1	5446	301.63	4.16	**392**
7	**5.37**	2.13	174	301.8	5.57	**362**
8	**12.43**	9.87	531	911.4	12.2	**1308**
9	**5.83**	2.12	52	355.17	4.68	**492**
10	**11.73**	9.06	727	147.0	2.91	**202**
11	**14.03**	5.98	1717	304.27	3.75	**414**
12	**23.3**	10.37	62	148.7	3.93	**216**

ProbCons Comparison - AMPSO

	Inserted Indels			Alignments		
Sequence Set	Mean	Std. Dev.	ProbCons	Mean	Std. Dev.	ProbCons
1	**33.17**	12.9	39	113.8	5.09	**146**
2	**60.9**	24.26	1608	245.0	4.98	**300**
3	**67.3**	29.11	6346	3003.7	14.64	**3399**
4	**29.33**	11.25	1460	**169.37**	5.57	162
5	**16.5**	6.76	277	**114.27**	2.32	112
6	**29.17**	16.06	5446	301.87	3.68	**392**
7	**6.23**	2.4	174	303.13	4.42	**362**
8	**21.13**	14.26	531	916.33	8.60	**1308**
9	**9.53**	4.58	52	357.83	8.79	**492**
10	**20.33**	11.71	727	148.47	2.76	**202**
11	**26.3**	12.16	1717	306.6	4.26	**414**
12	**40.93**	16.78	62	152.87	6	**216**

Summary

- Two forms of MSA problem were used: single-objective and multi-objective.
- AMPSO and BPSO are employed because they use a discrete search space, and were both respectively merged with MGPSO (AM-MGPSO and B-MGPSO).
- Introspective studies are carried out to inspect differences between parameter values, followed by inter-algorithm comparisons.
- AMPSO and AM-MGPSO are contrasted with ProbCons (state-of-the-art), as they were found superior from the inter-algorithm comparisons.
- Both AMPSO and AM-MGPSO are found to have lesser amounts of aligned characters than ProbCons.

References - 1

[1] Thompson, Julie D., et al. (1994)
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
Nucleic acids research 22.22, 4673-4680.

[2] Subramanian, Amarendran R., et al. (2008)
DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment.
Algorithms for Molecular Biology 3.1, 1-11.

[3] Lassmann, Timo, et al. (2005)
Kalign–an accurate and fast multiple sequence alignment algorithm.
BMC bioinformatics 6.1, 1-9.

[4] Katoh, Kazutaka, et al. (2013)
MAFFT multiple sequence alignment software version 7: improvements in performance and usability.
Molecular biology and evolution 30.4, 772-780.

[5] Edgar, Robert C. (2004)
MUSCLE: multiple sequence alignment with high accuracy and high throughput.
Nucleic acids research 32.5, 1792-1797.

References - 2


[6] Notredame, Cédric, et al. (2000)
T-Coffee: A novel method for fast and accurate multiple sequence alignment.
Journal of molecular biology 302.1, 205-217.


[7] Do, Chuong B., et al. (2005)
ProbCons: Probabilistic consistency-based multiple sequence alignment.
Genome research 15.2, 330-340.


[8] Thompson, Julie D., et al. (1999)
A comprehensive comparison of multiple sequence alignment programs.
Nucleic acids research 27.13, 2682-2690.


[9] Thompson, Julie D., et al. (2011)
A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives.
PloS one 6.3, e18093.


[10] Notredame, Cédric. (2002)
Recent progress in multiple sequence alignment: a survey.
Pharmacogenomics 3.1, 131-144.

References - 3

[11] Lalwani, Soniya, et al. (2013)
A review on particle swarm optimization variants and their applications to multiple sequence alignment.
Journal of Applied Mathematics and Bioinformatics 3.2, 87.

[12] Kennedy, James, and Russell Eberhart. (1995)
Particle swarm optimization.
Proceedings of ICNN'95-international conference on neural networks. Vol. 4. IEEE.

[13] Kennedy, James, and Russell C. Eberhart. (1997)
A discrete binary version of the particle swarm algorithm.
International conference on systems, man, and cybernetics. Computational cybernetics and simulation. Vol. 5. IEEE.

[14] Pampara, Gary, et al. (2005)
Combining particle swarm optimisation with angle modulation to solve binary problems.
2005 IEEE congress on evolutionary computation. Vol. 1. IEEE.

References - 4

[15] Scheepers, Christiaan. (2017)
Multi-guided particle swarm optimization: A multi-objective particle swarm optimizer.
Dissertation: University of Pretoria.

[16] Thompson, Julie D., et al. (2005)
BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark.
Proteins: Structure, Function, and Bioinformatics 61.1. 127-136.