Discovering DNA Motifs Using Particle Swarm Optimization with Sequence Preprocessing

NCSU Genomic Sciences Functional Genomics | Bioinformatics
NC STATE UNIVERSITY Genomic Sciences Graduate Program

Grant Billings and Karthik Sanka | CSC 530: Group 2

NC STATE UNIVERSITY

1. Introduction

Sequence motifs arise from evolutionary processes and are known to have many biological functions, like transcription factor-binding sites. Finding motifs given a set of DNA sequences is challenging due to the large search space and imperfect conservation of nucleotides at each position of the motif. We overcome this challenge by using particle swarm optimization.

2. Algorithm

- PSO uses a particle's current state, personal best state, global best state, and a random state to determine its future state as proposed by Lei and Ruan (2010).
- The particle states represent proposed consensus motifs, which are searched against the input DNA sequences for the best match in each sequence.
- A check-shift operation is used to search the neighbourhood of a position to avoid low scores due to partial overlap with a high scoring motif.
- In this project, we improved upon the biggest weakness of existing PSO algorithms in motif discovery, which is the memory consumption and time associated with searching for best matches in each DNA sequence (**Figure 1**).

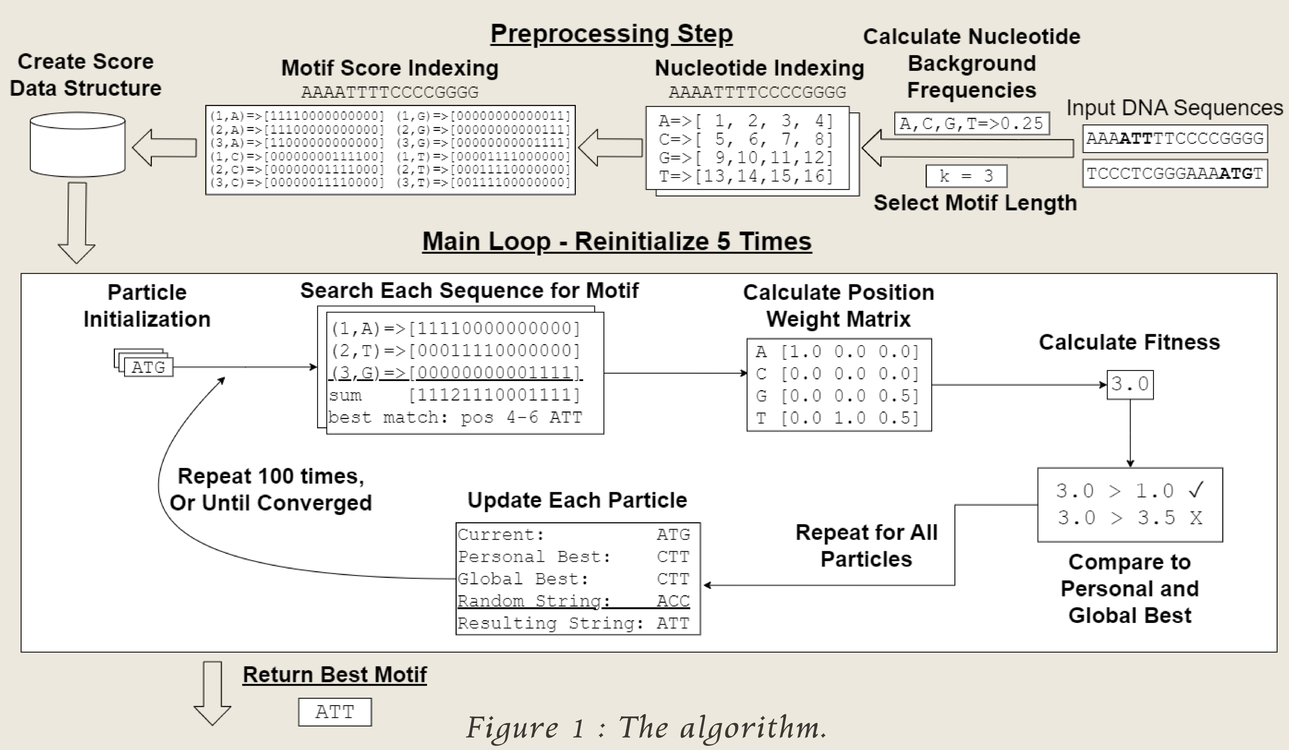


Figure 1 : The algorithm.

3. Experiments

- Synthetic Data was used to run experiments and analyze the performance of the algorithm.:
  - 5 restarts with 100 particles and 100 updates; length of motif = 14, number of mutations = 2, length of DNA sequences = 500, number of DNA sequences = 100
- Real data from Ando et al (2021), representing 1000-bp upstream of the transcription start site of 314 cotton fiber genes, was used to test the algorithm.

4. Results

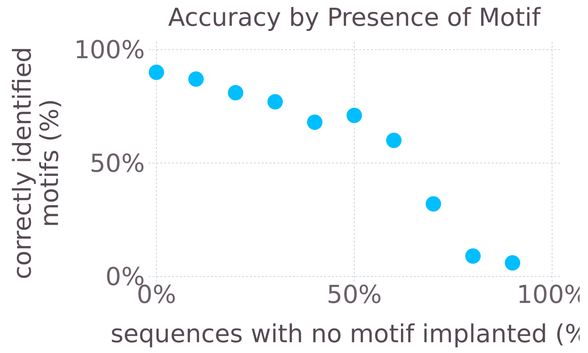


Figure 2: Accuracy of PSO with some sequences lacking the implanted motif.

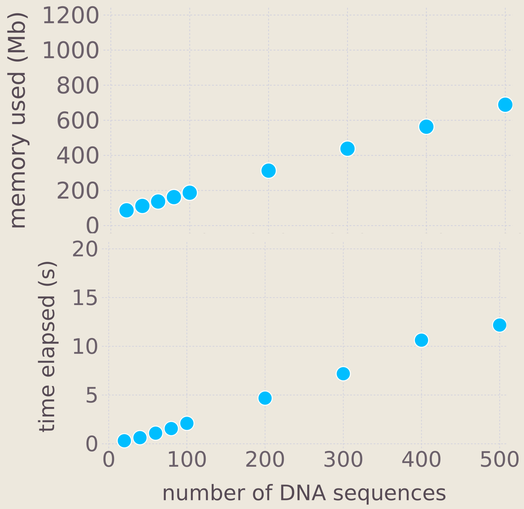
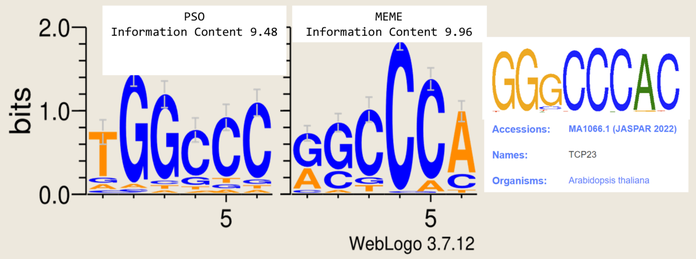


Figure 3: Memory usage (top) and run time (bottom) of PSO.



	accuracy (%)	time (s)	mem (Mb)
PSO	76%	2.2	187.2
MEME	100%	4.9	12.6

Figure 4: Comparison of PSO to MEME. Motifs discovered in cotton data (left) and run time with simulate data (right).

5. Discussion and Conclusions

On simulated data, our PSO implementation was fast, memory efficient, and usually produced the expected output (**Figure 2** and **3**). Testing showed that 5 reinitializations of 100 particles, each with 100 updates), had a good balance of speed and accuracy. The most surprising finding was that PSO continued to work well, even if most sequences lacked the implanted motif. This suggests the applicability of the algorithm to real datasets, where all provided sequences may not necessarily contain the motif of interests.

We tested our implementation of PSO against the standard package MEME on real cotton promoter data (**Figure 4**). A search for the motif sequence against a promoter motif database found TCP23 was a match in *Arabidopsis*. The TCP transcription factors are a biologically plausible candidate for regulating the differentiation of cotton seed hairs into spinnable fibers.

PSO is an effective approach for identifying DNA sequence motifs. Our introduction of a pre-processing step is a significant advance in DNA motif discovery which could be used to make many heuristics that require identification of best matches run considerably faster. Future research in this area could focus on:

1. Identifying weights that improve convergence rate without sacrificing accuracy;
2. Using the pre-processed data structures directly for choosing better initialization states;
3. Doing additional, in-depth comparison with other heuristic algorithms; and
4. Adding the ability to parallelize particle simulations across multiple threads or cores simultaneously.