Bees? DNA Motif Discovery with Alternating Global-Local Search

- CSC 530: Group 2 Project Proposal -

Grant Billings and Karthik Sanka

09/29/2022

Abbreviations

(l,d): an implanted motif of length l with d random changes; **DNA**: Deoxyribonucleic Acid; **HMC**: Hamiltonian Monte Carlo; **MEME**: Multiple Expectation Maximization for Motif Elicitation; **PSO**: Particle Swarm Optimization;

1 Executive Summary

Living organisms have genomes that evolve randomly over time, with natural selection working to increase the frequency of functionally beneficial sequences over generations. Motifs are non-random nucleotide sequences that in many cases have been shown to have biological function in gene regulation. Detection of motifs in sets of sequences is challenging because random mutations make exact matching of sequences ineffective, and brute force methods are very slow. The most popular software package for motif discovery is MEME, which works well but slows down significantly if many query sequences are provided. Motif discovery across large data sets has become an important step in genome analysis. Software that can efficiently mine these sequences for motifs is needed.

Nature-inspired algorithms have promise for DNA motif discovery since they broadly allow for efficient exploration of potential motifs while allowing good solutions to learn from each other. We propose use of Particle Swarm Optimization with Hamiltonian Monte Carlo (PSO-HMC) in alternating cycles of global and local search to quickly find motifs. Our algorithm will be tuned using implanted motifs in simulated data, tested on previously characterized benchmarking data, and finally applied to detect sequence motifs in a cotton genome dataset. Sensitivity, specificity, and running time will be used to compare performance between our software and other widely used alternatives. At the end of the semester, we will present our findings in a poster. We will also create an animation showing the algorithm running and share it upload it to Wikipedia so others can gain a visual intuition for how PSO-HMC works. The PSO-HMC method is a promising tool for DNA motif discovery.

2 Abstract

3 Prior Work

Example citation [1]

4 Project Description

4.1 Data

- Simulated (l,d) implanted motifs

- Actual annotated *cis* element plant dataset from Sabastian and Contreras-Moreira, Bioinformatics, Nov 2013
- Upstream sequence from fiber-specific genes in Ando et al, BMC Genomics, April 2021

4.2 The Algorithm

4.3 Implementation

- Implementation in Julia
- Particle Swarm Optimization with *inertia* and *social attraction* parameters
- Hamiltonian Monte Carlo

4.4 Experiments

4.5 Evaluation and Statistics

4.6 Deliverables

- Poster
- Graphical demonstration of algorithm
- Public availability of code on GitHub

4.7 Anticipated Problems and Solutions

5 Timeline

References

[1] Iztok Fister Jr et al. “A brief review of nature-inspired algorithms for optimization”. In: *arXiv preprint arXiv:1307.4186* (2013).