COMP 293J: Motif Discovery Project Report

1. Project Topic and Biological Context

My project focuses on discovering DNA motifs in the upstream regions of DosR-regulated genes in Mycobacterium tuberculosis. The DosR regulon plays a crucial role in bacterial dormancy and survival under stress. Identifying conserved motifs in the promoter regions of these genes can reveal potential regulatory elements and enhance our understanding of gene expression control in M. tuberculosis.

We aimed to detect consensus motifs using a custom Genetic Algorithm (GA) and compare its performance with traditional bioinformatics methods like Greedy Search, Randomized Motif Search, and Gibbs Sampling.

2. Initial Plan

The initial plan was to:

- Use upstream sequences of DosR-regulated genes as input.
- Implement a genetic algorithm to evolve k-mers based on fitness.
- Compare GA results with existing motif-finding techniques.
- Extend findings by scanning the full genome for the discovered motifs.

Tools and technologies: Python (custom scripts), DosR.txt for sequences, and M. tuberculosis genome file for full scanning.

3. Challenges and How They Were Addressed

Challenge 1: Slowness and Parameter Tuning in GA

- Solution: Iteratively tested combinations of parameters (e.g., mutation rate: 0.05, population size: 100, generations: 1000). Logging intermediate results helped in monitoring convergence trends.

Challenge 2: Measuring Motif Fitness

- Solution: Used the inverse of Hamming distance summed over all sequences as the fitness score, ensuring biologically relevant similarity measurement.

Challenge 3: Variability in Randomized Methods

- Solution: Repeated each method multiple times (up to 1000 for GA and 50 outer loops for Gibbs) to extract the best-performing motif sets.

4. Approach and Justification of Design Decisions

Genetic Algorithm:

- Representation: Each k-mer (length = 15) was encoded as a chromosome.
- Fitness: Defined as 1 / (total Hamming distance +) from all DNA sequences.
- Operators: Selection (Roulette Wheel), Single-point Crossover, 5% Mutation rate.
- Population: 100 chromosomes, 1000 generations.

Classical Methods:

- Greedy Motif Search: Profile matrix construction.
- Randomized Motif Search: Profile-guided updates from random start.
- Gibbs Sampling: Profile-based probabilistic re-sampling over 2000 inner steps and 50 repeats.

5. Results

Best GA Motif: ACTTCCGGCCCTAAC (Fitness: 0.02439)

GA Motif Score (Consensus Set): 88

Greedy, Randomized, Gibbs Scores: 35

Motif Comparisons:

- Traditional algorithms converged on motifs like GGACTTCAGGCCCTA with identical motif set scores, confirming the robustness of the motif signal.
- The Genetic Algorithm discovered novel motifs that retained the same core structure, demonstrating that the underlying biological signal was preserved.
- Both traditional and GA methods consistently discovered motifs from **conserved regions**, indicating biologically meaningful convergence.
- Genomic Occurrence: GA consensus motif GGGACTTTCGGCCCT appeared at index 3304, 88117, etc., highlighting potential regulatory hotspots.

6. Conclusions

The Genetic Algorithm successfully identified biologically meaningful motifs that align with results from traditional methods.

Despite their exploratory and stochastic nature, GA results consistently retained the core motif signal seen in deterministic algorithms. This indicates that **key regulatory signals were preserved** across approaches.

Additionally, convergence of motifs to **similar conserved upstream regions** in both traditional and GA approaches supports the biological relevance of the discovered patterns.

While GA offers broader motif diversity and optimization potential, it is more sensitive to parameter tuning and computationally intensive.

7. Future Directions

- Apply multi-objective GA to balance fitness and motif frequency across conditions.
- Explore IUPAC-degenerate motifs to generalize patterns.
- Validate motifs against ChIP-Seq DosR data.
- Extend framework to other transcription factors across Mycobacterium species.
- Experiment with different selection mechanisms (e.g., tournament selection), crossover strategies (multi-point), and adaptive mutation rates.
- Incorporate profile matrices or weight matrices to probabilistically generate offspring during the mutation step for biologically-informed evolution.