BRAUDE College of Engineering, Karmiel

Capstone Project Phase B

A new generation of DNA hidden repeats detection algorithm and its application for isochore research

Project Number: 25-2-R-8

A Statistical Approach to Hidden DNA Repeat Detection

Students: Fatmeh Zoabi, Khalil Mansour || Supervisor: Dr. Zakharia Frenkel

Background & Motivation

Hidden tandem and periodic DNA repeats often escape classical tools, especially when weak or noisy.
These patterns are biologically meaningful and linked to local composition (e.g., GC content).

Goal:

Develop a robust, statistically grounded algorithm to detect hidden DNA repeats under noise, mismatches, and local background variation.

System Objectives

- Detect hidden, low-complexity DNA repeats.
- Quantify repeat strength statistically.
- Combine statistical and consensus detection.
- Validate repeats with a composition-preserving null model.
- Produce interpretable genomic outputs.
- Scale to long genomic sequences.

Challenges and Solutions

- Weak or mutated repeats → consensus repeat-words.
- Unknown period → test multiple periods.
- Varying composition → composition-preserving null model.
- Random false positives → statistical significance tests.
- Genome-scale analysis → segment-based processing.

Core Idea

We segment DNA into fixed-length windows and search for a representative word (k-mer) that best explains the segment.

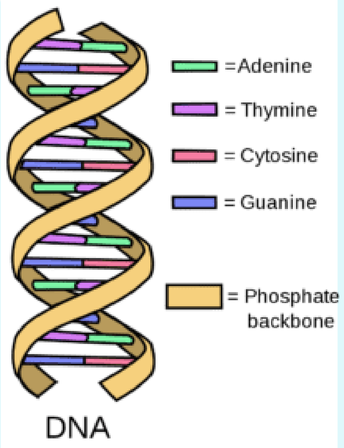
Two complementary pipelines are proposed:

- **AVG pipeline** – selects the representative word by maximizing average positional matches.
- **P-value pipeline** – selects the representative word by minimizing a statistically combined p-value.

Both pipelines are compared against carefully constructed null models.

Key Algorithmic Innovations

1. Representative Word with Mismatches
   - Allow mismatches when scoring k-mers.
   - Improves robustness beyond exact periodicity.
2. Canonical Rotation Invariance
   - Normalize repeat words by canonical rotation.
   - Enables shift-invariant detection.
3. Statistical Significance (P-values)
   - Use binomial tests per position.
   - Combine evidence via Fisher's method.
4. Composition-Preserving Null Model
   - Generate local-composition-matched random DNA.
   - Compare real vs. null using identical pipelines.



Experimental Results

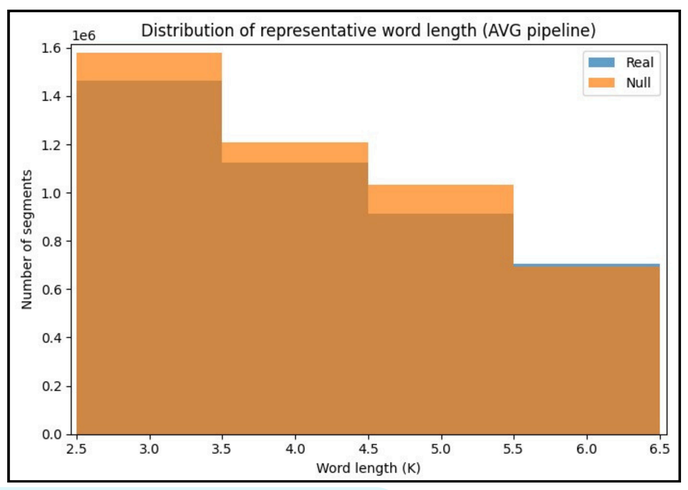
1. Statistical Significance
   - Real DNA shows stronger significance than null.
   - Clear separation in distributions.
2. AVG vs. P-value Pipelines
   - AVG captures smooth repeat patterns.
   - P-value detects sharp, extreme repeats.
   - Methods are complementary.
3. Word Length (k)
   - AVG favors longer k-mers.
   - P-value adapts to optimal k.
4. Agreement Analysis
   - Exact word matches are rare.
   - Count-based agreement remains high.

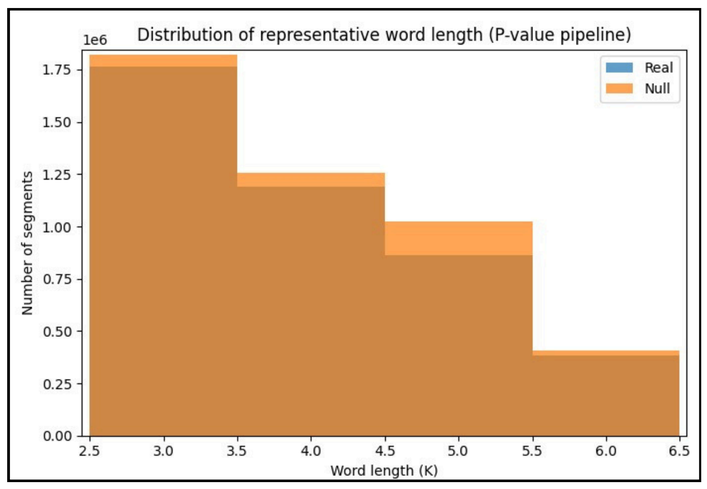
Project Success Metrics

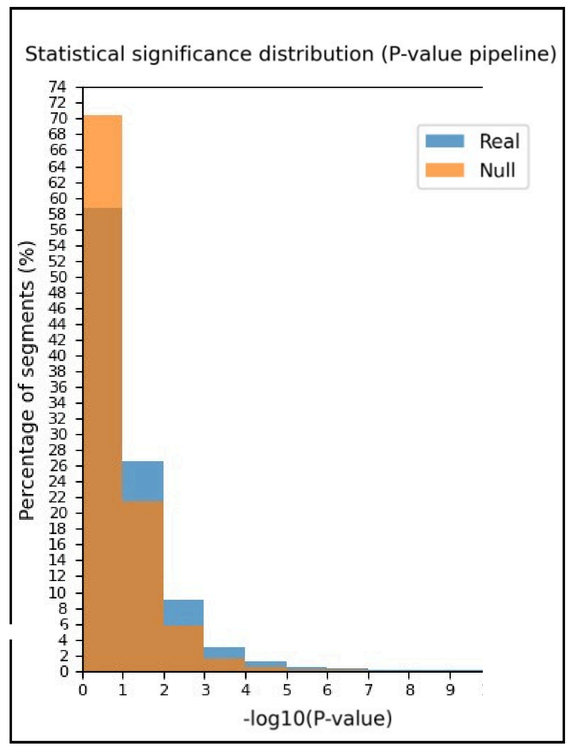
- Real DNA shows stronger signals.
- Repeats clearly separated from noise.
- Genome-scale scalability.

Technologies & Tools

- C++ (VS Code)
- Python
- Other:







Conclusions

- Compared to earlier approaches, the proposed model improves robustness and repeat sensitivity.
- Reveals repeats missed by classical methods.
- Scalable, interpretable, and biologically relevant.

Future Work

- Adaptive segmentation lengths.
- Improve P-value and AVG using mismatch tolerance.
- Refine the segment merging strategy.
- Improve computational complexity and efficiency.