Finding Transposable Elements in Genomic Data by Frequent Subsequent Detection

Chia-Hung Yang

Mining Frequent Patterns in DNA

- **Transposable Elements (TEs):**
  - Chunks of DNA sequence that are able to copy and move themselves to new positions in genomes.

- **Frequent Sequential Patterns Detection:**
  - Find sequential patterns from a group a sequences whose supports are not less than a pre-specified threshold.

- **Can we find TEs by mining frequent patterns in genomes?**

Existing Approaches

- **A-priori**
  - *Monotonicity*: If a sequential pattern is frequent, so are its prefix and suffix.
- **Position-based Method [1]**
  - If a prefix and a suffix instance are adjacent, it implies an instance of the joint sequential pattern.

ACTGGATTC

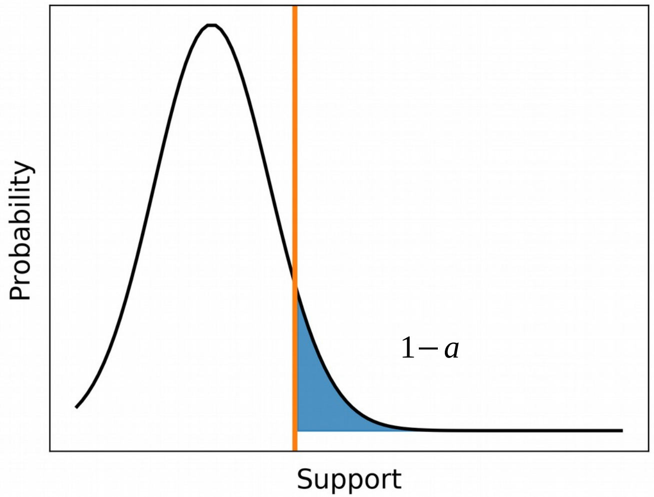
ACTGGATT
CTGGATTC

[1] Zerin *et al.*, IETE Tech. Rev. (2011)

Support Threshold of Significance

- **Null Model:**
  - Sequential patterns are drawn uniformly from all possibilities of the same length.

- **Hyperparameters:**
  - Number of base pairs ***n***
  - Minimum pattern length $\boldsymbol{l_{min}}$
  - Support lower bound $\boldsymbol{s_{min}}$
  - Confidence level ***a***



$$s_l = min\left\{ m \;\middle|\; \sum_{k=1}^{m} \binom{n-l}{k} p_l^{-k} (1-p_l)^{n-l-k} \geq a \right\}$$

$$p_l = 4^{-l}$$

Implemetation of Position-based Method

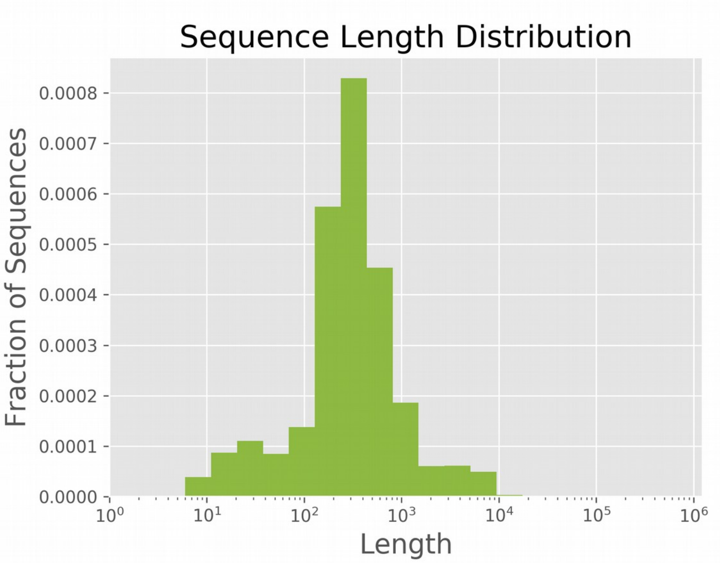
1. **Obtain support thresholds of significance and thus the range of pattern length of interests.**
2. **Find sequential patterns with supports not less than the lower bound:**
   - A) Obtain instances of short patterns and order them by position [2].
   - B) For each pattern length ***l***:
     - i. Find length-***l*** candidates from frequent patterns of length ***l-1***.
     - ii. Merge adjacent length-(***l-1***) instances if they form a candidate pattern, and calculate the supports.

[2] Tanvee *et al.*, Int. Journal of Comp. App. (2013)

Genomic Data

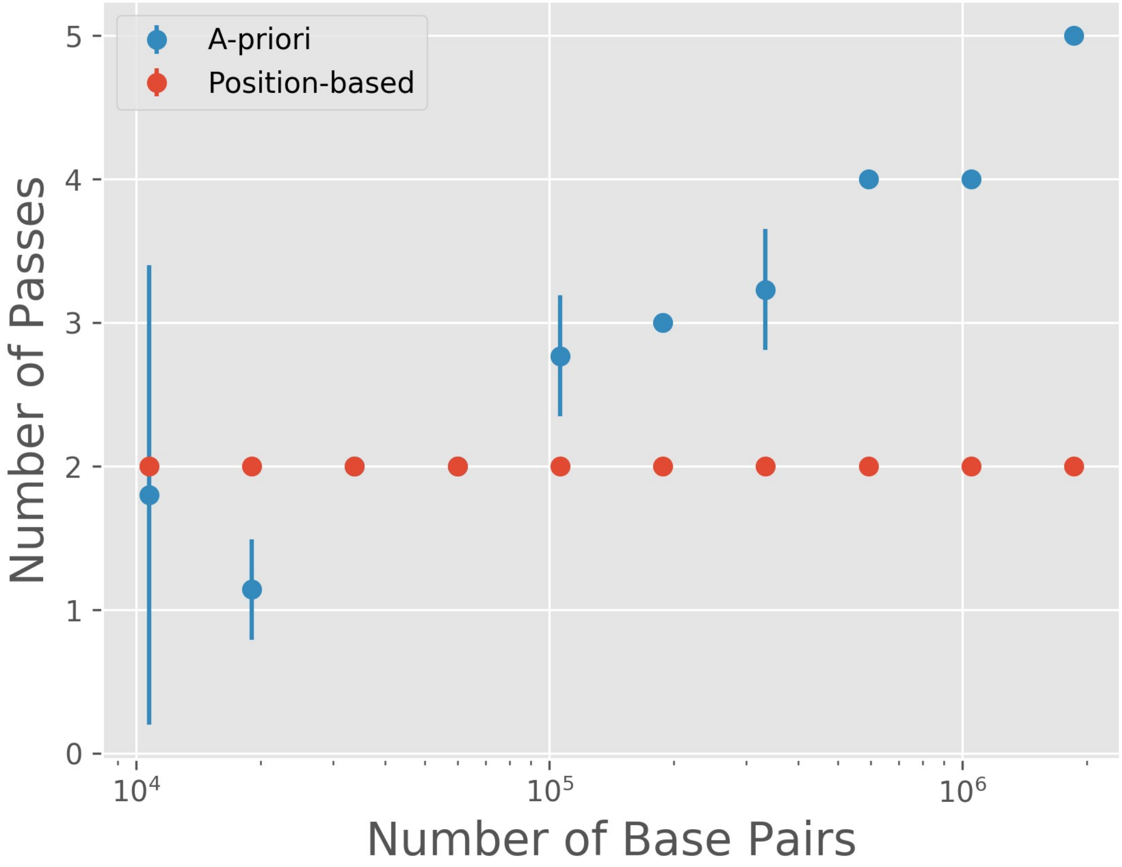
- **DNA sequences of *Arabdopsis thiliana* [3]**
  - ~120K sequences
  - Sequence length range from 6 to ~700K
  - Total ~340M base pairs

- **TE database: RepBase [4]**
  - ~500 sequences

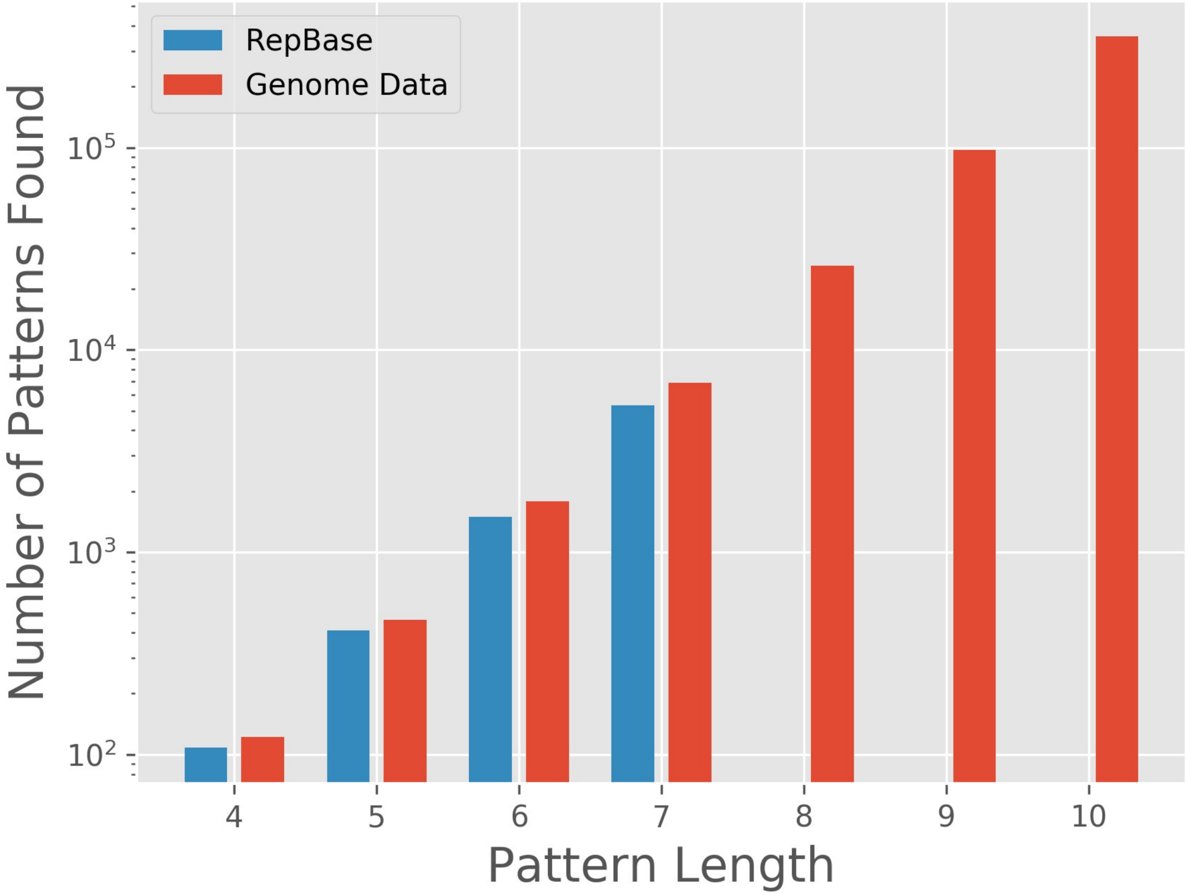


[3] Debladis *et al.*, BMC Genomics (2017)
[4] Bao, *et al.*, Mob DNA (2015)

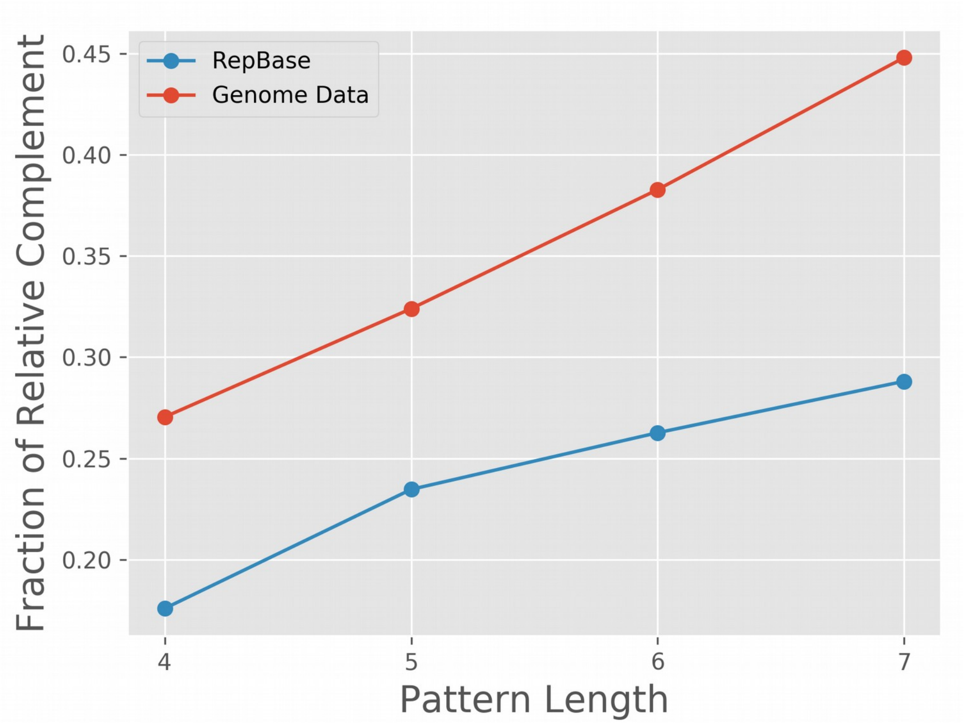
Computational Costs

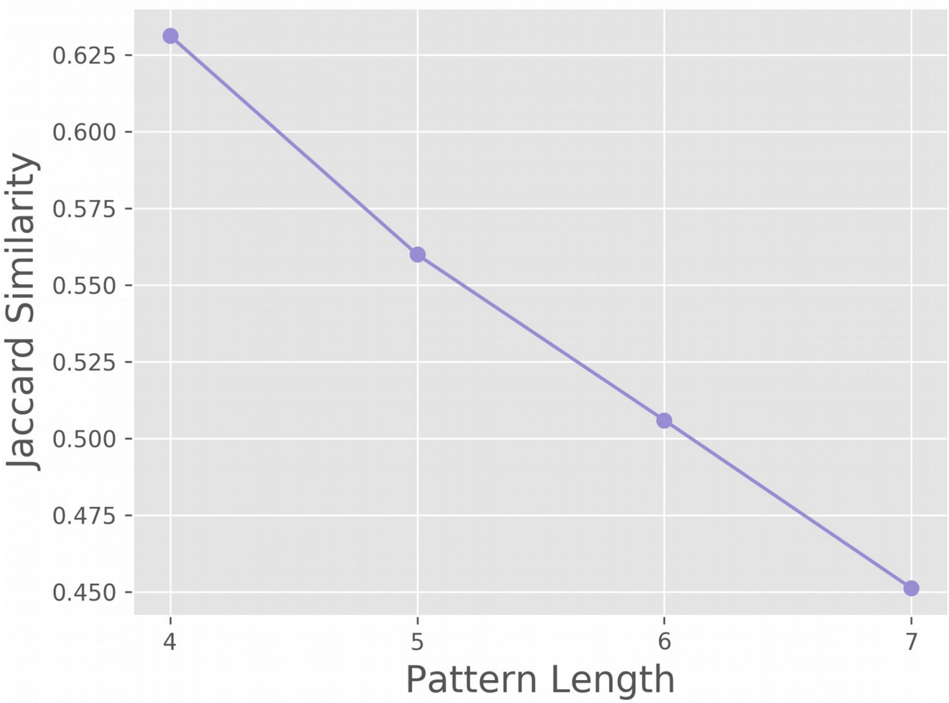


Frequent Patterns in Genome and TEs



Frequent Patterns in Geome and TEs





Summary

- Frequent sequential patterns detection through genome data recovers short patterns in transposable elements.
- Practical usage of mining frequent patterns in bioinformatics remains unclear, compared to popular probabilistic methods.