Optimizing Seed Selection in DNA Read Mapping

Sunny Nahar

Carnegie Mellon University

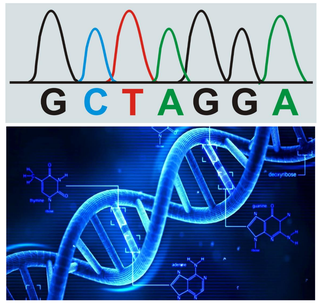
SAFARI

What is DNA Sequencing?

Genomes are composed of **base pairs**:

- *Adenine, Guanine, Cytosine, Thymine*.

The goal is to find the **sequence** of base pairs which compose the genome.

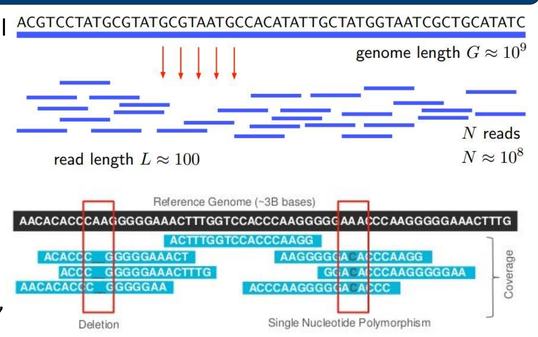


Why is this useful?

- Tracing evolution.
- Correlating genes with diseases.
- Forensics and identification.

How is Sequencing done?

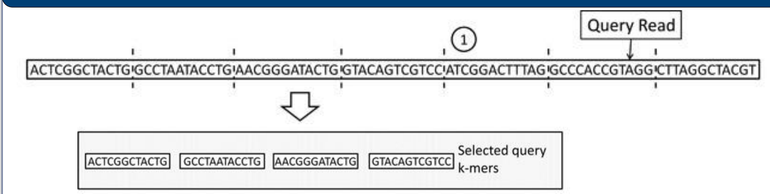
- DNA is split into small pieces called **reads**.
- Using a **reference genome**, reads are mapped to potential locations.
- Must account for errors in reads: **insertions**, **deletions**, and **substitutions**.



This problem is very **computationally challenging**:

- Billions of reads.
- Fuzzy string matching.
- Multiple mapping locations per read.
- Needs to work on commodity machines.

Past Research and Research Question



A method for mapping is **seed-and-extend:**

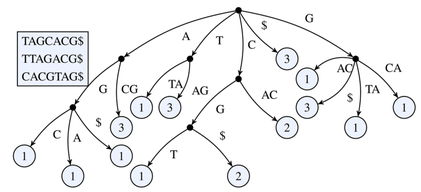
- Search **exactly** for small substrings (**seeds**) of the read.
- Use the seeds to find valid locations for the overall read.
- Complexity is frequency of seeds in reference.
- Used by leading mappers such as **Hobbes** and **FastHash**.

Question: Can we reduce seed frequency to improve speed?

- Develop new heuristics to choose seeds with the least frequency.
- Low complexity, memory efficient, cache efficient.

Hobbes Mapper

- Uses a **prefix trie** of the reference which stores seed frequency.
- Queries are **costly**.
- Finds set of N **equal length seeds** with lowest frequency.
- Limited as seed frequencies have large variation.

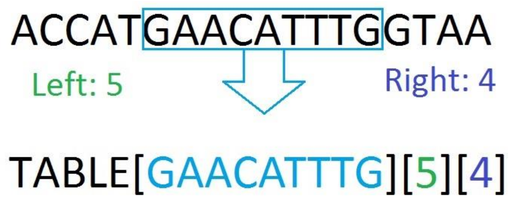


4 seeds of length 7:

Bidirectional Frequency Predictor

Idea: Create a data structure to **predict** frequency of seeds:

- Used to **minimize** reference trie queries.
- Predicts frequency given base seed, left and right extension.



Hobbes + Frequency Predictor

Combine Hobbes' seed selection and the frequency predictor:

- Called Bidirectional Hobbes (BDH).
- Use Hobbes to get initial fixed-length seeds.
- Use the predictor to modify seeds and lower frequency.

Results (4.5 Million Read Set)

Conclusion and Future Work

Combining Hobbes and the frequency predictor significantly reduced the sum of the frequencies of the reads:

- BDH **reduced the frequency sum about 7%** compared to Hobbes consistently for different length reads.
- BDH **performed much better** than other seed selection algorithms like Cheap K-mer selection and Threshold selection.
- The increase in seed selection time compared to Hobbes was only 12%, and this can be reduced further.

The predicted frequency from the Bidirectional Frequency Predictor was within 16% of the exact frequency, so using this is a viable method.

Future Work:

- Investigate additional methods to improve frequency.
- Improve runtime using parallel frameworks.

Acknowledgements

I would like to thank Prof. Onur Mutlu and Hongyi Xin for their help and advice on the project and for enabling this research.