LTR retrotransposon detection via deterministic finite automata

Author: Lucie Klímová
Supervisor: doc. Mgr. Lukáš Holík Ph.D.

MOTIVATION

- LTR retrotransposons make up a significant part of the human genome (8.3%)
- They can influence gene expression (the amount of protein that is synthesized)
- They are highly nested and therefore hard to detect

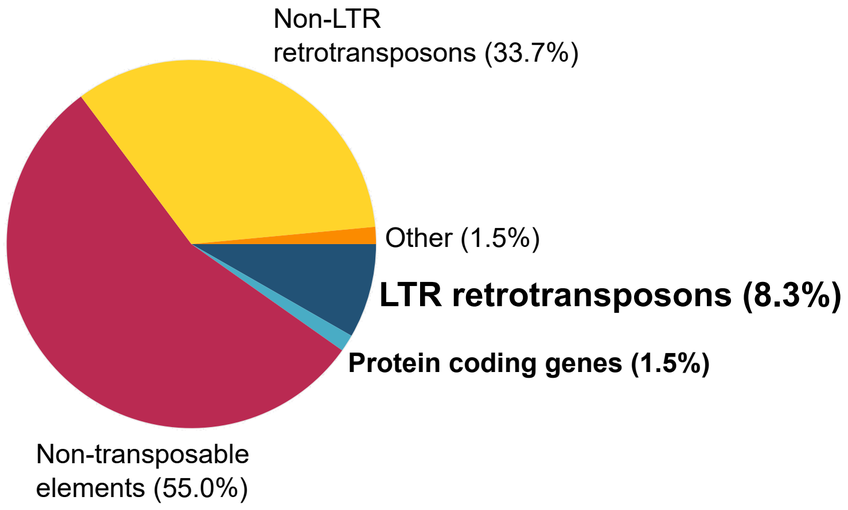

Figure 1 - Proportional representation of LTR retrotransposons in the human genome

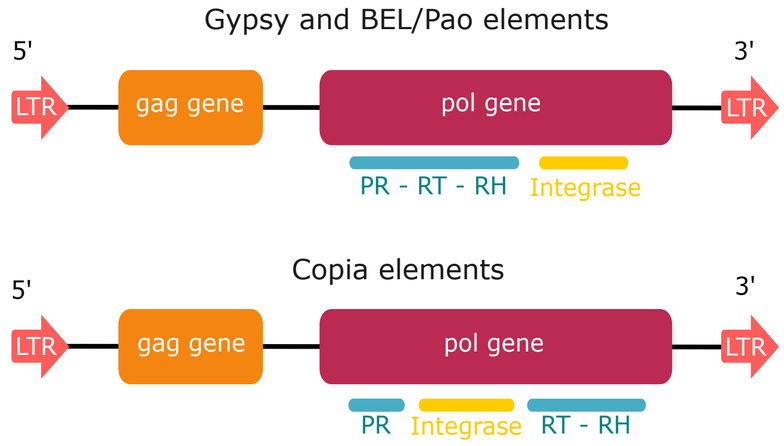

Figure 2 - Schematic structure of LTR retrotransposons

TE-GREEDY NESTER

- Detects even highly nested LTR retrotransposons
- Recursively removes the best matching LTR elements
- Due to the recursion, the algorithm appears relatively slow (80% of its runtime is taken up by BLASTX)

Figure 3 - Ilustration of transposon nesting

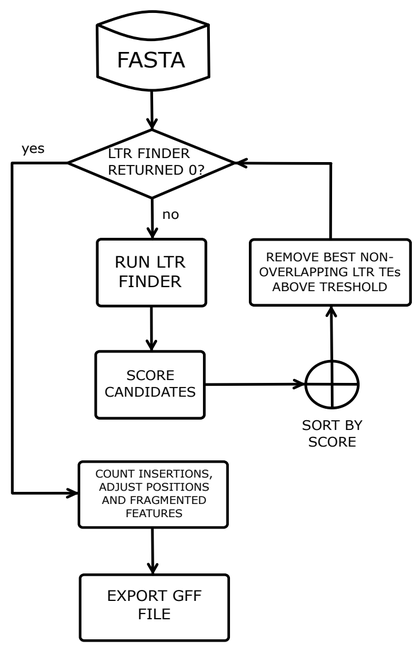

Figure 4 - Representation of the TE-greedy nester algorithm

PROFILE HMM

- Precisely models the desired sequence using match, insert and delete states
- Non-deterministic, time complexity $O(LM^2)$

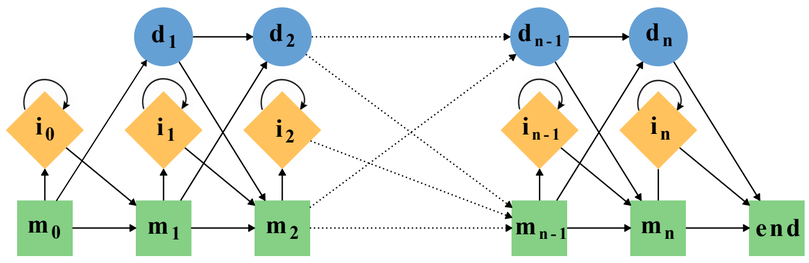

Figure 5 - Profile hidden Markov model structure

DOMAIN SEARCH USING DFAs

- The main idea is to transform a profile HMM into a deterministic model while minimizing the loss of accuracy
- Direct determinization is unfeasible, so the model must be simplified
- A profile HMM can be converted into a bounded counting automaton (BCA), which uses a counter to determine how many consecutive transitions below a threshold t were taken

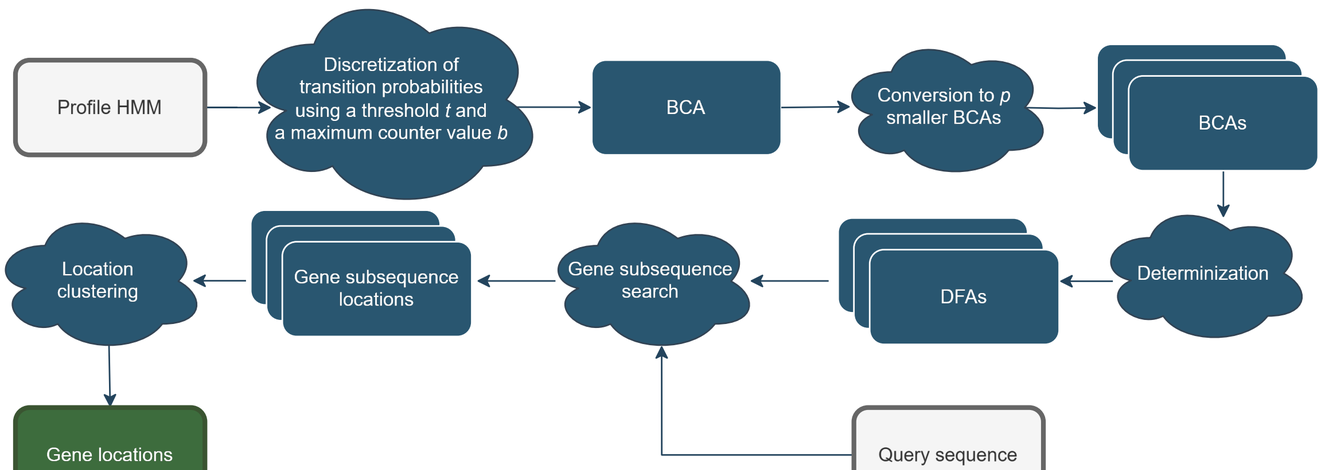

Figure 6 - Process of generating the deterministic model

RESULTS

Figure 7 - Execution time comparison

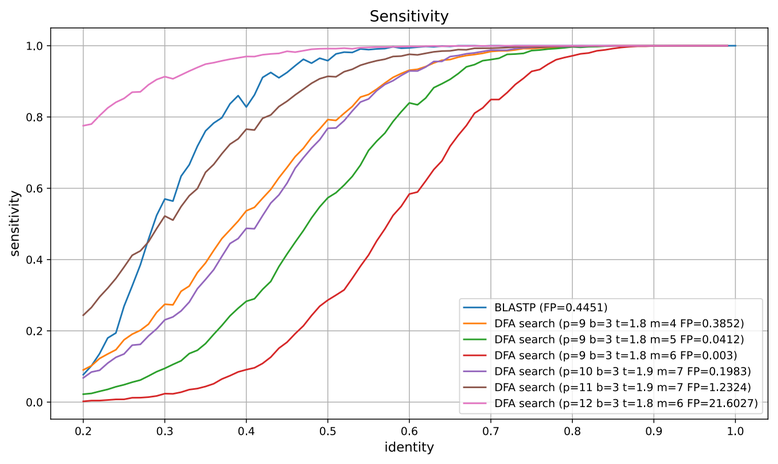

Figure 8 - Sensitivity comparison