LTR retrotransposons detection via Probabilistic 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 syntetized)
- They are higly nested and therefore hard to detect

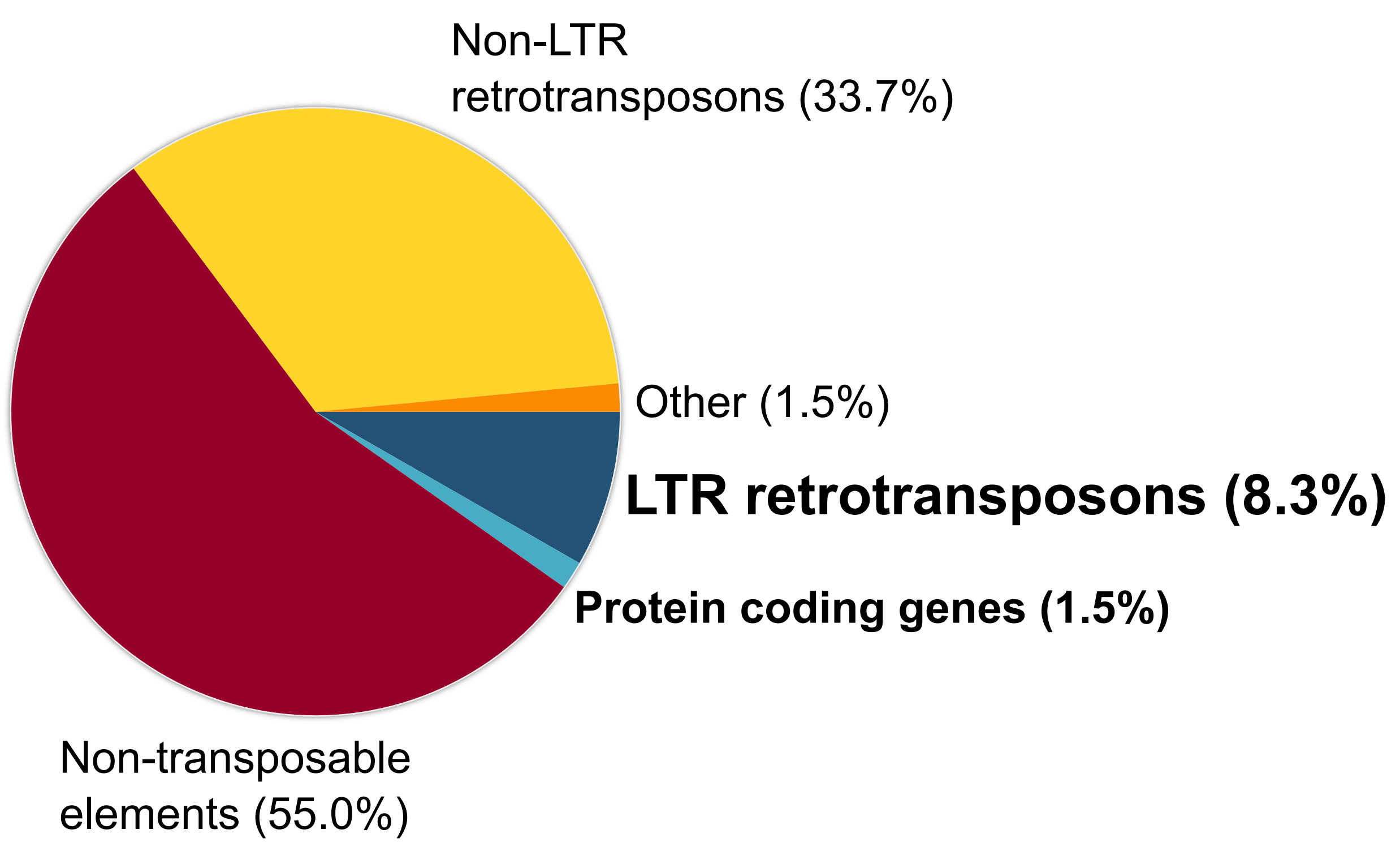

Figure 1 - Proportional representation of LTR retrotransposons in human genome

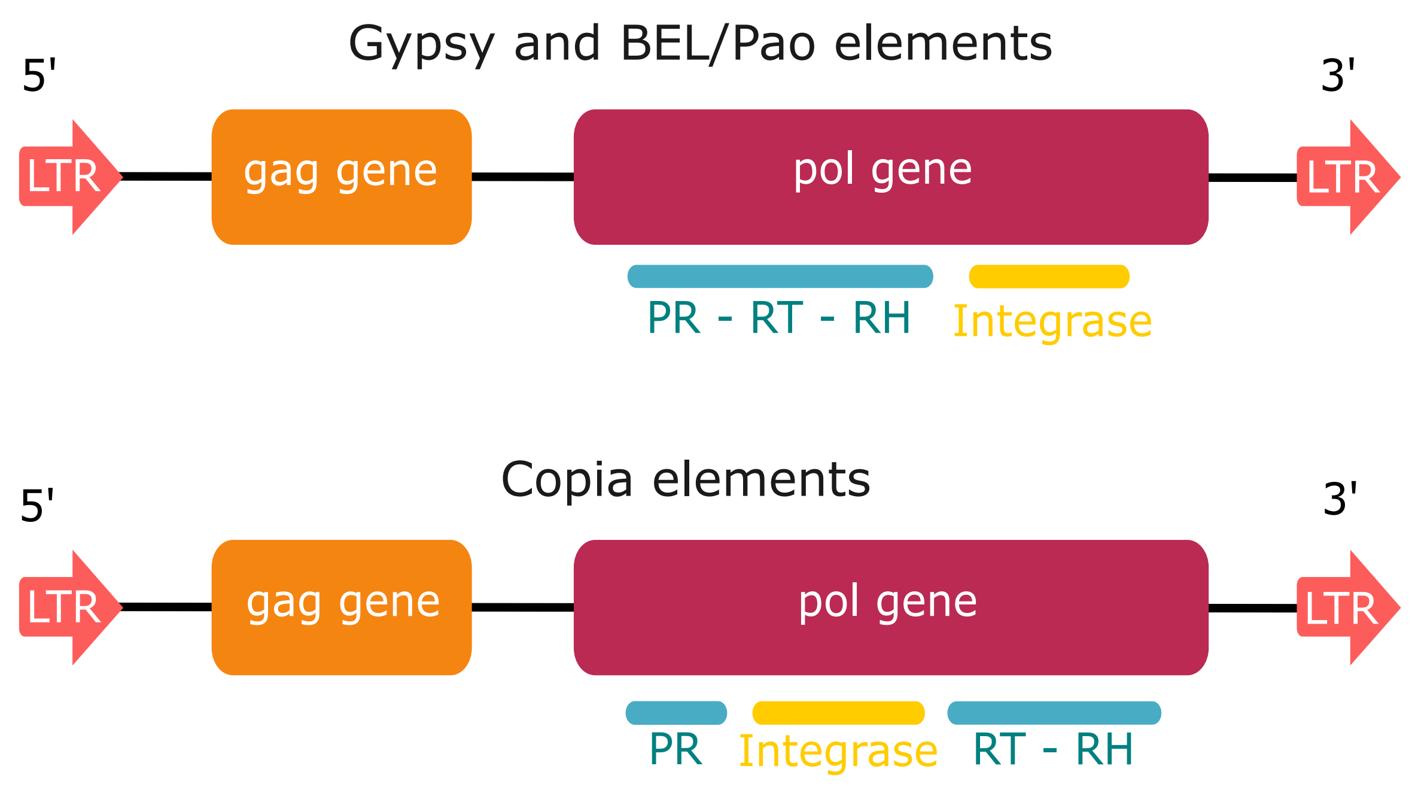

Figure 2 - Schematic structure of LTR retrotransposons

TE Greedy Nester

- Detects even higly nested LTR retrotransposons
- Recursively removes best matching LTR elements
- Due to the recursive calls appears relatively slow

Figure 3 - Ilustration of transposon nesting

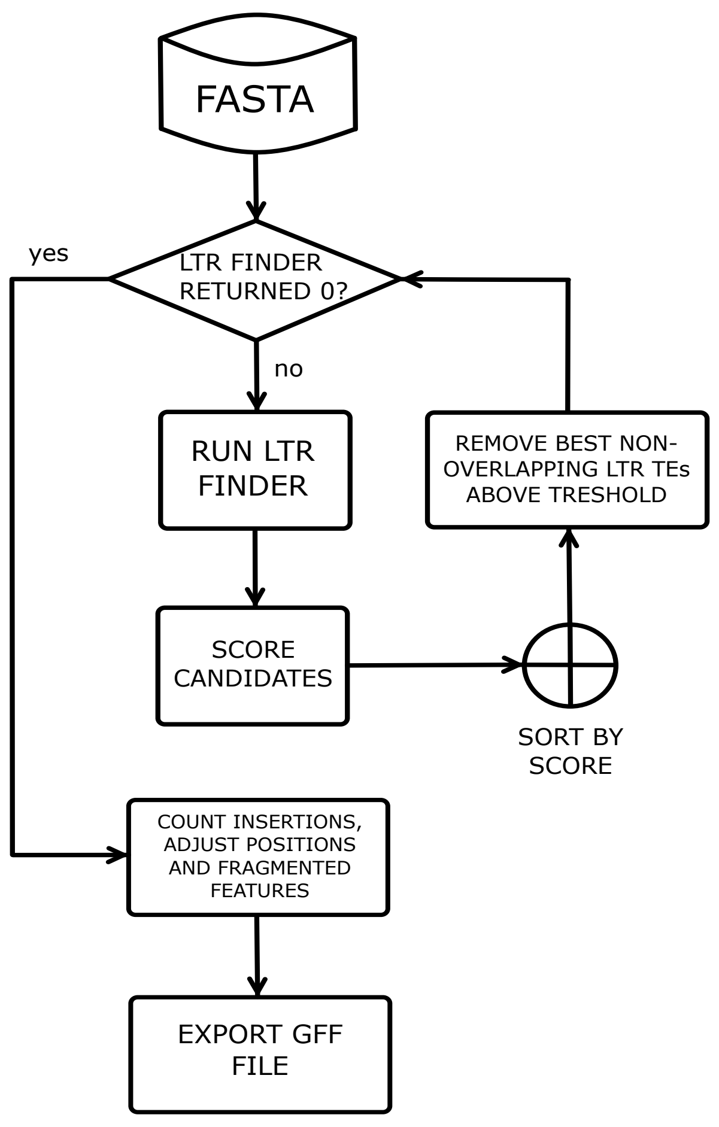
FASTA
LTR FINDER RETURNED 0?
yes
no
RUN LTR FINDER
SCORE CANDIDATES
SORT BY SCORE
REMOVE BEST NON-OVERLAPPING LTR TEs ABOVE TRESHOLD
COUNT INSERTIONS, ADJUST POSITIONS AND FRAGMENTED FEATURES
EXPORT GFF FILE

Figure 4 - Representation of the TE-greedy nester algorithm

Probabilistic Finite Automata

- Gene can be described by a pattern ➡ Finite Automata
- Randomness, simulating mutations, is introduced by merging similar states
- ALERGIA algorihm is used

$$\left|\frac{f_1}{n_1} - \frac{f_2}{n_2}\right| < \left(\sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_2}}\right) \cdot \sqrt{\frac{1}{2} \cdot log\left(\frac{2}{\alpha}\right)}$$

Inequation 1 - Used to determine whether to merge two states

Translation of amino acids

	Acidic	Neutral	Basic
POLAR	Asp Glu	Asn Ser Tyr Cys Gln Thr	Arg His Lys
NON-POLAR	Ala Ile Leu Pro Trp Val Gly Met Phe		

Table 1 - An example of amino acids equivalence classes

- The substitution of an amino acid with another amino acid from the same group may not lead to a significant change in the protein structure
- These mutations are much more common

Dealing with non-determinism

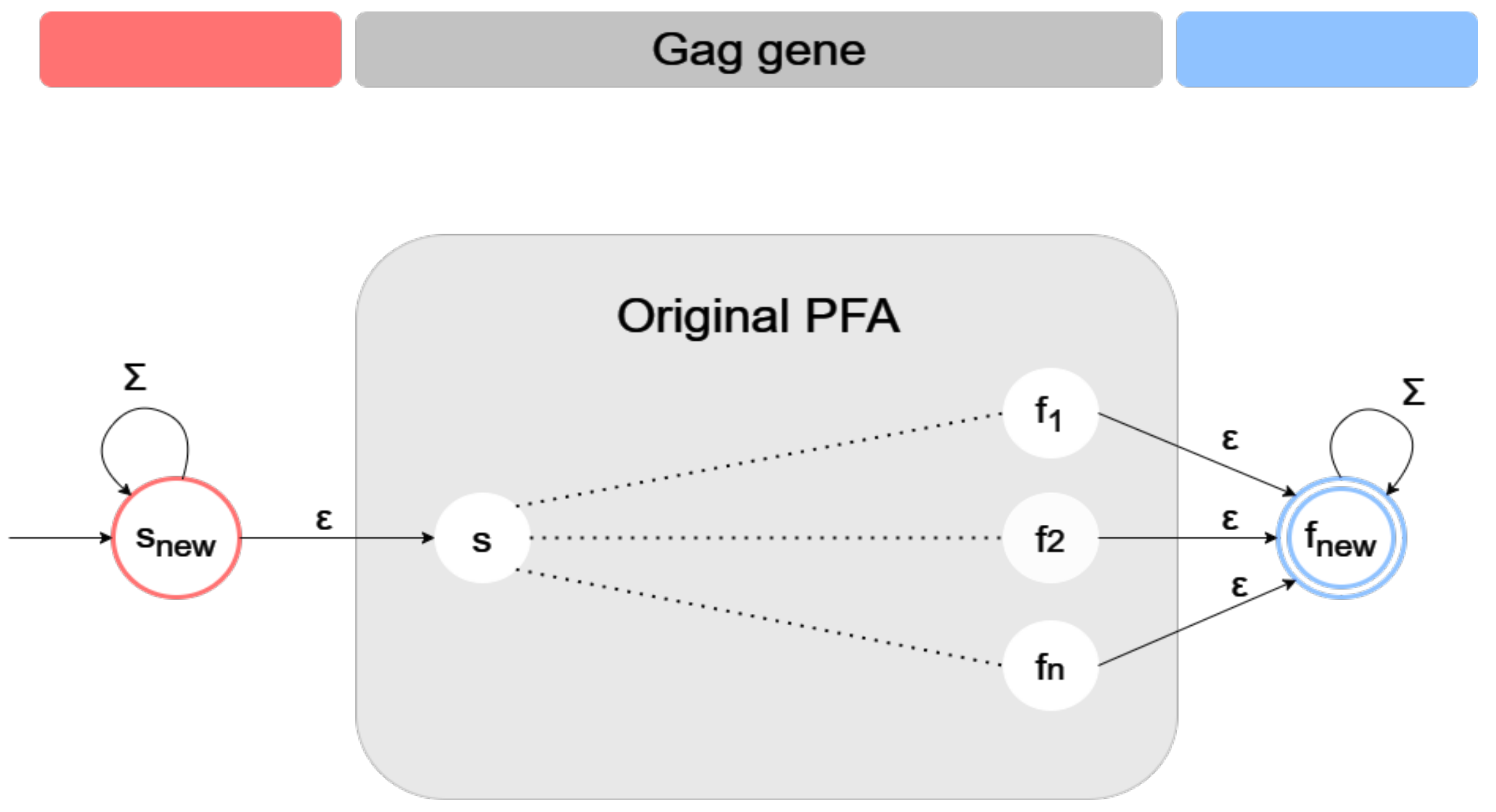

Figure 5 - Modified PFA. The sequence preceding the Gag gene and the corresponding part of the automaton are shown in red, the sequence following it in blue.

- Automaton generated by the ALERGIA algorithm has to be modified to find gene as a substring of a longer sequence
- Automaton is no longer determinictic
- We need to continuously store configurations of the automaton and select those that has the highest probability

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024