LTRFIND: A NOVEL ALGORITHM FOR HUMAN DE NOVO LTR RETROTRANSPOSONS IDENTIFICATION

Karel Sedlář

Doctoral Degree Programme (1), FEEC BUT

E-mail: sedlar@feec.vutbr.cz

Supervised by: Ivo Provazník

E-mail: provaznik@feec.vutbr.cz

Abstract: Modern LTR retrotransposons identification is mainly based on similarity searching using databases of known retroviruses. Nowadays, when various human genomes are available, this approach is not fast enough. Unlike the other species, human LTR retrotransposons are inactive and modified so similarity searches can hardly find these strongly mutated or previously unknown retroelements. In this paper, we present a novel algorithm for de novo identification of human LTR retrotransposons. Considering features of the human genome, we designed heuristic algorithm based on identification of long terminal repeats. Employ of exact sting match seed technique brings a very efficient search with reasonable sensitivity.

Keywords: retrotransposons, LTR, human genome, exact string match

1. INTRODUCTION

Discovering new LTR (long terminal repeats) retrotransposons and searching for known types is important part of human genome research. Nowadays, when whole human genome can be sequenced in a one day, the necessity of a new very fast data mining algorithm became topical. However LTR retrotransposons remains in human DNA inactive and methylated, they can be reactivated during pathological processes or environmental stress [1]. The analysis of these elements in various human genomes can uncover genotypes that are prone to serious diseases, e.g. Hodgkin's lymphoma, renal cell carcinoma or autoimmune disorders. Fast de novo identification of LTR retrotransposons in newly sequenced genomes is important part of the whole genome association studies. Also databases for retroelements already exist, e.g. Czech database HERVd: database of human endogenous retroviruses (http://herv.img.cas.cz/) [2].

Identification of retroelements in new genomes is routinely based on similarity searches against these databases of already known types. This technique works reliably for most genomes and it usually finds other types of repetitive elements too. The most commonly used software RepeatMasker [3] also uses this technique. On the other hand, these techniques suffer from 2 basic problems. Firstly, only previously described types of elements, that are stored in a database, can be found. Secondly, a human genome contains inactive and mutated elements that frequently lost substantial part of their sequence. Hence, they are unable to be detected by similarity search.

Only 2 algorithms for de novo LTR retroelements identification have been presented. LTR_STRUC [4] was the first algorithm for automatic detection. It was developed before next-generation sequencing platforms, so it's optimized for smaller amount of data. Newer de novo algorithm [5] is more effective. Unfortunately, it is optimized for genomes of organisms with still active retroelements.

2. RETROTRANSPOSONS

Retrotransposons belongs to non-coding, also repetitive or "junk" DNA. They are formed during process called transposition, which is "jumping" of a DNA segment from one place in the genome to another [6]. These elements thus expand (in quantity) by duplication mechanism (copy and paste).

2.1. LTR RETROTRANSPOSONS

LTR retrotransposons are also called endogenous retroviruses because they are very similar to proviruses of real viruses. They contain long terminal repeats at both ends, and *gag*, *pol*, *env*, *prt* genes. 5' LTR and 3' LTR parts have the same or highly similar sequence. Unlike the real viruses, at least one of the genes is missing. Thus, endogenous retroviruses are not able to assemble infectious particle and they can move only inside cells. *Pol* is the most important gene for lifecycle of retrotransposon because it consists of parts that are necessary for sequence duplication [7]. Reverse transcriptase (*rvt*) is responsible for DNA synthesis, ribonuclease H (*RNase H*) splits DNA/RNA hybrid and integrase (*int*) can split DNA and ligase retrovirus to the position. The schema of typical LTR retrotransposon is shown in Figure 1.

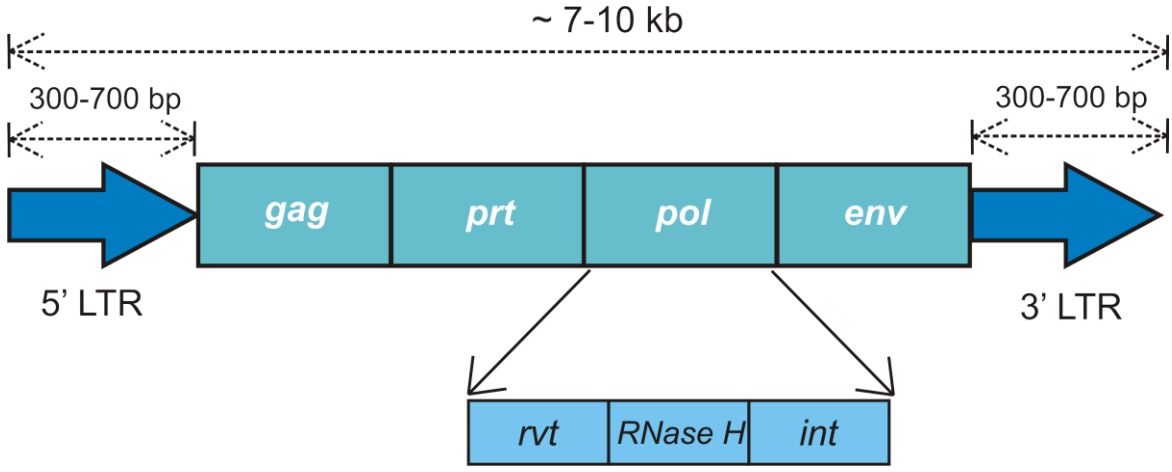


Figure 1: structure of LTR retrotransposon

The sequence of *rvt* is usually used as a barcode for retroelements identification based on similarity searches [5]. Thus, these techniques fail in identification of retrotransposons with missing *rvt*.

2.2. HUMAN LTRS

Unlike the other species, including other *Primates*, human genome contains only inactive LTR retrotransposons. Their length is very often within a range from 7 to 10 kb, but can be longer or shorter [8]. 5' LTR and 3' LTR ends are usually from 350 to 700 bases long and their similarity can fall below 80% due to amount of accumulated mutations during human evolution [9]. Human LTR retrotransposons can be divided into 2 groups.

Human endogenous retroviruses (HERVs) form a larger group that accounting for ~4.6% of the human genome [1]. Their sequence contains reverse transcriptase that is methylated. Thus, these elements are replication-incompetent under standard conditions. However, they can be reactivated by hypomethylation that is common in stress environment, e.g. during tumor proliferation.

Mammalian apparent LTR retrotransposons (MaLRs) are accounting for ~3.65% of the human genome. These elements lack of *pol* gene and so reverse transcriptase. Thus, they are unable to duplicate even in stress conditions. However, their LTR sequence contains a TATA box and transcription regulatory sequences which govern ubiquitous or tissue-specific gene expression. They potentially provide extra enhancer-promoter sequences and initiation sites for neighboring cellular genes. Connection between MaLRs and Hodking's lymphoma was recently proven [1].

3. ALGORITHM

Basic principle for de novo identification of LTR retrotransposons is analysis of their 5' LTR and 3' LTR ends.

3.1. CURRENT ALGORITHMS

Both of the currently mostly used algorithms are designed to do automatically several steps that would be otherwise done by person. However, the analysis using automation is matter of hours, unlike the analysis done manually which is matter of days; it is still not fast enough to process more genomes in a batch.

A sequence of length more than 100 bp (precise length is matter of specific genome) is taken and algorithm searches for matches in the area that potentially contains 3' LTR sequence [4]. Searching has to be done natively because similarity of each alignment is evaluated. When 40 nucleotides in a row are the same and similarity of whole sequence is greater than 70%, the sequence is used as the core of new local alignment in which LTR ends are specified. A bottleneck of this method is the string searching that is *O(mn)* algorithm, where *n* is length of the sequence taken and *m* length of area of interest.

3.2. PROPOSED ALGORITHM

The proposed algorithm is heuristic. It can possibly miss some of LTR transposons during a phase of searching for LTR ends, but the amount of time consumed for the first step is several times lower. Parameters were optimized for human genome and were based on empirical data by comparing the results with presumed values.

In the first step, potential LTR subsequences are marked. A seed of length 22 bp floats along the sequence (in position *i*) with step of 80 bp. The algorithm searches for exact string matches for seed in the range from *i*+6000 bp to *i*+12000 bp. Boyer-Moore string search algorithm could be used because exact string match technique without similarity evaluating of each position was chosen. The worst-case running time for the algorithm is *O(3m)* [10], but it is approaching to *O(n)* in most cases. When a match is found, subsequences of seed and its match are extended for 500 bp from both sides. In the third step, a local alignment of these subsequences is employed for LTR end identification. When score of alignment starts to fall, alignment is cut and borders of 5' LTR and 3' LTR are defined.

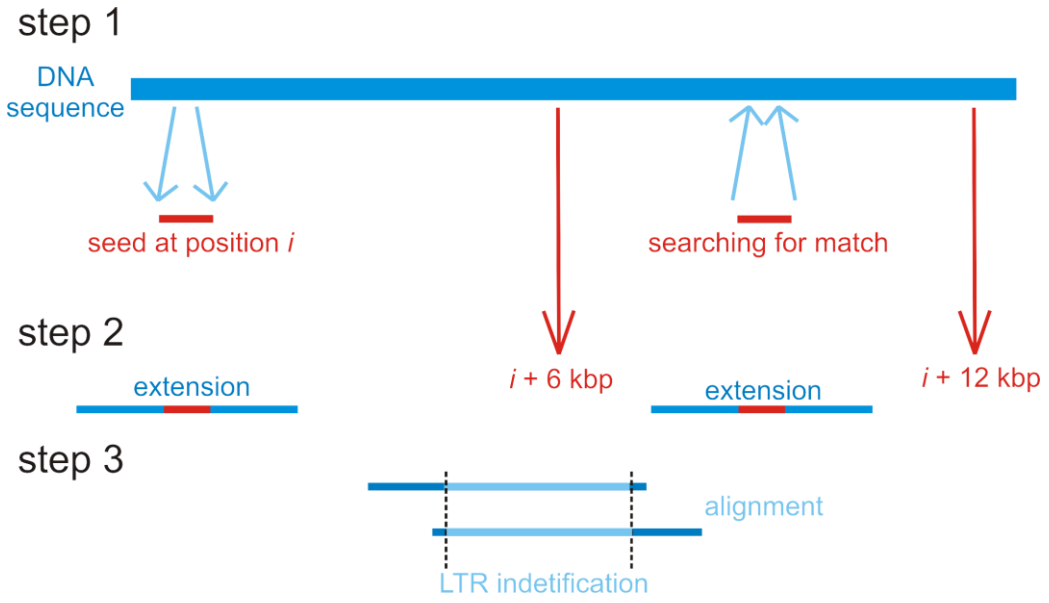


Figure 2: schema for ltrfind algorithm

3.3. DATA ANALYSIS

Data for analysis were obtained from GenBank database (http://www.ncbi.nlm.nih.gov/genbank/) at NCBI. Human genome project with accession No. PRJNA168 was the source for whole chromosome sequences. The assembly GRCh38 was used.

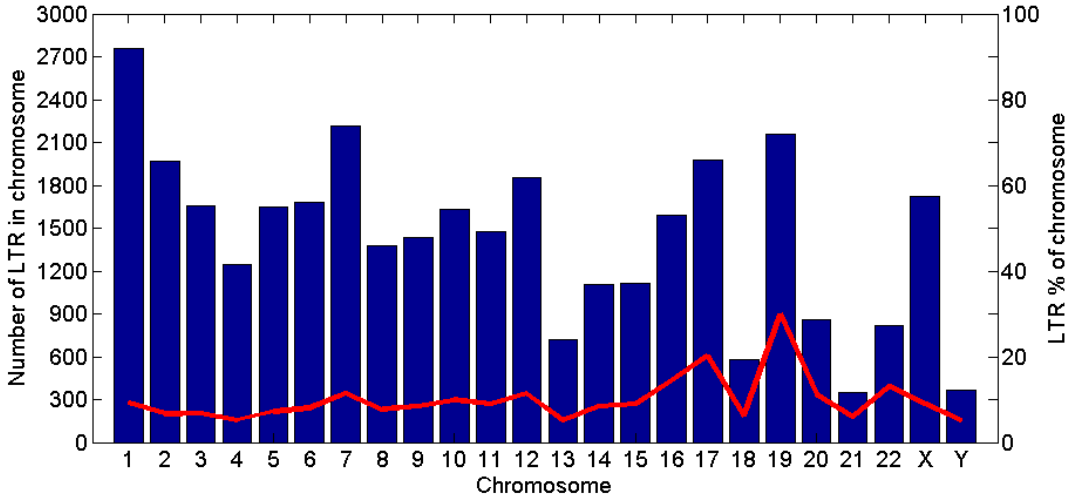
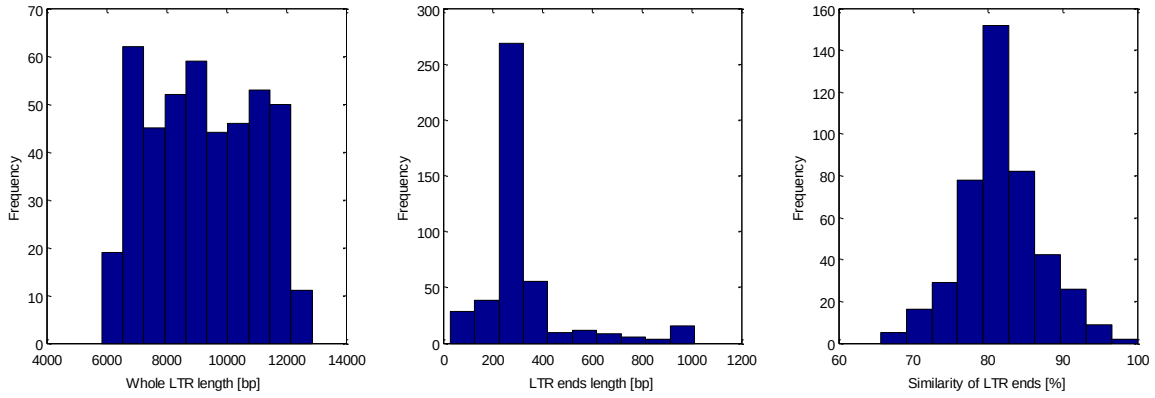


Figure 3: abundance of LTR retrotransposons in chromosomes (blue bars) and percentage content of LTR retrotransposons in each chromosome (red line)

The results from Figure 3 show that in most cases LTR retrotransposons found by algorithm form between 5 and 10% of chromosome sequence, however, two extremes were found. Both chromosomes 17 and 19 consist of more than 20% of retrotransposons. These two chromosomes contain great percentage of tandem repeats [11]. LTR retrotransposon content in whole genome was calculated as 8.7%. Further analysis for chromosome 21 as an example is provided in Figure 4.



(a) **(b)** **(c)**

Figure 4: histograms of (a) whole LTR length, (b) LTR ends length, (c) similarity of LTR ends

Results for chromosome 21 confirm theoretical assumptions that were specified in previous chapter. Lengths of whole LTR retrotransposons are within common range. Lengths of their ends reach various numbers, but the most common value falls to range from 200 to 400 bp. Algorithm was able to find even very old transposons with similarity of their ends below 70%.

4. CONCLUSION

In this paper, we present a novel algorithm for LTR retrotransposons identification in human genome. Due to use of heuristic technique, algorithm implemented in MATLAB is several times fast-

er than previous programs. Thus, whole human genome can be processed in 20 minutes using common PC which meets the current requirements, when sequencing of a genome takes only a day.

Algorithm is currently available from the author. Unlike the older algorithms, it can manage also "N" characters. These are omitted from main analysis, but their position is counted. Thus, whole chromosome sequence from DDBJ/EMBL/GenBank can be processed. Program output is gff file containing information about all LTR retrotransposons found in a sequence. Thus, it can be uploaded to any genome browser for further analysis.

However, verification of method is complicated because not all of LTR retrotransposons are already described, our program show very promising results and ability to find even very old elements with low LTR ends similarity. Algorithm counted LTR retrotransposon content in whole genome as 8.7% which is very close to predicted value of 8.25%.

ACKNOWLEDGEMENT

Supported by European Regional Development Fund - Project FNUSA-ICRC (No. CZ.1.05/1.1.00/02.0123) and by the grant project GACR P102/11/1068 NanoBioTECell.

REFERENCES

[1] Katoh, I., Kurata, S. I.: Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders. Frontiers in Oncology, 3, 1-8 (2013)

[2] Paces J, Pavlicek A, Paces V. HERVd: database of human endogenous retroviruses. Nucleic Acids Res., 30(1), 205-206 (2002)

[3] Smit, A. F. A., Hubley, R., Green, P.: RepeatMasker Open-3.0. http://www.repeatmasker.org (1996-2010)

[4] McCarthy, E. M., McDonald, J. F.: LTR STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics, 19(3), 362-367 (2002)

[5] Rho, M., Choi, J., Kim, S., Lynch, M., Tang, H.: De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*, 8(1), 90-106 (2007)

[6] Deininger, P., Batzer, M. A.: Mammalian Retroelements. *Genome Research*, 12(10), 1455-1465 (2002)

[7] Bannert, N., Kurth, R.: Retroelements and the human genome. *Proceedings of the National Academy of Sciences*, 101(2), 14572-14579. (2004)

[8] Polavarapu, N., Bowen, N., McDonald, J.: Newly Identified Families of Human Endogenous Retroviruses. *Journal of Virology*, 80(9), 4640-4642 (2006)

[9] Schon, U., Diem, O., Leitner, L., Gunzburg, W., Mager, D., Salmons, B.: Human Endogenous Retroviral Long Terminal Repeat Sequences as Cell Type-Specific Promoters in Retroviral Vectors. *Journal of Virology*, 83(23), 12643-12650 (2009)

[10] Cole, R.: Tight bounds on the complexity of the Boyer-Moore string matching algorithm. Proceedings of the 2nd annual ACM-SIAM symposium on Discrete algorithms, 224–233 (1991)

[11] Warburton, P. E., Willard, H. F.: Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. Journal of Mol Evol., 41(6), 1006–15 (1995)