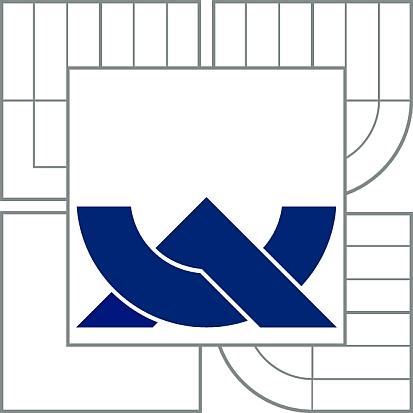
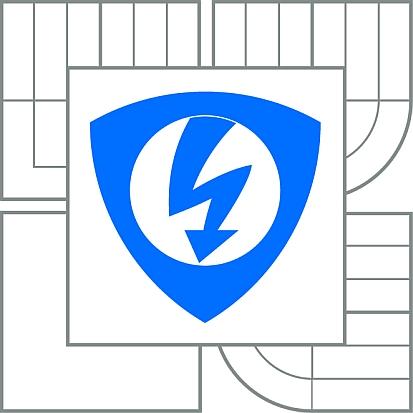
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ



BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ



ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ LTR RETROTRANSPOZONŮ V LIDSKÉM GENOMU

IDENTIFICATION OF LTR RETROTRANSPOSONS IN HUMAN GENOME

SEMESTRÁLNÍ PRÁCE

SEMESTRAL THESIS

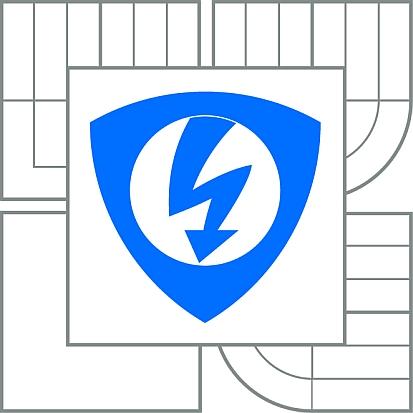
AUTOR PRÁCE EDUARD TROTT

AUTHOR

VEDOUCÍ PRÁCE Ing. KAREL SEDLÁŘ

SUPERVISOR

BRNO 2014

**VYSOKÉ UČENÍ**

**TECHNICKÉ V BRNĚ**

**Fakulta elektrotechniky**

**a komunikačních technologií**

**Ústav biomedicínského inženýrství**

**Semestrální práce**

bakalářský studijní obor

**Biomedicínská technika a bioinformatika**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Student:*** | Eduard Trott | ***ID:*** | 155615 |
| ***Ročník:*** | 3 | ***Akademický rok:*** | 2014/2015 |

**NÁZEV TÉMATU:**

**Vyhledávání LTR retrotranspozonů v lidském genomu**

**POKYNY PRO VYPRACOVÁNÍ:**

1) Zpracujte literární rešerši metod pro vyhledávání LTR retrotranspozonů v DNA, zaměřte se především na metody de novo. 2) Popište jednotlivé části retrotranspozonu a rodiny typické pro lidský genom, včetně jejich možných spojení s onemocněními. 3) Navrhněte a v jazyce R/Bioconductor realizujte nástroj pro vyhledávání LTR retrotranspozonů s vhodným výstupem (gff soubor). Funkčnost ověřte na sekvencích nejnovější dostupné verze lidského genomu. 4) Nástroj doplňte o možnost nalezené elementy rozdělit do rodin a tyto rodiny identifikovat pomocí vhodné referenční databáze. 5) Zhodnoťte úspěšnost vyhledávání pomocí již dostupné anotace, například z genomového prohlížeče UCSC. 6) Výsledky statisticky vyhodnoťte a diskutujte.

Pro splnění semestrálního projektu je nutné vypracování bodů 1) až 3).

**DOPORUČENÁ LITERATURA:**

1. RHO, Mina, Jeong-Hyeon CHOI, Sun KIM, Michael LYNCH a Haixu TANG. De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics. vol. 8, issue 1, s. 90-.
2. KATOH, Iyoko a Shun-ichi KURATA. Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders. Frontiers in Oncology. 2013, vol. 3.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Termín zadání:*** | 22.9.2014 | ***Termín odevzdání:*** | 5.1.2015 |
| ***Vedoucí práce:*** | Ing. Karel Sedlář |  |  |
| ***Konzultanti semestrální práce:*** | |  |  |
|  | **prof. Ing. Ivo Provazník, Ph.D.** |  |  |
| **UPOZORNĚNÍ:** | *Předseda oborové rady* |  |  |

Autor semestrální práce nesmí při vytváření semestrální práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

# 

**ABSTRACT**

The goal of this thesis is a literary background research on the topic Search LTR retrotransposons in the human genome. It is necessary to characterize the potential problems related to a given topic and to implement appropriate searching algorithm, the result of which is the GFF file that contains all found LTRs.

**Keywords:** long terminal repeat, LTR, retrotransposon, de novo

Abstrakt

Cílem této semestrální práce je zpracování literárního rešerši o tématu Vyhledávání LTR retrotranspozonů v lidském genomu. Je potřeba popsat možné problematiky navazující na danou tématu a implementovat vhodný algoritm vyhledávání, výsledkem kterého je GFF soubor, který obsahuje všechny nalezeny LTRs.

**Klíčová slova:** LTR,retrotranspozony

TROTT, E. *Vyhledávání LTR retrotranspozonů v lidském genomu.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 19 s. Vedoucí semestrální práce Ing. Karel Sedlář.

Prohlášení

Prohlašuji, že svou semestrální práci na téma Vyhledávání LTR retrotranspozonů v lidském genomu jsem vypracoval samostatně pod vedením vedoucího semestrální práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené semestrální práce dále prohlašuji, že v souvislosti s vytvořením této semestrální práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a~jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne .............................. ....................................

(podpis autora)

Poděkování

Děkuji vedoucímu semestrální práce Ing. Karlu Sedláři za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé semestrální práce.

V Brně dne .............................. ....................................

(podpis autora)

TABLE OF CONTENTS

[Introduction 2](#_Toc409222524)

[1 Theoretical background 3](#_Toc409222525)

[1.1 Genetics 3](#_Toc409222526)

[1.2 Transposable elements 3](#_Toc409222527)

[1.2.1 *Class I* (retrotransposons) 3](#_Toc409222528)

[1.2.2 *Class II* (DNA transposons) 4](#_Toc409222529)

[1.3 Retrotransposons 5](#_Toc409222530)

[1.3.1 LINEs (long interspersed elements): 6](#_Toc409222531)

[1.3.2 SINEs (short interspersed elements): 6](#_Toc409222532)

[1.3.3 LTR retrotransposons: 6](#_Toc409222533)

[2 LTR Retrotransposons 7](#_Toc409222534)

[2.1 Participation retrotransposons in human pathogenesis 8](#_Toc409222535)

[3 Algorithms of searching LTRs 9](#_Toc409222536)

[3.1 Алгоритмы основанные на сравнение с референчными датабазами 9](#_Toc409222537)

[3.2 De novo algorithms 10](#_Toc409222538)

[4 Implementation 11](#_Toc409222539)

[4.1 Python 11](#_Toc409222540)

[4.2 De novo search algorithm of LTRs 12](#_Toc409222541)

[4.2.1 The first step 12](#_Toc409222542)

[KMP algorithm 13](#_Toc409222543)

[Long Common Prefix with suffix array 13](#_Toc409222544)

[Search algorithm by Pythons built-in functions: 13](#_Toc409222545)

[4.2.2 The second step 13](#_Toc409222546)

[4.2.3 The third step 13](#_Toc409222547)

[4.3 Python toolboxes 14](#_Toc409222548)

[Results 15](#_Toc409222549)

[References 16](#_Toc409222550)

[List of abbreviations 17](#_Toc409222551)

[List of Figures 9](#_Toc409222552)

[List of attachments 18](#_Toc409222553)

Introduction

Human LTR elements are endogenous retroviruses which account for ~8% of the human genome. Now most human endogenous retroviruses (HERVs) are traces of viruses, which have been integrated millions of years ago. However HERVs and solitary LTR retrotransposons, not involved in the direct biological processes, may act as additional transcription apparatuses of genes by reactivation in generations or individuals. De novo approaches of searching retrotransposons in the human genome may later lead to finding new retroelements responsible for the cellular biological processes in the causes of which people have not understood at this time.

# Theoretical background

## Genetics

Since 1953, when Watson and Crick has discovered and deciphered the structure of DNA, started a new era of genetics, a lot of research areas was manifested. For example, molecular genetics, which reveals the chemical nature of heredity, or genetic engineering, dealing with genetic manipulation and introducing them to other organisms, or population genetics, archaeogenetics and many others. Also appeared direction, now representing the field of medicine, which identifies, examines and treats hereditary diseases - a medical genetics.

Big breakthrough of medical genetics is the possibility of sequencing the genome of an individual. This was made possible by the development of high-performance sequencing. The cost of genome sequencing is decreasing every year by an exponential scale, which provides opportunities of personal genomics for more people.

Given all of this, research in the field of genetics got a large value. Carried out a number of projects aimed at studying the human genome, such as The International HapMap Project or 1000 Genomes Project. Such projects are a key resource for researchers to find genetic mutations affecting health, and subsequently to consider options for their treatment.

## Transposable elements

Transposable elements (TEs), also known as transposons or "jumping genes", are discrete pieces of DNA sequence that can move in the genome from one location to another. Transposons represent one of types of mobile genetic elements. TEs are allocated to one of two classes, depending on their mechanism of transposition.

### *Class I* (retrotransposons)

### Retrotransposons - are mobile genetic elements that use the method of "copy and paste" for propagation in the genome of animals. At least 45% of the human genome is retrotransposons and their derivatives (FIG 1.2). The characteristics of retrotransposons are very similar to retroviruses.

More detail in the next Capitol.

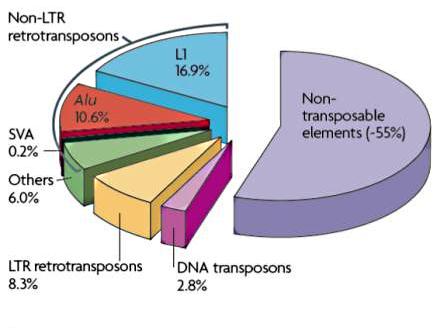
### *Class II* (DNA transposons)

DNA transposons to move inside genome use method "cut and paste" due to the complex enzyme called transposase [1]. Information on the amino acid sequence of the protein encoded transposase to the transposon sequences. Further, this piece of DNA may contain others related to transposon sequence, such genes or their parts.

Most DNA transposons have partial sequence. These transposons are not autonomous and move around in the genome due to the transposase, which is encoded by another, complete DNA transposon.

On the ends of the DNA-transposon regions located inverted repeats that are specific transposase recognition sites, thereby distinguishing it from the rest of the genome. Transposase is capable of doing double-stranded DNA sections, cut and paste into the target DNA transposon [13]

Different type of transposable elements use a variety mechanisms for their evolutionary survival. LINEs and SINEs rely on vertical transmission within the host genome. DNA transposons are more disorderly, requiring relatively frequent horizontal transfer. LTR retroposons use both of previous type of transfer. [2].



**Figure 1.2 The transposable element content of the human genome**

## Retrotransposons

Retrotransposons usually consist of three sub-types (FIG 1.3):

* LINEs(L1): encode reverse transcriptase, and are transcribed by RNA polymerase II
* SINEs(Alu): transcribed by RNA polymerase III
* LTRs(TEs with long terminal repeats): encode reverse transcriptase, similar to retroviruses

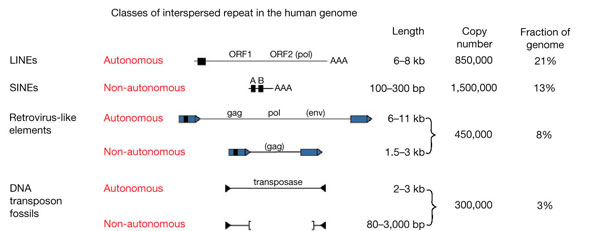


Figure 1.3 Classes of interspersed repeat in the human genome

### 

### LINEs (long interspersed elements)

One of the most ancient element in eukaryotic genomes. These transposons are about 6 kb long (FIG 1.3), encode two ORF and contain an internal polymerase II promoter. After translation, a LINE RNA assembles with its own encoded proteins and moves to the nucleus. An endonuclease activity makes a single-stranded nick and the reverse transcriptase uses the nicked DNA to prime reverse transcription from the 3′ end of the LINE RNA. [15]. The LINE machinery is probably responsible for most reverse transcription in the genome, including the creation of processed pseudogenes and the retrotransposition of the non-autonomous SINEs. There are three LINE families have been found in the human genome: LINE1, LINE2 and LINE3, of which LINE1 is still active. [2].

### SINEs (short interspersed elements)

Because of SINEs are non-autonomous they are require the presence of LINE elements to move. They are short (about 100–300 bp (FIG 1.3)), contain an internal polymerase III promoter and encode no proteins. Indeed, most SINEs ‘live’ by sharing the 3′ end with a resident LINE element. The promoter regions of all known SINEs are derived from tRNA sequences, with the exception of a single monophyletic family derived from the signal recognition particle component 7SL. This family includes the only active SINE in the human genome: the Alu element. The human genome contains three distinct monophyletic families of SINEs: the active Alu, and the inactive MIR and Ther2/MIR3. [2].

### LTR retrotransposons

LTR retroposons are surrounded by long terminal repeats that contain all of the necessary transcriptional regulatory elements. Exogenous retroviruses seem to have arisen from endogenous retrotransposons by acquisition of a cellular envelope gene (env). Transposition takes place through the retroviral mechanism with reverse transcription. Although a variety of LTR retrotransposons exist, only the vertebrate-specific endogenous retroviruses (ERVs) appear to have been active in the mammalian genome. Mammalian retroviruses fall into three classes (I–III), each comprising many families with independent origins. [2].

# LTR Retrotransposons

LTR retrotransposons are divided into three subclasses (FIG. 2.1):

* Ty1-copia-like (Pseudoviridae)
* Ty3-gypsy-like (Metaviridae)
* Endogenous retroviruses (ERV)

## C:\Users\Эдуард\Documents\GitHub\VUT_project\Auton_LTRs.jpg

Figure 2.1 The genomic organization of different types of LTR retrotransposons [12]

<http://genomebiology.com/2004/5/6/225>

Autonomous LTR retroelements

Ty1/Copia

<http://gydb.org/index.php/Ty1/Copia>

/Ty3/Gypsy

<http://gydb.org/index.php/Ty3/Gypsy>

Bel/Pao

<http://gydb.org/index.php/Bel/Pao>

Retroviridae

<http://gydb.org/index.php/Retroviridae>

Human LTR elements are endogenous retroviruses which account for ~8% of the human genome. [1]. Retroviruses can transform into LTR retrotransposons by inactivation or by disposal of structures responsible for the extracellular mobility. If a retrovirus infects and then embeds itself into the genome in germ line cells, it can become an Endogenous Retrovirus (ERV). Therefore exogenous retroviruses arose from the acquisition of endogenous retrotransposons cellular envelope gene. [3].

In general, most (85%) of the LTR retrotransposon-derived parts consist only of an isolated LTR, with the internal sequence having been lost by homologous recombination between the flanking LTRs. [2].

## Participation retrotransposons in human pathogenesis

Over 25 experimentally characterized cellular genes show LTR-mediated evolutionary changes in which are embedded LTRs alternative promoters to provide a new tissue-specificity, play as the major promoters, or promote only minor effects. [4]. For example, A HERV-K(HML-5) LTR plays as the major promoter of INSL4, a insulin-like growth factor gene expressed in placenta. [5]. A HERV-E family LTR plays as an alternative tissue-specific promoter of the endothelin B receptor (EDNRB) gene, by which the gene expression increased ∼15% in placenta. [6]. LTR-derived promoters often increase placenta-specific gene expression, despite the fact that in general the effect of the LTR insertions moderately manifested in many cases.

Recent studies have shown that HERV-encoded peptide as a tumor-specific antigen is involved in the hematopoietic stem cell transplantation for the therapy of renal cell carcinoma (RCC) [7]

A pioneering study investigate that HERV-E is activated in RCC and that it encodes an overexpressed immunogenic antigen, therefore providing a potential target for cellular immunity [7]. The tumor antigen, CT-RCC-1, recognized by RCC-specific CD8+ T cells is encoded by novel spliced variants of the HERV-E.

A study on tumorigenesis of Hodgkin’s lymphoma provided evidence that aberrant LTR activation contributes to lineage-inappropriate gene expression in transformed human cells and that such gene expression is central for tumor cell survival. They show that B cell–derived Hodgkin's lymphoma cells depend on the activity of the non-B, myeloid-specific proto-oncogene colony-stimulating factor 1 receptor (CSF1R). CSF1R transcription in these cells initiates at an aberrantly activated endogenous LTR of the MaLR family (THE1B). They conclude that LTR derepression is involved in the pathogenesis of human lymphomas. [8]

Human endogenous retroviruses are remnant forms of infectious retroviruses that integrated into the chromosomal DNA of germ-line cells of human ancestors, increased their copy numbers and have been inherited by present-day humans. Most HERVs are merely traces of original viruses, having first integrated millions of years ago. Within the published human genome sequence, there are over 98,000 human endogenous retroviruses (HERVs), but all are defective, containing nonsense mutations or major deletions. No replication-competent HERVs have been identified to date. [9] However, solitary LTRs derived from HERVs and MaLRs dominate the provirus forms in the copy numbers, and can serve as redundant enhancer-promoter sequences for nearby cellular genes. When the DNA methylation-mediated suppression system becomes compromised, HERVs and LTRs MaLR LTR in Hodgkin’s lymphoma and RCC-specific novel HERV-E antigen expression facilitating the immunotherapy. Future researches in oncology and immunogenetics will unveil more details about the endogenous LTR functions in human pathogenesis. [10].

По состоянию на 2012 год задокументировано 96 различных заболеваний человека, причиной которых является *de novo* внедрение мобильных генетических элементов[[22]](https://ru.wikipedia.org/wiki/%D0%A2%D1%80%D0%B0%D0%BD%D1%81%D0%BF%D0%BE%D0%B7%D0%BE%D0%BD%D1%8B#cite_note-HancksKazazian2012-22). Alu-повторы часто вызывают [хромосомные аберрации](https://ru.wikipedia.org/wiki/%D0%A5%D1%80%D0%BE%D0%BC%D0%BE%D1%81%D0%BE%D0%BC%D0%BD%D1%8B%D0%B5_%D0%BF%D0%B5%D1%80%D0%B5%D1%81%D1%82%D1%80%D0%BE%D0%B9%D0%BA%D0%B8) и являются причиной 50 разновидностей заболеваний [14]

<http://www.cell.com/cell-reports/fulltext/S2211-1247(14)01015-8>

<http://elementy.ru/news/432135>

<http://jvi.asm.org/content/80/21/10752.short>

<http://www.pnas.org/content/110/48/19472.abstract>

# Algorithms of searching LTRs

Все алгоритмы для поиска мобильных генетических элементов делятся на два больших класса:

## Алгоритмы основанные на сравнение с референчными датабазами

RepeatMasker is a popular software tool widely used in computational genomics to

identify, classify, and mask repetitive elements, including low-complexity sequences and

interspersed repeats. RepeatMasker searches for repetitive sequence by aligning the input

genome sequence against a library of known repeats, such as Repbase. Here, we describe

two Basic Protocols that provide detailed guidelines on how to use RepeatMasker, either

via the Web interface or command-line Unix/Linux system, to analyze repetitive elements

in genomic sequences. Sequence comparisons in RepeatMasker are usually performed by

the alignment program cross match, which requires significant processing time for larger

sequences. An Alternate Protocol describes how to reduce the processing time using an

alternative alignment program, such as WU-BLAST. Further, the advantages, limitations,

and known bugs of the software are discussed. Finally, guidelines for understanding the

results are provided

The conventional approach to annotating MGEs in genomic sequences is based upon homology searching against a well-updated library of known MGEs, e.g. Repbase , using a fast searching program, e.g. RepeatMasker. This approach, however, is limited to annotating those known MGE families, and thus cannot identify new elements. Furthermore, it sometimes even overlooks known elements, because the repetitive nature of MGE elements may confuse the statistical methods (e.g. E-values) that are commonly used in genome annotation. [1].

## De novo algorithms

LTR\_STRUC – широко распространенная программа для LTR retroelements

RECON

 RepeatScout (http://bioinformatics.oxfordjournals.org/content/21/suppl\_1/i351.long)

PILER (http://bioinformatics.oxfordjournals.org/content/21/suppl\_1/i152.long)

All these methods described above, however, attempted to identify repeat elements based on their copy numbers in a genome, thus facilitating identification of general repeat elements. Many MGEs indeed appear high copies in the host genome because of their transposition activity. But some MGE families have low copy numbers in some genomes. Furthermore, there exist other types of repeat elements than MGEs. For example, many low copy repeats (LCRs) in mammalian genomes are induced by segmental duplications [[10](http://www.biomedcentral.com/1471-2164/8/90#B10)]. Although these LCRs follow a completely different duplication mechanism from MGEs, there is often no clear distinction in copy numbers between these two classes of repeats. As a result, successful identification of new MGEs by these bioinformatics approaches requires subsequent manual inspection and experimental validation [[11](http://www.biomedcentral.com/1471-2164/8/90#B11)]. Recently, a new computational method was proposed that identified genome-specifically inserted sequences using multiple alignment of closely related genomes [[12](http://www.biomedcentral.com/1471-2164/8/90#B12)]. This new method does not rely on the copy number of the repeat elements to identify them, but does not attempt to distinguish different classes of repeats either.

(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1361722/)

# Implementation

## Python

Rewrite

Because scientists have long relied on the open availability of each other's research results, it was only natural that they would turn to Open Source software when it came time to apply computer processes to the study of biological processes. One of the first Open Source languages to gain popularity among biologists was Perl. Perl gained a foothold in bioinformatics based on its strong text processing facilities, which were ideally suited to analyzing early sequence data. To its credit, Perl has a history of successful use in bioinformatics and is still a very useful tool for biological research.

In comparison to Perl, Python is a relative newcomer to bioinformatics, but is steadily gaining in popularity. A few of the reasons for this popularity are the:

* Readability of Python code
* Ability to development applications quickly
* Powerful standard library of functionality
* Scalability from very small to very large programs

The Python language was designed to be as simple and accessible as possible, without giving up any of the power needed to develop sophisticated applications. Python's clean, consistent syntax leaves it free from the subtleties and nuances that can make other languages difficult to learn and programs written in those languages difficult to comprehend.

Python's dynamic nature adds to its accessibility. For example, Python doesn't require you to declare variables before you use them, and the same variable can refer to objects of different types over the course of its existence. Python can be also be used interactively, allowing you to familiarize yourself with the language of any Python modules in an interactive session where each command produces immediate results.

Python also has excellent support for the object-oriented style of programming. We'll show an example of this capability at the end of this article, but the basic idea is that object-orientation often provides a better way to organize the data and functionality within your programs. As the data and analytical techniques used in bioinformatics have become more complex, the value of object-oriented language features has risen.

In addition, Python integrates well with systems written in other languages, such as C, C++, Java and Fortran. One of the main benefits of C is speed. When a programmer needs an algorithm to run as fast as possible, they can code it in C or C++ and make it available to Python as an extension module. To the programmer, these are indistinguishable from pure Python modules. Similar utilities exist that make the large body of scientific algorithms coded in Fortran accessible to Python programs.

Java has become popular as a cross-platform and Web development language. The Python interpreter is now available in two variations: one version written in C, and the other version, known as Jython, written in Java. Jython allows Java programmers to write programs using the Python syntax and dynamic language features, and it allows Python programmers to use existing code developed in Java. These are just a few examples of the many ways Python is able to leverage and extend existing code written in other languages.

So while Perl is more well established in the bioinformatics community, many biologists and bioinformaticians are also turning to Python as it gains in popularity. To get a better sense of what Python has to offer, we'll look at examples of Python code that highlight some of its features. But first, we need to cover some of the basic biology that we'll touch on in the examples.

## De novo search algorithm of LTRs

This approach don't use reference database with found LTR, what allows find new sites of LTR. Finding young LTR retrotransposons is divided into three stages:

### The first step

is to find identical segments of the length of 40 bp limited by minimum (1000bp - minimal length of entire of LTR) and maximum (10 kbp maximum length of entire of LTR) distance between them.

For this step was implemented several algorithms for finding repeating patterns: KnuthMorrisPratt, Binary Search with LCP and Suffix array, searching with regular expressions and searching by built-in functions in Python. Turned out to be the fastest search functions embedded Python because it is C-implementation, when others are written in Python.

The following figure shows the difference in execution time of different algorithms. Original algorithm is implemented by built-in Python functions for pattern search.

Figure 4.2 Comparison of the time of the algorithms work

KMP algorithm

Long Common Prefix with suffix array

Search algorithm by Pythons built-in functions:

output = []

idx = 0

while idx < len(seq) - (min\_pattern\_len \* 2 + min\_distance):

pattern = seq[idx:idx + min\_pattern\_len]

if not 'N' in pattern:

text = seq[idx + min\_pattern\_len + min\_distance:idx + min\_pattern\_len + min\_distance + max\_distance]

if pattern in text:

output.append([idx, text.index(pattern) + idx + min\_pattern\_len + min\_distance])

idx += min\_pattern\_len

else:

idx += min\_pattern\_len

idx += 1

### The second step

is the formation of repetitive sequences in the groups associated with the individual LTR elements. These groups are based on the fact that the structure must meet the structure of LTR retrotransposons. Further, these groups are written in form of 4 indexes are responsible for: leading the beginning of LTR, the leading end of the LTR, the beginning of the trailing LTR, end of trailing LTR. The algorithm takes into account the gene amplification, minimal and maximal length of LTR parts.

### The third step

is the calculation of identity LTR for each set of indices obtained in the second stage and formation GFF file. GFF file contains the beginning and end of each LTR parts of retrotransposon and percentage identity between these LTRs.

Example of first several LTRs from GFF file for chrX ( Feb. 2009 GRCh37/hg19 ):

##gff-version 3

##sequence-region chrX 1 155270560

chrX ltrfind SO:0000186 458782 462270 . + . ID=UnknownLTR\_1;Note=identity 96.1538 %

chrX ltrfind SO:0000186 2120124 2130102 . + . ID=UnknownLTR\_2;Note=identity 93.2203 %

chrX ltrfind SO:0000186 2349094 2359925 . + . ID=UnknownLTR\_3;Note=identity 100 %

chrX ltrfind SO:0000186 2859399 2874732 . + . ID=UnknownLTR\_4;Note=identity 97.5 %

chrX ltrfind SO:0000186 3271550 3287635 . + . ID=UnknownLTR\_5;Note=identity 100 %

## Python toolboxes

For this project used the following Python libraries:

* Bio [1]
* BCBio
* time
* shelve

Biopython ( Bio ) is a set of freely available tools for biological computation written in Python

Time toolbox used to calculate the time of the pattern search algorithms.

Shelve is a persistent, Python dictionary-like object, used to store some intermediate data.

Results

The vast majority of detected LTR retrotransposons have identities more than 80%. This shows the flexibility of the algorithm to search for the young LTRs. This algorithm is aimed only to search young LTRs, what is a small number of elements relative to already found (references database), there is still room for improvement, both from the increase in speed, and by increasing the number of found retroelements. In continuation of this semestral thesis will be considered possible improvements of the search process, taking into account both structural component LTRs and their location in the genome and to each other.

References

[1] RHO, Mina, Jeong-Hyeon CHOI, Sun KIM, Michael LYNCH a Haixu TANG. De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*. vol. 8, issue 1, s. 90-. DOI: 10.1186/1471-2164-8-90. Dostupné z: <http://www.biomedcentral.com/1471-2164/8/90>

[2] LANDER, Eric S. et al.. Initial sequencing and analysis of the human genome. *Nature*. 2001-2-15, vol. 409, issue 6822, s. 860-921. DOI: 10.1038/35057062. Dostupné z: <http://www.nature.com/doifinder/10.1038/35057062>

[3] MALIK, H. S. Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. *Genome Research*. vol. 10, issue 9, s. 1307-1318. DOI: 10.1101/gr.145000. Dostupné z: <http://www.genome.org/cgi/doi/10.1101/gr.145000>

[4] COHEN, Carla J., Wynne M. LOCK a Dixie L. MAGER. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. *Gene*. 2009, vol. 448, issue 2, s. 105-114. DOI: 10.1016/j.gene.2009.06.020. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0378111909003540>

[5] BIECHE, I. Placenta-Specific INSL4 Expression Is Mediated by a Human Endogenous Retrovirus Element. *Biology of Reproduction*. 2002, vol. 68, issue 4, s. 1422-1429. DOI: 10.1095/biolreprod.102.010322.

[6] LANDRY, J.-R. a D. L. MAGER. Functional Analysis of the Endogenous Retroviral Promoter of the Human Endothelin B Receptor Gene. *Journal of Virology*. 2003, vol. 77, issue 13, s. 7459-7466. DOI: 10.1128/jvi.77.13.7459-7466.2003.

[7] TAKAHASHI, Yoshiyuki et al. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. *Journal of Clinical Investigation*. s. -. DOI: 10.1172/JCI34409. Dostupné z: <http://www.jci.org/articles/view/34409>

[8]  LAMPRECHT, Björn et al. . Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Medicine*. 2010-5-2, vol. 16, issue 5, s. 571-579. DOI: 10.1038/nm.2129. Dostupné z: <http://www.nature.com/doifinder/10.1038/nm.2129>

[9] BELSHAW, R., A. L. A. DAWSON, J. WOOLVEN-ALLEN, J. REDDING, A. BURT a M. TRISTEM. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. *Journal of Virology*. 2005-09-13, vol. 79, issue 19, s. 12507-12514. DOI: 10.1128/JVI.79.19.12507-12514.2005. Dostupné z: <http://jvi.asm.org/cgi/doi/10.1128/JVI.79.19.12507-12514.2005>

[10] KATOH, Iyoko a Shun-ichi KURATA. Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders. *Frontiers in Oncology*. 2013, vol. 3, s. -. DOI: 10.3389/fonc.2013.00234. Dostupné z:

<http://journal.frontiersin.org/Journal/10.3389/fonc.2013.00234/full>

[11] CORDAUX, Richard a Mark A. BATZER. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*. 2009, vol. 10, issue 10, s. 691-703. DOI: 10.1038/nrg2640. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg2640>

[12] LLORENS, C., R. FUTAMI et al.. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Research*. 2010-12-22, vol. 39, Database, D70-D74. DOI: 10.1093/nar/gkq1061. Dostupné z: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq1061>

[13] VAN OPIJNEN, Tim a Andrew CAMILLI. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nature Reviews Microbiology*. 2013-5-28, vol. 11, issue 7, s. 435-442. DOI: 10.1038/nrmicro3033. Dostupné z:

<http://www.nature.com/doifinder/10.1038/nrmicro3033>

[14] HANCKS, Dustin C a Haig H KAZAZIAN. Active human retrotransposons: variation and disease. *Current Opinion in Genetics*. 2012, vol. 22, issue 3, s. 191-203. DOI: 10.1016/j.gde.2012.02.006. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0959437X12000172>

[15] REILLY, M. T., G. J. FAULKNER, J. DUBNAU, I. PONOMAREV a F. H. GAGE. The Role of Transposable Elements in Health and Diseases of the Central Nervous System. *Journal of Neuroscience*. 2013-11-06, vol. 33, issue 45, s. 17577-17586. DOI: 10.1523/JNEUROSCI.3369-13.2013. Dostupné z: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3369-13.2013>

List of abbreviations

*DNA*

*ERVs*

*GFF*

*HERVs*

*IN*

*LARDs*

*LCP*

*LINE*

*LTR*

*MaLRs*

*MGE*

*ORF*

*PBS*

*PPT*

*PR*

*RCC*

*RNA*

*RT*

*SINE*

*TEs*

*TRIMs*

Deoxyribonucleic acid

Endogenous retroviruses

Gene-finding format, generic feature format

Human endogenous retroviruses

Integrase

Large retrotransposons derivates

Long common prefix

Long interspersed elements

Long terminal repeats

Mammalian apparent LTR retrotransposons

Mobile genetic elements

Open reading frame

Primer-binding site

Polypurine tract

Protease

Renal cell carcinoma

Ribonucleic acid

Reverse transcriptase

Short interspersed elements

Transposable elements

Terminal-repeat retrotransposons in miniature

List of Figures

**Figure 1.2 The transposable element content of the human genome [11]**

Figure 4.2 Comparison of the time of the algorithms work

Figure 1.3 Classes of interspersed repeat in the human genome [2]

Figure 2.1 The genomic organization of different types of LTR retrotransposons [12]

List of attachments

AttachmentA - source code ( .py )

AttachmentB – GFF file ( .gff )