VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

BIOINFORMATICKÝ NÁSTROJ PRO ANOTACI TRANSPOZONŮ

BIOINFORMATICS TOOL FOR TRANSPOSON ANNOTATION

DIPLOMOVÁ PRÁCE
DIPLOMA THESIS

AUTOR PRÁCE Bc. MICHAL JENČO
AUTHOR

VEDOUCÍ PRÁCE Ing. JANKA PUTEROVÁ
SUPERVISOR

BRNO 2017

Zadání diplomové práce/19833/2016/xjenco00

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav informačních systémů Akademický rok 2016/2017

Zadání diplomové práce

Řešitel: **Jenčo Michal, Bc.**
Obor: Bioinformatika a biocomputing
Téma: **Bioinformatický nástroj pro anotaci transposonů**
Bioinformatics Tool for Transposons Annotation
Kategorie: Bioinformatika

Pokyny:

1. Nastudujte základy molekulární biologie a dále se zaměřte na transposony.
2. Nastudujte literaturu z dané oblasti a vytvořte přehled dostupných nástrojů a metod pro anotaci transposonů.
3. Navrhněte metodu na hledání strukturních rysů transposonů, zaměřte se na tandemové repetice, určování věku transposonů a přídatné geny (miRNA, eORF aj.).
4. Po konzultaci s vedoucí implementujte navržený algoritmus a otestujte na vhodně zvoleném vzorku dat.
5. Diskutujte dosažené výsledky a další možné pokračování projektu.

Literatura:

- Kennedy RC, Unger MF, Christley S, Collins FH, Madey GR. 2011. An automated homology-based approach for identifying transposable elements.
- Wicker T et al. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8:973-982. doi: 10.1038/nrg2165-c4.
- Xu Z, Wang H. 2007. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35:265-268. doi: 10.1093/nar/gkm286.
- Dále dle doporučení vedoucího.

Při obhajobě semestrální části projektu je požadováno:

- Splnení bodů 1 až 3.

Podrobné závazné pokyny pro vypracování diplomové práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva diplomové práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap, které byly vyřešeny v rámci dřívějších projektů (30 až 40% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Puterová Janka, Ing.**, UIFS FIT VUT
Konzultant: Kejnovský Eduard, doc. RNDr., CSc., PřF MUNI
Datum zadání: 1. listopadu 2016
Datum odevzdání: 24. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav informačních systémů
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Dušan Kolář
vedoucí ústavu

Abstrakt

Táto práca poskytuje teoretické východiská pre návrh nového bioinformatického nástroja pre anotáciu transpozónov so zameraním na ich prídavné štruktúrne prvky. Sú v nej z biologického hľadiska popísané transpozóny, mobilné elementy v DNA, ich rozdelenie a vnútorná štruktúra. Ďalej sa zaoberá prehľadom a rozdelením dostupných bioinformatických nástrojov na identifikáciu a anotáciu transpozónov, popisom funkcie a implementácie vybraných z nich. Následne je popísaný návrh a implementácia nového bioinformatického nástroja na vyhľadávanie a anotáciu LTR retrotranspozónov so zameraním na extra ORF a tandemové repetície. Funkcionalita nástroja bola testovaná na genóme *A. thaliana*. Bolo identifikovaných 95 skupín konzervovaných extra ORF a 10 skupín konzervovaných tandemových repetícií.

Klíčová slova

transpozóny, satelitná DNA, bioinformatika, anotácia, extra ORF, tandemové repetície, LTR retrotranspozóny

Abstract

This thesis provides theoretical resources for the design of a new bioinformatics tool for transposon annotation with focus on their additional structural elements. There is a biological description of transposons, the mobile elements in DNA, their classification and structure. It further deals with the overview and classification of available transposon identification and annotation bioinformatics tools, description of function and implementation of a select few. Next we state the scheme of a new bioinformatics tool for LTR retrotransposon identification and annotation with a focus on extra ORFs and tandem repeats. The functionality of this new tool was tested on the *A. thaliana* genome. We identified 95 groups of conserved extra ORFs and 10 groups of conserved tandem repeats.

Keywords

transposons, satellite DNA, bioinformatics, annotation, extra ORF, tandem repeats, LTR retrotransposons

Citace

JENČO Michal: *Bioinformatický nástroj pro anotaci transpozonů*. Brno, 2017. 58 s. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Puterová Janka.

Bioinformatický nástroj pro anotaci transpozonů

Prehlásenie

Prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod vedením Ing. Janky Puterovej. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

……
Michal Jenčo
20. mája 2017

Poďakovanie

Rád by som sa poďakoval vedúcej diplomovej práce Ing. Janke Puterovej za odborné a motivačné vedenie.

Obsah

Úvod **7**

1 Transpozóny **8**

1.1 Transpozóny triedy I (retrotranspozóny) 8

1.2 Transpozóny triedy II (DNA transpozóny) 10

1.3 Ďalšie štruktúrne rysy transpozónov 12

2 Prehľad nástrojov a prístupov pre identifikáciu a anotáciu transpozónov **15**

2.1 *De novo* identifikácia 16

2.2 Prístup založený na homológii 16

2.3 Prístup založený na vnútornej štruktúre 17

2.4 Metóda porovnávania genómov 17

2.5 Existujúce nástroje na vyhľadávanie a anotáciu retrotranspozónov v zostavených genómoch 17

2.6 Výber nástroja pre vyhľadávanie LTR retrotranspozónov 21

3 Návrh nového nástroja pre anotáciu transpozónov **27**

3.1 Návrh riešenia 27

4 Implementácia **30**

4.1 Štruktúra zdrojového textu 30

4.2 Parametre nástroja 31

4.3 Vyhľadávanie retrotranspozónov 32

4.4 Vyhľadávanie ORF 32

4.5 Vyhľadávanie tandemových repetícií 32

4.6 Identifikácia proteínových domén a rodín 33

4.7 Identifikácia konzervovaných eORF 33

4.8 Identifikácia konzervovaných tandemových repetícií 35

5 Testovanie **38**

5.1 Vyhľadávanie extra ORF 38

5.2 Identifikácia konzervovaných extra ORF 41

5.3 Vyhľadávanie tandemových repetícií 43

5.4 Identifikácia konzervovaných tandemových repetícií 43

6 Záver **45**

Literatúra **46**

Príloha 1 – popis parametrov a výstupov použitých nástrojov **50**

Príloha 2 – špecifikácia výstupného formátu nástroja.....56
Príloha 3 – obsah CD a návod na sprevádzkovanie nástroja.....58

Úvod

Transpozóny sú DNA sekvencie, ktoré sú narozdiel od iných štruktúrnych prvkov DNA (napr. gény, satelity) schopné meniť svoju polohu v rámci genómu. Ich prítomnosť je mimoriadne častá u eukaryotických aj prokaryotických organizmov, kde tvoria významnú časť genómu, napríklad 20% u *D. melanogaster* [1], 50% u *H. sapiens* [2], alebo až 85% u *Z. mays* [3]. Zmena polohy transpozónu v rámci genómu sa môže udiať buď vystrihnutím už existujúceho transpozónu a jeho vložením na nové miesto, alebo vytvorením kópie už existujúceho transpozónu; tento druhý spôsob zmeny polohy výrazne prispieva ku zväčšovaniu genómu (Petrov 2001). Transpozóny často regulujú expresiu blízko ležiacich génov [4].

Rôzne typy transpozónov obsahujú rozličné štruktúrne prvky. Niektoré z nich sú nevyhnutné pre ich mobilitu, napríklad proteínové domény kódované v génoch obsiahnutých vnútri transpozónov. Ďalšie sú menej nutné a často dôvod ich prítomnosti a funkcia nie sú známe; napríklad prídavné čítacie rámce, alebo tandemové repetície. Práve na identifikáciu a anotáciu týchto bude zameraný nový bioinformatický nástroj, ktorého implementácia a testovanie bude náplňou tejto diplomovej práce. Na základe výskumu, ktorého sa zúčasťňujem na Biofyzikálnom ústave Akadémie vied Českej Republiky, bude tento nástroj zameraný výlučne na LTR retrotranspozóny.

Prvá kapitola sa zaoberá LTR retrotranspozónmi z biologického hľadiska. Je v nej popísané ich rozdelenie na základe mechanizmu pohybu v genóme a podrobnejšie delenie podľa vnútornej štruktúry. Kapitola obsahuje detailnejší popis hlavných a prídavných štruktúrnych prvkov transpozónov.

Druhá kapitola obsahuje prehľad využívaných prístupov pre identifikáciu a anotáciu LTR retrotranspozónov, popis funkcie a implementácie niektorých vybraných bioinformatických programov a testovanie dvoch nástrojov na vyhľadávanie LTR retrotranspozónov.

Tretia kapitola obsahuje motiváciu a návrh nového nástroja pre anotáciu transpozónov zameraného na anotáciu prídavných štruktúrnych prvkov, konkrétne tandemových repetícií a extra ORF.

Štvrtá kapitola obsahuje popis implementácie navrhnutého nástroja a piata kapitola obsahuje popis a výsledky testovania implementovaného nástroja na genóme rastliny *A. thaliana*.

1 Transpozóny

Transpozóny, tiež známe ako „skákajúce gény“, boli prvýkrát identifikované Barbarou McClintockovou v roku 1948 pri výskume DNA kukurice. Jej zistenie, že niektoré funkčné oblasti genómu sú schopné meniť pozíciu, priamo odporovalo v tom čase zaužívanému presvedčeniu, a tak bolo z väčšej časti ignorované až do prelomu 60. a 70. rokov 20. storočia, kedy boli transpozóny „znovuobjavené“ u baktérií. Za objav jej bola v roku 1983 ako doposiaľ jedinej žene udelená nezdieľaná Nobelova cena za fyziológiu a medicínu.

Postupne sa zistilo, že transpozóny nielenže existujú a sú schopné „skákať“, ale sú prítomné takmer u všetkých organizmov, a to typicky vo výraznom počte. Vďaka schopnosti pohybovať sa môžu spôsobiť v genóme organizmu výrazné zmeny, pretože vo fáze transpozície a spätného vloženia do genómu môžu nastať viaceré mutácie, ako napríklad inzercia, delécia alebo duplikácia [5]. Transpozícia, jav, podľa ktorého transpozóny dostali svoje pomenovanie, označuje zmenu pozície štruktúrnej jednotky DNA (napríklad génu) v rámci genómu.

Dnes je známe, že existuje veľké množstvo rôznych typov transpozónov a vyvinulo sa množstvo spôsobov ich kategorizácie. Jedna z najčastejších kategorizácií delí transpozóny do dvoch skupín podľa nutnosti reverznej transkripcie (prepis RNA do DNA) pri ich transpozícii. Transpozóny triedy I využívajú reverznú transkripciu a nazývajú sa retrotranspozóny alebo retroelementy. U transpozónov triedy II prebieha transpozícia pomocou DNA a nazývajú sa DNA transpozóny.

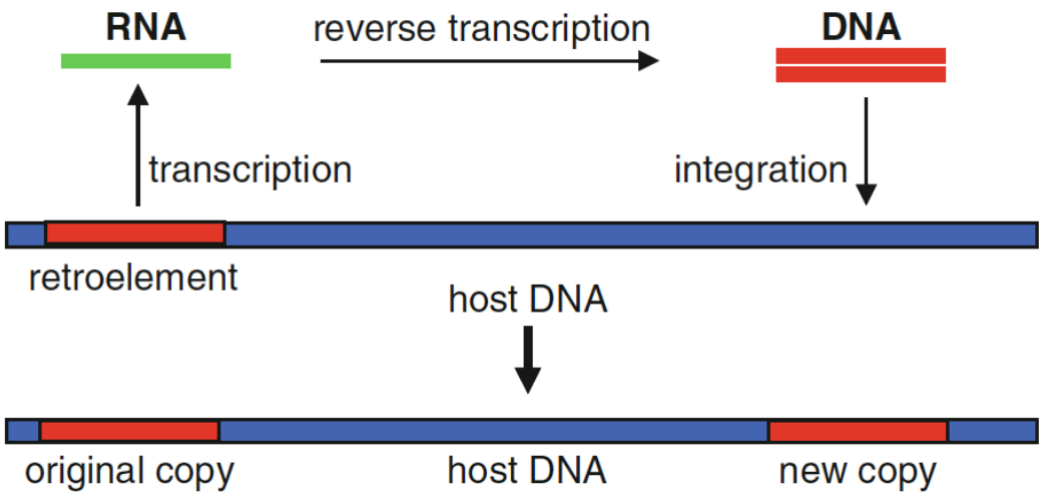
Obe tieto triedy obsahujú autonómne aj neautonómne elementy. Autonómne elementy obsahujú otvorené čítacie rámce (ORF), v ktorých sú zakódované gény nutné pre transpozíciu. Neautonómne transpozóny (často vznikajúce z autonómnych postupným zbieraním mutácií) potrebné gény neobsahujú a vyžadujú na svoj presun prítomnosť iného transpozónu, od ktorého si „požičajú“ potrebné proteíny.

1.1 Transpozóny triedy I (retrotranspozóny)

Transpozóny triedy I sa v genóme pohybujú mechanizmom „copy-and-paste“, znázorneným na obrázku 1.1. To znamená, že každá udalosť ich transpozície spôsobí vytvorenie novej kópie a teda predĺženie hostiteľského genómu. Princíp ich pohybu pozostáva z dvoch krokov. Prvým je transkripcia (prepis) z DNA do RNA, druhým je reverzná transkripcia novovzniknutej RNA kópie transpozónu naspäť do DNA. Táto kópia sa vloží do genómu na novú pozíciu. Reverzná transkripcia je umožnená enzýmom reverzná transkriptáza (RT), ktorý je u autonómnych retroelementov zakódovaný priamo v ich tele. Táto skupina má podobné charakteristiky ako retrovírusy, napríklad HIV. Retrotranspozóny sa ďalej zvyčajne delia na nasledujúce skupiny:

- **LTR retrotranspozóny** – elementy, ktorých telo je obklopené z oboch strán „dlhými koncovými repetíciami“ (Long Terminal Repeats, LTRs). Tieto elementy obsahujú gén pre reverznú transkriptázu.

- **Dlhé rozptýlené jadrové elementy** (Long Interspersed Nuclear Elements, **LINEs**). Obsahujú gén pre reverznú transkriptázu a sú prepisované pomocou RNA polymerázy II.
- **Krátke rozptýlené jadrové elementy** (Short Interspersed Nuclear Elements, **SINEs**). Nekódujú reverznú transkriptázu a ich prepis prebieha pomocou RNA polymerázy III.



Obrázok 1.1: Znázornenie „copy-and-paste“ mechanizmu transpozície LTR retrotranspozónov, prebraté z [6].

1.1.1 LTR retrotranspozóny

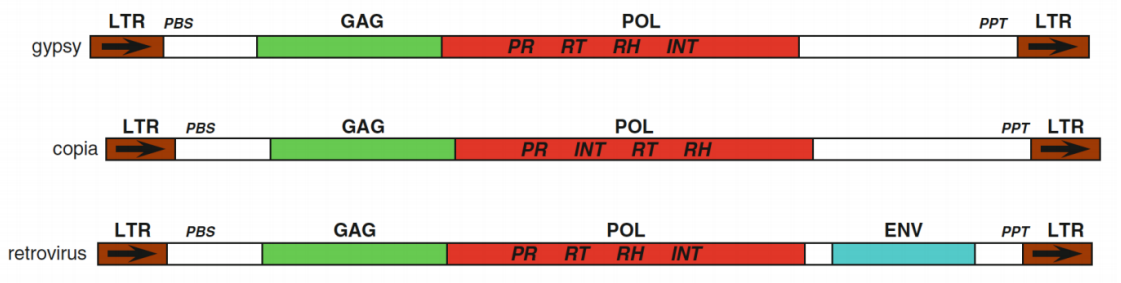
Delia sa na dve hlavné rodiny, a to *Ty3/gypsy* (často označovaná iba *gypsy*) a *Ty1/copia* (často iba *copia*). Elementy oboch rodín obsahujú rovnaké gény, konkrétne *gag* a *pol*. Gén *pol* kóduje viaceré proteíny, a práve ich poradie odlišuje tieto dve rodiny (obrázok 2.2). Gén *gag* kóduje štrukturálny proteín, ktorý vytvára ochranné prostredie, vnútri ktorého prebieha reverzná transkripcia. Gén *pol* kóduje nasledujúce proteíny:

- Proteáza (PR)
- Reverzná transkriptáza (RT)
- Ribonukleáza H (RH, prípadne RNaseH)
- Integráza (INT)

Tieto proteíny zabezpečujú proces reverznej transkripcie a integrácie novovytvorenej kópie na nové miesto v genóme. Dĺžka LTR retrotranspozónov sa pohybuje v rozmedzí niekoľkých stoviek párov báz (base pairs, bp) až po 25 kbp.

LTR sú sekvencie DNA o dĺžke od niekoľkých stoviek po niekoľko tisíc bp, ktoré sa nachádzajú na oboch koncoch LTR retrotranspozónov. V momente vloženia kópie retrotranspozónu na nové miesto v genóme sú obe jeho LTR identické. V priebehu „života“ transpozónu však zbierajú mutácie nezávisle, a tak sa ich podobnosť znižuje. Podľa divergencie

LTR sekvencií u LTR retrotranspozónu je možné odhadnúť jeho vek, teda moment jeho vloženia na aktuálnu pozíciu v genóme. Typický LTR retrotranspozón obsahuje štruktúru zvanú TG..CA box, čo značí, že na 5' konci 5'LTR sa nachádza dinukleotid TG a na 3' konci 3'LTR sa nachádza dinukleotid CA.



Obrázok 1.2: Porovnanie štruktúry dvoch hlavných rodín LTR retrotranspozónov a retrovírusu. Doména INT kódujúca integrázu je u rodiny *gypsy* na poslednom mieste v géne *pol*, zatiaľčo u rodiny *copia* je na druhom mieste. PBS označuje *primer binding site*, PPT označuje *poly-purine tract*, prebraté z [6].

1.1.2 LINE elementy

Často označované skrátene L1, nemajú LTR sekvencie, ale obsahujú poly(A) sekvenciu na 3' konci. Nachádzajú sa vo vysokom počte v eukaryotických genómoch, napríklad tvoria 17% ľudského genómu (avšak 99.9% z nich je považovaných za neschopné retrotranspozície) [7]. Dĺžka LINE elementov sa pohybuje v jednotkách kilobáz. Každý element obsahuje dva čítacie rámce; ORF1 obsahuje *gag* proteín, ORF2 obsahuje endonukleázu a reverznú transkriptázu. Tieto spolu umožňujú autonómnu transpozíciu LINE elementov.

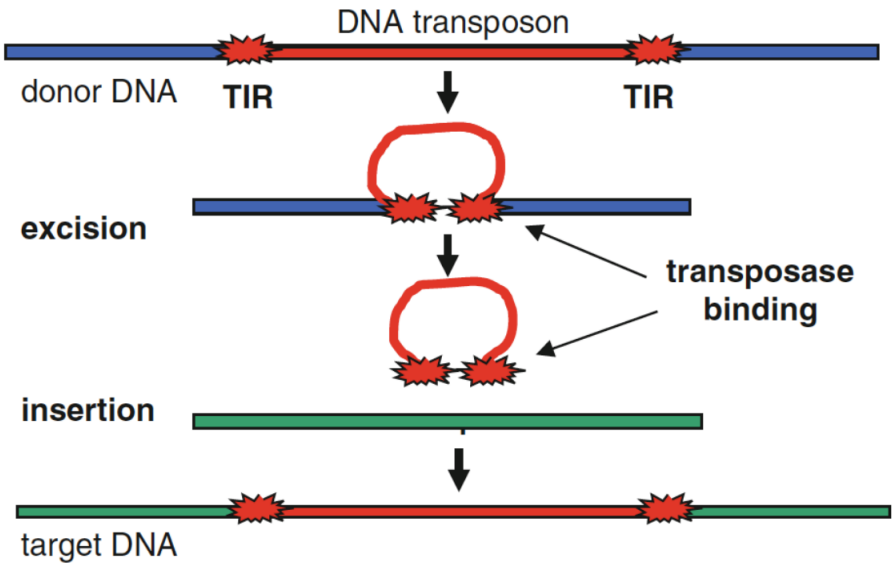
1.1.3 SINE elementy

Relatívne krátke transpozóny, s dĺžkou do 500 bp. Mechanizmus ich retrotranspozície je komplikovanejší, než u LINE elementov, a menej závislý na proteínoch, ktoré kódujú. Neobsahujú gén pre reverznú transkriptázu a spoliehajú sa pri transpozícii na iné mobilné elementy. Tvoria približne 11% ľudského genómu [8].

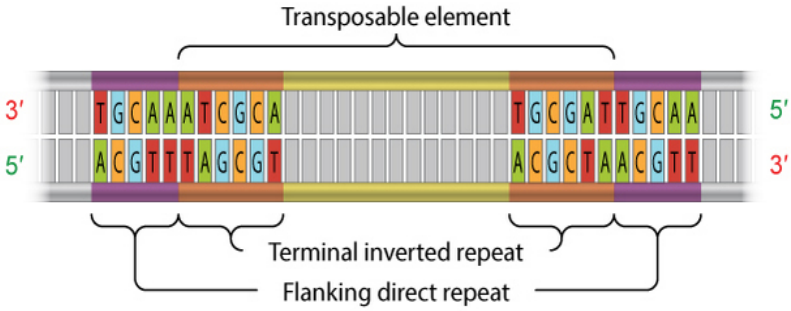
1.2 Transpozóny triedy II (DNA transpozóny)

DNA transpozóny sa v genóme pohybujú mechanizmom „cut-and-paste“ (obrázok 1.3), ktorý nevyžaduje RNA prostredníka. Zjednodušene sa dá povedať, že DNA transpozón sa vystrihne z aktuálnej pozície v genóme a vloží sa na nové miesto.

Telo DNA transpozónov sa skladá z jedného čítacieho rámca, ktorý kóduje enzým transpozáza. Rámec je zvyčajne obklopený obrátenými koncovými repetíciami (Terminal Inverted Repeats, TIR; vyskytuje sa aj varianta ITR), a na úplných koncoch DNA transpozónov sa nachádzajú krátke priame repetície, označované TSDs (Target Site Duplications, obrázok 1.4); tieto štruktúrne prvky však nie sú prítomné vždy. Obrátené koncové repetície sú reverzne komplementárne.



Obrázok 1.3: Znázornenie „cut-and-paste“ mechanizmu transpozície DNA transpozónov, prebraté z [6].

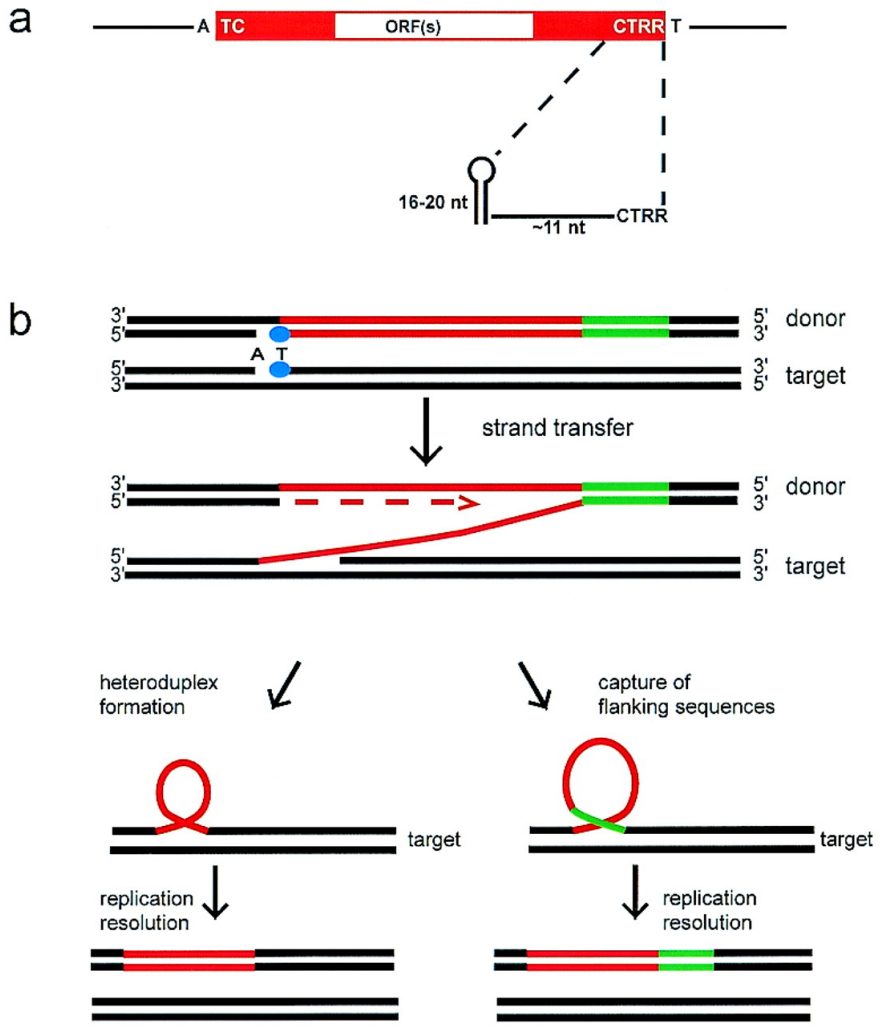


Obrázok 1.4: Znázornenie invertovaných koncových repetícií (TIR) a duplikácií cieľovej sekvencie (TSD) u DNA transpozónu, prebraté z [9].

Obrátené koncové repetície sú rozpoznané transpozázou a následne sa telo DNA transpozónu vystrihne a vloží na nové miesto v genóme. Po vložení sa zduplikuje sekvencia v mieste vloženia, čím vzniknú TSD, ktoré predstavujú charakteristickú črtu každého DNA transpozónu. DNA transpozóny sú klasifikované do dvoch podtried podľa toho, koľko vláken DNA je nutné rozstrihnúť počas transpozície:

- **Podtrieda I**
  - najznámejšími zástupcami sú rodiny *Tc1/mariner*, *PIF/Harbinger*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *piggyBac* a *CACTA*.
- **Podtrieda II**
  - rodiny *Helitron* a *Maverick*.

Jednotlivé rodiny sa líšia sekvenciou tela transpozónu, koncových invertovaných repetícií (TIR) a TSD. Členovia podtriedy II sa od ostatných líšia tým, že sú replikovaní a majú odlišný spôsob inzercie na nové miesto v genóme. Obrázok 1.5 znázorňuje mechanizmus transpozície helitrónov.



Obrázok 1.5: Znázornenie mechanizmu rolling circle, ktorým prebieha transpozícia helitrónov, prebraté z [12].

1.3 Ďalšie štruktúrne rysy transpozónov

Okrem zatiaľ popísaných štruktúrnych prvkov transpozónov, ktoré buď priamo definujú jednotlivé skupiny, alebo sú nutné pre ich životný cyklus (napr. LTR u LTR retrotranspozónov, TIR u DNA transpozónov, gény *gag* a *pol*), sa u transpozónov môžu nachádzať aj iné charakteristické sekvencie.

1.3.1 Tandemové repetície

Tandemové repetície sú po transpozónoch druhým hlavným typom repetitívnych DNA sekvencií. Narozdiel od transpozónov, pre ktoré je typická roztrúsenosť jednotlivých elementov v genóme, sú typické priamo za sebou nasledujúcimi neprerušenými opakovaniami tej istej sekvencie. Jednotka opakovanej sekvencie sa nazýva monomér; napríklad u repetície AACTAACTAACT je monomérom sekvencia AACT, ktorá sa trikrát opakuje. Monomér môže mať dĺžku od jedného nukleotidu až po niekoľko tisíc; dĺžky celých neprerušených repetícií môžu dosahovať až milióny nukleotidov.

Podľa dĺžky monoméru sa tandemové repetície rozdeľujú na mikrosatelity (monomér dĺžky od 1 po typicky 6 až 10 nukleotidov), minisatelity (monomér dlhý až 60 nukleotidov) a satelity (monomér dlhší, než 60 nukleotidov); tieto hranice však nie sú ostré a niekedy sa používajú mierne odlišné rozsahy.

Existuje viacero štúdií zaoberajúcich sa tandemovými repetíciami vnútri transpozónov. Napríklad Macas a kolektív [13] objavili repetitívnu sekvenciu (nazvanú PisTR-A) v genóme hrášku (*Pisum sativum*), ktorá je prítomná ako krátke rozptýlené opakovania, aj ako dlhé neprerušené satelitné sekvencie na rozličných miestach v genóme. Pomocou ďalších analýz zistili, že zatiaľčo druhý prípad predstavoval štandardné satelitné repetície, krátke rozptýlené repetície boli súčasťou *gypsy* LTR retrotranspozónov z rodiny Ogre. Tieto LTR retrotranspozóny obsahovali PisTR-A sekvencie vo svojich 3' UTR (untranslated region; neprekladaný úsek DNA, ktorý nie je súčasťou čítacieho rámca), ktoré sa nachádzajú medzi *gag-pol* regiónom a 3' LTR. Ukázali, že tento región je u Ogre elementov u hrášku vysoko variabilný, a okrem alebo namiesto PisTR-A sekvencií obsahuje aj rozličné iné tandemové repetície. Ďalšie analýzy ukázali, že častá prítomnosť rozličných tandemových repetícií v 3' UTR je typickou vlastnosťou u Tat línie rastlinných retrotranspozónov. Porovnanie týchto repetícií so známymi sekvenciami rastlinných tandemových repetícií ukázalo ďalšie dva prípady so sekvenčnou podobnosťou 3' UTR regiónov u Tat retrotranspozónov. Tieto pozorovania podľa Macasa a kolektívu naznačujú, že niektoré retrotranspozóny môžu výrazne prispievať k evolúcií satelitnej DNA vytváraním „knižnice“ krátkych opakovaní. Tie môžu byť neskôr rozptýlené pozdĺž genómu a eventuálne amplifikované a homogenizované do nových satelitných repetícií.

Hoci sa transpozóny a tandemové repetície odlišujú v štruktúre, mechanizme šírenia a v evolučnej dynamike, môžu vykazovať podobnosť sekvencií a organizácie v genóme, čo naznačuje možnosť komplexných vzájomných vzťahov. Tie, aj vzhľadom na vysoký podiel repetícií v genóme, môžu hrať významnú rolu vo vývoji a stavbe genómu ako takého.

1.3.2 Extra ORF

Keďže doterajší výskum extra čítacích rámcov (extra ORF, eORF) u transpozónov je zameraný na LTR retrotranspozóny u rastlín, táto podkapitola bude zameraná iba na ne.

Transpozícia LTR retrotranspozónov je zabezpečovaná proteínmi kódovanými v génoch *gag* a *pol*, ktoré sú spoločné pre všetky autonómne elementy. Translácia *gag-pol* regiónu prebieha vo väčšine prípadov v rámci jedného čítacieho rámca; hoci v niektorých skupinách elementov môže *gag-pol* región obsahovať viaceré prekrývajúce sa, alebo nadväzujúce ORF.

Napriek tomu, že proteíny kódované v génoch *gag* a *pol* sú považované za dostatočné pre replikáciu a transpozícu LTR retrotranspozónov, boli u niektorých LTR retrotranspozónov nájdené ďalšie ORF, nachádzajúce sa buď pred, alebo za *gag-pol* regiónom, ktoré môžu kódovať prídavné proteíny [14]. Konkrétne sa našli napríklad fragmenty ATPázy, endohydrolázy a glukanázy. V niektorých prípadoch sa tieto ďalšie ORF spájali s ORF obsahujúcim *gag-pol* a vytvárali ORF so skrátenou verziou *gag*, v mnohých však boli úplne separátne [15]. Vyskytujú sa v dvoch zo siedmich evolučných línií Ty1/*copia* elementov definovaných v [16] a sú ešte častejšie a rozličnejšie u Ty3/*gypsy* elementov. Líšia sa polohou voči *gag-pol* (buď 3' alebo 5') a orientáciou (niekedy je rovnaká, ako u *gag-pol*, niekedy opačná).

Väčšina identifikovaných eORF však nebola spojená so žiadnymi známymi génmi. Zároveň je zrejmé, že viac, než ich konkrétne sekvencie, ktoré sú veľmi heterogénne, je konzervovaná ich poloha v rámci tela elementu. Ich opakovaný výskyt však vyvoláva otázky o ich pôvode a úlohe v evolúcii genómu.

2 Prehľad nástrojov a prístupov pre identifikáciu a anotáciu transpozónov

Repetitívna povaha transpozónov komplikuje mnohé typy štúdií, ako sú predikcie génov, alebo genómové zarovnanie. Na druhej strane ich mobilita a repetitívny charakter majú potenciál prispieť k rozličným typom biologických problémov a činností, napríklad ku chorobám, evolúcii genómov, vývoja organizmov, alebo génovej regulácií. Okrem výrazného vplyvu na veľkosť, štruktúru a variáciu genómu a údržbu centromér a telomér [17], transpozóny sú tiež zdrojom „surového materiálu“ pre evolučnú inováciu, ako napríklad tvorbu nových génov [18] a nekódujúcej RNA [19]. S mimoriadne rýchlo rastúcim množstvom genomických dát narastá aj potreba výskumníkov čo najrýchlejšie, najsprávnejšie a automaticky identifikovať transpozóny v DNA sekvenciách.

Správna detekcia a anotácia transpozónov je náročná kvôli ich veľkej inter a intragenómovej diverzite. Existuje množstvo typov transpozónov, ako bolo popísané v predchádzajúcej kapitole, ktoré sa odlišujú mnohými rôznymi atribútmi, od vnútornej štruktúry, cez dĺžku, množstvo kópií, distribúciu pozdĺž chromozómov až po mechanizmus transpozície. Navyše, kým „nedávno“ (čo môžu byť v genomickej časovej škále stovky tisíc rokov) vložené elementy majú relatívne nízku variabilitu v rámci rodiny, v priebehu času jednotlivé elementy zbierajú mutácie a divergujú, čím sa ich detekcia komplikuje. Predpokladá sa, že veľká časť DNA, ktorej pôvod v súčasnosti u niektorých organizmov (napríklad aj u človeka) nie je známy, môže pozostávať zo silne rozpadnutých pozostatkov transpozónov [20]. Kvôli tejto vysokej diverzite transpozónov vnútri a medzi genómami sa u jednotlivých organizmov líšia aj hlavné prekážky ich presnej anotácie. Rozličné organizmy totiž majú odlišné mechanizmy umlčovania transpozónov a potenciálne „zažili“ odlišné vzory transpozónovej aktivity a premeny. U niektorých organizmov (napr. u človeka [21]) je napríklad väčšina transpozónov pozostatkami pradávnych udalostí veľmi vysokej aktivity iba niekoľkých rodín, a teda ich anotácia je obmedzovaná hlavne silnou divergenciou starých a rozpadnutých elementov. Iné genómy (napr. kukurica [22]) obsahujú veľké množstvo rozdielnych nedávno aktívnych elementov; hlavnou výzvou u týchto prípadov je špecifikácia komplexných štruktúr tvorených skupinami transpozónov, ako napríklad vnorené inzercie (kópia transpozónu sa môže pri vložení na nové miesto v genóme ocitnúť vnútri iného transpozónu).

Navyše, hoci knižnice už známych transpozónov existujú a sú užitočné, transpozónové rodiny prítomné aj u blízko príbuzných genómov sa môžu výrazne líšiť [23], čo obmedzuje užitočnosť týchto knižníc pri anotácií novo osekvenovaných genómov. Ďalšími výzvami presnej anotácie sú napríklad génové rodiny, ktorých gény sú prítomné v mnohých kópiách, a segmentové duplikácie (všeobecné duplikácie úsekov DNA bez špecifickej funkcie alebo štruktúry); obe tieto skupiny sekvencií sa kvôli svojej repetitívnosti podobajú na transpozóny. Regióny nízkej komplexity a jednoduché repetície môžu byť taktiež zdrojom falošných detekcií.

V tejto kapitole sa zaoberám programami určenými na detekciu a anotáciu transpozónov v zostavených genómoch. Na túto činnosť bolo vyvinuté množstvo výpočtových prístupov. Nasleduje popis štyroch najpoužívanejších prístupov podľa [24].

2.1 *De novo* identifikácia

De novo prístup (z latinčiny, „od začiatku“) využíva opakovaný výskyt transpozónov v poskytnutej DNA sekvencii, typicky bez predošlej znalosti štruktúry alebo podobnosti so známymi sekvenciami transpozónov. Najčastejšou stratégiou je detekcia párov podobných sekvencií na rozličných pozíciách a následné zhlukovanie nájdených párov, čím sa získajú rodiny repetícií. Keďže tento prístup nie je špecifický čiste pre vyhľadávanie transpozónov, medzi výsledkami sa obvykle nachádzajú aj repetície vytvorené úplne odlišnými procesmi, napríklad tandemové repetície alebo segmentové duplikácie.

Väčšina implementácií metódy *de novo* využíva klasické výpočtové stratégie na detekciu repetitívnych regiónov v genóme, ako je suffixový strom, alebo vyhľadávanie párovej podobnosti.

Hoci je prístup *de novo* výpočtovo náročný, umožňuje identifikovať doposiaľ neznáme transpozóny a je najefektívnejší vo vyhľadávaní transpozónov prítomných v mnohých kópiách. *De novo* techniky sú typicky neefektívne pri vyhľadávaní degradovaných transpozónov.

2.2 Prístup založený na homológii

Najrozšírenejší prístup ku identifikácii nových rodín transpozónov je založený na detekcii homológie voči známym transpozónovým sekvenciám. Výhodou tohoto prístupu oproti metóde *de novo* je, že ťažia z predošlých znalostí zachytených vo veľkom množstve známych transpozónov. Vďaka tomu majú vyššiu šancu identifikovať transpozóny prítomné v nízkom počte (vrátane jedinej kópie) a sú schopné poskytnúť klasifikáciu (do triedy I/II, alebo do transpozónovej rodiny) nového transpozónu na základe jeho podobnosti k už známym a anotovaným transpozónom. Tieto metódy sú typicky aplikované na zostavené genómy.

Väčšina metód založených na homológií používa verziu rýchleho zarovnávacieho algoritmu (napr. z programu `BLAST`). Alternatívou je použitie skrytých Markovovych modelov (HMM). Tie totiž umožnia zachytiť variabilitu elementu a tým spresniť vyhľadávanie a zachytiť aj ich menej zachované kópie. Vo všeobecnosti detekcia transpozónov založená na homológií vyžaduje následnú analýzu štrukturálnych vlastností, kým je možné získať referenčnú sekvenciu plnej dĺžky. Hoci týchto nástrojov je málo a nie sú schopné identifikovať transpozóny nepríbuzné známym elementom, sú zvyčajne najpresnejšie v identifikácii známych a degradovaných transpozónov. Najpoužívanejším nástrojom založeným na tomto prístupe je `RepeatMasker` [25].

2.3 Prístup založený na vnútornej štruktúre

Tretia skupina nástrojov využíva dopredu známe znalosti o spoločných štruktúrnych prvkoch zdieľaných rôznymi transpozónmi, ktoré sú nevyhnutné pre proces transpozície, ako napríklad LTR. Sú teda vhodné pri hľadaní kompletných transpozónov tých typov, ktoré si udržiavajú typickú štruktúru, napríklad LTR retrotranspozóny. Narozdiel od metód založených na homológií sú menej ovplyvnené sekvenčnou podobnosťou so známymi elementami. Ich limitáciou je potreba separátnych modelov pre každý štrukturálny typ transpozónov, a fakt, že miera štruktúrovanosti sa u jednotlivých typov transpozónov líši. Príkladom programu tohoto typu je `LTR_FINDER` [28], ktorého princíp a funkcionalita sú podrobne popísané v podkapitole 2.6.

2.4 Metóda porovnávania genómov

Tento relatívne inovatívny prístup využíva fakt, že transpozóny tvoria dlhé inzercie, ktoré je možné identifikovať v mnohonásobnom zarovnaní genómov príbuzných organizmov. Prvým krokom je vytvorenie mnohonásobného zarovnania viacerých genómov. V ňom sa následne vyhľadávajú oblasti (insertion regions, IR), v ktorých sú u jedného alebo viacerých genómov zarovnania prerušené dlhou inzerciou (> 200bp). Po filtrácii jednoduchých opakovaní a konkatenácií nájdených IR sú tieto navzájom lokálne zarovnané, čím sa identifikujú regióny vloženia repetícií. Táto metóda je schopná identifikovať úplne nové transpozóny a zároveň datovať ich inzerciu na fylogenetickom strome.

Úspešnosť tejto metódy je výrazne závislá na kvalite zarovnania celých genómov. Ďalej tiež závisí na aktivite transpozónov v skúmaných genómoch; ak všetky inzercie pochádzajú z predkov skúmaných organizmov, nebudú identifikované žiadne transpozóny. Napriek tomu je tento prístup vzhľadom na rastúcu dostupnosť sekvencií kompletných genómov príbuzných organizmov veľmi perspektívny. Táto metóda je navyše výpočtovo veľmi náročná a úplne závislá na dostupnosti genómov mnohých príbuzných druhov, čo ju robí v súčasnosti veľmi špecializovanou a najmenej často použiteľnou.

2.5 Existujúce nástroje na vyhľadávanie a anotáciu retrotranspozónov v zostavených genómoch

V tejto podkapitole popíšem niektoré najaktuálnejšie a najčastejšie používané bioinformatické nástroje na vyhľadávanie a anotáciu retrotranspozónov.

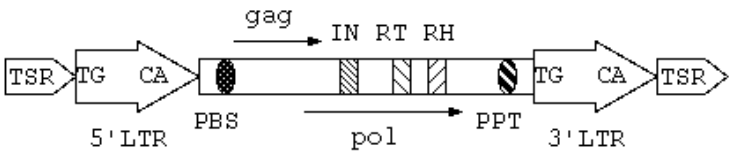
2.5.1 LTR_FINDER

`LTR_FINDER` [28] je nástroj určený pre identifikáciu full-length LTR retrotranspozónov pomocou ich štruktúrnych vlastností. Vstupom programu je DNA sekvencia vo formáte FASTA alebo multi-FASTA. Program je pre DNA sekvencie obmedzenej dĺžky (súbor do veľkosti 50Mb) dostupný cez webové rozhranie, pre potreby náročnejšieho výskumu je na vyžiadanie od autorov dostupná aj binárna verzia spustiteľná na systémoch s operačným systémom Linux

umožňujúca spracovanie DNA sekvencií ľubovoľnej dĺžky (v dokumentácii sa neuvádza obmedzenie, z osobnej skúsenosti zrejme závisí iba na množstve dostupnej operačnej pamäte).

Program je okrem samotnej identifikácie LTR retrotranspozónov schopný aj ich pomerne extenzívnej anotácie. Tá zahŕňa nasledujúce štruktúrne prvky (znázornené aj na obrázku 2.1 nižšie):

- **LTR**: Dlhé koncové opakovania. Dve LTR jedného LTR retrotranspozónu sú si navzájom podobné, čo je hlavnou informáciou používanou pri vyhľadávaní LTR retrotranspozónov.
- **TSR**: Repetícia na mieste vloženia LTR retrotranspozónu (Target Site Repeat). Je to 4-6 bp dlhá priama repetícia obklopujúca oba konce LTR retrotranspozónu. Je identifikačným znakom vloženia transpozónu.
- **PBS**: Primer Binding Site, región DNA dlhý ~18bp, na ktorý sa na začiatku replikácie viaže primer, čo je prvým krokom reverznej transkripcie. Nachádza sa blízko 3' konca 5'LTR.
- **PPT**: Polypurínový trakt je krátky, na puríny (nukleotidy adenín a guanín) bohatý segment o dĺžke 11-15bp. Podobne ako PBS je to kľúčová oblasť pre proces reverznej transkripcie.
- **Proteínové domény**: Typický autonómny LTR retrotranspozón obsahuje dva gény – *gag* a *pol*; *pol* ďalej obsahuje proteínové domény PR, IN, RT a RH.



Obrázok 2.1: Znázornenie štruktúrnych prvkov LTR retrotranspozónu, ktoré je `LTR_FINDER` [28] schopný anotovať, prebraté z [28].

Program najprv zostrojí všetky presne sa zhodujúce páry sekvencií (tie predstavujú predpokladané LTR) pomocou algoritmu založeného na poli suffixov a predĺži ich na dlhé vysoko podobné páry. Následne použije Smith-Watermanov algoritmus a upraví konce kandidátnych párov LTR, čím získa hranice zarovnania. Tieto hranice sú ďalej podrobené úprave na základe podporných informácií ako TG..CA box a TSR a tým sa vyberú dôveryhodné LTR. Ďalej sa `LTR_FINDER` pokúsi identifikovať PBS, PPT a RT v sekvenciách medzi pármi LTR pomocou zabudovaných zarovnávacích modulov. Identifikácia RT domény zahŕňa dynamické programovanie, ktoré spracúva posun čítacieho rámca. Čo sa týka ostatných domén, `LTR_FINDER` používa program ps_scan [1]. Na základe informácií z neho sa zostavia ORF. Výstupom programu `LTR_FINDER` sú anotované sekvencie LTR retrotranspozónov s uvedenými úrovňami dôveryhodnosti podľa množstva štruktúrnych prvkov, ktoré sa podarilo

1 z balíka PROSITE, http://www.expasy.org/prosite/

identifikovať. LTR_FINDER dosiahol pri testoch autorov na krátkom genóme kvasinky senzitivitu 100% a špecificitu 96%.

2.5.2 LTR_harvest

LTR_harvest [36] je zameraný na rýchlu detekciu LTR retrotranspozónov vo veľkých genómoch. Vstupom je sekvencia vo formáte FASTA, prípadne viacero sekvencií v jednom súbore vo formáte multi-FASTA.

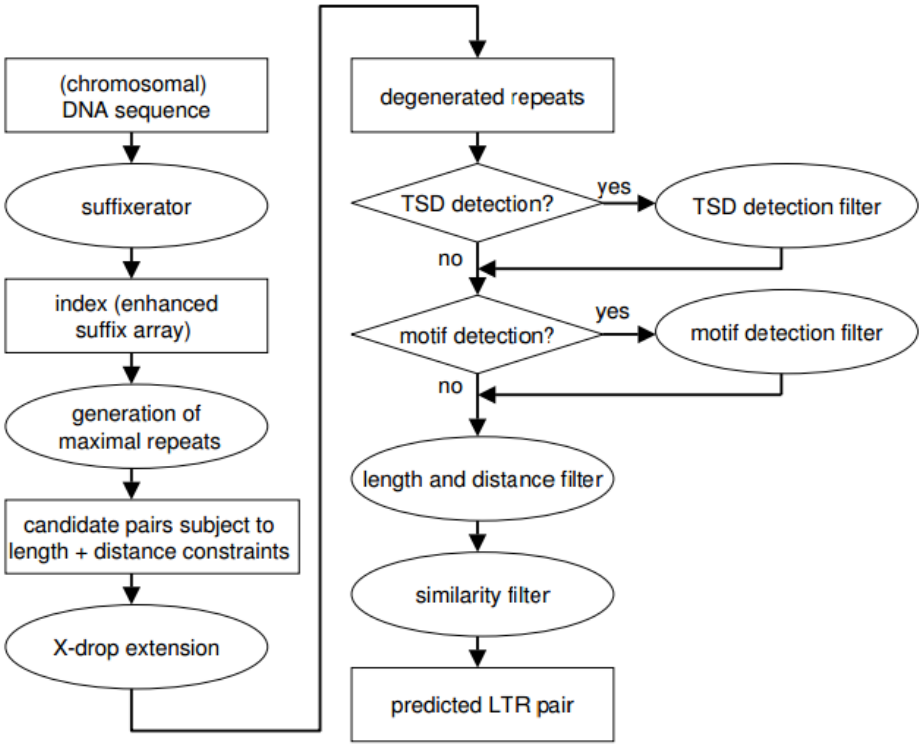
Prvým krokom programu je vytvorenie poľa suffixov zadanej sekvencie. Na to používa program suffixerator, ktorý je súčasťou balíka GenomeTools [37]. Tvorba tohoto poľa je časovo najnáročnejšou predspracovacou fázou vyhľadávania LTR retrotranspozónov. Pole vyžaduje 5n bytov pamäti, kde n je dĺžka vstupnej sekvencie. Narozdiel od iných programov pre vyhľadávanie LTR retrotranspozónov sa pole ukladá do súboru, čím je dosiahnuté oddelenie tvorby poľa a samotnej detekcie LTR retrotranspozónov. To umožní ušetriť výrazné množstvo času pri potrebe mnohonásobného spracovania vstupnej sekvencie s odlišnými parametrami – vytvoriť pole suffixov je potrebné iba raz. Ďalším krokom je vyhľadanie najdlhších bezchybných repetícií (maximal exact repeats) pomocou algoritmu špecifikovaného v [38]. Ich minimálnu dĺžku môže špecifikovať užívateľ. Najdlhšie repetície spĺňajúce užívateľskú požiadavku sú ďalej spracované a určia sa degenerované repetície, nazývané kandidátnymi pármi. Tie sú použité v detekcii špecifických znakov LTR retrotranspozónov, hlavne TSD a LTR. Ďalej nasleduje filtrácia podľa dĺžky, vzdialenosti a podobnosti kandidátnych LTR sekvencií. Ak sú splnené všetky obmedzenia, kandidátny pár je označený za LTR retrotranspozón. LTR_harvest označí jeho pozíciu vo vstupnej sekvencii a niektoré štruktúrne detaily (napr. pozíciu TSD sekvencií, podobnosť LTR) v tabuľkovom formáte. Obrázok 2.2 znázorňuje všetky kroky fungovania programu LTR_harvest.

Pri testovaní podľa autorov programu dosiahol LTR_harvest pri detekcii LTR retrotranspozónov senzitivitu na úrovni 90% na genóme *S. cerevisiae* a viac, než 96% na genóme *D. melanogaster*. Čo sa týka nekompletných, prípadne vnorených LTR retrotranspozónov, LTR_harvest ich nedokáže identifikovať.

2.5.3 LTR Annotator

LTR Annotator [39] je automatizovaný nástroj schopný vykonať identifikáciu a anotáciu LTR retrotranspozónov (princíp fungovania je znázornený na obrázku 2.3). Používa prístup zreťazeného spracovania (pipeline).

Prvým krokom je vyhľadanie transpozónov pomocou programov LTR_FINDER (využíva jeho nízku mieru falošne pozitívnych výsledkov) a LTR_harvest (uňho využíva vysokú citlivosť). Užívateľ má možnosť pomocou parametrov použiť iba jeden z nich, alebo oba; v druhom prípade sú z identifikovaných LTR retrotranspozónov zvolení unikátni kandidáti a zrušené duplicitné elementy.



Obrázok 2.2: Znázornenie funkcie nástroja `LTR_harvest`, prebraté z [36].

Druhý krok pozostáva z vyhľadávania nových LTR retrotranspozónov pomocou prístupu založeného na homológii. Na tento účel používa dve databázy známych LTR retrotranspozónov, a to *MIPS-REdat v9.0* [1] a *TREP* [2]. Databáza *MIPS-REdat* vo verzii 9.0 obsahuje viac, než 42000 transpozónov, a to ako zástupcov triedy I (retrotranspozóny), tak zástupcov triedy II (DNA transpozóny). Na začiatku sa všetky kandidátne LTR sekvencie (identifikované v prvom kroku) podrobia testu na homológiu voči spomenutým databázam pomocou programu `BLASTn` s hranicou e-value na úrovni 1e-30. Každej kandidátnej sekvencii LTR sú po tomto kroku priradené anotácie elementu z databázy podľa najlepšej zhody. LTR kandidáti bez identifikovanej zhody v referenčných databázach sú považované za „nové“.

Tretím krokom je detekcia a odstránenie falošne pozitívnych LTR. Podľa autorov programu sú ich hlavnými zdrojmi duplikátne gény a tandemovo opakované DNA transpozóny. Detekcia a odstránenie prebieha v dvoch krokoch. Prvým je odstránenie tandemových opakovaní DNA transpozónov, druhý vykonáva modul berúci do úvahy počet kópií a homológiu ku známym génom. Wickeret a kolektív [40] navrhli, že potenciálny transpozón by sa mal v genóme vyskytovať aspoň v 5 kópiách. Ak je počet výskytov LTR kandidáta patriaceho do neznámej superrodiny nájdeného pomocou prístupu založeného na homológií menší než 5, alebo ak sú dve kandidátne LTR sekvencie anotované ako členovia jednej génovej rodiny, sú tieto LTR považované za falošne pozitívne. Keďže je však evolúcia LTR dynamickým procesom, a nie vždy je vstupom programu kompletný genóm, minimálny počet kópií LTR je

1 http://mips.helmholtz-muenchen.de/plant/recat/

2 Triticeae repeats, http://wheat.pw.usda.gov/ITMI/repeats

možné špecifikovať aj parametrom. Kandidáti zostávajúci po odstránení falošne pozitívnych výsledkov v tomto kroku sú považovaní za správne identifikované LTR; tým je dokončená fáza identifikácie LTR.

Ďalšou fázou je anotácia LTR, ktorá pozostáva z analýzy domén, analýzy susedstva LTR sekvencií, ďalšieho odhadu počtu kópií, klasifikácie do rodín, ďalší krok detekcie falošných pozitívnych výsledkov a analýza divergencie. Na konci prebehne sumarizácia a prezentácia výsledkov.

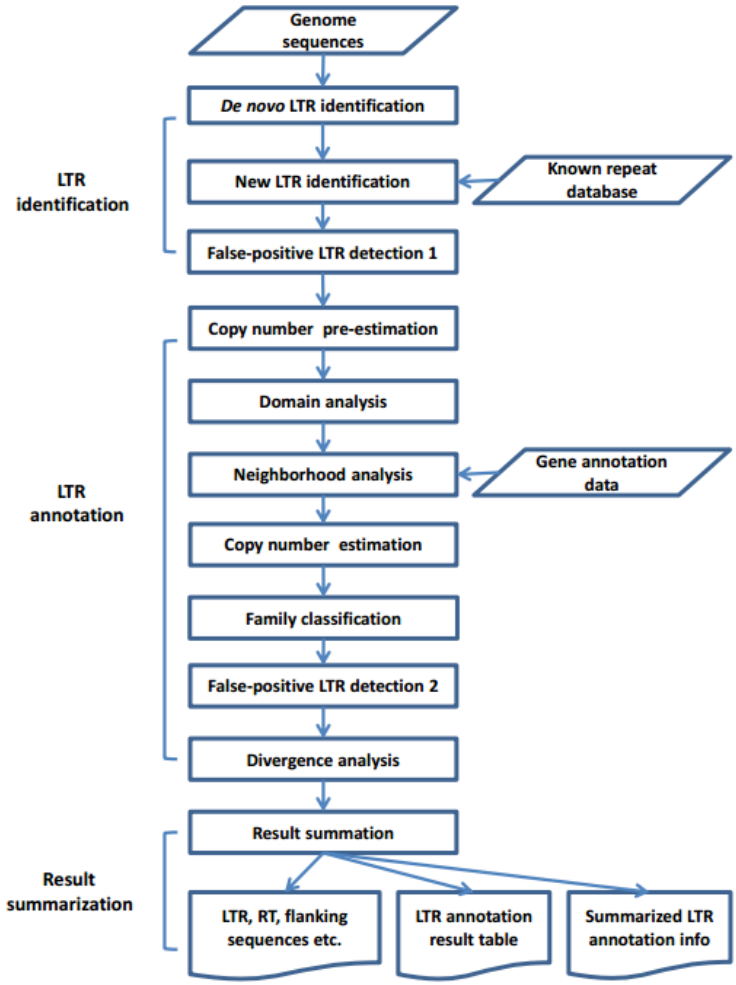
2.6 Výber nástroja pre vyhľadávanie LTR retrotranspozónov

V tejto podkapitole sa zameriam na porovnanie dvoch vybraných bioinformatických nástrojov určených na vyhľadávanie LTR retrotranspozónov v DNA sekvencii. Porovnávať budem nástroje `LTR_FINDER` a `LTR_harvest`. Túto dvojicu som zvolil na základe výsledkov štúdie [42]. Nástroj `LTR_harvest` v nej identifikoval najväčší počet referenčných LTR retrotranspozónov zo všetkých porovnávaných nástrojov a tiež bol schopný nájsť druhý najvyšší počet nereferenčných LTR retrotranspozónov. Nástroj `LTR_FINDER` relatívne veľké množstvo referenčných LTR retrotranspozónov nenašiel, no výrazne sa v úspešnosti neodlišuje od ostatných porovnávaných nástrojov. Zároveň je jedným z najpoužívanejších nástrojov na vyhľadávanie LTR retrotranspozónov (podľa štúdie [42] bol v roku 2014 použitý v 6 štúdiách, čím sa nachádza na rovnakej úrovni používanosti, ako ostatné najpoužívanejšie nástroje).

2.6.1 Testovacia sada

Za účelom testovania úspešnosti vyhľadávania LTR retrotranspozónov je potrebné mať k dispozícii už známe sekvencie transpozónov. Rozhodol som sa použiť sekvencie z *MIPS Repeat Element Database* (mips-REdat)[1] vo verzii 9.3, ktorá obsahuje 61730 sekvencií LTR retrotranspozónov rôznych rastlinných druhov. Táto databáza je kompiláciou verejne dostupných databáz (*TREP*, *TIGR Repeats*, *PlantSat*, *Genbank*) obohatená o *de-novo* detekované LTR retrotranspozóny (pomocou programu `LTR_STRUC`) v genómoch v databáze PGSB PlantsDB. Pre vylúčenie redundancie boli tieto sekvencie autormi databázy zhlukované s identitou ≥95% na ≥95% dĺžky.

Z tejto databázy som vzhľadom na náš aktuálne prebiehajúci výskum na Biofyzikálnom ústave AVČR vybral LTR retrotranspozóny nasledujúcich rodov: *Arabidopsis, Brachypodium, Glycine, Gossypioides, Gossypium, Hordeum, Lotus, Lycopersicon, Medicago, Oryza, Physcomitrella, Solanum, Sorghum, Triticum, Zea*. Následne som pre ďalšie vylúčenie redundancie aplikoval zhlukovanie s identitou ≥88%. Výsledný fasta súbor obsahujúci 32177 sekvencií LTR retrotranspozónov sa nachádza v zložke s nástrojom pod názvom `mipsREdatSelectedTEs.fa`. V tabuľke 2.1 sa nachádza prehľad počtu zástupcov jednotlivých rodov usporiadaný podľa početnosti.



Obrázok 2.3: Znázornenie funkcie programu `LTR Annotator`, prebraté z [39].

Rod	Počet zástupcov
Zea	7208
Gossypium	5378
Sorghum	4007
Glycine	3499
Hordeum	2041
Triticum	1874
Oryza	1613
Lycopersicon	1332
Solanum	1267
Arabidopsis	1031

Physcomitrella	953
Medicago	607
Brachypodium	575
Lotus	560
Gossypioides	232

Tabuľka 2.1: Prehľad počtu LTR retrotranspozónov podľa jednotlivých rodov v testovacej sade zoradený podľa početnosti.

2.6.2 Úspešnosť vyhľadávania v databáze známych LTR retrotranspozónov

V tejto podkapitole popíšem testovanie oboch zvolených nástrojov pri identifikácii LTR retrotranspozónov vo fasta súbore, ktorý obsahoval iba sekvencie známych LTR retrotranspozónov. Pomocným skriptom som vygeneroval fasta súbor obsahujúci rádovo 1000 náhodne vybraných sekvencií z pracovnej sady. Pre `LTR_FINDER` som použil nasledujúce parametre:

```
-S0 -D30000 -d200 -o2 -p15 -g80 -G5 -T1 -B0.3 -b0.25 -r12
```

Pre `LTR_harvest` som použil nasledujúce parametre:

```
-minlenltr 200 -maxdistltr 30000 -similar 70 -overlaps best
```

Parametre boli zvolené tak, aby bolo vyhľadávanie čo najmenej prísne (napríklad hodnota parametra -p u programu `LTR_FINDER` je na spodnej hranici odporúčania autorov programu), čím by sa mala maximalizovať detekčná schopnosť oboch programov, a zároveň aby sa tieto dva experimenty dali medzi sebou porovnať (parameter -d200 u nástroja `LTR_FINDER` zodpovedá parametru -minlenltr 200 u nástroja `LTR_harvest`, podobne -D30000 zodpovedá parametru -maxdistltr 30000). Tento experiment som zopakoval päťkrát. Výsledky týchto piatich experimentov sú uvedené v tabuľke 2.2. Priemerná senzitivita programu `LTR_FINDER` pri vyhľadávaní LTR retrotranspozónov v sekvenciách známych LTR retrotranspozónov dosiahla 82,91%, zatiaľčo priemerná úspešnosť nástroja `LTR_harvest` dosiahla 85,91%.

Beh	Počet prítomných LTR retrotranspozónov	Senzitivita LTR_FINDER [%]	Senzitivita LTR_harvest [%]
1.	1000	82,60	87,20
2.	1028	83,95	83,46
3.	1027	83,93	88,61
4.	958	82,36	83,30
5.	967	81,70	87,00

Tabuľka 2.2: Výsledky vyhľadávania LTR retrotranspozónov programami `LTR_FINDER` a `LTR_harvest` v náhodne vybranej sade ~1000 LTR retrotranspozónov z testovacej sady.

Tieto výsledky potvrdzujú závery štúdie [42] – `LTR_harvest` je skutočne schopný identifikovať LTR retrotranspozóny vo väčšom množstve prípadov, hoci rozdiel v tomto experimente nie je veľký (iba 3%). Ako výrazná slabina nástroja `LTR_FINDER` sa však ukázal fakt, že v sekvencii obsahujúcej jediný LTR retrotranspozón často identifikuje viacero rôznych alternatív, ktoré sa líšia dĺžkou, skóre, divergenciou LTR sekvencií a ďalšími vlastnosťami. Vo výstupe programu takýto prípad vyzerá nasledovne:

```
>Sequence: RLG_161363_Hordeum Len:16393
index SeqID Location LTR len  Inserted element len TSR   PBS   PPT
   RT IN (core)   IN (c-term) RH Strand   Score Sharpness   Similarity
[ 1]  RLG_161363_Hordeum   1-16393  412,412  16393 N  N-N   15967-
15981 747-1253 14871-15179 N-N   N-N   +  19 1,1   0.983
[ 2]  RLG_161363_Hordeum   3238-11704  877,878  8467  CCCGA N-N
   10807-10821 4320-4397   N-N   N-N   N-N   +  11 0.443,0.529 0.967
[ 3]  RLG_161363_Hordeum   5418-13870  2088,2083   8453  N  N-N
   7537-7551   11013-11099 N-N   N-N   N-N   -  6.5   0.471,0.486
   0.974
```

Je vidieť, že v sekvencii obsahujúcej jediný LTR retrotranspozón (RLG_161363) `LTR_FINDER` označil tri alternatívy jeho sekvencie. V prípade, akým bol napríklad tento experiment, kde vieme, že každá sekvencia obsahuje iba jeden LTR retrotranspozón, to nepredstavuje problém (môžeme napríklad vybrať ten, ktorý pokrýva najväčšiu časť danej sekvencie). Avšak pri každom inom scenári by táto vlastnosť komplikovala spracovanie výsledkov, keďže by po behu nástroja bolo nutné ďalej spracovať jeho výstup tak, aby sa z každej skupiny prekrývajúcich sa identifikovaných LTR retrotranspozónov vybral iba jeden, najlepší zástupca. Bez tohoto kroku by nebolo možné povedať ani to, koľko LTR retrotranspozónov vlastne `LTR_FINDER` našiel. Hoci som sa pokúšal optimalizovať parametre (hlavne -g, -j a -J, ktoré ovládajú rozširovanie potenciálnych LTR sekvencií) tak, aby sa vo výstupe nachádzala maximálne 1 varianta každého LTR retrotranspozónu, nepodarilo sa mi to. Prípadne ak sa to podarilo, bolo to za cenu neprípustného zníženia citlivosti

identifikácie LTR retrotranspozónov. Kvôli tomuto dôvodu, a tiež kvôli celkovo nižšej úspešnosti vyhľadávania, som sa rozhodol z testovanej dvojice vo finálnom nástroji použiť na identifikáciu LTR retrotranspozónov nástroj `LTR_harvest` a ďalej testovať iba jeho vlastnosti.

2.6.3 Úspešnosť vyhľadávania v umelo vytvorenom genóme

V tomto experimente testujem úspešnosť programu `LTR_harvest` pri vyhľadávaní LTR retrotranspozónov v umelo vygenerovanom genóme. Pomocou pomocného skriptu `randomGenome.py` v programovacom jazyku `python` generujem DNA sekvenciu dlhú 10 Mbp. Tá pozostáva z náhodne vybratých LTR retrotranspozónov z databázy mips-REdat, oddelených od seba náhodne vygenerovanými DNA sekvenciami (každý nukleotid sa na každej pozícii vyskytuje s pravdepodobnosťou 25%) o dĺžke 0-100 Kbp (s uniformnou distribúciou dĺžky). Sekvencia sa generuje, až kým nepresiahne dĺžku 10 Mbp, takže počet LTR retrotranspozónov aj presná dĺžka náhodného genómu sú medzi jednotlivými behmi experimentu premenlivé. Samozrejme, presný počet LTR retrotranspozónov prítomný v sekvencii je mi známy, takže som schopný zhodnotiť úspešnosť vyhľadávania. V ideálnom prípade by sa nemala líšiť od úspešnosti v predchádzajúcom experimente (vyhľadávanie v sekvenciách známych LTR retrotranspozónov). Výsledky jednotlivých behov tohoto experimentu sa nachádzajú v tabuľke 2.3.

Priemerná úspešnosť vyhľadávania LTR retrotranspozónov v tomto experimente dosiahla 83,79%, čo je mierne horší výsledok, než pri predchádzajúcom experimente.

2.6.4 Úspešnosť vyhľadávania v reálnom genóme

Tento experiment slúži na vyhodnotenie úspešnosti programu `LTR_harvest` pri vyhľadávaní LTR retrotranspozónov v skutočnom genóme. Ako testovací organizmus som si vybral *Arabidopsis thaliana*, čo je rastlina, ktorá sa kvôli svojmu rýchlemu životnému cyklu a relatívne krátkemu genómu veľmi často používa v bioinformatickom výskume; je to tiež rastlina, ktorej genóm bol osekvenovaný ako prvý. Sekvenciu genómu tejto rastliny som získal z databázy *The Arabidopsis Information Resource* (TAIR)[1]. Genóm je zložený z 5 chromozómov a má dĺžku ~120Mbp.

Databáza *mips-REdat* obsahuje 1322 LTR retrotranspozónov identifikovaných v genóme *A. thaliana*. Tento počet budem považovať za 100% LTR retrotranspozónov identifikovateľných programom `LTR_harvest`. Pri vyhľadávaní som použil nasledujúce parametre: `-maxlenltr 8000 -maxdistltr 30000 -similar 70 -overlaps best`. `LTR_harvest` identifikoval 2045 LTR retrotranspozónov. Je to možné vysvetliť tým, že mnohé z nich sú si natoľko podobné, že pri zhlukovaní databázy *mips-REdat* autormi tejto databázy sú v nej reprezentované iba jedným zástupcom, zatiaľčo `LTR_harvest` pochopiteľne našiel všetky individuálne kópie.

1 Genóm dostupný na https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_chromosome_files

Beh	Počet prítomných LTR retrotranspozónov	Počet identifikovaných LTR retrotranspozónov	Senzitivita [%]
1.	179	163	91,06
2.	173	144	83,24
3.	182	151	82,97
4.	160	130	81,25
5.	173	147	84,97
6.	160	134	83,75
7.	177	144	81,36
8.	179	151	84,36
9.	163	143	87,73
10.	171	136	79,53

Tabuľka 2.3: Výsledky vyhľadávania LTR retrotranspozónov programom `LTR_harvest` v náhodne vygenerovanom genóme obsahujúcom známy počet LTR retrotranspozónov náhodne vybratých z databázy *mips-REdat*.

3 Návrh nového nástroja pre anotáciu transpozónov

Z predchádzajúcej kapitoly je možné vidieť, že bioinformatických nástrojov zaoberajúcich sa vyhľadávaním a anotáciou transpozónov existuje pomerne veľké množstvo. Niektoré sú schopné nájsť doposiaľ neznáme kópie iba podľa ich vnútornej štruktúry, pre iné vnútorná štruktúra nie je smerodajná a hľadia viac na podobnosť sekvencií v skúmanom genóme s už známymi transpozónmi. Viacero z dostupných nástrojov je tiež schopných okrem identifikácie transpozónových sekvencií tieto aj anotovať, teda vyhľadať a prezentovať jednotky ich vnútornej štruktúry, ako napríklad dlhé koncové repetície u LTR retrotranspozónov, proteínové domény, PBS alebo PPT. Taktiež sú niektoré nástroje schopné vnútorné štruktúry aj ďalej analyzovať, napríklad poskytnúť informáciu o divergencii LTR a tým aj relatívny odhad času vloženia daného elementu na aktuálne miesto v genóme.

Čo sa však týka vyhľadávania a anotácie prídavných štruktúrnych prvkov transpozónov pomocou existujúcich nástrojov, situácia je iná. Súčasné nástroje na vyhľadávanie a anotáciu transpozónov neposkytujú možnosť prídavné štruktúry vyhľadať a anotovať. Samozrejme, existujú samostatné nástroje na vyhľadávanie tandemových repetícií, alebo ORF v sekvenciách DNA. Tie je možné manuálne použiť na sekvencie transpozónov. Avšak bolo by vítaným vylepšením mať k dispozícii jeden nástroj schopný anotácie všetkých žiadaných prídavných štruktúrnych prvkov automaticky.

Vzhľadom na aktuálne prebiehajúci výskum na Biofyzikálnom ústave Akadémie vied Českej Republiky v Brne (BFÚ AVČR), ktorý je zameraný špecificky na LTR retrotranspozóny v rastlinných druhoch, bude tento nástroj taktiež zameraný výlučne na LTR retrotranspozóny. Výstup tejto diplomovej práce tak bude priamo použiteľný v praxi.

3.1 Návrh riešenia

Úlohou implementovaného nástroja bude anotácia vnútornej štruktúry LTR retrotranspozónov so zameraním na prídavné štruktúry (tandemové repetície, eORF). Vstupom programu budú DNA sekvencie uložené v súborovom formáte *FASTA*. Výstupom bude anotačný súbor v užívateľom špecifikovanom formáte obsahujúci anotácie identifikovaných LTR retrotranspozónov a ich štruktúrnych prvkov a súbory vo formáte *csv* obsahujúce skupiny konzervovaných štruktúr.

3.1.1 Tandemové repetície vnútri LTR retrotranspozónov

Ako bolo spomenuté v kapitole 2, transpozóny a tandemové repetície sú dva základné typy repetitívnych sekvencií v DNA organizmov. Napriek tomu je relatívne málo známych faktov o ich vzájomnej koexistencii a interakcii, napríklad čo sa týka tandemových repetícií vnútri transpozónov. Existuje množstvo programov, ktoré dokážu v DNA sekvenciách identifikovať tandemové repetície. Pri rozhodovaní o výbere som vychádzal z výsledkov štúdie [43], v ktorej

autori porovnávali najčastejšie používané nástroje na vyhľadávanie tandemových repetícií. Z nich vyšiel konzistentne najlepšie nástroj `Mreps`, tesne nasledovaný nástrojom `Tandem Repeats Finder` (TRF) [41]. Autori však uvádzajú, že rozdiely v úspešnosti nie sú natoľko veľké, aby bol jeden nástroj lepší vo všeobecnom prípade použitia; vždy je potrebné hľadieť na účel daného výskumu. Nástroj Mref je napríklad schopný pracovať s repetíciami obsahujúcimi iba mismatche, a nie už inzercie a delécie, zatiaľčo `Tandem Repeats Finder` je schopný pracovať so všetkými spomenutými. Pri skúmaní tandemových repetícií vnútri transpozónov je možné očakávať prítomnosť inzercií a delécií, preto som sa rozhodol na vyhľadávanie tandemových repetícií v implementovanom nástroji použiť `Tandem Repeats Finder`.

3.1.2 Vyhľadávanie eORF

Aktívne LTR retrotranspozóny obsahujú vo svojej sekvencii proteínové domény, ktoré kódujú jednotlivé funkcie, ako napríklad schopnosť retrotranspozície. Tieto proteínové domény sa nachádzajú vnútri čítacích rámcov (ORF). Okrem čítacích rámcov obsahujúcich známe proteínové domény (popísané v predchádzajúcej kapitole) však sekvencie LTR retrotranspozónov často obsahujú aj ďalšie ORF. Je možné, že niektoré z nich kódujú doposiaľ neznámu funkciu. Za účelom anotácie otvorených čítacích rámcov som sa rozhodol využiť program Open Reading Frame Finder (`ORF Finder`) vyvinutý Národným centrom pre biotechnologické informácie (National Center for Biotechnology Information, NCBI). Keďže vyhľadávanie ORF je triviálna úloha, neexistujú ani štúdie porovnávajúce jednotlivé vyhľadávače ORF. Tento nástroj som teda zvolil kvôli dôveryhodnosti zdroja – činnosť NCBI je v podstate bioinformatickým štandardom.

3.1.3 Identifikácia proteínových domén

Nástroj `ORF Finder` je schopný čítacie rámce identifikovať, ale nehovorí nič o tom, aká (ak vôbec nejaká) proteínová doména sa nachádza v ich vnútri. Hoci primárnym účelom implementovaného nástroja je anotácia prídavných štruktúr LTR retrotranspozónov, za účelom zistenia, aké funkcie jednotlivé ORF kódujú, je potrebné vo všetkých identifikovať proteínové domény. Okrem anotácie prídavných ORF bude teda nástroj poskytovať aj anotáciu obsahu všetkých identifikovaných ORF. Na tento účel je potrebné použiť ďalší nástroj. Štandardným nástrojom na vzájomné porovnávanie biologických sekvencií (DNA, proteínové sekvencie) je balík `BLAST`. Vzájomné porovnávanie preto, lebo proteínové domény v ORF sa najčastejšie identifikujú tak, že sa sekvencie ORF porovnávajú s anotovanou databázou známych proteínových sekvencií. Presne takto fungujú programy z balíčka `BLAST`. Balíček obsahuje (okrem iného) nástroje na porovnávanie nukleotidových sekvencií (`BLASTn`), proteínových sekvencií (`BLASTp`), nukleotidových sekvencií voči proteínovým sekvenciám (`BLASTx`) a proteínových sekvencií voči nukleotidovým sekvenciám (`tBLASTn`). Prípady použitia konkrétnych programov z balíka na identifikáciu obsahu jednotlivých otvorených čítacích rámcov sú popísané v ďalšej kapitole zaoberajúcej sa implementáciou.

3.1.4 Vyhľadávanie konzervovaných tandemových repetícií a extra ORF

Po samotnej identifikácii a anotácii jednotlivých prídavných štruktúr LTR retrotranspozónov je hlavne zaujímavé zistiť, či sú niektoré z týchto štruktúr prítomné u viacerých kópií LTR retrotranspozónov z jednej podrodiny, a to prípadne aj na rovnakých, alebo podobných, pozíciách. Je oprávnené očakávať, že v mnohých LTR retrotranspozónoch sa budú nachádzať rôznorodé tandemové repetície a otvorené čítacie rámce neobsahujúce žiadnu známu proteínovú doménu. Zaujímavejšie však bude, ak sa ukáže, že niektoré z takýchto štruktúr sú u niektorých LTR retrotranspozónov stabilne prítomné, teda konzervované. To by naznačovalo možnosť ich konkrétnej funkcie, alebo prinajmenšom spoločného pôvodu (pretože jednotlivé kópie pôvodného LTR retrotranspozónu si zachovajú štruktúrne prvky, napríklad extra ORF, aj napriek tomu, že im neposkytuje žiadnu funkciu). V nástroji teda implementujem aj túto funkcionalitu.

3.1.5 Formát vstupného súboru

Nástroj bude akceptovať vstupný súbor vo formáte fasta. Jeho obsahom môže byť jedna alebo viacero sekvencií DNA. Každá sekvencia vo vstupnom súbore je uvedená unikátnym identifikátorom. Identifikátor, ktorý sa musí nachádzať na jednom riadku priamo pred samotnou nukleotidovou sekvenciou, začína symbolom „>“, ktorý je priamo nasledovaný identifikačným reťazcom neobsahujúcim medzery. Nasleduje príklad správneho formátu sekvencie vo vstupnom súbore:

```
>RLG_3398
aagctttcatggtgtagcgaaagtccgtatgagtctttggctttgtatgttctaacaaggaaacactgct
taggcctataagatcgggttgcggtttaagttcttatactctgatatgtctataatttgcatgatttagg
cattcattcttcaccactttgtcattgtttcatccacatttcatctcattaggagtcgttttagtgtgtt
ttacatacttaggacgattttgcattaggattgcatgtgcatggcatatttgg
```

3.1.6 Výstup nástroja

Výstupom nástroja budú textové súbory obsahujúce anotačné informácie o jednotlivých identifikovaných štruktúrach. Tieto sa budú nachádzať v súboroch štandardného bioinformatického formátu (*BED* alebo *GFF*). Ďalej budú výstupom dva súbory vo formáte *csv*, ktoré budú obsahovať identifikované skupiny konzervovaných extra ORF a tandemových repetícií. Nástroj taktiež užívateľovi ponechá výstupy jednotlivých externých nástrojov, ktoré pre svoju činnosť používa, takže užívateľ bude mať prístup ku všetkým čiastkovým výsledkom.

4 Implementácia

Za implementačnú formu nástroja som zvolil skript v programovacom jazyku `Python` vo verzii 2.7.9, a to hlavne kvôli jednoduchej práci s textovými súbormi (keďže časť práce nástroja spočíva v analýze textových výstupov existujúcich nástrojov) a so zložitými štruktúrami, hlavne so slovníkmi. Nástroj je určený pre operačné systémy typu Unix. Implementácia a testovanie prebiehalo v linuxovom prostredí s distribúciou `Debian` vo verzii `3.16.36-1` na výpočtovom centre `MetaCentrum`[1]. Na identifikáciu transpozónov, otvorených čítacích rámcov, tandemových repetícií a proteínových domén tento nástroj využíva existujúce bioinformatické nástroje (`LTR_harvest`, `ORFFinder`, `Tandem Repeats Finder`, `BLASTp`, v tomto poradí).

Keďže výstupom tejto diplomovej práce je bioinformatický nástroj, ktorý pre svoje fungovanie používa iné bioinformatické nástroje, pre vyhnutie sa mnohoznačnosti v texte budem slovom „nástroj“ bez ďalšej špecifikácie vždy myslieť mnou implementovaný nástroj, zatiaľčo pri spomínaní ostatných programov budem vždy špecifikovať, o ktorý sa jedná.

Hoci to nie je všeobecnou podmienkou súborového formátu fasta, kvôli využitiu programu `ORF Finder` identifikátor nesmie obsahovať znak „|“. Identifikátor taktiež nesmie obsahovať biele znaky. Tieto podmienky kontroluje funkcia `checkInputFile`, nachádzajúca sa v zdrojovom súbore `func.py`.

Výstup nástroja obsahujúci anotácie jednotlivých identifikovaných základných a prídavných štruktúr je možné vytvoriť vo formáte *BED* alebo *GFF*. Detailný popis výstupného formátu sa nachádza v prílohe 2. Diagram detailne znázorňujúci princíp fungovania nástroja sa nachádza na obrázku 4.1 na konci tejto kapitoly.

4.1 Štruktúra zdrojového textu

Kvôli prehľadnosti je implementácia nástroja rozdelená do viacerých súborov so zdrojovými textami, ktoré sa nachádzajú v jednom adresári. Jedná sa o nasledujúce súbory:

- **dp.py** - hlavný súbor, ktorým sa celý nástroj spúšťa. Obsahuje spracovanie a kontrolu parametrov z príkazového riadku pomocou knižnice `argparse`, postupne spúšťa jednotlivé nástroje a podprogramy na analýzu ich výstupov a ich načítanie do štruktúr v pamäti. Nakoniec sa tu nachádza spúšťanie podprogramov vytvárajúcich výstupné anotačné súbory vo formáte *BED* alebo *GFF*.
- **parsers.py** - nachádzajú sa tu funkcie určené na načítanie a spracovanie dát z výstupných súborov jednotlivých nástrojov, ktoré sú spúšťané v hlavnom súbore dp.py. Konkrétne sú to funkcie `parseTRFOutput`, `parseLTR_harvestOutput`, `parseORFFinderOutput`, `parseBlastpOutput`.
- **createOutput.py** - obsahuje funkcie `createGFF` a `createBED`, ktoré umožňujú vytvoriť anotačné súbory v týchto dvoch fotmátoch obsahujúce užívateľom vybrané identifikované štruktúry.

1 https://metavo.metacentrum.cz/

- **classes.py** - obsahuje definície dátových štruktúr pre vyhľadávané biologické štruktúry, konkrétne sú to štruktúry `Domain`, `Cluster`, `ORF`, `TandemRepeat` a `Transposon_ltrharvest`.
- **runCommands.py** - v tomto súbore sa nachádza konštrukcia príkazov príkazového riadku a spúšťanie jednotlivých nástrojov, a tiež funkcie `identifyConservedExtraORFs` a `identifyConservedSatellites`.
- **func.py** - Obsahuje ďalšie pomocné funkcie, napríklad načítanie parametrov používaných bioinformatických nástrojov z textového súboru a vytváranie rôznych pomocných súborov.

4.2 Parametre nástroja

Tento nástroj využíva viacero externých bioinformatických nástrojov, pričom každý z nich má vlastnú sadu parametrov, niekedy s neprázdnym prienikom s ostatnými sadami. Preto by nebolo praktické uvádzať všetky parametre pri spúšťaní nástroja na príkazovom riadku. Zvolil som riešenie, kde parametre pre jednotlivé nástroje budú uvedené v súbore `params.txt`, nachádzajúcom sa v rovnakej zložke, ako zdrojové súbory nástroja.

Parametre pre vyhľadávanie ORF sa nachádzajú na riadku začínajúcom reťazcom `#orf`, parametre pre vyhľadávanie LTR retrotranspozónov na riadku začínajúcom reťazcom `#ltr`, a parametre pre vyhľadávanie tandemových repetícií na riadku začínajúcom reťazcom `#trf`. Jednotlivé parametre na riadkoch musia byť oddelené bielymi znakmi (napr. medzera, tabulátor). Poradie riadkov môže byť ľubovoľné. Na konci súboru sa musí nachádzať jeden prázdny riadok.

Niektoré (pre daný bioinformatický nástroj potenciálne voliteľné) parametre nástroj na svoj beh nutne potrebuje (napr. pre špecifikáciu formátu výstupu); takéto parametre nie je potrebné zadávať v súbore `params.txt`, pretože sú napevno pridané ku príslušnému reťazcu parametrov priamo v zdrojovom kóde nástroja. Konkrétne prípady sú popísané v prílohe.

Parametre nástroja, ktoré sa zadávajú pri spustení na príkazový riadok sú nasledujúce (všetky sú povinné):

- **-i, --inputFile**: cesta ku vstupnému súboru.
- **-o, --outputFile**: názov výstupného súboru.
- **-outfmt, --outputFormat**: formát výstupného súboru; povolené hodnoty sú „bed“ a „gff“.
- **-outfeat, --outputFeature**: špecifikácia štruktúr, pre ktoré má byť vytvorený výstupný súbor; povolené hodnoty sú „orf“, „sat“, „te“ a „all“.
- **-ctnt, --clusteringThresholdNT**: hodnota prahu zhlukovania nukleotidových sekvencií, v rozmedzí 0-1.
- **-ctaa, --clusteringThresholdAA**: hodnota prahu zhlukovania proteínových sekvencií, v rozmedzí 0-1.
- **-et, --evalueThreshold**: prah hodnoty e-value pre identifikáciu proteínových domén, platné hodnoty sú ≥ 0.

- **-mc, --minNumCopies**: minimálny počet kópií monoméru v tandemovej repetícii, zadaná hodnota musí byť kladné celé číslo.

4.3 Vyhľadávanie retrotranspozónov

Na vyhľadávanie LTR retroranspozónov vo vstupnej sekvencii nástroj využíva program `LTR_harvest` vo verzii 1.5.9, ktorý je súčasťou balíka `GenomeTools`. Celý tento balík sa nachádza v zložke `genometools-1.5.9/`, ktorá sa nachádza v hlavnej zložke nástroja. Užívateľ bude parametre vyhľadávania LTR retrotranspozónov uvádzať do súboru `params.txt` na riadok začínajúci identifikátorom `#ltr`.

4.4 Vyhľadávanie ORF

Ako jeho vstup sa použije hlavný vstupný súbor popísaný vyššie. V ňom obsiahnuté identifikátory sekvencií nesmú obsahovať znak „|“ kvôli tomu, že v takom prípade sa vo výstupnom súbore programu `ORF Finder` tento identifikátor neobjaví. Napríklad, pri vyhľadaní ORF v sekvencii s identifikátorom „>RLX_56183“ sa vo výstupe ORF Finderu budú nachádzať identifikátory ORF vo formáte „>lcl|ORF1_RLX_56183:2507:2148“, zatiaľčo pri vyhľadaní ORF v sekvencii s identifikátorom „>RLX_56183|abcd“ sa vo výstupnom súbore programu `ORF Finder` budú nachádzať identifikátory vo formáte „>lcl|ORF1_1:2507:2148“. Pôvodný identifikátor sa teda stratil, a bol nahradený číslom „1“, ktoré označuje poradové číslo sekvencie s týmto identifikátorom vo vstupnom súbore. Pre túto komplikáciu môj nástroj neumožňuje pracovať so vstupnou sekvenciou obsahujúcou znaky „|“ v identifikátoroch.

4.4.1 Parametre vyhľadávania ORF

Užívateľ môže v súbore `params.txt` špecifikovať nasledujúce parametre behu programu ORF Finder. Podrobnejší prehľad ostatných parametrov je dostupný v prílohe 1.

- **-ml** <Integer>: minimálna dĺžka ORF (v počte nukleotidov); pri zadaní hodnoty menšej, než 30, sa prah automaticky nastaví na 30, implicitná hodnota je 75.
- **-n** <Boolean>: ignorovanie kompletne vnorených ORF; implicitná hodnota je „false“.
- **-strand** <String>: špecifikácia vlákna, na ktorom sa majú vyhľadávať čítacie rámce; platné hodnoty sú „both“, „plus“ a „minus“, implicitná hodnota je „both“.

4.5 Vyhľadávanie tandemových repetícií

Na vyhľadávanie tandemových repetícií nástroj využíva binárnu verziu programu `Tandem Repeats Finder` vo verzii 4.09. Spustiteľný súbor má názov `trf409.linux64` a nachádza sa v hlavnej zložke, ktorá obsahuje všetky zdrojové súbory nástroja. Užívateľ musí špecifikovať parametre programu `Tandem Repeats Finder` v súbore `params.txt` na riadok začínajúci reťazcom `#trf`. Detailný popis týchto parametrov sa nachádza v prílohe 1. Parametre programu `Tandem Repeats Finder` sú v súbore `params.txt` pre užívateľa prednastavené na reťazec „2 7 7 80 10 50 2000 -h“. Tieto parametre boli zvolené podľa implicitných parametrov webovej

verzie nástroja `Tandem Repeats Finder`. Prepínač „-h“ je prednastavený, aby potenciálne veľké množstvo html súborov bolo vytvorené až v prípade, že to užívateľ explicitne potrebuje.

4.6 Identifikácia proteínových domén a rodín

Pri identifikácii proteínových domén nástroj využíva prístup založený na vyhľadávaní v databáze proteínových sekvencií pomocou programu `BLASTp` vo verzii 2.6.0. Zložka s nástrojom obsahuje binárnu verziu programu `BLASTp`, a to v zložke `ncbi-blast-2.6.0+/bin`. Vstupom sú jednotlivé proteínové sekvencie otvorených čítacích rámcov, vyhľadané nástrojom `ORF Finder`, nachádzajúce sa v súbore `ORF.out`. Parametre a formát výstupu nástroja `BLASTp` použité v implementovanom nástroji sú podrobne popísané v prílohe 1. Nástroj pre identifikáciu proteínových domén využíva vlastnú databázu, ktorú som vytvoril z anotovaných sekvencií všetkých proteínových domén dostupných v databáze `Gypsy Database`, súbor dostupný z [1]. Databázu vo formáte, ktorý program blastp potrebuje, som vytvoril príkazom „`makeblastdb -in cores-database -parse_seqids -dbtype prot`“. Databáza obsahuje sekvencie všetkých základných retrotranspozónových domén (RT, INT, RNaseH, GAG), a okrem nich množstvo ďalších. Program `makeblastdb` je súčasťou balíčka `BLAST` a nachádza sa v zložke `ncbi-blast-2.6.0+/bin`. Databáza vytvorená týmto príkazom sa nachádza v zložke `Cores`. V tejto zložke sa nachádza aj súbor `Cores-readme`, ktorý obsahuje okrem iného zoznam a popis skratiek používaných pre jednotlivé proteínové domény.

Pôvodná databáza je vo formáte fasta. Každý záznam v nej má identifikátor rozdelený na dve časti znakom „_“. Prvá časť je skratkou proteínovej domény, ktorej sekvenciu tento záznam obsahuje, a druhá časť identifikuje LTR retrotranspozón, z ktorého sekvencia pochádza. Uvediem príklad takéhoto záznamu:

```
>RT_Ogre
WDAGFLAVTSYPPWMANIVPVPKKDGKVRMCVDYRDLNRASPKDDFPLPHIDVLVDNTAQSSVFSFMDGF
SGYNQIKMAPEDMEKTTFITPWGTFCYKVMPFGLKNAGATYQRAMTTLFHDMMHKEIEVYVDDMIAKSQT
EEEHLVNLQKLFDRLRKFKLRLNPNKCTFGVRSGKLLGFIVSEKGIEVDPAKVKAIQEMPEPKTEKQVRG
FLGRLNYIARFISHLT
```

Tento záznam obsahuje proteínovú sekvenciu jadra RT domény z LTR retrotranspozónu patriaceho do rodiny Ogre. V tomto kroku tak zároveň s identifikáciou proteínovej domény obsiahnutej v danom ORF nástroj zistí aj pravdepodobnú retrotranspozónovú rodinu, do ktorej skúmaný LTR retrotranspozón patrí.

4.7 Identifikácia konzervovaných eORF

Hlavným prínosom tejto diplomovej práce na poli bioinformatických nástrojov je automatizovaná identifikácia konzervovaných extra ORF (a konzervovaných tandemových repetícií, popísané v ďalšej podkapitole). Po predchádzajúcom kroku, identifikácii proteínových

1 http://gydb.org/images/1/1a/Cores.zip

domén v rozpoznaných ORF, sa nástroj zameria na tie ORF, v ktorých nebola rozpoznaná žiadna doména (vrátane tých ORF, kde doména síce rozpoznaná bola, ale s vyššou hodnotou e-value, než bol užívateľom zadaný prah).

Niekedy sa stane, že u jedného LTR retrotranspozónu sú niektoré proteínové domény identifikované vo viac, než jednom ORF. V takom prípade sa pomocou funkcie `consolidateTEs` štruktúra obsahujúca záznamy o všetkých identifikovaných LTR retrotranspozónoch upraví tak, že u každého ostane z viacerých prípadných variánt jedného druhu domény záznam iba o tej, ktorá má najnižšiu hodnotu e-value. Po tomto kroku sa pomocou funkcie `createFastaOfRTDomains` vygeneruje fasta súbor s proteínovými sekvenciami všetkých ORF, v ktorých bola identifikovaná RT doména. Ten sa následne použije ako vstup bioinformatického nástroja `CD-HIT`. Tento program vykonáva zhlukovanie proteínových sekvencií na základe užívateľom zadaného prahu, ktorý predstavuje hodnotu identity pri lokálnom zarovnaní. Na všetkých sekvenciách vo vstupnom súbore (v našom prípade sú to ORF obsahujúce RT domény) sa vykoná lokálne zarovnanie, a sekvencie, ktoré majú identitu vyššiu, než zadaný prah, sa označia za jeden zhluk (cluster). Nasleduje ukážka výstupu programu `CD-HIT` (hviezdička na konci záznamu označuje, že daná sekvencia sa považuje za reprezentanta daného zhluku):

```
>Cluster 1
0183aa, >1_TE376_ORF2... *
1183aa, >5_TE1726_ORF2... at 98.36%
2183aa, >4_TE1525_ORF2... at 99.45%
>Cluster 2
0116aa, >1_TE376_ORF1... *
1116aa, >5_TE1726_ORF1... at 99.14%
2116aa, >5_TE1997_ORF1... at 100.00%
3116aa, >4_TE1525_ORF1... at 99.14%
```

Zhlukovanie RT domén sa vykonáva preto, že na jeho základe určíme v skúmanom genóme skupiny príbuzných LTR retrotranspozónov, u ktorých je väčšia šanca, že okrem konzervovanej RT domény v nich sú konzervované aj prípadné extra ORF. Pri vyhľadávaní konzervovaných eORF bez tohoto prvotného zhlukovania by sa výrazne predĺžilo zhlukovanie pomocou programu `CD-HIT`, kvôli lokálnemu zarovnávaniu tisícok ORF sekvencií (len v relatívne krátkom genóme *A. thaliana* nástroj vo všetkých nájdených LTR retrotranspozónoch identifikoval dohromady 5233 ORF).

Po tomto kroku teda máme k dispozícii zhluky RT domén, ktoré vlastne predstavujú skupiny LTR retrotranspozónov, ktoré zdieľajú podobné RT domény. Následne sa vyberú zhluky obsahujúce 2 a viac LTR retrotranspozónov (mnoho RT domén vytvára iba jednoprvkové zhluky, a s týmito sa ďalej nepracuje) a zistí sa, či obsahujú ORF, v ktorých neboli identifikované žiadne proteínové domény (pokiaľ takéto ORF neobsahujú, žiadne konzervované extra ORF nenájdeme). Pokiaľ sa v LTR retrotranspozónoch v danom zhluku nachádzajú

dohromady aspoň 2 ORF, v ktorých nebola identifikovaná žiadna proteínová doména, nasleduje ďalší krok.

Tým je zhlukovanie vybraných ORF domén neobsahujúcich žiadnu identifikovanú proteínovú doménu. Vytvorí sa pomocný fasta súbor obsahujúci ich sekvencie a následne sa na ňom spustí `CD-HIT`. Ten vytvorí zhluky ORF (neobsahujúcich žiadnu identifikovanú proteínovú doménu) vybraných zo zhluku LTR retrotranspozónov zdieľajúcich podobnú RT doménu. V prípade, že niektorý zhluk obsahuje viac, než jeden ORF, podarilo sa nám identifikovať konzervované extra ORF – teda ORF, ktoré síce neobsahujú žiadnu známu proteínovú doménu, no napriek tomu sa vyskytujú vo viacerých LTR retrotranspozónoch. Táto analýza je implementovaná vo funkcii `identifyConservedExtraORFs`.

Výstupom tejto analýzy je súbor `conservedExtraORFs.csv`, ktorý obsahuje záznamy o jednotlivých zhlukoch vo formáte csv. Prvá položka na každom riadku označuje poradové číslo zhluku, nasledujú identifikátory jednotlivých eORF, ktoré do daného zhluku patria. Hviezdička na konci jedného zo záznamov označuje ten eORF, ktorý program `CD-HIT` zvolil za reprezentanta zhluku. Nasleduje príklad záznamu o jednom zhluku:

```
Cluster2,1_TE225_ORF4*,3_TE1030_ORF8,3_TE1045_ORF6,2_TE662_ORF1,
```

4.8 Identifikácia konzervovaných tandemových repetícií

Identifikácia konzervovaných tandemových repetícií prebieha podobne, ako identifikácia konzervovaných eORF. Použije sa súbor obsahujúci zhluky podobných RT domén vygenerovaný pre identifikáciu konzervovaných eORF. Súbor sa prechádza po jednotlivých zhlukoch, a ak aktuálne skúmaný zhluk obsahuje aspoň dve sekvencie, pracuje sa s ním ďalej. Vytvorí sa zoznam LTR retrotranspozónov, z ktorých RT domény pochádzajú a následne zoznam všetkých tandemových repetícií, ktoré boli v týchto LTR retrotranspozónoch identifikované. Ďalej sa pokračuje iba v prípade, že zoznam obsahuje aspoň 2 tandemové repetície.

V ďalšom kroku sa vytvorí dočasný fasta súbor obsahujúci sekvencie tandemových repetícií zo zoznamu a pomocou programu `CD-HIT-EST` (varianta `CD-HIT` slúžiaca na zhlukovanie nukleotidových sekvencií) sa tieto sekvencie pozhlukujú. Nasleduje ukážka výstupu programu `CD-HIT-EST` (jeho výstup sa od výstupu programu `CD-HIT` líši iba tým, že dĺžka sekvencie je uvedená v „nt“, nie v „aa“):

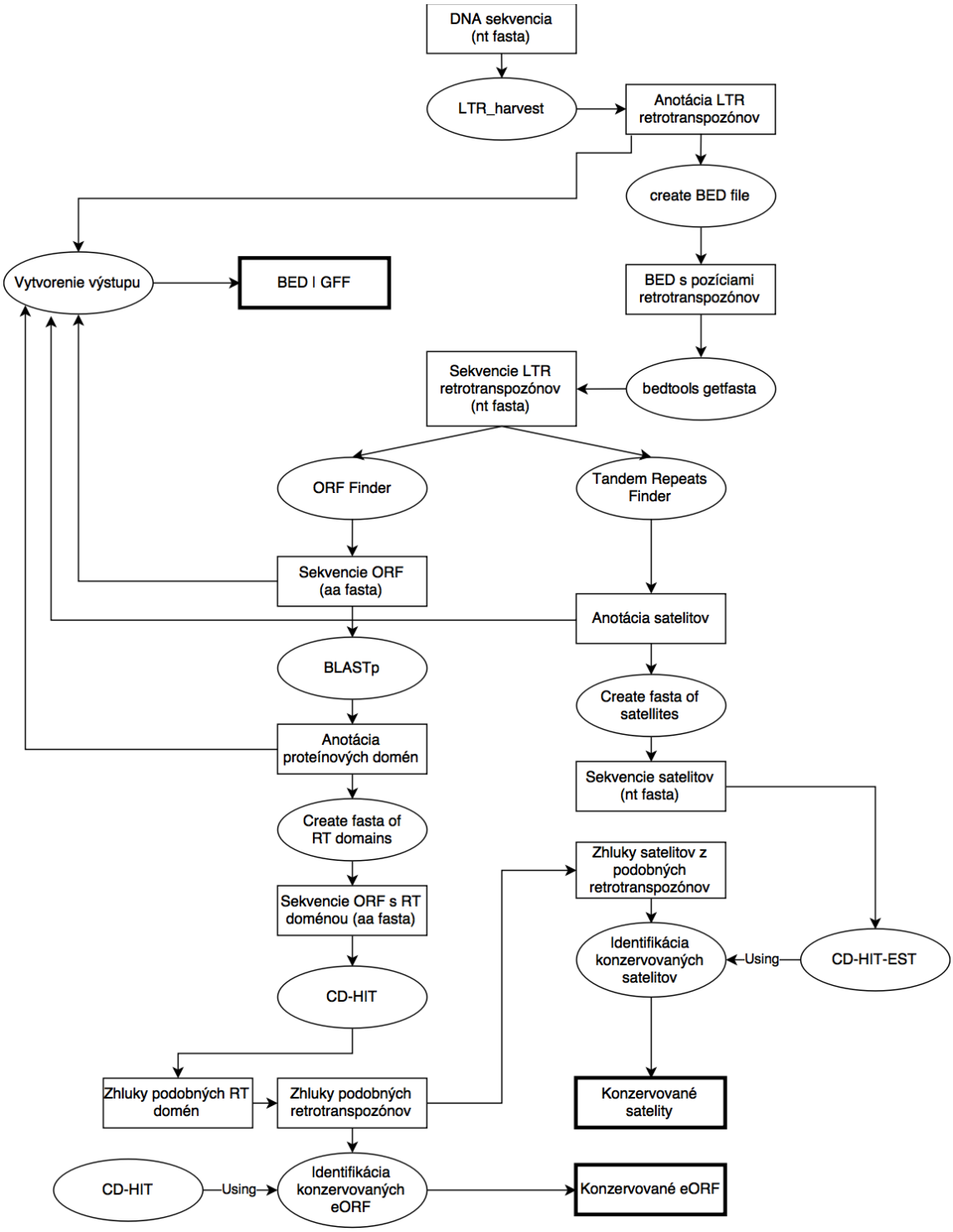
```
>Cluster 0
0 121nt, >5_TE2013:328:448:29... *
1 92nt, >5_TE2013:4899:4990:29... at +/100.00%
2 109nt, >4_TE1414:4707:4815:29... at -/100.00%
```

Pokiaľ po tomto kroku dostaneme zhluky, ktoré obsahujú viac, než jednu tandemovú repetíciu, a zároveň to sú tandemové repetície pochádzajúce z viacerých rozličných LTR retrotranspozónov (často sa stáva, že jedna tandemová repetícia je prítomná v jednom LTR

retrotranspozóne viackrát; takýto prípad, hoci je potenciálne zaujímavý, nepovažujem za konzervovanú tandemovú repetíciu), podarilo sa nám identifikovať konzervované tandemové repetície. Táto analýza je implementovaná vo funkcii `identifyConservedSatellites`.

Výstupom tejto analýzy je súbor `conservedTandemRepeats.csv`, ktorý obsahuje záznamy o jednotlivých zhlukoch vo formáte csv. Prvá položka na každom riadku označuje poradové číslo zhluku, nasledujú identifikátory jednotlivých tandemových repetícií, ktoré do daného zhluku patria. Hviezdička na konci jedného zo záznamov označuje tú tandemovú repetíciu, ktorý program `CD-HIT-EST` zvolil za reprezentanta zhluku. Nasleduje príklad záznamu o jednom zhluku:

```
Cluster3,5_TE2013:328:448:29*,5_TE2013:4899:4990:29,4_TE1414:4707:4815
:29,
```



Obrázok 4.1: Diagram ilustrujúci princíp fungovania implementovaného nástroja.

5 Testovanie

V tejto kapitole predstavím a popíšem experimentálne výsledky nasledujúcich aspektov funkcionality implementovaného nástroja:

- Vyhľadávanie extra ORF vnútri LTR retrotranspozónov
- Vyhľadávanie tandemových repetícií vnútri LTR retrotranspozónov
- Vyhľadávanie konzervovaných extra ORF
- Vyhľadávanie konzervovaných tandemových repetícií

Za testovací organizmus som zvolil *A. thaliana*. Genóm tejto rastliny som získal z databázy `The Arabidopsis Information Resource` (TAIR) vo verzii 10. Genóm sa skladá z 5 chromozómov a má celkovú dĺžku približne 120 Mbp. Vyhľadávanie LTR retrotranspozónov bolo vykonané s parametrami `-maxlenltr 8000 -maxdistltr 30000 -similar 70 -overlaps best`. Tabuľka 5.1 obsahuje počty LTR retrotranspozónov identifikovaných v jednotlivých chromozómoch.

Chromozóm	Počet identifikovaných LTR retrotranspozónov
1	500
2	351
3	418
4	353
5	423
Dohromady	**2045**

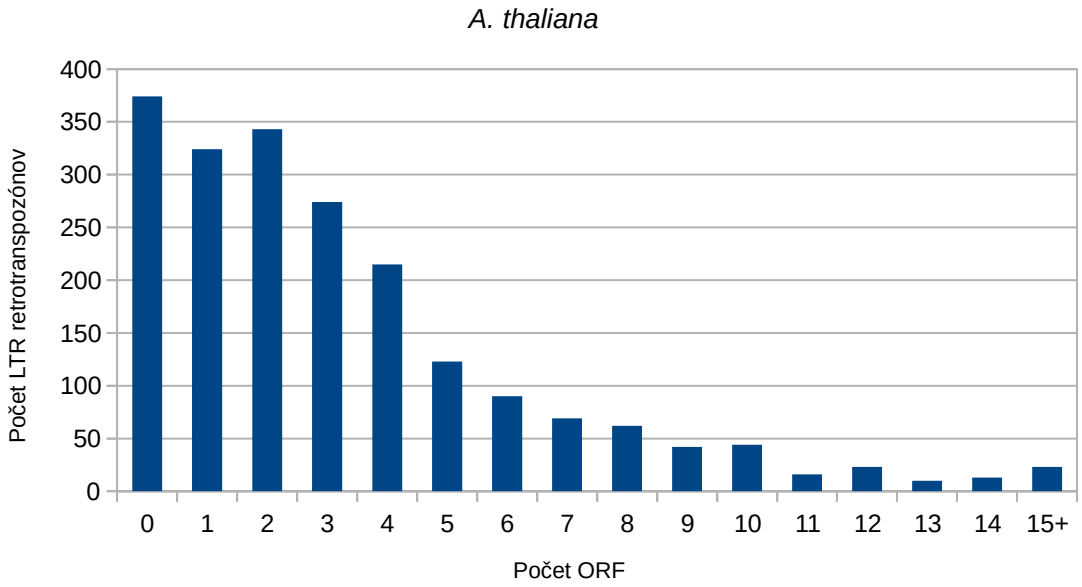
Tabuľka 5.1: Počty identifikovaných LTR retrotranspozónov v genóme *A. thaliana*.

5.1 Vyhľadávanie extra ORF

V tomto experimente som sa zameral na ORF vnútri identifikovaných LTR retrotranspozónov. Dohromady bolo v 2045 LTR retrotranspozónoch nástrojom nájdených 6822 ORF s minimálnou dĺžkou 300bp. Priemerná dĺžka RT domény v LTR retrotranspozónoch sa pohybuje v rozmedzí 650-700bp (800-850bp pre INT, 350-450bp pre RnaseH, podľa databázy Cores), takže minimálnu dĺžku som zvolil s dostatočnou rezervou. Celkovo sa žiadnu proteínovú doménu nepodarilo identifikovať v 76,19% identifikovaných ORF. Údaje o experimente sa nachádzajú v tabuľke 5.2 a v obrázkoch 5.1-5.3.

Chromozóm	Počet ORF vnútri LTR retrotranspozónov	Počet eORF	Pomer eORF [%]
1	1532	1240	80,94
2	1198	878	73,29
3	1557	1163	74,70
4	1131	829	73,30
5	1404	1088	77,50
	6822	**5198**	**76,19**

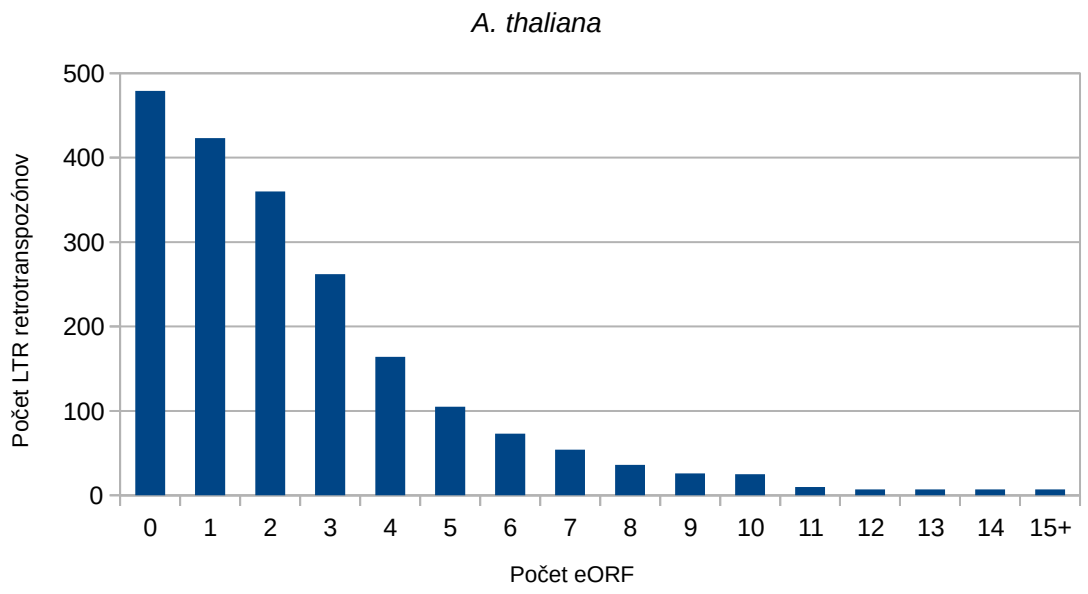
Tabuľka 5.2: Výsledky vyhľadávania ORF v LTR retrotranspozónoch identifikovaných v genóme *A. thaliana*. Zo 6822 celkovo identifikovaných ORF sa v 5198 nepodarilo identifikovať žiadnu známu proteínovú doménu.



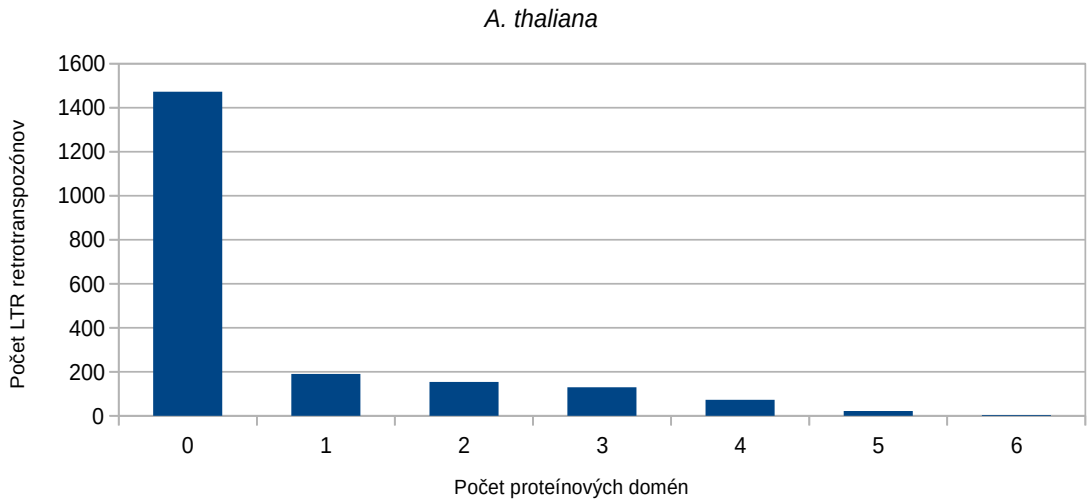
Obrázok 5.1: Histogram počtu ORF v LTR retrotranspozónoch, ktoré nástroj identifikoval v genóme *A. thaliana*.

Z grafu na obrázku 5.1 je vidieť, že v najväčšom počte LTR retrotranspozónov (374) nebol identifikovaný žiadny ORF. Ďalej počet LTR retrotranspozónov v závislosti na počte identifikovaných ORF klesá.

Graf na obrázku 5.2 má podobný priebeh; v najväčšom počte LTR retrotranspozónov nebol identifikovaný žiadny extra ORF, čo je samozrejme spôsobené tým, že v 374 LTR retrotranspozónoch nebol identifikovaný ani jeden ORF. Žiaden extra ORF však nebol identifikovaný v 479 LTR retrotranspozónoch, a keďže iba 374 ich neobsahuje žiadny ORF, viac než 100 LTR retrotranspozónov obsahujúcich aspoň jeden ORF s identifikovanou proteínovou doménou neobsahuje žiaden extra ORF.



Obrázok 5.2: Histogram počtu extra ORF v LTR retrotranspozónoch, ktoré nástroj identifikoval v genóme *A. thaliana*.



Obrázok 5.3: Histogram počtu proteínových domén v LTR retrotranspozónoch, ktoré nástroj identifikoval v genóme *A. thaliana*.

V grafe na obrázku 5.3 je vidieť, že výrazná väčšina identifikovaných LTR retrotranspozónov (1473, 72%) neobsahuje žiadnu identifikovateľnú proteínovú doménu. Iba 228 LTR retrotranspozónov (11%) obsahuje 3 a viac proteínových domén.

5.2 Identifikácia konzervovaných extra ORF

Po identifikovaní všetkých extra ORF v genóme *A. thaliana* nasleduje experiment zisťujúci, či sú niektoré z nich konzervované, teda či sa vyskytujú vo viac, než jednom LTR retrotranspozóne. Zhlukovanie RT domén a následné zhlukovanie potenciálne konzervovaných extra ORF som vykonal s prahom podobnosti 0,4.

Nástroj identifikoval celkovo 95 zhlukov konzervovaných extra ORF. Popis ich veľkostí sa nachádza v tabuľke 5.3, zoznam LTR retrotranspozónov v jednotlivých zhlukoch sa nachádza v tabuľke 5.4.

Veľkosť zhluku	Počet zhlukov s touto veľkosťou
2	45
3	16
4	13
5	2
6	5
7	4
8	2
9	2
10	2
11	2
12	1
13	1

Tabuľka 5.3: Popis identifikovaných zhlukov konzervovaných extra ORF identifikovaných v genóme *A. thaliana*.

Číslo zhluku	LTR retrotranspozóny v zhluku
Cluster 0	3_TE1255_ORF3,3_TE1255_ORF4*,4_TE1409_ORF7,
Cluster 2	3_TE1008_ORF2*,2_TE666_ORF8,5_TE1767_ORF4,
Cluster 9	1_TE236_ORF1,5_TE1856_ORF4*,4_TE1310_ORF5,
Cluster 12	1_TE235_ORF7*,3_TE1026_ORF9,3_TE1102_ORF4,
Cluster 14	1_TE235_ORF8,1_TE251_ORF3,1_TE264_ORF5,3_TE1118_ORF4,5_TE1783_ORF8*,4_TE1334_ORF6,4_TE1354_ORF3,
Cluster 20	1_TE264_ORF6*,1_TE264_ORF7,3_TE1102_ORF10,4_TE1354_ORF9,
Cluster 27	1_TE219_ORF10,3_TE1045_ORF4*,3_TE1049_ORF10,5_TE1733_ORF4,5_TE1837_ORF8,4_TE1327_ORF9,4_TE1327_ORF10,
Cluster 29	1_TE219_ORF3,3_TE1030_ORF6,3_TE1049_ORF2,3_TE1152_ORF4,2_TE672_ORF2,5_TE1837_ORF2,4_TE1290_ORF6,4_TE1327_ORF1,4_TE13

	96_ORF1*,
Cluster 32	1_TE275_ORF2*,3_TE1030_ORF8,3_TE1045_ORF8,3_TE1049_ORF3,2_TE662_ORF1,5_TE1733_ORF8,5_TE1806_ORF8,4_TE1290_ORF9,4_TE1327_ORF2,
Cluster 41	5_TE1788_ORF3*,5_TE1808_ORF2,

Tabuľka 5.4: Popis 10 vybraných identifikovaných zhlukov konzervovaných extra ORF identifikovaných v genóme *A. thaliana*. Identifikátory sú vo formáte <fasta id sekvencie>_TE<poradové číslo vo výstupe LTR_harvest>_ORF<poradové číslo vo výstupe ORF Finder>. Hviezdička označuje LTR retrotranspozón, ktorý bol pri zhlukovaní zvolený za reprezentanta zhluku. V tomto prípade fasta id sekvencie označuje chromozóm, z ktorého pochádza daný LTR retrotranspozón.

V jednotlivých zhlukoch sa nachádzajú LTR retrotranspozóny z rovnakých aj odlišných chromozómov (prvé číslo v identifikátore v tomto prípade označuje chromozóm). Nasledujú dve ukážky globálneho zarovnania dvoch vybraných extra ORF z jednotlivých zhlukov (ORF2_1_TE376 a ORF2_4_TE1525, ORF10_3_TE1049 a ORF10_4_TE1327). Identita zarovnania u prvého páru dosahuje 98%, u druhého páru síce iba 46,5%, ale je možné vidieť, že od istej pozície sú tieto dve ORF takmer totožné. Podarilo sa teda skutočne identifikovať konzervované extra ORF.

Prvá zarovnaná dvojica:

```
1   MHIHLMYMPITGRSHSQKQTHRAHFCNRSKSFIIIHAFFLTITQDYQTSF     50
    ||||||||||||||||||||||||||||||||||||||||||||||||||
1   MHIHLMYMPITGRSHSQKQTHRAHFCNRSKSFIIIHAFFLTITQDYQTSF     50

51  ISLYRPIRSILDFVYPFVAKSPLSWRQIYNREGLINSQRFNLSRHSLSPS    100
    ||||||||||||||||||||||||||||||||||||||||||||||||||
51  ISLYRPIRSILDFVYPFVAKSPLSWRQIYNREGLINSQRFNLSRHSLSPS    100

101 FVNHSLVVSCWFHSSSDSCEKSLVRRSESMAIHVFLNWIQRGGRSLRGNR    150
    ||||||||||||||||:||||||||||||||||||||||||||||||||
101 FVNHSLVVSCWFHSSSNSCEKSLVRRSESMAIHVFLNWIQRGGRSLRGNR    150

151 FNTHSIGDKILQLHRFLCLARTTSKLYWRRCCR    183
    |||||||||||||||||||||||||||||||||
151 FNTHSIGDKILQLHRFLCLARTTSKLYWRRCCR    183
```

Druhá zarovnaná dvojica (začiatok zarovnania som vynechal):

```
251 RDSKFTSAFWRAFQGEMGTKVQMSTAYHPQTDGQSERTIQTLEDMLRMCV    300
                    ||||:||||||||||||||||||||||||||||||
1   ----------------MGTKLQMSTAYHPQTDGQSERTIQTLEDMLRMCV     34

301 LDRGGHWADHLSLVEFAYNNSYQASIRMAPFEALYGRPCRTPLCWTQVGE    350
    ||.||||||||||||||||:|.|||.||||||||||||||||||||||||
35  LDWGGHWADHLSLVEFAYNNNYHASIGMAPFEALYGRPCRTPLCWTQVGE     84
```

```
351 RSIYGADYVLETTERIRVLKLNMKEAQDRQRSYADKRRRELEFEVGDRVY    400
    |||||||||.|:|||||||||||:|||||||||||.||||||||||||:|||
85  RSIYGADYVQESTERIRVLKLNIKEAQDRQRSYVDKRRRELEFEVGNRVY    134

401 LKMAMLRGPNRSISETKLTPRYMGPFRIVERVGPVAYRLELPDVMRAFHK    450
    ||||||:.|||||||||::||||||||||||||||||||:||||||||||
135 LKMAMLQCPNRSISETKVSPRYMGPFRIVERVGPVAYRLQLPDVMRAFHK    184

451 VFHVSMLRKCLHKDDEVLAKILEDLQPNMTLEARPVRILERRIKELRRKK    500
    ||||||||||||||||||||||.:|||||||||||||||:||||||||||||
185 VFHVSMLRKCLHKDDEVLAKIPKDLQPNMTLEARPVRVLERRIKELRRKK    234

501 IPLIKVLWNCDGVTEETWEPEARMKASFKKWFEKQVAA    538
    ||||||||:||||||||||||:||||||.|||.|||||||
235 IPLIKVLWDCDGVTEETWEPDARMKARFKKCFEKQVAA    272
```

5.3 Vyhľadávanie tandemových repetícií

V tomto experimente som skúmal tandemové repetície vnútri LTR retrotranspozónov. Dohromady nástroj v LTR retrotranspozónoch v genóme *A. thaliana* identifikoval 1591 tandemových repetícií. Väčšina LTR retrotranspozónov (1416, 69,24%) neobsahovala žiadnu tandemovú repetíciu. Všetky údaje o experimente sa nachádzajú v tabuľke 5.5 a v grafe na obrázku 5.4.

Chromozóm	**Počet tandemových repetícií vnútri LTR retrotranspozónov**
1	305
2	341
3	391
4	289
5	265
	1591

Tabuľka 5.5: Počty tandemových repetícií identifikovaných v LTR retrotranspozónoch nachádzajúcich sa v jednotlivých chromozómoch *A. thaliana*.

5.4 Identifikácia konzervovaných tandemových repetícií

Po identifikovaní všetkých tandemových repetícií vnútri LTR retrotranspozónov v genóme *A. thaliana* nasleduje experiment zisťujúci, či sú niektoré z nich konzervované, teda či sa vyskytujú vo viac, než jednom LTR retrotranspozóne. Zhlukovanie RT domén (pomocou `CD-HIT`) som vykonal s prahom podobnosti 0,4 a následné zhlukovanie potenciálne konzervovaných tandemových repetícií (pomocou `CD-HIT-EST`) som vykonal s prahom podobnosti 0,8. Nástroj v celom genóme identifikoval 10 zhlukov konzervovaných tandemových repetícií. Dve dvojice z dvoch vybraných zhlukov som zarovnal pomocou lokálneho zarovnania. Prvá dvojica je 100% zhodná, druhá dvojica dosiahla skóre zarovnania

96,3%. Podarilo sa teda identifikovať aj konzervované tandemové repetície. Nasledujú ukážky zarovnania:

Zarovnanie tandemových repetícií 4_TE1290:529:555:6 a 3_TE1152:1439:1465:6

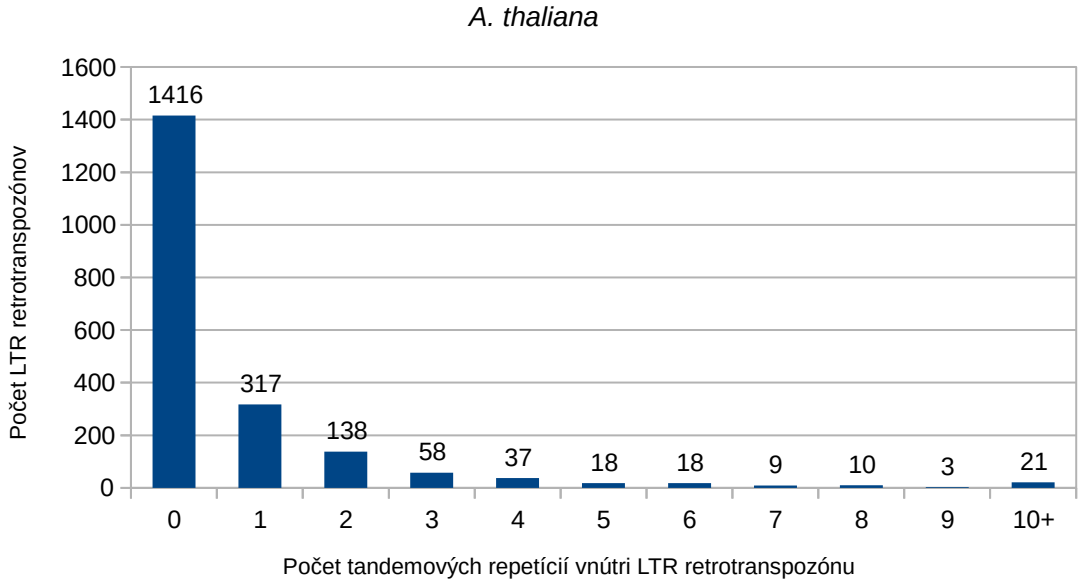
```
1 TCGTGGTCGTGGTCGTGGTCGTGGTCG     27
  ||||||||||||||||||||||||||||
1 TCGTGGTCGTGGTCGTGGTCGTGGTCG     27
```

Zarovnanie tandemových repetícií 4_TE1414:4707:4815:29 a 3_TE970:4764:4950:29

```
1   CGAGCTGACTAAAGTCATAACACGTGAGGCGAGCTGACTAAAGTCATAAC     50
    ||||||||||||||||||||||.|||.|||||||||||||||.||||||||
21  CGAGCTGACTAAAGTCATAACATGTGCGGCGAGCTGACTAAGGTCATAAC     70

51  ACGTGAGGCGAGCTGACTAAAGTCATAACACGTGAGGCGAGCTGACTAAA    100
    ||||||||||||||||||||.||||||||||||||||||||||||||||||
71  ACGTGAGGCGAGCTGACTAAGGTCATAACACGTGAGGCGAGCTGACTAAA    120

101 GTCATAACA    109
    |||||||||
121 GTCATAACA    129
```



Obrázok 5.4: Histogram počtu tandemových repetícií vnútri identifikovaných LTR retrotranspozónov v genóme *A. thaliana*. Väčšina LTR retrotranspozónov neobsahuje žiadnu tandemovú repetíciu.

6 Záver

V tejto diplomovej práci som sa zaoberal transpozónmi, mobilnými sekvenciami DNA. Transpozóny sú prítomné u takmer všetkých organizmov a často tvoria výraznú časť ich genómu (jednotky až vysoké desiatky percent). Vďaka svojej schopnosti kopírovať sa a pohybovať sa v rámci genómu môžu spôsobiť v DNA výrazné zmeny a a ich aktivita je jednou z hlavných príčin narastania veľkosti genómu u mnohých organizmov, hlavne rastlín.

V texte som popísal rozdelenie transpozónov do skupín podľa mechanizmu, ktorým sa pohybujú a podľa ich štruktúrnych charakteristík, a podrobnejšie som opísal niektoré dôležité rodiny. Ďalej som sa hlbšie venoval štruktúrnym rysom transpozónov, a to jednak tým, ktoré ich definujú a sú nevyhnutné pre ich existenciu, a následne tým, ktoré sa považujú za prídavné a ich funkcia a vznik často ešte nie sú uspokojivo objasnené. Popis štruktúry a mechanizmu transpozície som ilustroval obrázkami.

Následne som vytvoril prehľad dostupných a často používaných bioinformatických nástrojov umožňujúcich identifikáciu a anotáciu transpozónov v sekvenciách DNA. Nástroje som rozdelil podľa prístupu, ktorý na identifikáciu používajú a popísal som hlavné klady a zápory jednotlivých prístupov. Potom som vybral niektoré konkrétne programy a podrobnejšie popísal ich funkciu, motiváciu a implementáciu použitých metód.

Na základe nadobutnutých znalostí som implementoval nový bioinformatický nástroj zameraný na vyhľadávanie a anotáciu extra ORF a tandemových repetícií v LTR retrotranspozónoch. Nástroj je taktiež schopný vyhľadať konzervované extra ORF a tandemové repetície. Výstupom nástroja sú aj informácie o proteínových doménach identifikovaných vo všetkých ORF vyskytujúcich sa v nájdených LTR retrotranspozónoch, a taktiež zoznam všetkých proteínových domén identifikovaných v jednotlivých LTR retrotranspozónoch.

Najväčším prínosom implementovaného nástroja voči doterajšiemu stavu v tejto oblasti je to, že dokáže vyhľadať konzervované extra ORF a tandemové repetície. Táto funkcionalita je zaujímavá, pretože práve tieto štruktúry sú stále predmetom aktívneho výskumu a ich funkcia nie je zatiaľ plne vysvetlená. Vo fáze prieskumu existujúcich nástrojov som nenašiel žiadny s touto funkcionalitou.

Čo sa týka ďalšieho vývoja nástroja, bolo by prínosné skombinovať pri vyhľadávaní LTR retrotranspozónov výstupy viacerých bioinformatických nástrojov, ktoré by sa navzájom dopĺňali. Ďalej by bolo možné pridať schopnosť anotácie ďalších relatívne málo preskúmaných prídavných štruktúr prítomných v LTR retrotranspozónoch, ako napríklad miRNA, alebo triplexová, či kvadruplexová DNA.

Literatúra

[1] Bergman, C. M.; Quesneville, H.; Anxolabéhère, D.; Ashburner, M.: Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biology*, ročník 7, č. 11, 2006: s. R112.

[2] Landerer, E. S.; a kolektív: Initial sequencing and analysis of the human genome. *Nature*, ročník 409, č. 6822, 2001: s. 860-921.

[3] Schnable, P. S.; a kolektív: The B73 maize genome: complexity, diversity, and dynamics. *Science*, ročník 326, č. 5956, 2009: s. 1112–1115.

[4] Feschotte, C.: Transposable elements and the evolution of regulatory networks. *Nature Review Genetics*, ročník 9, č. 5, 2008: s. 397–405.

[5] Kidwell, M. G.: Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, ročník 115, č. 1, 2002: s. 49–63.

[6] Kejnovsky, E.; Hawkins, J. S.; Feschotte, C.: Plant transposable elements. *Plant Genome Diversity*, zväzok 1, 2012: s. 17.

[7] Doucet, A. J. a kolektív.: Characterization of LINE-1 ribonucleoprotein particles. *PLOS Genetics,* ročník 6, č. 10, 2010: s. e1001150.

[8] Cordaux, R.; Batzer, M.: The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics,* ročník 10, č. 10, 2009: s. 691–703.

[9] Nature Education Adapted from Pierce, Benjamin. Genetics: A Conceptual Approach, druhá edícia.

[10] Robertson, H. M.: The mariner transposable element is widespread in insects. *Nature,* č. 6417, 1993: s. 241–245.

[11] Leroy, H.; Castagnone-Sereno, P.; Renault, S.; Auge-Gouillou, C.; Bigot, Y.; Abad, P.: Characterization of Mcmar1, a mariner-like element with large inverted terminal repeats (ITRs) from the phytoparasitic nematode Meloidogyne chitwoodi. *Genetics*, č. 304, 2003: s. 35–41.

[12] Feschotte, C.; Wessler, S. R.: Treasures in the attic. *Proceedings of the National Academy of Sciences of the United States of America*, ročník 98, č. 16, 2001: s. 8923.

[13] Macas, J.; Koblížková, A.; Navrátilová, A.; Neumann, P.: Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Genetics*, ročník 44, č. 2, 2009: s. 198–206.

[14] Elrouby, N.; Bureau, T. E.: A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. *The Journal of Biological Chemistry, ročník 276, č. 45,* 2001: s. 41963-41968.

[15] Steinbauerová, V.; Neumann, P.; Novák, P.; Macas, J. A.: Widespread occurrence of extra open reading frames in plant Ty3/gypsy retrotransposons. *Genetica*, ročník 139, č. 11–12, 2011: s. 1543–1555.

[16] Wicker, T.; Keller, B.: Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Research*, ročník 17, č. 7, 2007: s. 1072–1081.

[17] Hoen, D. R.; Bureau T. E.: Plant transposable elements. *Springer Berlin Heidelberg*, č. 24, 2012: s. 219–251.

[18] Volff, J. N.: Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*, č. 28, 2006: s. 913–922.

[19] Li, Y.; Li, C.; Xia, J.; Jin, Y.: Domestication of transposable elements into MicroRNA genes in plants. *PLoS One*, č. 6, 2011: s. e19212.

[20] de Koning, A. P. J.; Gu, W.; Castoe, T. A.; Batzer, M. A.; Pollock, D. D.: Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genetics*, č. 7, 2011: s. e1002384.

[21] Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; a kolektív.: Initial sequencing and analysis of the human genome. *Nature*, č. 409, 2001: s. 860–921.

[22] Schnable, P. S.; Ware, D.; Fulton, R. S.; Stein, J.C.; Wei, F.; Pasternak, S.; a kolektív.: The B73 maize genome: complexity, diversity, and dynamics. *Science*, č. 326, 2009: s. 1112–1115.

[23] Hu, T. T.; Pattyn, P.; Bakker, E. G.; Cao, J.; Cheng, J-F.; Clark, R. M.; a kolektív.: The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, č. 43, 2011: s. 476–481.

[24] Bergman, C. M.; Quesneville, H.: Discovering and detecting transposable elements in genome sequences. *Brief Bioinformatics*, č. 8, 2007: s. 382–392.

[25] Smit, A.; Hubley, R.; Green, P.: 1996–2010. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

[26] Smit, A.; Hubley, R.: RepeatModeler Open-1.0. Repeat Masker Website (2010) na <http://www.repeatmasker.org>.

[27] Price, A. L.; Jones, N. C.; Pevzner, P. A.: De novo identification of repeat families in large genomes. *Bioinformatics*, 2005: suplement 1, s. i351–358.

[28] Xu, Z.; Wang, H.: LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research,* ročník 35, 2007: s. 265–268.

[29] McCarthy, E. M.; McDonald, J. F.: LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, ročník 19, č. 3, 2003: s. 362-367.

[30] Feschotte, C.; Keswani, U.; Ranganathan, N.; Guibotsy, M. L.; Levine, D.: Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biology Evolution*, ročník 2009, č. 1: s. 205–220.

[31] Jurka, J.; Kapitonov, V. V.; Pavlicek, A.; Klonowski, P.; Kohany, O.; Walichiewicz, J.: Repbase update, a database of eukaryotic repetitive elements. *Cytogenetics and Genome Research*, ročník 110, č. 1-4, 2005: s. 462–467.

[32] Wheeler, T. J.; Clements, J.; Eddy, S. R.; Hubley, R.; Jones, T. A.; Jurka, J.; a kolektív: Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*, ročník 41, 2013: s. D70–D82.

[33] Green, P.: Cross_match. <http://www.phrap.org/phredphrapconsed.html>.

[34] Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; a kolektív: BLAST+: architecture and applications. *BMC Bioinformatics*, ročník 10, č. 1, 2009: s. 421.

[35] Wheeler, T. J.; Eddy, S. R.: nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, ročník 29, č. 19, 2013: s. 2487–2489.

[36] Ellinghaus, D.; Kurtz, S.; Willhoeft, U.: LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, ročník 9, 2008: s. 18.

[37] GenomeTools [http://genometools.org/]

[38] Abouelhoda, M. I.; Kurtz, S.; Ohlebusch, E.: Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, ročník 2, č. 2, 2004: s. 53–86.

[39] You, F. M.; Cloutier, S.; Shan, Y. F.; Ragupathy, R.: LTR Annotator: Automated Identification and Annotation of LTR Retrotransposons in Plant Genomes. *International Journal of Bioscience, Biochemistry and Bioinformatics.* 5(3), s. 165-174.

[40] Wicker, T.; a kolektív: A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, ročník 8, č. 12, 2007: s. 973-982.

[41] Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, ročník 27, č. 2, 1999: s. 573–580.

[42] Hoen, D. G.: A call for benchmarking transposable element annotation methods. *Mobile DNA*, ročník 6, č. 13, 2015: s. ?.

[43] Lim, K. G.; Kwoh, C. K.; Hsu, L. Y.; Wirawan, A.: Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Briefings in Bioinformatics*, ročník 14, č. 1, 2013: s. 67-81.

Príloha 1 – popis parametrov a výstupov použitých nástrojov

Parametre nástroja ORF Finder

Parametre, ktoré sú nastavené napevno (nesmú sa zadávať do `params.txt`), sú nasledujúce.

- **-out** <File_Out>: cesta ku výstupnému súboru, napevno nastavená na „ORF.out“, pôvodný výstup ORF Finderu sa teda po každom behu nástroja bude nachádzať v tomto súbore.
- **-in** <File_In>: cesta ku vstupnému súboru; použije sa vstupný súbor pre celý nástroj, špecifikovaný na príkazovom riadku pri spustení skriptu dp.py.
- -outfmt <Integer>: formát výstupného súboru, vždy nastavený na hodnotu „0“, čo značí fasta súbor s proteínovými sekvenciami nájdených čítacích rámcov.
- **-b** <Integer>: index počiatočnej pozície vo vstupnej sekvencii, od ktorej sa má vyhľadávať; použije sa implicitná hodnota „1“.

Voliteľné parametre, ktoré je možné uviesť do súboru params.txt.

- **-e** <Integer>: index konečnej pozície vo vstupnej sekvencii, do ktorej sa má vyhľadávať; použije sa implicitná hodnota „0“, čo znamená „až do konca sekvencie“.
- **-c** <Boolean>: udáva cirkularitu sekvencie; použije sa implicitná hodnota „false“.
- **-g** <Integer>: špecifikuje genetický kód; použije sa implicitná hodnota „1“, ktorá označuje štandardný kód.
- **-s** <Integer>: špecifikuje použitý štart kodón; použije sa implicitná hodnota „1“, ktorá označuje „ATG“ a alternatívne štart kodóny.

Formát výstupného súboru programu ORF Finder

`ORF Finder` ukladá identifikované ORF v súbore formátu fasta s proteínovými sekvenciami. Identifikátory týchto sekvencií obsahujú pôvodný identifikátor DNA sekvencie, v ktorej sa vyhľadávalo, a tiež ďalšie údaje, ktoré uvediem na príklade.

Povedzme, že vyhľadávame ORF v DNA sekvencii s identifikátorom „RLX_56183“. Identifikátor prvého nájdeného ORF v tejto sekvencii môže vyzerať nasledovne.

```
>lcl|ORF1_RLX_56183:2507:2148 unnamed protein product
```

- Reťazec „lcl|“ neobsahuje žiadnu informáciu
- Reťazec „ORF1“ označuje, že táto sekvencia je v poradí prvým ORF identifikovaným v DNA sekvencii. Táto položka má formát obsahujúci reťazec „ORF“, nasledovaný poradovým číslom tohoto ORF. Od nasledujúcej položky je oddelená znakom „_“.
- Reťazec „RLX_56183“ označuje identifikátor DNA sekvencie, v ktorej sa vyhľadávalo. Od nasledujúcej položky je oddelený dvojbodkou.
- Nasledujúce dve čísla oddelené dvojbodkou, „2507:2148“, označujú pozíciu začiatku a konca ORF v DNA sekvencii. Prípad, v ktorom index začiatku ORF je väčší, než index konca ORF, značí, že daný ORF sa nachádza v opačnom „-“ vlákne DNA.
- Reťazec „unnamed protein product“ bude v tomto nástroji vždy prítomný v tejto podobe a neobsahuje žiadnu užitočnú informáciu. Na jeho mieste by sa nachádzala identifikácia kódujúcej sekvencie potenciálne prítomnej v identifikovanom ORF. Toto však bude nástroj skúmať v inom kroku, preto nás to na tomto mieste nezaujíma.

Nasleduje ukážka ORF identifikovaného na pozíciach 7932-9332 v sekvencii RLX_56086_Arabidopsis:

```
>lcl|ORF2_RLX_56086_Arabidopsis:7932:9332
MKLVKTQIHSVSSSAPNIDRILEESRNTPFTKRISETTMSNLGKFKIDSYNGSTDPKCHIKSFVISVARA
RFKPGEKDAGLFLLFVEHLKGPALEWFSRLEQNSIDSFDELSTLFLKQYSVLINPGTSDADLWSLSQQPT
EPLRDFLTTFKSTLAKVEGFTDVAALSALKKALWYKSEFRKELNLSKRTTIRDALHQASDFVAHEEEMAL
LAKRHEPTKQAPRAEKTEKAQATPPVQKKTRESGTYTHHEGRNFSGYHNYQIDTPRGRGRGRGRGRGRGR
ESFTWTRDQPPSNDQEYCEHHKIFGHHTSRCQSLGARLATKFLAGELGTNVNLKDLEPEPEPEQTNPGRD
PEPRNQESQKRTRGSEDKDRDGTRQKVLTIMGGSPYCPDTVAAIKAYQRRAETPSNWTRRFDRPNDVITF
EESETNGLDMPHNDPIVVTLAIGDHDIYQVLIDTGSTIDVIFRETL
```

Parametre nástroja Tandem Repeats Finder

- **Váha zhody** (match weight): jediná povolená hodnota je 2.
- **Váha nezhody** (mismatch weight): povolené hodnoty sú 3, 5 a 7; hodnota 3 značí vyššiu toleranciu ku nezhode pri zarovnaní, hodnota 7 značí nižšiu toleranciu.
- **Váha inzercie a delécie** (indel weight): význam a rozsah totožný s váhou nezhody.
- **Pravdepodobnosť zhody** (match probability) <Integer>: určuje priemerné percento zhody nukleotidov medzi jednotlivými monomérmi satelitu; rozsah 0-100.
- **Pravdepodobnosť inzercie a delécie** (indel probability) <Integer>: určuj priemerné percento inzercií a delécií medzi jednotlivými monomérmi satelitu; rozsah 0-100.
- **Prah skóre zarovnania** (minimum alignment score) <Integer>: číslo v rozsahu 0-100; skóre zarovnania identifikovanej repetície musí byť vyššie, než udaná hodnota.
- **Maximálna dĺžka periódy** (maximum period size) <Integer>: udáva maximálnu dĺžku periódy repetície, ktorá bude nahlásená, povolený rozsah je 1-2000.
- **-ngs**: zapnutie kompaktnejšieho formátu výstupu, vhodného pre vstupné súbory obsahujúce veľké množstvo sekvencií. Výstup programu Tandem Repeats Finder je v tomto prípade implicitne smerovaný na štandardný výstup a nástroj ho presmeruje

do súboru „TRF.out“. Tento prepínač užívateľ nemusí zadávať (a naopak, nemôže ho nezadať); keďže je jeho použitie pre nástroj nutné kvôli spracúvaniu výstupu Tandem Repeats Findera práve v tomto formáte, je v zdrojovom kóde napevno pridaný za užívateľom špecifikovaný reťazec parametrov.

Užívateľ môže špecifikovať nasledujúce nepovinné parametre. Pre samotný beh nástroja nie je potrebný ani jeden z nich, ale je možné ich využiť, pokiaľ ich užívateľ potrebuje na iné analýzy:

- **-m**: vytvorenie kópie vstupného súboru s DNA sekvenciou, v ktorej je každý nukleotid, ktorý sa vyskytol v tandemovej repetícií, zmenený na písmeno „N“; nástroj umožňuje tento prepínač použiť, pokiaľ ho užívateľ potrebuje na ďalšie analýzy, avšak jeho výstup sa ďalej nepoužíva.
- **-f**: lemujúca sekvencia (flanking sequence); tento prepínač zapne zaznamenanie 500 nukleotidov, nachádzajúcich sa tesne pred začiatkom tandemovej repetície, a 500 nukleotidov, nachádzajúcich sa priamo za koncom tandemovej repetície, do súboru so zarovnaním (alignment file).
- **-d**: (data file), vytvorí textový súbor, ktorý obsahuje rovnaké informácie a v rovnakom poradí, ako tabuľkový súbor, a navyše obsahuje sekvencie konsenzuálneho monoméru a celú sekvenciu tandemovej repetície. Tento súbor neobsahuje žiadne popisy a je vhodný predovšetkým na ďalšie automatizované spracovanie užívateľskými skriptami; nástroj tento súbor nevyužíva a nepotrebuje, takže uvedenie tohoto prepínača je na vlastnej iniciatíve užívateľa.
- **-h**: vypnutie výstupu vo formáte html (suppress html output); bez uvedenia tohoto prepínača Tandem Repeats Finder vytvorí výstup vo formáte html stránok, ktoré je možné prezerať vo webovom prehliadači. Pre beh nástroja html výstup nie je potrebný, takže tento prepínač by v prípade, že užívateľ html výstup nepotrebuje pre ďalšie analýzy, mal ostať zapnutý (v prípade vstupného súboru s veľkým množstvom DNA sekvencií môže byť vytvorené veľmi veľké množstvo jednotlivých html súborov).
- **-r**: vypnutie eliminácie redundantných výsledkov (no redundancy elimination). Pre nástroj je vhodnejšie tento prepínač nepoužiť, pretože eliminácia redundantných výsledkov štandardne prinesie kvalitnejší výstup.
- **-l** <Integer>: najväčšia očakávaná dĺžka tandemovej repetície, v miliónoch bp.

Formát výstupného súboru programu Tandem Repeats Finder

`Tandem Repeats Finder` pri použití prepínača „-ngs“ produkuje výstup vo vlastnom špecifickom formáte. Sled nájdených tandemových repetícií v každej zo vstupných sekvencií je uvedený pôvodným identifikátorom tejto sekvencie, predchádzaným znakom „@“, na samostatnom riadku (napr. sled tandemových repetícií, identifikovaných v sekvencii s identifikátorom „RLX_56183“, bude uvedený riadkom „@RLX_56183“). Po tomto

identifikátore nasledujú záznamy o jednotlivých tandemových repetíciách nájdených v tejto sekvencii. Záznam pre jednu tandemovú repetíciu sa nachádza na jednom riadku a obsahuje nasledujúce položky, v tomto poradí, oddelené jednou medzerou. Na konci každej odrážky uvádzam dátový typ záznamu.

- Index začiatku satelitu v pôvodnej sekvencii; <integer>.
- Index konca satelitu v pôvodnej sekvencii; <integer>.
- Dĺžka periódy (monoméru) satelitu v počte párov bází; <integer>.
- Počet opakovaní monoméru zarovnaných s konsenzuálnou sekvenciou; <float>.
- Dĺžka konsenzuálnej sekvencie; nemusí byť totožná s dĺžkou periódy, ale v drvivej väčšine prípadov je totožná; <integer>.
- Percento zhôd nukleotidov medzi jednotlivými monomérmi satelitu; <integer>.
- Percento inzercií a delécií medzi jednotlivými monomérmi satelitu; <integer>.
- Skóre zarovnania, v rozsahu 0-100; <integer>.
- Koľko percent sekvencie satelitu tvorí nukleotid A; <integer>.
- Koľko percent sekvencie satelitu tvorí nukleotid C; <integer>.
- Koľko percent sekvencie satelitu tvorí nukleotid G; <integer>.
- Koľko percent sekvencie satelitu tvorí nukleotid T; <integer>.
- Entropia založená na percentuálnom pomere jednotlivých nukleotidov; <float>.
- Konsenzuálna sekvencia monoméru; <string>.
- Celková sekvencia identifikovaného satelitu; <string>.
- Sekvencia o dĺžke 50bp priamo predchádzajúca satelit; <string>.
- Sekvencia o dĺžke 50bp priamo nasledujúca satelit; <string>.

Nasleduje ukážka záznamu o jednej identifikovanej tandemovej repetícii z výstupného súboru programu `Tandem Repeats Finder`:

```
14542 14635 21 4.5 21 78 0 89 12 43 22 21 1.86 CCGCCTCGGCAATCCTGGCAT
CCGCCTCGGCAATCTTGGCCTCAGCCTCGGCACTCTTGGCATCCGCCTCGACTATCCTGGCATCGGCCTC
AGCAATCCTGACCTCCGCCTCGGC
CAACGCTTCGGCTTTCCGCCTGGCCTCATCTGCCTCGGCAATTTTGGTCG
TCGAGAAGTCTCGTATTTCCGATTCGAAGAGTCGTAGGCTCGAACCATGC
```

Parametre nástroja blastp pre identifikáciu proteínových domén

Program blastp pri identifikácii proteínových domén LTR retrotranspozónov nástroj spúšťa s nasledujúcimi parametrami.

- **-query:** špecifikuje vstupný súbor obsahujúci proteínové sekvencie, v ktorých chceme identifikovať domény. Hodnota parametra je nastavená na „ORF.out“

- **-max_target_seqs:** určuje počet vrátených výsledkov identifikácie proteínových domén. Tie sú zoradené podľa podobnosti od najvyššej po najnižšiu. Hodnota tohoto parametra je napevno nastavená na „1“, pretože nás zaujíma iba doména s najvyššou podobnosťou.
- **-outfmt:** formát výstupného súboru. Hodnota tohoto parametra je nastavená na hodnotu „7 stitle evalue“. Hodnota 7 určuje tabuľkový formát výstupu, subjekty „stitle“ a „evalue“ určujú dva stĺpce, ktoré nástroj potrebuje - „stitle“ označuje reťazec obsahujúci slovný popis identifikovanej proteínovej domény, „evalue“ hodnotu pravdepodobnosti výskytu takéhoto zarovnania v náhodnej sekvencii.
- **-db:** špecifikuje databázu, v ktorej sa má vyhľadávať. Hodnota tohoto parametra je nastavená na „Cores/cores-database“.
- Výstup je operátorom „>“ presmerovaný do súboru „BLASTPdomains.out“

Formát výstupného súboru s proteínovými doménami

Program `BLASTp` pri rozpoznávaní proteínových domén v ORF produkuje textový výstup, v ktorom záznam pre jeden ORF má nasledujúcu podobu (po jednotlivých riadkoch).

- Riadok obsahujúci identifikáciu použitého programu a jeho verziu.
- Identifikátor vstupnej sekvencie prebratý zo vstupného súboru `ORF.out`.
- Identifikácia použitej databázy, v našom prípade `Cores/cores-database`.
- Nasledujúci riadok sa líši podľa toho, či bola identifikovaná zhoda s proteínovou sekvenciou v databáze, alebo nie. V prípade, že nie, obsahuje riadok reťazec „# 0 hits found“ a končí sa ním záznam o jednom ORF. V prípade, že áno, obsahuje riadok názvy stĺpcov, ktoré sa budú nachádzať v zázname o identifikovanej proteínovej doméne.
- Počet nájdených výsledkov.
- Záznam o identifikovanej proteínovej doméne, rozdelený znakom tabulátor na stĺpce špecifikované o dva riadky vyššie.

Nasleduje príklad záznamu z identifikácie proteínovej domény v ORF číslo 3 v sekvencii RLG_48181_Zea:

```
# BLASTP 2.6.0+
# Query: lcl|ORF3_RLG_48181_Zea:4989:6164
# Database: Cores/cores-database
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 1 hits found
lcl|ORF3_RLG_48181_Zea:4989:6164 RT_RIRE2 71.111   225   65 0  45 269
1  225   8.44e-122   348
```

Formát výstupného súboru programu LTR_harvest

Výstupný súbor programu `LTR_harvest` spracúvaný nástrojom má tabuľkový formát, kde sa záznam o každom identifikovanom LTR retrotranspozóne nachádza na samostatnom riadku. Na začiatku súboru sa nachádzajú riadky s komentárom, ktoré začínajú znakom „#“. Záznam o jednom LTR retrotranspozóne má nasledujúci formát:

- index začiatku LTR retrotranspozónu, <integer>
- index konca LTR retrotranspozónu, <integer>
- dĺžka identifikovaného LTR retrotranspozónu, <integer>
- index začiatku ľavej sekvencie LTR, <integer>
- index konca ľavej sekvencie LTR, <integer>
- dĺžka ľavej sekvencie LTR, <integer>
- index začiatku pravej sekvencie LTR, <integer>
- index konca pravej sekvencie LTR, <integer>
- dĺžka pravej sekvencie LTR, <integer>
- similarita LTR (percento zhodných nukleotidov medzi pravou a ľavou sekvenciou LTR), <float>
- číslo sekvencie, v ktorej bol LTR retrotranspozón identifikovaný (0-based index označujúci jednotlivé identifikátory vo vstupnom súbore), <integer>

Jednotlivé položky v zázname o jednom LTR retrotranspozóne sú oddelené dvoma medzerami. Nasleduje príklad jedného záznamu:

```
7332  12194  4863  7332  8556  1225  10969  12194  1226  99.59  9
```

Príloha 2 – špecifikácia výstupného formátu nástroja

Špecifikácia formátu *BED*

Súborový formát *BED* (Browser Extensible Data) je jedným zo základných formátov určených na ukladanie bioinformatických anotačných dát. Každý riadok obsahuje záznam o jednom ryse. Jednotlivé riadky majú 3 povinné a ďalších 9 nepovinných stĺpcov. Počet stĺpcov na riadkoch musí byť v rámci jedného súboru konzistentný (napríklad nie je možné, aby jeden riadok obsahoval 3 stĺpce, a ďalší 5). Jednotlivé stĺpce musia byt oddelené tabulátorom. Poradie povinných a nepovinných stĺpcov je záväzné a v poradí vyššie stĺpce je možné použiť iba v prípade, že sú použité všetky v poradí nižšie stĺpce. Súborový formát *BED* obsahuje nasledujúce povinné stĺpce:

- **chrom:** názov chromozómu alebo sekvencie, v ktorej bola daná štruktúra identifikovaná.
- **chromStart:** index začiatku štruktúry v danej sekvencii, 0-based.
- **ChromEnd:** index konca štruktúry v danej sekvencii (v skutočnosti index prvého nukleotidu nasledujúceho priamo po konci danej štruktúry). Príklad – prvých 100 bází chromozómu sa zapíše ako chromStart=0, chromEnd = 100.

A nasledujúce nepovinné stĺpce:

- **name:** definuje názov riadku. Môj nástroj v tejto položke uvádza typ štruktúry (orf, satellite, transposon).
- **score:** hodnota skóre danej štruktúry. Formát *BED* očakáva túto hodnotu v rozsahu 0-1000 a podľa nej priraďuje danej štruktúre farebnú hodnotu pri vizualizácii súboru. Môj nástroj do tejto položky vkladá hodnotu skóre podľa výstupu konkrétneho bioinformatického nástroja, ktorý štruktúru identifikoval. `ORF Finder` ani `LTR_harvest` žiadne skóre neposkytujú, takže záznamy ORF a Transposon majú tento stĺpec nastavený na hodnotu „0“. `Tandem Repeats Finder` poskytuje hodnotu skóre.
- **strand:** definuje vlákno, na ktorom bola daná štruktúra identifikovaná. Prípustné hodnoty sú „.“ (nedefinované vlákno), „+“ a „-“.

Ďalšie nepovinné stĺpce nástroj nevyužíva.

Špecifikácia formátu *GFF*

Súborový formát *GFF* (General Feature Format) je ďalším z najčastejšie používaných formátov pre ukladanie bioinformatických anotačných dát. Záznam pre jednu anotovanú štruktúru sa nachádza na jednom riadku. Záznam na jednom riadku je rozdelený na 9 povinných stĺpcov.

Tie sú oddelené znakom tabulátor. V prípade, že údaj pre nejaký stĺpec nie je dostupný, uvedie sa namiesto neho hodnota „.“. Formát *GFF* používa nasledujúce stĺpce:

- **seqname:** názov sekvencie (chromozóm, scaffold), v ktorej bola daná štruktúra identifikovaná.
- **source:** názov programu, ktorý danú štruktúru identifikoval (v tomto nástroji `ORFFinder`, `Tandem Repeats Finder`, alebo `LTR_harvest`).
- **feature:** typ štruktúry (orf, satellite, transposon).
- **start:** index začiatku štruktúry v sekvencii (1-based).
- **end:** index konca štruktúry v sekvencii.
- **score:** použije sa hodnota skóre z jednotlivých použitých nástrojov (iba Tandem Repeats Finder). `ORF Finder` ani `LTR_harvest` skóre neudáva, takže pre záznamy typu „orf“ a „transposon“ sa uvedie v tomto stĺpci „.“.
- **strand:** vlákno, „+“ alebo „-“.
- **frame:** indikuje posun čítacieho rámca voči začiatku štruktúry. Prípustné hodnoty sú „0“ (prvá báza štruktúry je prvou bázou kodónu), „1“ (druhá báza štruktúry je prvou bázou kodónu) a „2“(tretia báza štruktúry je prvou bázou kodónu). V štruktúrach, ktoré nástroj anotuje, je toto pole nepoužiteľné, takže obsahuje vždy hodnotu „.“.
- **attribute**: zoznam párov vlastnosť-hodnota vlastnosti oddelených bodkočiarkami. Pre záznamy typu „orf“ nástroj do tohto stĺpca uvedie názov identifikovanej domény, e-value a rodinu (napríklad: domain=GAG; evalue=4.01e-155; family=Athila4-1;). Pre záznamy typu „transposon“ nástroj do tohto stĺpca uvedie zoznam identifikovaných proteínových domén (napríklad: GAG=1.07e-31; RNaseH=1.54e-39; INT=2.39e-64; RT=3.52e-58; AP=8.34e-34;). V prípade, že v ORF nebola identifikovaná žiadna proteínová doména, nástroj do tohto stĺpca uvedie reťazec „domain=none“. V prípade, že v LTR retrotranspozóne neboli identifikované žiadne proteínové domény, nástroj do tohto stĺpca uvedie reťazec „domains=none“.

Príloha 3 – obsah CD a návod na sprevádzkovanie nástroja

Obsah CD

Cores/	- zložka so sekvenciami proteínových domén
suffixerator/	- prázdna zložka, ktorú používa `LTR_harvest`
genomes/	- zložka obsahujúca testovací genóm
diagram/	- zložka obsahujúca diagram nástroja vo formáte `png`
bedtools-2.25.0.tar	- archív s bioinformatickým balíkom `bedtools`
genometools-1.5.9.tar	- archív s bioinformatickým balíkom `genometools`
ncbi-blast-2.6.0+-x64-linux.tar	- archív s bioinformatickým balíkom `BLAST`
cdhit-master.zip	- archív s bioinformatickým balíkom `CD-HIT`
ORFfinder	- spustiteľná verzia nástroja `ORF Finder`
trf409.linux64	- spustiteľná verzia nástroja `Tandem Repeats Finder`
params.txt	- textový súbor obsahujúci parametre použitých nástrojov
zdrojové kódy nástroja	- `dp.py`, `runCommands.py`, `createOutput.py`, `func.py`, `parsers.py`, `classes.py`

Návod na sprevádzkovanie nástroja

Najprv je potrebné rozbaliť archívy `bedtools`-2.25.0.tar, genometools-1.5.9.tar, ncbi-blast-2.6.0+-x64-linux.tar a cdhit-master.zip. Následne je možné nástroj spustiť, napríklad pomocou príkazu „`python dp.py -i genomes/TAIR10_chr_all.fasta -outfmt gff -o TAIR10_annotations -outfeat all -ctnt 0.8 -ctaa 0.4 -et 0.001 -mc 3`“.