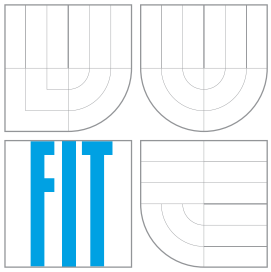
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

REKONSTRUKCE TRANSPOSONŮ

RECONSTRUCTION OF TRANSPOSONS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

OLIVER ŠURINA

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. JANKA PUTEROVÁ

BRNO 2016

Abstrakt

Táto práca sa zaoberá rekonštrukciou transpozónov na základe výstupu programu RepeatExplorer. V prvej časti práca predstavuje základy molekulárnej biológie so zameraním na transpozóny a ich štruktúru. Ďalej popisuje návrh a implementáciu aplikácie, ktorá rekonštruuje transpozóny na základe výstupov RepeatExploreru. K prezentácii dosiahnutých výsledkov bolo vytvorené interaktívne grafické uživateľské rozhranie. Aplikácia bola následne testovaná v realnom prostredí. Práca tiež prináša prehľad o existujúcich nástrojoch pre identifikáciu transpozónov.

Abstract

This thesis deals with the reconstruction of transposable elements based on an output of RepeaterExplorer. In the first section, this thesis introduces basics of molecular biology with focus on transposons and their structure. Then it describes design and implementation of application which reconstructs transposons based on an output of RepeatExplorer. Achieved results are presented by created interactive user interface. The application was then tesed in real enviroment. This thesis also delivers overview of existing tools for transposons identification.

Klíčová slova

DNA, transpozóny, LTR, RepeatExplorer, rekonstrukce transpozónov, GUI

Keywords

DNA, transposons, LTR, RepeatExplorer, transposons reconstruction, GUI

Citace

Oliver Šurina: Rekonstrukce transposonů, bakalářská práce, Brno, FIT VUT v Brně, 2016

Rekonstrukce transposonů

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Janky Puterovej. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Oliver Šurina
19. května 2016

Poděkování

Týmto by som chcel poďakovať vedúcej mojej bakalárskej práce Ing. Janke Puterovej za jej trpezlivosť, odborné vedenie a cenné rady, ktoré mi poskytla pri jej vypracovaní.

Obsah

Kapitola 1

Úvod

Informačné technológie v dnešnej dobe zasahujú do všetkých vedeckých oborov. Výnimkou nie je ani bioinformatika, ktorá sa okrem iného zaoberá porozumeniu fungovania mobilných DNA elementov - transpozónov. V súčasností je využívané široké spektrum nástrojov na analýzu DNA sekvencií ktoré využívajú rôzne metódy. Medzi takéto metódy patri napríklad porovnávanie s referenčnou databázou, zhlukovanie na základe podobnosti alebo zarovnávanie sekvencií. Táto práca je zameraná na širšie priblíženie metódy zhlukovania.

Metódu zhlukovania implementuje program RepeatExplorer. V súčasnej dobe existuje málo nástrojov, ktoré by vedeli spracovať zhluky vyprodukované týmto programom a v prijateľnej forme ich prezentovať užívateľovi. Cieľom práce je vytvorenie aplikácie, ktorá spracuje tento výstup a vytvorí zhluky vyššej úrovne. Následne overí ich vnútornú štruktúru a prezentuje dosiahnuté výsledky grafickou formou.

Pre priblíženie problematiky, sa v úvodnej kapitole 2 čitateľ zoznámi so základmi molekulárnej biológie a problematikou transpozónov. Na tieto základy potom naväzuje kapitola 3, ktorá sa zaoberá predstavením nástrojov LTR Finder, Tedna, RepeatMasker na identifikáciu transpozónov. Bližšie je rozobraný program RepeatExplorer a jeho fungovanie.

Nasledujúce kapitoly 4 a 5 sa venujú priblíženiu požiadavkov na výslednú aplikáciu a detaily jej implementácie. V tejto časti sú popísané použité knižnice, postupy a rozobratá koncepcia aplikácie.

Kapitola 6 obsahuje výsledky testovania funkčnej aplikácie a zhŕňa získané poznatky. Ich možné využitie v praxi a možnosti rozšírenia sú uvedené v závere práce 7.

Kapitola 2

Základy Molekulárnej Biológie

Gény sú základné jednotky potrebné k uchovaniu vlastnosti ako jedinca tak aj celého druhu. Genetická informácia je prenášaná z materských buniek do dcérskych, ako aj z generácie na generáciu pohlavnými bunkami. Vďaka experimentu Averyho-MacLeoda-McCartyho v roku 1944 sa podarilo zistiť, že nosičom genetickej informácie je pravdepodobne deoxyribonukleová kyselina, taktiež označovaná ako DNA. O objavenie štruktúry DNA sa v roku 1653 postarali vedci James Watson a Francis Crick. Navrhli, že je to dvojzávitnicová špirála, ktorej vlákna majú navzájom opačnú orientáciu. Na to aby mohla DNA plniť svoju funkciu musí byť schopná zdvojenia sa, čo sa uskutočňuje replikáciu - kopírovaním. Správnosť replikácie zaisťujú rôzne mechanizmy, niekedy však dochádza k poškodeniu a trvalej zmene DNA tzv. mutáciám. Tie môžu daný organizmus v rámci evolúcie zvýhodniť, alebo im uškodiť. Jedným z mutačných činiteľov môžu byť napríklad transpozóny, ktoré po presune na nové miesto ovplyvňujú genetickú informáciu. Zmeny v DNA z generácie na generáciu počas miliónov rokov viedla nielen k vzniku novým druhom, ale aj k odlišnostiam v rámci toho istého druhu.

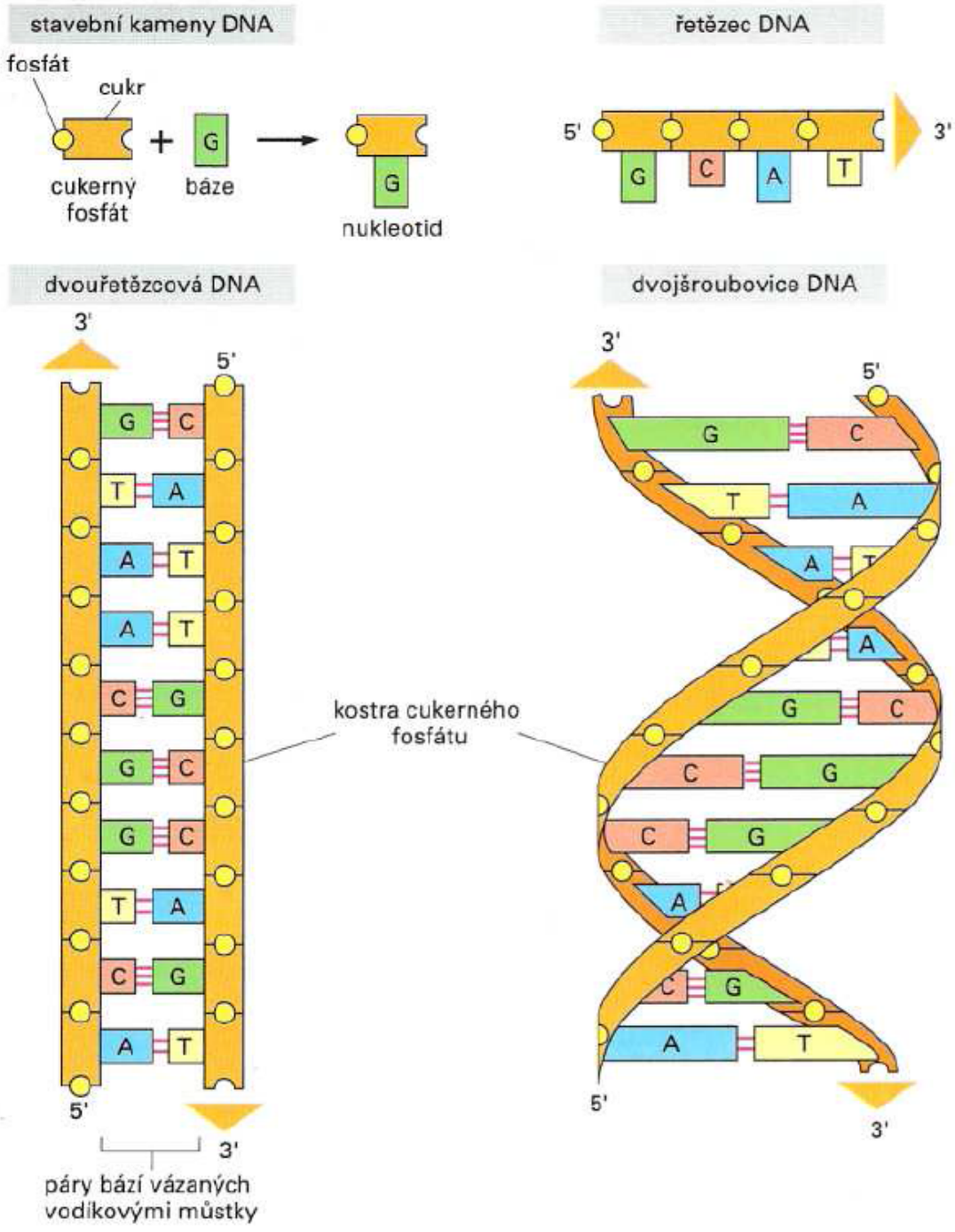
2.1 Štruktúra a funkcia DNA

Gény sa nachádzajú na chromozómoch v jadre bunky, tieto chromozómy sú zložené z DNA a proteínov. Molekula DNA sa skladá z dvoch dlhých polynukleotidových vlákien zložených štyrmi typmi nukleotidových podjednotiek. Obe tieto vlákna, taktiež nazývané reťazce DNA sú vzájomne prepojené vodíkovými mostíkmi naprieč neukleovými bázami.

Nukleotidy sú tvorené päťuholníkovým sacharidom, na ktorý je naviazaná dusíkatá báza a jedna alebo viacej fosfátových skupín. Báza môže byť adenin (A), cytosin (C), guanin (G), thymin (T). Primárnu štruktúru DNA tvorí reťazec nukleotidov prepojený kovelntnými väzbami.

Reťazec DNA môže končiť buď -OH skupinou sacharidu tzv. 3' koniec, alebo fosfátovou skupinou tzv. 5' koniec. Pri párovaní bázach sa vždy viaže bicyklciká báza s monocyklickou. Vznikajú tak väzby medzi adeninom a thyminom (A-T), alebo guaninom a cytosinom (G-C). Toto komplementárne párovanie báz umožňuje zaujať energeticky najvýhodnejšiu konformáciu vrámci dvojzávitnice.

Genetická informácia je zakódovaná v DNA ako poradie jednotlivých nukleotidov. Každá báza A,C,G a T sa môže považovať za jedno písmeno v abecede o veľkosti štyri. Gén je teda krátky úsek (sekvencia) DNA, ktorá kóduje informáciu pre tvorbu nejakého produktu, napríklad proteínu. Kompletná genetická informácia organizmu sa nazýva genóm [5].



Obrázek 2.1: DNA a jej stavebné podjednotky

2.2 Replikácia DNA

Pretože oba reťazce obsahujú sekvenciu nukleotidov, ktoré sú navzájom presne komplementárne, môžu obe vlákna poslúžiť ako templát - predloha pre syntézu nového vlákna. Toto

dovoľuje bunke replikovať svoje gény na dve dcérske bunky, ktoré sú jej verné kópie. Behom ôsmych hodín dokáže živočíšna bunka skopírovať cely svoj genóm s priemerne jednou či dvomi chybami. Túto prácu vykonáva súbor proteínov, ktorý dohromady vytvára replikačný aparát. Replikácia DNA dáva vznik dvom novým dvojzávitniciam, ktoré pochádzajú z materského hélixu.

K oddeleniu vlákien dochádza až pri teplote okolo stop stupňov Celzia, inak je dvojzávitnica veľmi stabilná. Vysoká teplota dodá dostatočnú tepelnú energiu potrebnú na prerušenie vodíkových mostíkov medzi bázami. Celý proces replikácie začínajú iniciačné proteíny, ktoré sa viažu na DNA a rozvíjajú jej štruktúru prerušovaním vodíkových mostíkov. Energia jedného vodíkového mostíku je malá, oddelenie krátkeho úseku DNA nevyžaduje veľa energie a môže prebiehať pri normálnej teplote. Vznikajú tak replikačné počiatky, tie väčšinou obsahujú viac A-T sekvencií, pretože adenin a thymín je viazaný iba dvoma vodíkovými mostíkmi. Hneď ako sa otvorí dvojzávitnica tak sa na replikačné počiatky viažu proteíny spolupracujúce na syntéze novej DNA.

V začiatkoch replikácie môžme pozorovať typické útvary v tvare Y, nazývajúc sa replikačné vidličky. Na každom replikačnom počiatku sa tvoria ľ vidličky, ktoré sa pohybujú smerom od seba a pomocou proteínu rozdeľujú dvojzávitnicu DNA a súčasne syntetizujú nový reťazec. Replikácia prebieha obomi smermi a preto je rýchla.

DNA polymeráza je jedna z najdôležitejších enzýmov pri polymerizácii a slúži k syntéze nového vlákna DNA. Nukleotidy dodávajú polymerizačnej reakcii energiu uvoľnenú pri hydrolýze. Aj keď je DNA polymeráza veľmi presným enzým, stáva sa, že pri syntéze sa objaví chyba. Syntetizovať sa môžu ako A-T a G-C, tak aj menej stabilné G-T a A-C. Pokiaľ by sa nesprávne replikovaný úsek neopravil, bude dochádzať k mutáciám, ktoré by mohli spôsobiť zánik organizmu. DNA polymerizácia má aj funkciu opravy. Ak je pripojený zlý nukleotid, DNA polymerizácia ho odštiepia naviaže nový. Pokiaľ je všetko v poriadku pokračuje sa ďalej v syntéze.

DNA polymerizácia však nedokáže syntetizovať úplne nové vlákno, iba pripojuje ďalšie nukleotidy k už spárovaným bázam. Úplne nové vlákno dokáže syntetizovať však iný enzým - primáza. Nevytvára DNA, ale RNA (ribonukleovú kyselinu), tá je dlhá približne 10 nukleotidov. Tento RNA úsek sa spáruje s templátovým reťazcom a poskytuje svoj 3' koniec, ktorý nejde ďalej nastavovať. Reťazec RNA je chemický dosť podobný jednovláknovej DNA, avšak je tvorený ribonukleotidmi namiesto deoxyribonukleotidmi, v ktorých je sacharid deoxyribózia nahradený ribózou. Ďalší rozdiel je použitie uracilu namiesto tymínu.

Pri syntéze je reťazec rozdelený na mnoho úsekov (Ozakiho fragmenty), ktoré je nutné spojiť dokopy. Ako prvé sú nueklázou odstránené RNA priméry, potom sú nahradené opravnou DNA polymerázou, a v konečnej fáze sú spojené všetky úseky spolu DNA ligázou.

Jednou zo základných zložiek replikačného aparátu je helikáza, enzým ktorý využíva energiu z hydrolýzy ATP k pohybu pozdĺž DNA a súčasne rozvíja dvojzávitnicovú štruktúru. Celý replikačný aparát je tvorený niekoľkými enzýmami (DNA polymerázou, primázou, helikázou). Taktiež obsahuje SSB-proteíny (z anglického single-strand binding proteins), ktoré ochraňujú jednovláknovú DNA uvoľnenú helikázou pred znovu spojením. Celý replikačný aparát je veľmi komplexný systém pohybujúci sa ako celok pozdĺž DNA. Jeho detailná štruktúra stále nieje úplne známa [5].

2.3 Prepis DNA na RNA

Transkripcia (prepis) a translácia (preklad) sú procesy, ktorými bunka realizuje svoje genetické inštrukcie - svoje gény. Na základe jedného génu môžu vzniknúť viacero kópií RNA

a jedna RNA môže dať vznik niekoľkým identickým molekulám proteínu. Tento obecný mechanizmus sa nazýva ústredná dogma molekulárnej biológie [5].

2.3.1 Transkripcia

Pri prepise DNA na RNA inak transkripcii sa tvorí reťazec RNA komplementárny k danému úseku DNA. Transkripcia začína rozvinutím krátkeho úseku dvojzávitnice DNA, jeden z reťazcov sa potom stane templátom pre syntézu RNA. O rozvinutie sa stará RNA-polymeráza pohybujúca sa po DNA v smere 3'->5', tá však nemá oproti DNA-polymeráze korektársku schopnosť a preto obsahuje viac chyb. Energia na pohyb sa získava z hydrolýzy. RNA nezostáva spojená s templátovou DNA, namiesto toho dochádza hneď za miestom kde bol pridaný ribonukleotid k obnoveniu dvojzávitnicovej štruktúre DNA a vytesneniu vlákna RNA. Preto sú molekuly RNA jednovláknové [5].

V bunke vzniká niekoľko typov RNA, a každá má svoju úlohu.

mRNA - mediátorová RNA, ktorej zásadná funkcia odovzdať genetickú informáciu

tRNA - transferová RNA prenáša jednotlivé aminokyseliny na miesto tvorby proteínov v ribozómoch

rRNA - ribozomálna RNA vytvára spolu s ďalšími proteínmi ribozómy

2.3.2 Translácia

Translácia je preklad genetickej informácie z poradia nukleotidov v mRNA do poradia aminokyselín v polypeptidovom reťazci prostredníctvom genetického kódu. Aminokyseliny alebo aminokarboxylové kyseliny sú organické zlúčeniny obsahujúce v molekule aminoskupinu a karboxylovú skupinu. Na skladanie proteínov sa používa 21 základných aminokyselín. Rastliny dokážu syntetizovať všetky aminokyseliny z anorganických látok. Naproti tomu živočíchy si dokážu tvoriť len niektoré aminokyseliny, a iné musia dostávať hotové z potravy.

Z transkripcie sa dostáva mRNA, ktorá sa napája na malú podjednotku ribozómu. Sekvencia RNA je postupne čítaná, nie ako jednotlivé bázy, ale ako ich trojice. Každá skupina troch nukleotidov sa nazýva kodón a určuje jednu aminokyselinu. Translácia mRNA k proteínu závisí na adaptorových molekulách, ktoré sú schopné jednou časťou molekuly rozpoznať a spárovať sa s kodónom v mRNA a inou častou naviazať aminokyselinu. Spoločne sa nazývajú tRNA, majú dĺžku okolo 80 nukleotidov.

Ribozóm je zložený z dvoch podjednotiek a zachytáva komplementárne molekuly tRNA a zapojuje aminokyseliny z tRNA na rastúci proteínový reťazec. Malá jednotka sa stará o spojenie tRNA a mRNA, veľká jednotka katalyzuje spojenie aminokyseliny s reťazcom proteínu. Ribozóm obsahuje 4 väzbové miesta, jedno pre mRNA a ostatné pre tRNA.

Translácia začína na inicializačnom (AUG) kodóne a končí stop-kodónom (UAA, UAG alebo UGA), ku ktorým nie je priradená žiadna aminokyselina a na stop-kodón sa namiesto tRNA viažu proteíny tzv. terminačné faktory, ktoré naviažu vody na uvoľnenie hotového reťazca [5].

2.4 Transpozóny

Transpozóny alebo transponovateľné elementy sú úseky DNA, ktoré sa môžu transpozíciou premiestňovať v rámci genómu z jedného miesta na druhé, a sú tak zdrojom genetickej

rozmanitosti. Ich veľkosť sa pohybuje od niekoľko stoviek až po desiatky tisíc párov báz [7]. V ľudskom genóme tvoria okolo 50% a v prípade kukurice je to až 90% [10].

Vieme, že existujú rôzne typy transpozónov, ako aj viac možností ako ich kategorizovať. Jedno zo základných rozdelení je podľa toho či vyžadujú reverznú transkriptázu k tomu aby schopné transponovaniu alebo nie. Tie čo áno sú známe ako retrotranspozóny alebo trieda 1 TEs, kde tie druhé sú DNA transpozóny alebo trieda 2 TEs [7].

Každá skupina transpozónov obsahuje autonómne a neautonómne prvky. Autonómne elementy majú ORF (z angl. Open Reading Frame), ktorý kóduje produkty potrebné k transpozícii. Na druhej strane, neautonómne elementy nekódujú transpozičné proteíny, ale sú schopné transpozície pretože si ponechávajú cis sekvencie potrebnej k transpozícii[11].

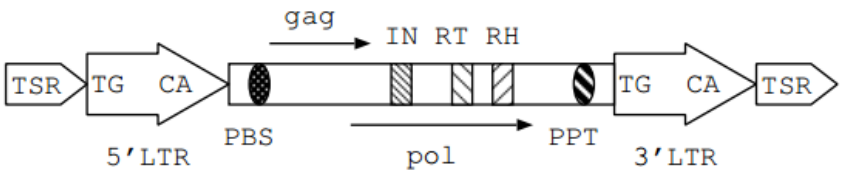
2.4.1 Retrotranspozóny

Sú najzložitejšie nevírové retroelementy dostatočne sa vyskytujúc u všetkých skupín eukaryotických buniek. Dlhé sú niekoľko kilobáz a obsahujú dlhé koncové repetície a gény gag a pol. Pri retropozícii sa RNA používa ako intermediát RNA[1]. Najskôr sa vytvorí RNA transkripciou DNA, a potom pri reverznej transkripcií dochádza k prepisu genetickej informácii do DNA. Tento mechanizmus sa tiež označuje ako *"copy and paste"*, pretože pri každom cykle replikácie sa vytvára jedna nová kópia.

Retrotranspozóny sa delia ďalej podľa toho či obsahujú dlhé ukončovacie repetície - LTR (z angl. Long Terminal Repeats), ktoré obklopujú telo retroelementu na **LTR retrotranspozóny** a **Non-LTR retrotranspozóny**[7].

LTR retrotranspozóny

Sú charakteristické tým, že pri premiestnení zanechajú na mieste svoju pôvodnú kópiu. Dochádza teda k ich replikácii a vďaka tomu môžu genóm zahltiť veľkým počtom svojich kópii [6].



Obrázek 2.2: Štruktúra LTR retrotranspozónu

Štruktúra plnej dĺžky LTR retrotranspozónu typicky môže obsahovať tieto časti:

LTR región : 5'LTR a 3'LTR sú dva podobné regióny. Sú totožne v dobe vloženia transpozónu do hosťovateľského génomu, ale potom sa vyvíjajú nezávisle. Mutácie sa vyskytujú často.

TSR región (z angl. target site repeat) je 4-6bp krátky opakujúci úsek lemujúci konce elementu. Je to príznak vloženia transpozónu.

PBS : Blízko 3' konca je 18bp dlhá sekvencia komplementárna k tRNA koncu. Je veľmi podstatná pretože vytvorenie väzby k tRNA je prvý krok inicializácii reverznej transkripcie.

PPT : Polypurínový trakt je krátky úsek okolo 11-15bp bohatý na purín. Tento región je dôležitý pre reverznú transkripciu.

Proteínové domény : V typickom vírusovom genóme sa nachádzajú tri polygény: *gag*, *pol* a *env*. Medzi nimi, pol je najviac udržovaný. Vrámci pol sú tri podstatné proteínové domény: IN(integráza), RT(reverzná transkripcia) a RH, čo sú enzýmy pre reverznú transkripciu a inzerciu. RT a IN sú považované ako nevyhnutné pre fungovanie LTR elementu [2].

Nachádzame ich hlavne v rastlinách. Medzi ne patria rodiny Gypsy a Copia (tiež označované ako Ty1 a Ty3). Najviac sa vyskytujú v obilninách. LTR TE sú dosť podobné retrovírusom, avšak stým rozdielom, že im chýba gen *env* (z angl. envelope, obálka), ktorý kóduje jedno zo zložiek vírovej kapsule umožňujúc retrovírusom opustiť bunku. Z tohto dôvodu je pravdepodobné, že sa LTR retrotranspozóny vyvinuli z endogénnych retrovírusov alebo sa naopak endogénne retrovírusy vyvinuli z transpozónov [5].

Non-LTR retrotranspozóny

Tieto transpozóny typicky obsahujú jedno alebo dva ORF. Ich premiestňovanie sa deje pomocou mechanizmu TPRT (z angl. Target-Primed Reverse Transcription). Najskôr je transpozón prepísaný polymerázou II do mRNA, tak ako pri LTR transpozónoch. Potom endonukleáza vytvorí na jednom reťazci hostiteľskej DNA zárez uvoľňujúci 3' koniec, ktorý sa dá využiť ako počiatok pre reverznú transkripciu. Mnohé non-LTR transpozóny sú na 5' konci poškodené z dôvodu predčasného ukončenia reverznej transkripcie, ktorá nie je tak efektívna ako transkripcia popredná.

2.4.2 DNA transpozóny

Pretože sa DNA transpozóny pohybujú tzv. *"cut and paste"* - vystrihni a vlož, počet ich kópii zostáva v génome konštantný. Mnoho DNA transpozónov lemuje na oboch koncoch obrátené repetície TIR (z angl. Terminal Inverted Repeat) o dĺžke 9 až 40 bp. DNA transpozóny kódujú enzým transopozázy, ktorá rozoznáva TIR repetície na ktoré sa naviaže rozštiepenú hostiteľskú DNA a integruje transpozón na cieľové miesto.

Kapitola 3

Nástroje na identifikáciu transpozónov

Na identifikáciu transpozónov sa používajú dva hlavné prístupy. *De novo* metóda vytvára predikcie len na základe výpočtového modelu, bez použitia vonkajších porovnávaní s už vopred existujúcimi dátami.

3.1 LTR Finder

Je to efektívny program na hľadanie plnej dĺžky LTR retrotranspozónov v genómových sekvenciách. LTR Finder ako prvé skonštruuje všetky exaktné páry na základe pripájajúceho algoritmu, ktorý ich rozšíri na dlhšie vysoko podobné páry. Potom je použitý Smith-Watermanov algoritmus k prispôsobeniu koncov LTR párových kandidátov na získanie zarovnavacích hraníc. Tieto hranice sú ešte posúvané na základe informácie o 5'-TG..CA-3' a TSR, ktoré LTR retrotranspozóny obsahujú. Ďalej sa snaží LTR Finder identifikovať PBS, PPT a RT vrámci vopred vyznačenými hranicami. Identifikácia reverznej transkripcie zahrňuje dynamický proces zohľadňujúci rámový posun. Pre ostatné proteínové domény využíva *ps_scan* (od PROSTITE[3]), aby lokalizoval jadrá podstatných enzýmov.

Nakoniec LTR Finder reportuje a nájdených výsledkoch kde rôzne retrotranspozóny sú zoradené podľa presnosti zásahu jednotlivých domén[2].

3.2 Tedna

Tedna (z angl. a transposable element de novo assembler) sa snaží poskladať výsledný model transpozónu z jeho viacero kópii, ktoré sa menili časom. Každá kópia naznačuje ako by mal výsledný transpozón vyzerať, ale iba v porovnaní s ostatnými kópiami. Tenda je prvý nástroj ktorý využíva de Brujiho Grafy k skladaniu transpozónov. Graf skladá na základe K-merov, čo sú podreťazce o dĺžke k [8].

3.3 RepeatMasker

Program, ktorý prehľadáva DNA reťazce a hľadá prekladané repetície. Porovnáva ich s vybranými databázami pomocou niektorých rozšírených vyhľadávačov: nhmmer, cross_match, ABBlast/WUBlast, RMBlast a Decypher. Taktiež využíva udržiavané knižnice repetícií,

aktuálne podporuje RepBase a Dfam. Ešte predtým, než RepeatMasker začne porovnávať sekvencie s databázami, umožni užívateľovi maskovať určité regióny. Nízko komplexné DNA, tandemové repetície, polypurín a regióny bohaté na AT prípadne CG môžu vyvolať falošnú zhodu. Predvolené sú maskované spolu s prekladanými repetíciami [4].

3.4 RepeatExplorer

RepeatExplorer je kolekcia softvérových nástrojov na charakterizáciu opakujúcich sa elementov. Kľúčová súčasť serveru je výpočtová rúra (z angl. pipeline), ktorá zamestnáva grafovo založený algoritmus zhlukovania uľahčujúc *de novo* identifikáciu repetíc, bez použitia referenčných databáz už známych elementov. Keďže algoritmus používa krátke sekvencie náhodne navzorkované z vloženého génomu, je ideálny na analýzu novej generácie sekvenčných readov.

Ďalšie nástroje sú poskytnuté na pomoc klasifikácie identifikovaných častí, investigáciu fylogenetických vzťahov retroelementov a vykonávajú komparatívnu analýzu nad zloženiami repetícií viacerých druhov. Server dovoľuje analyzovať niekoľko miliónov sekvenčných readov, čo typicky znamená identifikáciu väčšiny vysoko a stredne opakujúcich sekvenciách u vyšších rastlín.

RepeatExplrer je implementovaný v rámci prostredia Galaxy a poskytuje webové prostredie na jeho obsluhu.

3.4.1 Zhlukovanie

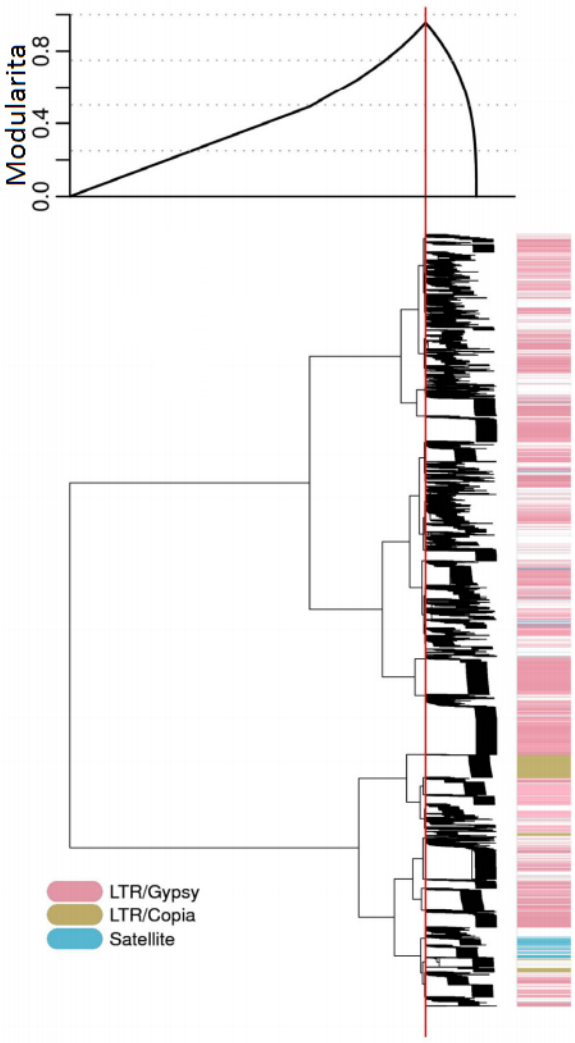
Analýza je vykonaná na všetkých vstupných readoch, ktoré reprezentujú krátke nukleotidové sekvencie náhodne vzorkované z analyzovaného génomu. Začína s identifikáciou podobností medzi vstupnými sekvenciami a to tak, že sa porovnáva každá s každou, a zaznačí sa ak sa prekrývajú nad určitú zadanú úroveň. Táto informácia je potom použitá ku konštrukcii grafu kde uzly korešpondujú s jednotlivými sekvenčnými readmi. Prekrývajúce ready sú prepojené hranami a ich skóre podobnosti je vyjadrené váhou hrany.

V prípade nízko-hĺbkového sekvencovania poskytujúceho menej ako polovicu pokrytia genómu, unikátne sekvencie sú riedko pokryté z čoho vyplývajú izolované uzly grafu, ktoré nemajú žiadne prepojenia s inými časťami. Na druhú stranu, opakujúce sa repetice konštruujú skupiny vzájomne prepojených uzlov, kvôli častému prekrývaniu readov.

Izolovaná skupina uzlov grafu, v ktorej akékoľvek dva vrcholy sú prepojené a nie je možne pridať ďalšie uzly alebo hrany sa označuje termínom *prepojený komponent* [9]. O identifikáciu prepojeného komponentu sa stará program tclust, ktorý bol predtým zamestnaný analýzou zhlukov.

Ideálne by takéto zhlukovanie malo byť dostatočné k oddeleniu rôznych rodín opakujúcich sa elementov, avšak stáva sa, že zhluky sú rozdelené prípade obsahujú viacero prvkov. Aby sa predišlo tomuto problému je vykonaná ďalšia analýza grafovej štruktúry za použitia hierarchického aglomeračného algoritmu, ktorý detekuje skupiny uzlov v grafe, ktoré sú medzi sebou prepojené hustejšie ako so zvyškom uzlov. Nazývame ich *komunity* [9].

K nájdeniu optimálneho rozdelenia grafu do komunít je použitý greedy algoritmus, aby našiel maximálnu modularitu, čo je náznak kvality pre grafový zhluk. Určuje frekvenciu prepojení uzlov vrámci jednej komunity, ale hľadí aj na náhodne prepojenia a zahŕňa ich v úvahu. Modularita o hodnote nula naznačuje, že prepojenia v komunite nie sú o moc lepšie ako náhodné, kde hodnota jedna indikuje silné väzby.



Obrázek 3.1: Hierarchická organizácia sekvencií *P.sativum*. Každá vetva stromu reprezentuje jednu sekvenciu. Tento strom korešponduje najväčšej prepojenej komunite. Červená vertikála ukazuje najlepšie miesto na rozdelenie, k zaisteniu najvyššej modularity.

Nakoniec ešte pospája jednotlivé zhluky to tzv. super zhlukov, kde je znázornená hierarchická organizácia. Prepojenie medzi dvoma zhlukmi musí spĺňať minimálnu úroveň nazývanú ako *cutoff* kde sila prepojenia k dvoch zhlukov x a y je definovaná ako

$$k_{x,y} = 2 * W/(n_x + n_y)$$

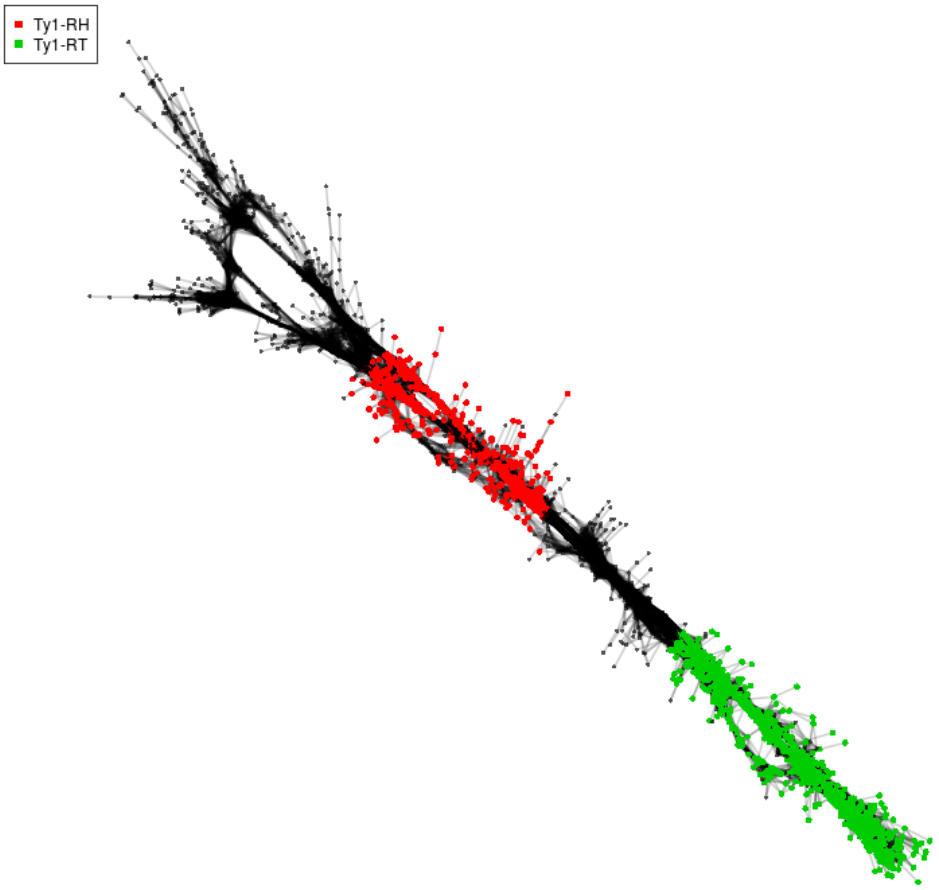
pričom W je počet zdieľaných readov medzi zhlukmi x a y, a n_x je počet readov v zhluku x s chýbajúcimi readmi vrámci toho istého zhluku. Táto medza sa označuje ako *cutoff* a RepeatExporer ju ponúka v troch hodnotách: 0.05, 0.1 a 0.2.

3.4.2 Výstupné dáta

Výstupom RepeatExploreru sú zhluky ktoré reprezentujú rôzne rodiny repetitívnych elementov. Zaznamenané sú buď ako HTML stránky alebo komprimovaný archív, ktorý obsa-

huje nasledovné priečinky a súbory:

- Adresár, ktorého obsah je použitý k prezentácii výsledkov pomocou HTML stránky. Obsahuje podadresáre pre každý zhluk, v ktorom sú dodatočné informácie.
- Adresár vstupných sekvencií, ktoré boli použité k analýze.
- Súbory obsahujúce tabuľky vyhľadávania podobnosti, alebo anotácie proteínových domén transpozónov. Napríklad blastx, blastn alebo RepeatMasker.
- Adresár obsahujúci súbory zo zostavovania readov do contigov [9].



Obrázek 3.2: Zhluk kódujúci časť transpozónu z rodiny Copia - reverznú transkriptázu (Ty1-RT) a ribonukleázu H (Ty1-RH)

Kapitola 4

Špecifikácia a analýza požiadavkov

Jeden z najlepších programov na identifikáciu repetitívnych transpozónov je RepeatExplorer, avšak jeho výstupné súbory obsahujú veľké množstvo dát, ktoré potom treba prehľadávať k nájdeniu správneho výsledku. Následná validácia týchto výsledkov je zdĺhavá. Bolo teda za potrebné vymyslieť spôsob ako uľahčiť a zautomatizovať tento krok. Problém bol identifikovaný na tri časti:

- vytvoriť graf zhlukov na základe podobnosti
- overiť vnútornú štruktúru podľa navrhnutých modelov
- prezentácia výsledkov

4.1 Vytvorenie super zhlukov

RepeatExplorer poskytuje možnosť vytvárania zhluku zhlukov (ďalej už len ako super zhluky z angl. *super cluster*), ale iba na základe fixných hodnôt. Stáva sa teda, že zhluky obsahujú aj časti ktoré by tam byť nemali. Preto bolo potrebné navrhnúť metódu spájania zhlukov v reálnom čase, ktorá by umožňovala experimentovanie s výsledkom.

4.1.1 Získanie dát o spojení zhlukov

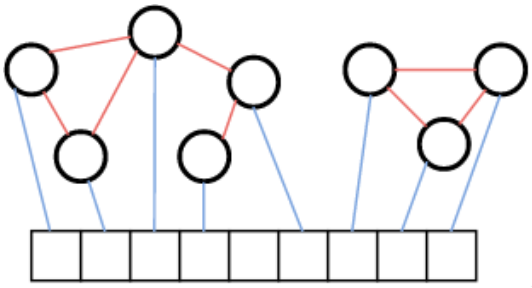
Na základe analýzy vstupných dát bol vybraný súbor **clusterConnections.txt**. Obsahuje záznam o prepojení jednotlivých zhlukov. Každý riadok predstavuje mieru podobnosti dvoch zhlukov a číslo na konci udáva jej veľkosť. Ak sa názvy zhlukov zhodujú, dané číslo znamená veľkosť zhluku alebo aj maximálnu možnú podobnosť.

4.1.2 Dátová reprezentácia super zhlukov

Po získaní dát bolo treba vymyslieť spôsob, akým budú uložené super zhluky, aby bol zaistený:

- rýchly prístup k jednotlivým zhlukom
- zachovaná štruktúra super zhluku

Tieto požiadavky splňuje pole uzlov kde každý uzol môže obsahovať 0 až N prepojení na iný uzol, tvoria teda neorientovaný graf. Pri prechádzaní grafu si musí uzol poznačiť či už



Obrázek 4.1: Grafické znázornenie navrhnutej dátovej štruktúry. Vidíme dva možné super zhluky o veľkosti 5 a 3. Krúžky znázorňujú zhluky, červené čiary prepojenia medzi zhlukmi a modré čiary referenciu na pole zhlukov.

bol navštívený, pretože v grafe môžu vznikať slučky. Indikácia návštevy vyznačuje začiatok a koniec grafu na poli uzlov.

4.1.3 Algoritmus vytvárania super zhlukov

Algoritmus sa skladá z dvoch fáz: V prvej fáze sa asynchrónne číta súbor s podobnosťami zhlukov. Na začiatku sa inicializuje pole uzlov P. Na každom riadku sa nachádza jeden záznam, ktorým sa rozumie trojica ($z1$, $z2$, x) pričom $z1$ a $z2$ su identifikátory zhlukov a x je veľkosť podobnosti.

Pre každý záznam následne vytvorí dvojica uzlov. Následne sa prehľadá pole všetkých uzlov. Ak sa v poli uzol $z1$ už nachádza, tak sa uzol $z2$ naviaže na nájdený uzol. Ak sa v poli nachádza $z2$, tak sa uzol $z1$ naviaže na nájdený uzol. Veľkosť väzby je určená parametrom x.

Prvá fáza končí spracovaním posledného záznamu. V poli P je zoznam všetkých uzlov, ktoré sú vzájomne previazané a tvoria grafy - super zhluky.

V druhej fáze sa inicializuje prázdne pole super zhlukov S, ktoré bude naplnené poľami uzlov, ktoré vytvárajú super zhluk. Prechádza sa pole uzlov P.

Pre každý uzol u nachádzajúci sa v poli P sa zisťuje, či nebol v druhej fáze navštívený. Ak uzol dosiaľ nebol navštívený, vytvorí sa pole X do ktorého sa vloží tento uzol. Rekurzívne je prechádzaný graf vytvorený fázou 1 začínajúc uzlom u. Pre každý uzol grafu je označené navštívenie a je pridaný do poľa X. Pole X je potom pridané do poľa super uzlov S.

Po skončení fáze 2 je pole S naplnené super zhlukmi.

4.2 Rekonštrukcia transpozónov

Aby LTR transpozón správne fungoval, musí kódovať určité funkčné prvky - proteínové domény.

Rekonštrukcia transpozónov prebieha na základe známosti ich vnútornej štruktúry, kde musí byť dodržané správne poradie. Toto poslúži ako model overenia správnosti. Z analýzy štruktúry LTR transpozónov boli použite dva nasledovné modely, rozdelené podľa rodiny.

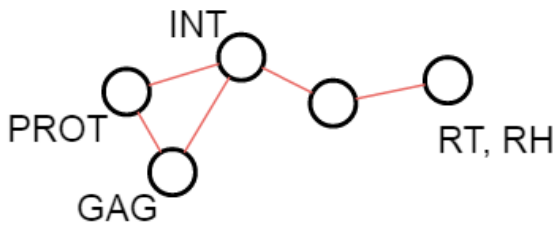
Copia - GAG, PROT, INT, RT a RH

Gypsy - GAG, PROT, RT, TH a INT

Model si možno predstaviť ako orientovaný graf, kde jeden uzol má maximálne jedno prichádzajúce a jedno odchádzajúce spojenie. Rekonštrukciou alebo validáciou sa potom rozumie hľadanie grafu modelu v grafe super zhluku, pričom dva uzly sú zhodné ak proteínové domény, ktoré označujú sú zhodné.

Jeden zhluk môže reprezentovať viacero proteinových domén, preto by výsledná zhoda uzlov, mala uvažovať aj nasledujúcu doménu validačného modelu.

Rekonštrukcia je úspešná, ak sa nájde graf modelu v grafe super zhluku.



Obrázek 4.2: Príklad možného super zhluky, ktorý by nevyhovel žiadnemu uvedenému modelu, pretože chýba priame prepojenie uzlov INT a RT,TH.

4.2.1 Mapovanie proteinových domén

Na to aby rekonštrukcia vôbec mohla začať je potrebné získať informácie o doménach jednotlivých zhlukov. Výstup RepeatExploreru poskytuje nasledovné

blastx

Tento adresár obsahuje výstup programu blastx rozdelený na jednotlivé zhluky. V súbore CL_*n*_blastx.csv pričom *n* označuje identifikátor zhluku, je priamy výstup obmedzený na sekvencie konkrétneho zhluku. Súbor CL_*n*_domains.csv zaznamenáva tieto dáta ale už v filtrovanej forme. Obsahuje informácie o názvu proteínovej domény, jej rodinu a počet zásahov v danom zhluku.

RM_output_tablesummary.csv

Tabuľka ktorá obsahuje záznam pre každý zhluk. Je poskladaná z výstupu programu RepeatMasker (viď 3.3). Obsahuje informácie o type zhluku.

Po skombinovaní informácii z uvedených súborov je možné s určitou presnosťou považovať, že sekvencie zhluku kódujú vybrané domény.

4.3 Zobrazenie super zhlukov

Grafické rozhranie programu RepeatExplorer zobrazuje každý super zhluk na vlastnej obrazovke, čím znemožňuje rýchle porovnávanie. Zobrazenie všetkých super zhlukov na jednej obrazovke by tento problém vyriešilo, avšak príliš veľká hustota zobrazovaných informácii stráca výpovednú hodnotu.

Použitím interaktívnej zobrazovacej plochy s možnosťou približovania umožňuje používateľovi sa zamerať na časť, ktorú považuje za podstatnú.

Zároveň je dôležité aby mal program jednoduché a intuitívne grafické prostredie a čo najviac tak uľahčil prácu užívateľovi.

Kapitola 5

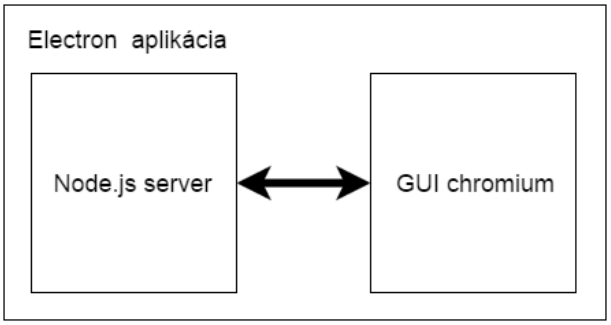
Návrh a implementácia

5.1 Použité technológie

K implantácii bol použitý slabo typovaný programovací jazyk JavaScript štandardu ECMAScript 6[1], vďaka veľkému množstvu už existujúcich knižníc, či už pre prácu s dátami alebo ich vizualizáciou.

Electron[2]

Architektúra programu electron je implementovaná vo forme modelu klient-server. Ako server slúži Node.js, ktorý používa na interpretáciu JavaScript kódu Chrome V8 engine[3]. Tento server spolu s grafickým užívateľským prostredím, ktoré vykresľuje Chromium[4] tvorí jadro aplikácie založenej na technológii Electron.



Obrázek 5.1: Model klient-server vrámci aplikácie Electron

Keďže kód pre klientsku časť a kód pre serverovú časť je napísaný v tom istom jazyku nieje potrebné zložito riešiť komunikáciu medzi týmito dvomi časťami. Jednou z výhod použitia Electronu je, že pobeží na rôznych operačných platformách.

[1] `http://java.ociweb.com/mark/STLJS/ES6.pdf`
[2] `http://electron.atom.io/`
[3] `https://nodejs.org/en/`
[4] `https://www.chromium.org/Home`

HTML[5]

Jazyk HTML patrí do rodiny značkovacích jazykov SGML (Standart Generalized Markup Language). Skratka HTML pochádza z anglického HyperText Makup Language. Jazyk je určený k vytváraniu statických dokumentov, ktoré obsahujú hypertextové odkazy a pokročilé formátovanie.

CSS[6]

Kaskádové štýly alebo CSS (skratka z angl. Cascading Style Sheets) je všeobecné rozšírenie HTML, ktoré si za úlohu dáva zmenu prezentácie dokumentu. Umožňuje oddelenie formátovania od obsahu dokumentu.

5.1.1 Použité externé knižnice

Underscore.js[7]

Poskytuje celú sadu užitočných pomocných funkcií na prácu s dátami. V aplikácii je táto knižnica použitá pri spracovaní a následnom mapovaní či filtrovaní vstupných tabuliek.

NodeCSV[8]

Poskytuje generovanie, parsovanie, transformáciu a serilizáciu CSV súborov pre Node.js. Využitá bola parsovacia časť, na získanie dát zo vstupných súborov v serverovej časti aplikácie.

ShellJS[9]

ShellJS je prenosná (Windows/Linux/OS X) implementácia príkazov Unixového shellu pomocou Node.js API.

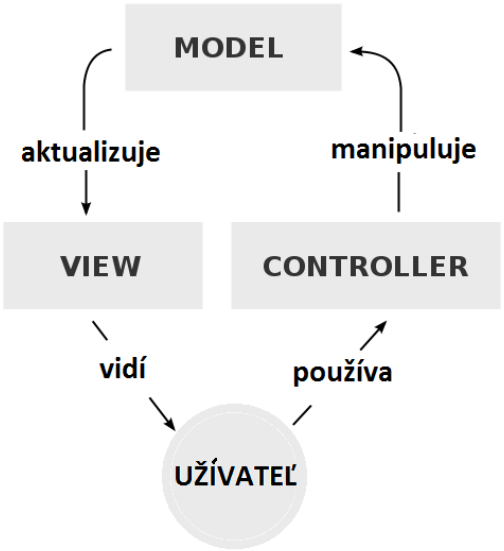
D3.js[10]

Javascriptová knižnica D3.js (D3 z anglického Data-Driven-Documents) sa používa na produkovanie dynamických, interaktívnych dátových vizualizácii. K tomu využíva štandardy SVG, HTML a CSS. D3 poskytuje veľkú kontrolu nad finálnou vizualizáciou.

Angular[11]

Komplexný framework, na vytváranie bohatého užívateľského rozhrania alebo stránok. Používa návrhový vzor MVC (Model View Controller), kde jednotlivé modely (Model) sú spracované dáta, zobrazovaciu vrstvu (View) tvoria chytré šablóny a v kontroléroch (z angl. Controller) je uložená logika aplikácie.

[5] `https://github.com/shelljs/shelljs`
[6] `http://http://underscorejs.org/`
[7] `http://underscorejs.org/`
[8] `https://github.com/wdavidw/node-csv`
[9] `https://github.com/shelljs/shelljs`
[10] `https://d3js.org/`
[11] `http://http://underscorejs.org/`



Obrázek 5.2: Znázornenie architektúry Model View Controller

5.2 Implementácia

Základ aplikácie je postavený použitím technológie *Electron* (viď 5.1). Tá po spustení vytvorí inštanciu *BrowserWindow* kde beží grafické užívateľské rozhranie.

5.2.1 Tvorba a validácia super zhlukov

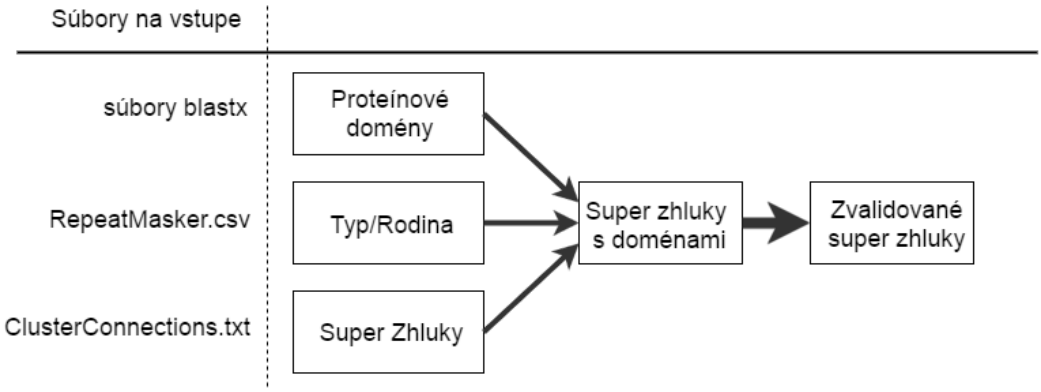
Na získanie ciest k jednotlivým súborom je použitá knižnica *ShellJS* (viď 5.1.1). Súbory sa vyhľadávajú pomocou funkcie *find* a jej výstup je následne filtrovaný aby odpovedal názvu danému súboru.

Tabuľky vo formáte *csv* sú získané s použitím knižnice *NodeCSV* (viď 5.1.1). Výstupom funkcie *parse* je pole objektov, kde každý objekt predstavuje jeden riadok vstupnej tabuľky. Tieto objekty tvoria vstupné dáta algoritmu vytvárania super zhlukov popísaný v 4.1.3. Algoritmus je implementovaný použitím jednoduchých javascriptových objektov a polí.

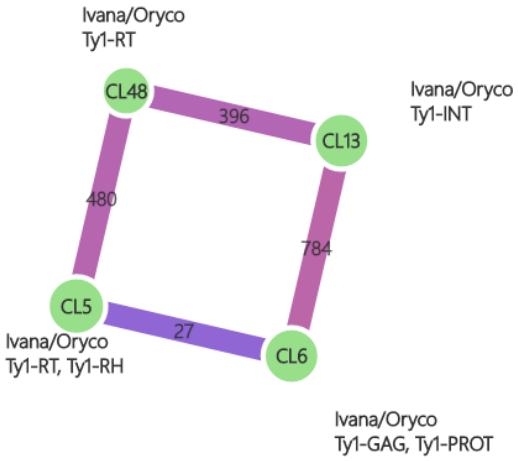
Všetky súbory sa čítajú asynchrónne a následne sa čaká na dokončenie spracovania všetkých vstupov. Potom sa mapujú informácie o rodine a typu sekvencie zhluku na super zhluky. LTR retrotranspozóny rodín *Gypsy* a *Copia* sú podrobené validácii. Ako model na overenie správnosti je použitý orientovaný graf unikátny pre každú rodinu, tvorený názvami proteinových domén. Algoritmus na základe daného modelu transpozónu prehľadáva graf super zhluku a hľadá v ňom zhodu s grafom modelovým. Niektoré zhluky môžu obsahovať viacero proteinových domén, preto prehľadávanie grafu počíta aj s touto možnosťou.

5.2.2 Vizualizácia

Jednotlivé grafy super zhlukov su vykreslené na spoločné SVG plátno s použitím d3.js Force Layout (viď 5.1.1). Veľkosť jednotlivých zhlukov je logaritmicky proporčná k maximálnej veľkosti prepojenia zhluku. Hrany grafu, ktoré reprezentujú veľkosť prepojenia sú odlíšené farbou. Vizualizovaný graf je interaktívny, možno v ňom približovať a jeho časti rôzne posúvať a prekladať, štruktúra grafu však ostane zachovaná.



Obrázek 5.3: Tok dát pri spracovaní vstupov



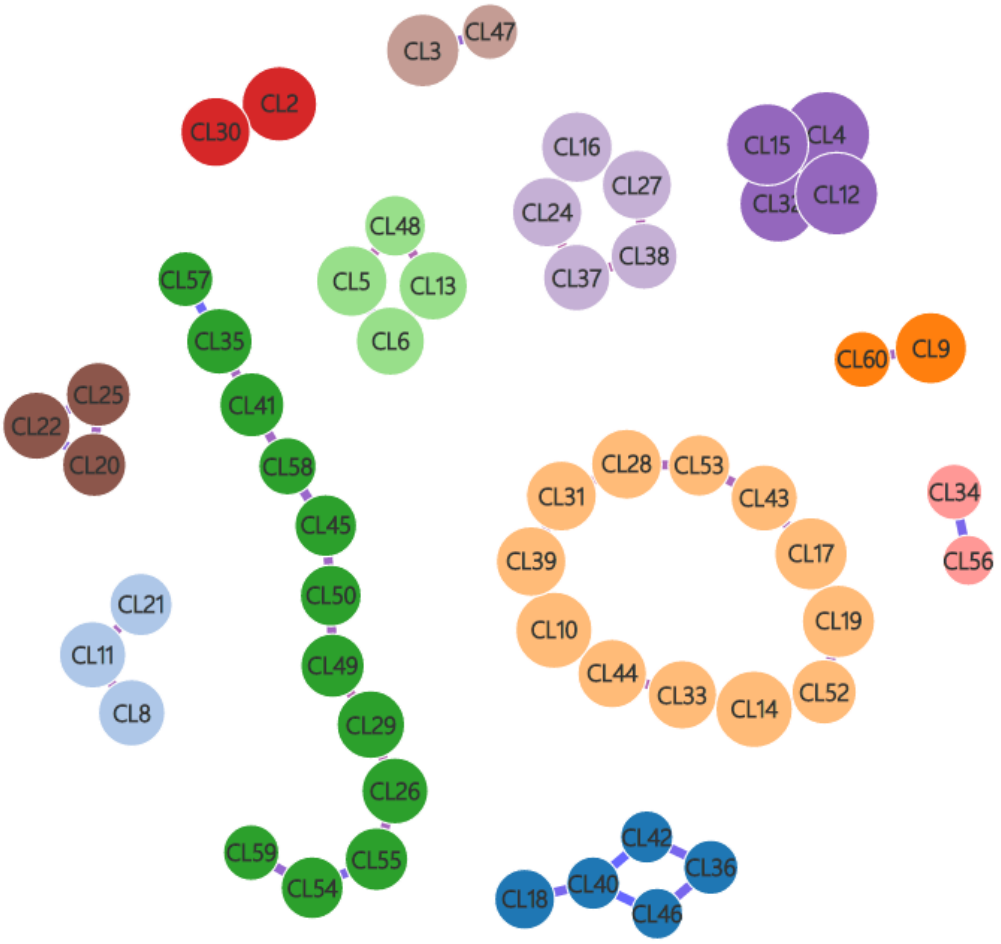
Obrázek 5.4: Zobrazenie detailu super zhluku. Je možné vidieť názov zhluku, jeho zaradenie do rodiny, proteínové domény a silu prepojenia.

5.2.3 Grafické rozhranie

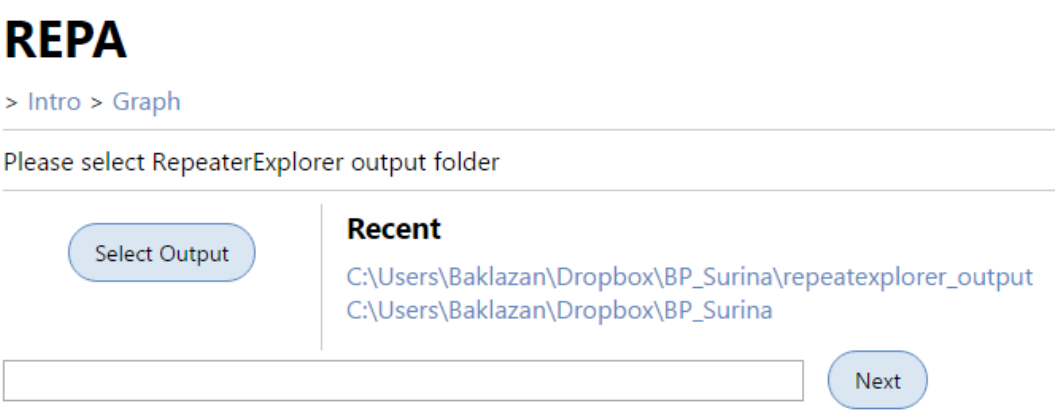
Zvyšok grafického rozhrania pomocou HTML, CSS a framewroku Angular2, kde jedna jeho komponenta je graf. Úvodná trieda *Index* definuje možné obrazovky a to:

Intro - úvodná obrazovka, nastavená ako predvolená. Obsahuje tlačidlo na výber cesty k výstupným súborom RepeatExploreru. Po každom úspešnom načítaní si túto cestu perzistentne uloží a užívateľ si ju môže zvoliť pri ďalšom spustení aplikácie. Uchováva sa posledných 5. Po kliknutí na tlačidlo next, nasleduje prechod na druhú obrazovku. Keďže spracovanie dát aplikáciou trvá pár sekúnd (3-20, záleží na veľkosti vstupných dát), užívateľ musí čakať. Táto obrazovka sa stamví a dá najavo pomocou textovej hlášky, že sa čaká na spracovanie vstupov.

Graph - Hlavným komponentom tejto obrazovky je graf, ďalej sa tu nachádzajú tabuľky vytvorené chytrými šablónami pomocou AngularJS. Tabuľka SuperClusters obsahuje zoznam vytvorených super zhlukov, ich veľkosť a klasifikáciu. V prípade LTR transpozónu rodiny Copia a Gypsy je zobrazená informácia o validácia. Buď tu je **valid**,



Obrázek 5.5: Zobrazenie všetkých super zhlukov.



Obrázek 5.6: Obsah úvodnej obrazovky

vtedy je validácia v poriadku, alebo je tu napísané prečo sa nepodarila.

SuperClusters

	Size	Classification	
SuperCluster1	2	LTR/Copia	**valid**
SuperCluster2	12	rDNA	
SuperCluster3	12	Organelle/plastid	
SuperCluster4	5	N/A	
SuperCluster5	5	rDNA/5S	
SuperCluster6	4	LTR/Copia	**valid**
SuperCluster7	3	LTR/Copia	**valid**
SuperCluster8	5	LTR/Gypsy	**valid**
SuperCluster9	3	DNA/En-Spm	
SuperCluster10	2	Satellite/TR11	
SuperCluster11	2	Low_complexity	
SuperCluster12	2	Low_complexity	

Obrázek 5.7: Tabuľka aktuálne vytvorených super zhlukov. Každá farba zodpovedá farbe zobrazenej v grafe.

Summary of RepeatMasker hits

Class.Family	hits	hits[%]
LTR/Copia	354	40.925

Total number of similarity hits for each lineage

Lineage	Domain	Hits
Ivana/Oryco	Ty1-RT	368
Alel/Retrofit	Ty1-RT	23

Obrázek 5.8: Tabuľka zobrazujúca podrobnejšie informácie o vybranom zhluku.

Kapitola 6

Testovanie

Keďže aplikácia je určená pre úzku skupinu používateľov, testovanie funkčnosti aplikácie prebiehalo predovšetkým v spolupráci s Biofyzikálnym ústavom Akedemie věd Českej Republiky. Užívatelia, ktorý testovali aplikáciu boli výskumníci z oddelenia vývojovej genetiky rastlín. Testovanie sa uskutočňovalo na reálnych dátach, pričom boli overené jednotlivé funkcie aplikácie.

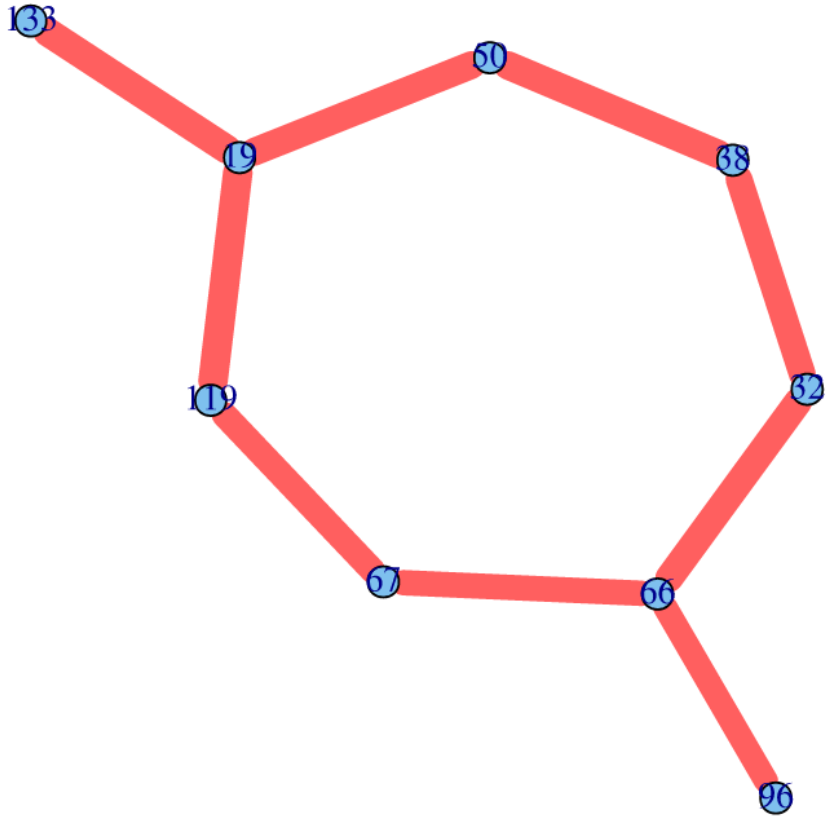
Porovnávali sme výstupy ktoré umožňuje RapidExplorer a výstupy mnou implementovanej aplikácie REPA. Ďalej sme testovali využiteľnosť novej funkcionality ako nastavovanie spojení, rozdeľovanie do super zhlukov a ich validáciu.

V prvej fáze testovania som prezentoval prototyp, ktorý umožňoval vytváranie super zhlukov a ich prezentáciu . Toto testovanie ukázalo, že algoritmus vytvárania super zhlukov bol nedostatočný, pretože spájal zhluky z rôznych rodín. Užívatelia považovali vizualizáciu za najväčší prínos oproti nástrojom, ktoré dovtedy používali. Ďalej boli odhalené rôzne implementačné chyby. Taktiež boli vznesené nasledovné požiadavky na doplnenie aplikácie:

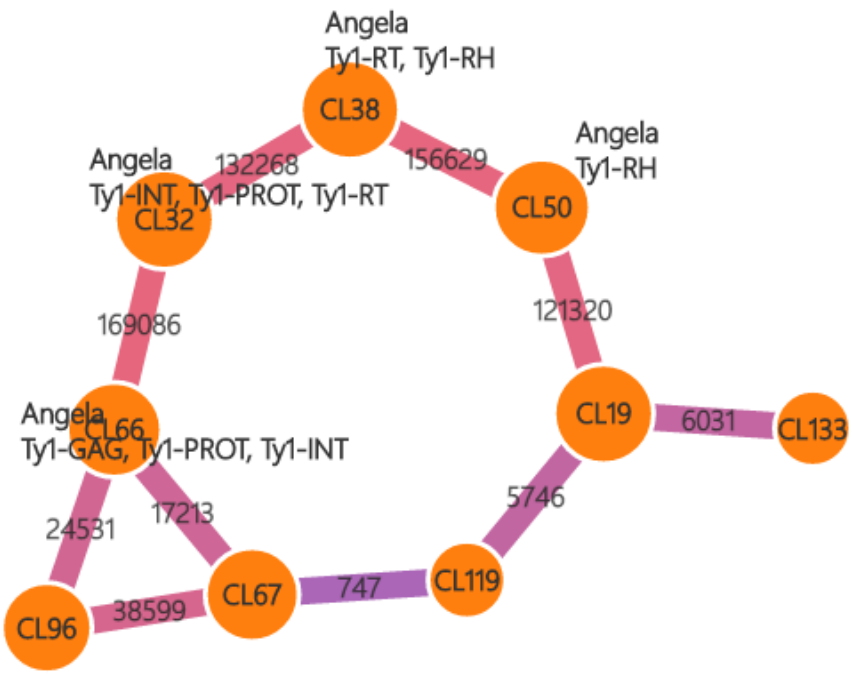
- možnosť experimentovania s vytváraním super zhlukov
- vizuálne rozlíšenie rôznych super zhlukov pridelením odlišných farieb
- možnosť približovania v grafe
- grafické znázornenie výskytu proteínových domén vrámci zhlukov
- možnosť uloženia vytvoreného výstupu do súboru

Väčšina týchto požiadaviek bola zahrnutá v implementácii výslednej aplikácie. Zo spomenutých problémov som neimplementoval jedine možnosť uloženia do súboru a zahrnul ju do možných rozšírení v budúcich verziách.

Testovanie výslednej aplikácie ukázalo prínos v kvalite zobrazenia zhlukov a možnosti nastavenia variabilného prahu algoritmu vytvárania super zhlukov. Na obrázkoch 6.1 a 6.2 je možné vidieť rozdiel v prezentácii transpozónu rodiny *Copia*. Pri aplikácii REPA vidí užívateľ aj vnútornú štruktúru a silu jednotlivých spojení. Taktiež sa ukazuje rozdiel algoritmov na vytváranie super zhluku.



Obrázek 6.1: Zobrazenie super zhluku programom RepeatExplorer, pri hodnote cutoff 0.5



Obrázek 6.2: Zobrazenie super zhluku aplikáciou REPA pri nastavenom prahu spojení 500

Kapitola 7

Záver

V tejto práci bol vytvorený bioinformatický nástroj, ktorý overuje štruktúru LTR transpozónov.

V priebehu vývoja som si rozšíril znalosti z molekulárnej biológie, konkrétne v oblasti genetiky so zameraním na transpozóny. Dozvedel som sa ako využívajú informačné technológie k porozumeniu biologických dát. Vyskúšal som si prácu s nástrojmi na ich detekciu a identifikáciu. Tieto znalosti mi potom pomohli navrhnúť vlastnú aplikáciu.

Navrhol a implementoval som funkčnú aplikáciu s názvom REPA (RepeatExplorer Post Analysis), ktorú možno považovať za nadstavbu programu RepeatExplorer. Aplikácia využíva výstupné dáta programu RepeatExplorer k vytvoreniu super zhlukov, z ktorých sa identifikujú LTR transpozóny. Následne aplikácia overuje štruktúru predtým identifikovaných transpozónov podľa modelov, ktoré boli vytvorené na základe znalostí ich štruktúry.

Výsledné dáta sú prezentované v užívateľsky prívetivej forme, ktorej hlavnú časť tvorí interaktívna zobrazovacia plocha. V nej sa nachádzajú spomínané super zhluky.

Testovanie ukázalo, že táto aplikácia ušetrí čas pri rekonštrukcii LTR transpozónov. Žiadne iné prakticky využiteľné rozšírenie pre RepeatExplorer v súčasnej dobe nie je implementované, takže táto aplikácia má dobrý potenciál na uplatnenie v praxi. Najbližšie bude využívaná výskumníkmi Biofyzikálneho ústavu Akademie věd České republiky.

Testovanie ukázalo niekoľko rozšírení, ktoré by mohli mať prínos pre užívateľov Napríklad algoritmus vytvárania super zhlukov by mohol brať na úvahu viac faktorov.

Literatura

[1] Evoluční genomika.
URL http://www.evolucnigenomika.cz/Skripta/Evolucni%20genomika%20skripta%202008.pdf

[2] LTR FINDER USER MANUAL.
URL http://tlife.fudan.edu.cn/ltr_finder/help/help.pdf

[3] New and continuing developments at PROSITE.
URL http://nar.oxfordjournals.org/content/early/2012/11/16/nar.gks1067.full.pdf

[4] RepeatMasker Open-4.0.
URL http://www.repeatmasker.org

[5] Bruce Alberts, A. J., Dennis Bray: *Základy buněčné biologie*. Espero Publishing, s.r.o, 1998, ISBN 80-902906-0-4.

[6] Han, J. S.: Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions.

[7] Leslie A. Pray, P.: Transposons: The Jumping Genes.

[8] Matthias Zytnicki, H. Q., Eduard Akhunov: Tedna: a transposable element de novo assembler.

[9] Petr Novák, P. N.; Macas, J.: MGerthaodpohlog-yb araticsleed clustering and characterization of repetitive sequences in next-generation sequencing data.

[10] Phillip SanMiguel, Y.-K. J. N. M. D. Z. A. M.-B. P. S. S. K. J. E. M. L. Z. A. J. L. B., Alexander Tikhonov: Nested Retrotransposons in the Intergenic Regions of the Maize Genome.

[11] Wessler, S. R.: Transposable elements and the evolution of eukaryotic genomes.

Přílohy

Seznam příloh

Příloha A

Obsah CD

Priložené cd obsahuje nasledovné adresáre:

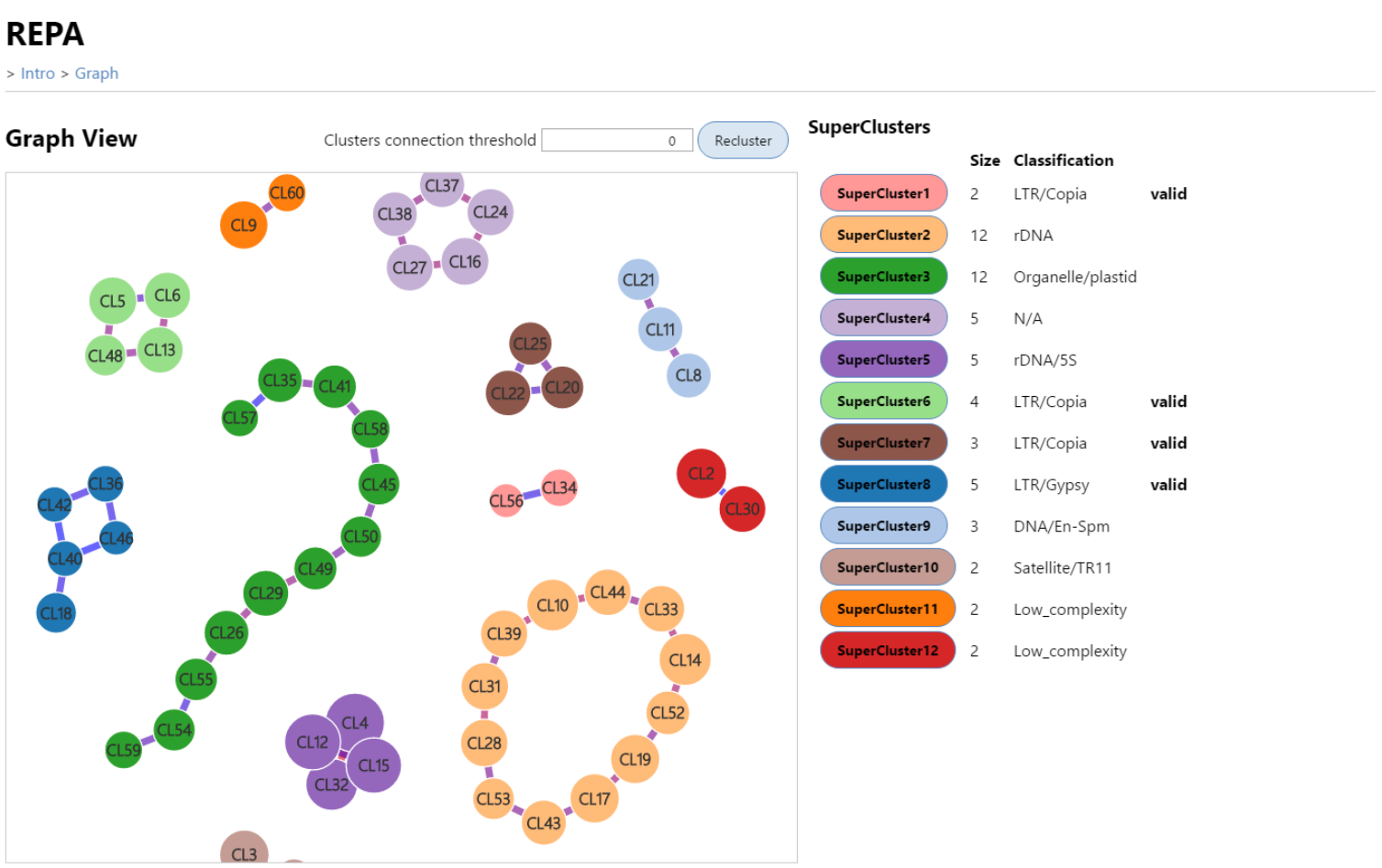
src zdrojové kódy aplikácie

bin spustiteľné súbory aplikácie

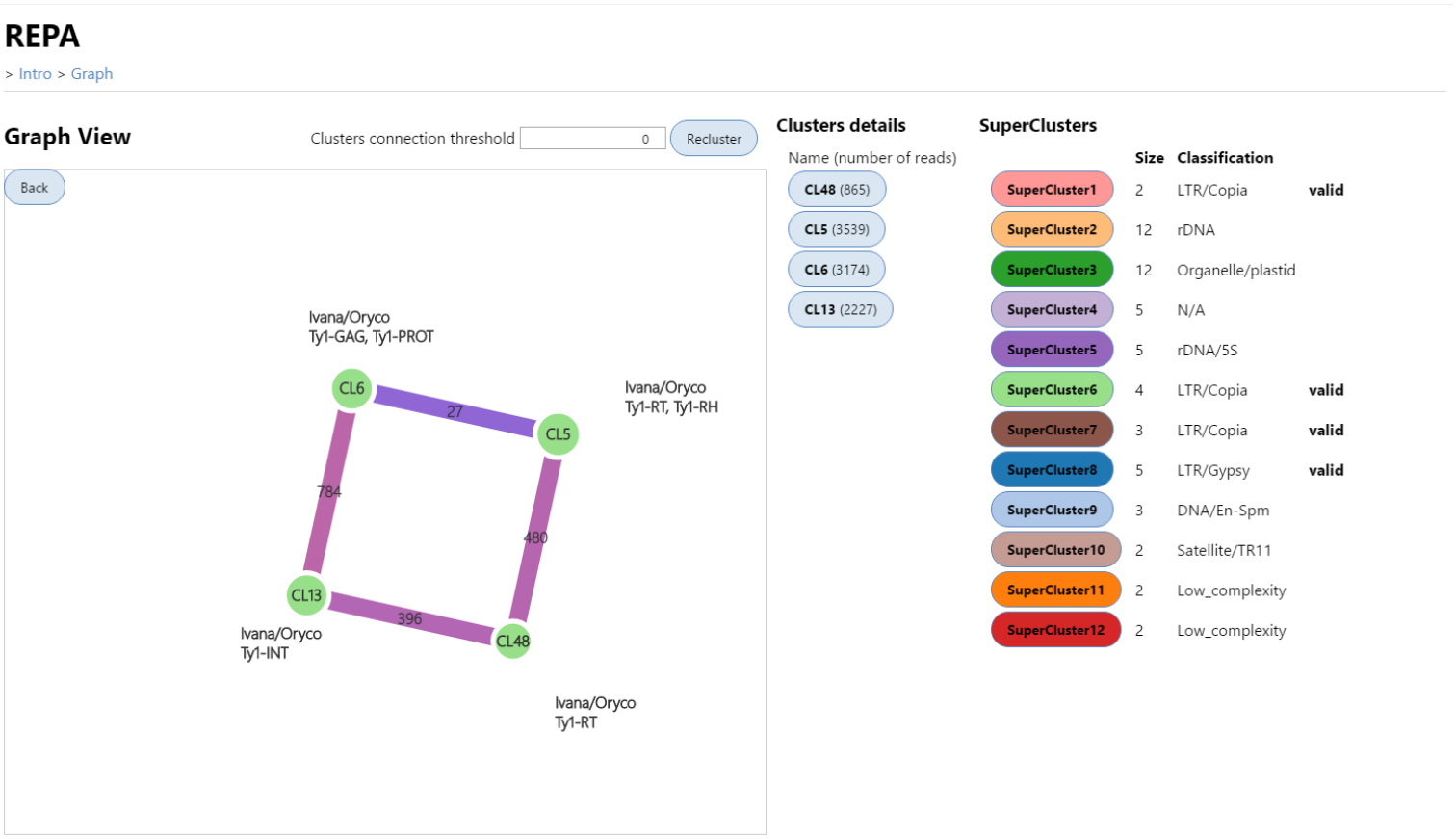
tex podklady tejto práce vo formáte LaTeX

Příloha B

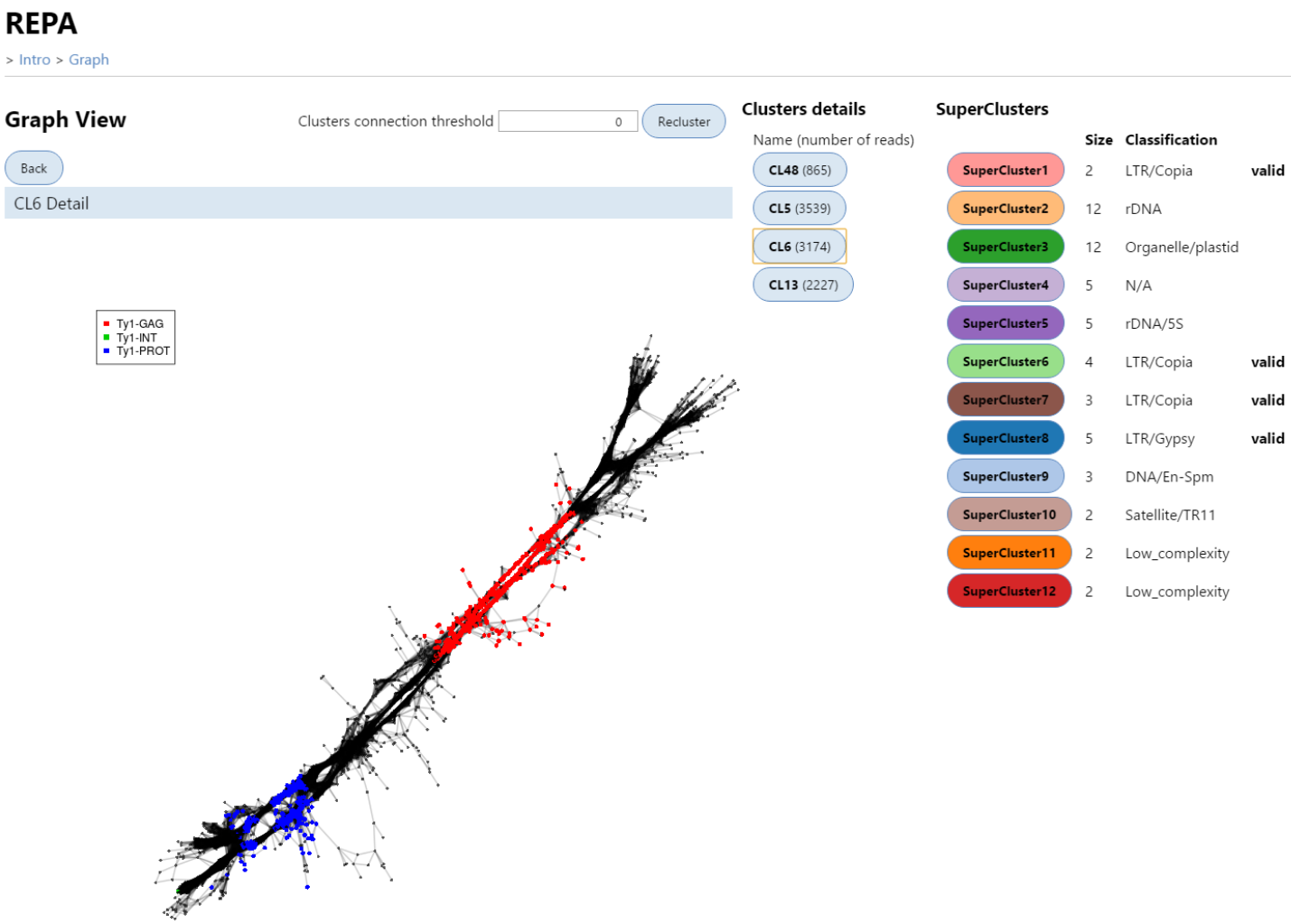
Obrazovky implementovanej aplikácie



Obrázek B.1: Zobrazenie všetkých super zhlukov



Obrázek B.2: Výber jedného super zhluku



Obrázek B.3: Detail zhluku

Summary

Summary of RepeatMasker hits

Class.Family	hits	hits[%]
LTR/Copia	354	40.925

Total number of similarity hits for each lineage

Lineage	Domain	Hits
Ivana/Oryco	Ty1-RT	368
AleI/Retrofit	Ty1-RT	23

Summary of TE domain hits from blastx

Domain	ID	Type	Lineage	Hits	MeanScore
Ty1-RT	Copia-11_SB-I	Ty1/copia	Ivana/Oryco	116	43.1025862068966
Ty1-RT	Copia-40_SB-I	Ty1/copia	Ivana/Oryco	80	45.57125
Ty1-RT	SC-8_I	Ty1/copia	Ivana/Oryco	44	52.3181818181818
Ty1-RT	Wicker_Kasia_102J11-1	Ty1/copia	Ivana/Oryco	41	52.8390243902439
Ty1-RT	SHACOP8_I_MT	Ty1/copia	Ivana/Oryco	40	45.3825
Ty1-RT	Llorens_Poco_AC210386.1	Ty1/copia	Ivana/Oryco	13	43.7538461538462
Ty1-RT	Copia30-PTR_I	Ty1/copia	AleI/Retrofit	12	51.275
Ty1-RT	Copia27-ZM_I	Ty1/copia	AleI/Retrofit	10	49.68
Ty1-RT	Wicker_HORPIA_AY661558-1	Ty1/copia	Ivana/Oryco	6	45.3166666666667
Ty1-RT	Copia-51_SB-I	Ty1/copia	Ivana/Oryco	6	49.75
Ty1-RT	Llorens_Vitico1-1_AM465428.1	Ty1/copia	Ivana/Oryco	6	39.9833333333333
Ty1-RT	Copia-46_SB-I	Ty1/copia	Ivana/Oryco	4	36.1
Ty1-RT	Copia-33_SB-I	Ty1/copia	Ivana/Oryco	4	41.3
Ty1-RT	Copia-8_SB-I	Ty1/copia	AleII	3	38.2333333333333
Ty1-RT	Copia41-PTR_I	Ty1/copia	Ivana/Oryco	2	36.6
Ty1-RT	Wicker_Beyla_A_Os_cons	Ty1/copia	Bianca	2	51.2
Ty1-RT	ZMCOPIA2_I	Ty1/copia	TAR	1	45.1
Ty1-RT	Copia4-ZM_I	Ty1/copia	Ivana/Oryco	1	33.5

Obrázek B.4: Detail zhluku