GENOMIC FEATURES VISUALIZATION OF TRANSPOSABLE ELEMENTS

Barbora Nétková

Bachelor Degree Programme (3), FIT BUT

E-mail: xnetko00@stud.fit.vutbr.cz

Supervised by: Ivan Vogel

E-mail: ivogel@fit.vutbr.cz

Abstract: Genomic feature format (GFF) is nowadays a de-facto standard format for genome description. Although there are several tools for GFF visualization, an open source application with advanced visualization features is missing. This work presents a design of such a tool. The application has a graphical user interface to simplify the user's work and combines the advantages of existing commercial products with free access. The application provides simple way to import user's biological data from a GFF file, creates hierarchical tree of individual elements including a detailed internal structure visualization in a subwindow.

Keywords: visualisation, transposons, gff

1. ÚVOD

Pomocí vizualizace genomických rysů transpozonů mohou biologové snadno nahlížet do jejich vnitřní struktury a usnadňuje to tak práci při jejich výzkumu. Pro tento účel již existuje několik nástrojů, které vizualizaci umožňují, ale buď nejsou pro uživatele volně dostupné - např. Geneious (http://www.geneious.com/) nebo práci s nimi značně ztěžuje nutnost pracovat pomocí příkazového řádku a instalace dodatečných grafických knihoven nutných k běhu aplikace - např. GenomeTools (http://genometools.org/annotationsketch.html).

Za tímto účelem je tvořena aplikace, která má spojit výhody komerčních produktů jako je například intuitivní ovládání, grafické uživatelské rozhraní nebo různé možnosti vizualizace s volnou dostupností.

2. MOLEKULÁRNÍ BIOLOGIE

K plnému pochopení práce je třeba nejprve se krátce seznámit s problematikou molekulární biologie, transpozony a způsobem uložení biologických informací.

Základem buněčné biologie je DNA, neboli deoxyribonukleová kyselina, složená z několika bází (adenin, thimin, cytosin a guanin), sacharidů a fosfátů. Struktura DNA je stabilní a tvoří ji dvě spojená vlákna stočená do dvojšroubovice. DNA řídí vznik proteinů, určuje aminokyselinové sekvence.

Většina typů buněk předává genetickou informaci nejprve z DNA do RNA a poté z RNA do proteinu. Tento obecný mechanismus je nazýván ústřední dogma molekulární biologie. RNA je ribonukleová kyselina, oproti DNA je většinou jednovláknová a místo thiminu obsahuje uracil [1].

2.1. TRANSPOZONY

Transpozony jsou úseky DNA schopné přesunu z jednoho místa na jiné - jsou schopné transpozice. Můžeme pozorovat dva typy transpozonů. DNA transpozony využívají k přesunu především mechanismu cut & paste (vyštěpí se ze svého místa a vloží se na jiné). Retrotranspozony kódující re-

verzní transkriptázu nejdříve vytvoří z RNA transkriptu DNA kopie, které se pak mohou vložit do různých míst genomu. Využívají mechanismus copy & paste [2].

Transpozony mohou způsobovat mutace, choroby nebo nefunkční geny. Na druhou stranu některým oraganismům mohou být i užitečné. Tím, že mění genetickou informaci, mohou přispět k lepší adaptaci organismu [3].

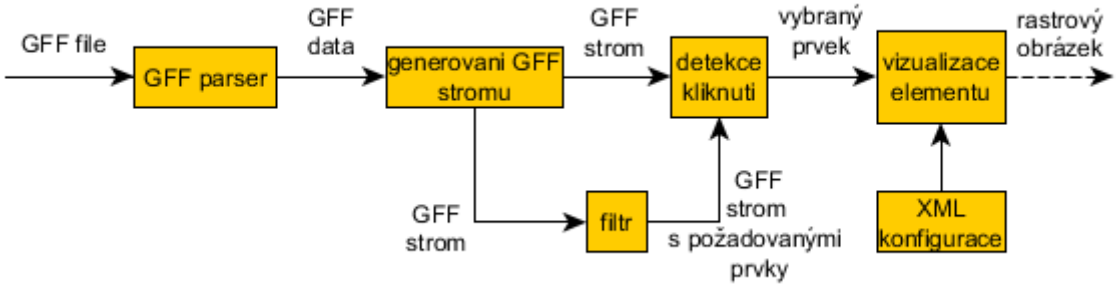
2.2. UKLÁDÁNÍ BIOLOGICKÝCH INFORMACÍ

Aby bylo možné biologická data zpracovávat bylo vytvořeno několik druhů souborů určených pro biologická data. Aplikace pro vizualizaci transpozonů pracují většinou se soubory typu GFF. Jejich struktura je pevně definovaná, tvoří ji 9 sloupců oddělených tabulátory a každý ze sloupců má daný význam [4]. Navíc informace v GFF souboru navazují na hierarchickou strukturu pojmů v projektu SequenceOntology, který definuje biologické sekvence a jejich závislosti [5].

3. SPECIFIKACE VÝSLEDNÉ APLIKACE

Návrh aplikace vychází z požadavku na intuitivní jednoduché ovládání a přehlednost. V komerčních produktech se také objevuje možnost různě přibližovat a oddalovat náhled na části transpozonu. Tato možnost byla inspirací pro vytvoření GFF stromu a následnou vizualizaci zvolené části.

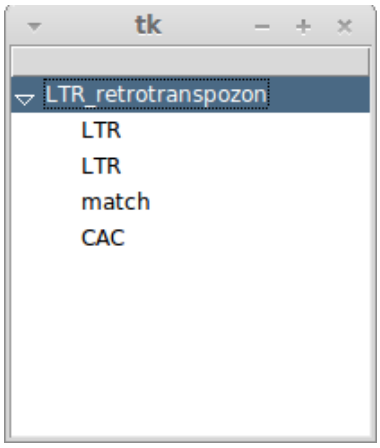
Aplikace je složena z několika základních částí - parser GFF souboru, hierarchický strom tvořený prvky ze souboru GFF a samotná vizualizace zvolené části GFF stromu. Dále bude program obsahovat filtr umožňující zobrazit jen vybrané prvky v GFF stromu, bude podporovat uživatelská nastavení (např. barvy vizualizace) a umožní uživateli vizualizace ukládat. Blokové schéma zobrazující strukturu je znázorněno na obrázku 1.



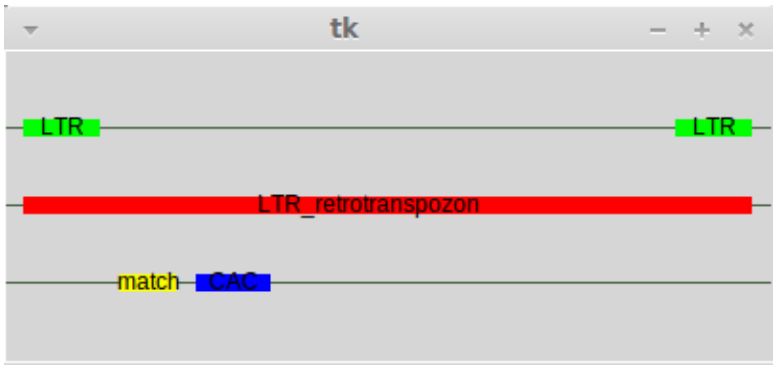
Obrázek 1: Blokové schéma aplikace.

Poté, co uživatel pomocí grafického uživatelského rozhraní vybere požadovaný soubor k vizualizaci, proběhne jeho rozparsovaní. Následně je z rozparsovaných částí souboru GFF vygenerován hierarchický strom, který se zobrazí v první části obrazovky aplikace (viz obrázek 2). Uživatel potom může ze stromu myší vybrat požadovaný prvek nebo jeho část, kterou chce zobrazit. Vizualice zvoleného prvku se pak objeví v druhé části obrazovky (viz obrázek 3).

K implementaci je použit jazyk Python a jeho rozšíření Biopython. Ke grafickému zobrazení hierarchického stromu prvků byla použita knihovna TkInter, konkrétně její podčást TreeView. Samotnou vizualizaci prvků zajišťuje stejná knihovna.



Obrázek 2: Náhled první části aplikace. Vygenerovaný GFF strom s vybraným prvkem LTR_retrotranspozon.



Obrázek 3: Přibližný náhled podoby druhé části aplikace. Vizualizace zvoleného prvku a jeho vnitřních částí.

4. ZÁVĚR

Navržená aplikace nabízí uživateli co nejjednodušší volně dostupný nástroj na vizualizaci transpozonů a jejich vnitřních částí. Zaměřuje se na snadné intuitivní ovládání a možnost zobrazit jednotlivé podčásti transpozonu a tím detailněji zkoumat strukturu. Aplikace bude po plné implementaci umožňovat nastavení barev vizualizace a uživateli bude také podstatně ulehčovat práci filtr na zobrazování jen vybraných elementů ve vygenerovaném GFF stromu. Úspěšně byla dokončena fáze návrhu aplikace na základě specifikovaných požadavků a v současnosti je ve vývoji základní verze aplikace s funkcionalitou popsanou v sekci 3. Za zvážení by jistě stála rozšíření a vylepšení aplikace v podobě dalších modulů. Například možnost importu FASTA souboru s informacemi o pořadí jednotlivých bází a jejich vizualizace při dostatečném přiblížení, nebo zobrazení tabulky informací o zvoleném prvku z GFF souboru.

REFERENCE

[1] ALBERTS, Bruce. Základy buněčné biologie: Úvod do molekulární biologie buňky. 2. vyd. Překlad Arnošt Kotyk, Bohumil Bouzek, Pavel Hozák. Ústí nad Labem: Espero, 1998, 1 sv. (různé stránkování). ISBN 80-902-9062-0

[2] FISCHER, Lukáš. Molekulární genetika rostlin: Transpozóny. Univerzita Karlova v Praze Přírodovědecká fakulta: Katedra experimentální biologie rostlin [online]. 2013 [cit. 2014-02-27]. Dostupné z: http://kfrserver.natur.cuni.cz/lide/lukasf/molgen/transpozony.ppt

[3] KEJNOVSKÝ, Eduard. Zkopíruj a ulož. Vesmír [online]. 2000, roč. 2000, č. 79 [cit. 2014-01-26]. Dostupné z: http://www.vesmir.cz/clanek/zkopiruj-a-uloz

[4] Generic Feature Format Version 3 (GFF3). SEQUENCE ONTOLOGY. Generic Feature Format Version 3 (GFF3) [online]. 2009 [cit. 2014-02-27]. Dostupné z: http://www.sequenceontology.org/gff3.shtml

[5] SEQUENCE ONTOLOGY. The Sequence Ontology [online]. 2009 [cit. 2014-02-27]. Dostupné z: http://www.sequenceontology.org/