TRIPLEX: AN R/BIOCONDUCTOR PACKAGE FOR IDENTIFICATION AND VISUALIZATION OF POTENTIAL INTRAMOLECULAR TRIPLEX PATTERNS IN DNA SEQUENCES

Jiří Hon
Bachelor Degree Programme (3), FIT BUT
E-mail: xhonji01@stud.fit.vutbr.cz

Supervised by: Tomáš Martínek
E-mail: martinto@fit.vutbr.cz

Abstract: Triplex-forming DNA sequences have been implicated as important players in several key processes, such as transcriptional regulation, DNA recombination and mutagenesis. Unfortunately there is no compact tool suitable for their identification and visualization. Existing implementations suffer from being independently developed and unintegrated in larger cooperating software ecosystem. This paper deals with transformation of such standalone tools into an *R/Bioconductor* package to simplify subsequent filtration and analysis of the results for molecular biologists and to enable cooperation with existing *Bioconductor* packages.

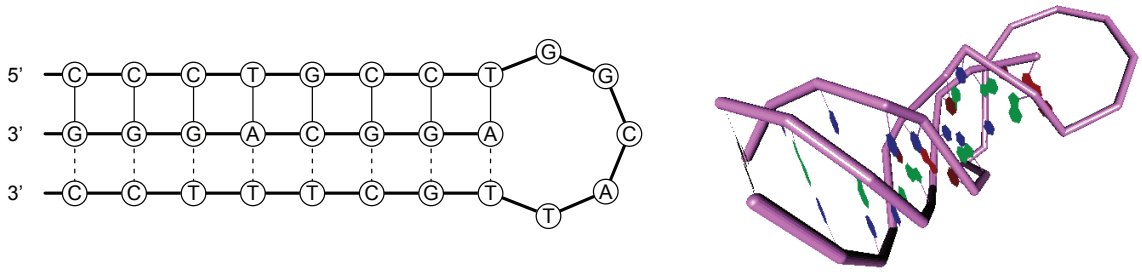
Keywords: package, R, Bioconductor, DNA, triplex

1 ÚVOD

Analýza DNA a její anotace jsou důležitými kroky k porozumění molekulárním základům života. Pozornost současných vědeckých studií se posunula od analýzy genů ke zkoumání mezigenové DNA a strukturních vlastností chromozomů a buněčného jádra. Pozoruhodné jsou zejména úseky DNA, které se v důsledku působení Watson-Crickových a Hogsteenových vazeb skládají do tvaru trojšroubovice – triplexu (viz obrázek 1). Tyto úseky mají výrazný vliv na regulaci transkripce, rekombinaci DNA nebo mutagenesi a prokazatelně se podílejí na vzniku geneticky podmíněných chorob [5].

Současné nástroje vhodné k identifikaci a vizualizaci sekvencí schopných tvořit triplexy jsou naneštěstí dostupné pouze v podobě samostatných nespolupracujících programů či knihoven mimo rozsáhlejší aplikační ekosystém. Cílem této práce proto bylo transformovat existující algoritmus pro vyhledávání triplexů [1], [2] spolu se zpětným zarovnáním nalezené sekvence [4] a její vizualizace [3] do podoby softwarového balíčku využitelného v prostředí jazyka *R*. Současný stav je nevyhovující, neboť klade příliš vysoké nároky na koncové uživatele, na jejich zkušenosti s kompilací programů a orientací v příkazové řádce. Formát vstupů i výstupů navíc komplikuje následné statistické analýzy i operace s výsledky.

Naproti tomu prostředí jazyka *R* ve spojení se softwarovými a anotačními balíčky dostupnými v *Bioconductoru* představuje pro molekulární biology mocný prostředek pro další analýzu výsledků vyhledávacího algoritmu, jejich vizualizaci a dávání do souvislostí s anotacemi z dalších vědeckých studií. Výraznou výhodou prostředí *R* je navíc propracovaná dokumentace i vysoká kvalita softwarových balíčků, které před zveřejněním prochází nezávislou revizí.



Obrázek 1: Ukázka 2D a 3D vizualizace triplexu.

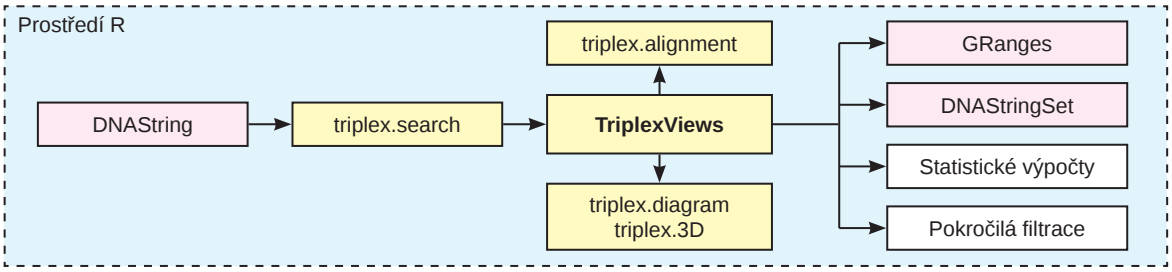
2 NÁVRH A ŘEŠENÍ TRANSFORMACE

Interaktivní prostředí interpretovaného jazyka *R* je specifické nejen svým způsobem použití především ve statistických analýzách, ale také svými pravidly pro vytváření rozšiřujících softwarových balíčků.

Prostředí *R* umožňuje nahrávání sdílených knihoven psaných v jazyce *C* nebo *Fortran*. Tohoto jsem s výhodou využil při transformaci zdrojových kódů výchozích algoritmů [1], [4] do podoby, ve které se dají spouštět přímo z jazyka *R*. Abych zajistil optimální funkčnost z hlediska spotřeby operační paměti, využil jsem rozšířeného rozhraní pro volání funkcí ze sdílených knihoven, které umožňuje předat objekty jazyka *R* přímo do jazyka *C* a tam s nimi pracovat jako se speciálním typem struktury. Díky transformaci do podoby sdílené knihovny pro jazyk *R* se uvedené algoritmy staly automaticky přenositelné na všechny platformy podporované tímto jazykem, mezi něž patří Windows, MacOS i Linux.

V důsledku nezávislého vývoje několika programů zabývajících se identifikací a vizualizací triplexů došlo k jejich vzájemné nekompatibilitě z hlediska vstupů a výstupů.

Jako první krok k řešení tohoto problému jsem implementoval třídu `TriplexViews` v jazyce *R*, která zajišťuje základní operace nad výsledky vyhledávacího algoritmu. Zároveň slouží jako vstup pro zarovnání a vizualizační funkce do 2D a 3D zobrazení (viz obrázek 2). Díky této třídě je možné provádět dodatečné filtrování a řazení podle různých vlastností nalezených triplexů – podle skóre, délky, počtu insercí, pozici v sekvenci, apod. Třída `TriplexViews` je navržena tak, aby umožňovala snadný export do formátů *GFF3* a *FASTA*. V druhém kroku jsem sjednotil výstup algoritmu pro zpětné zarovnání nalezené sekvence se vstupem vizualizačních funkcí.



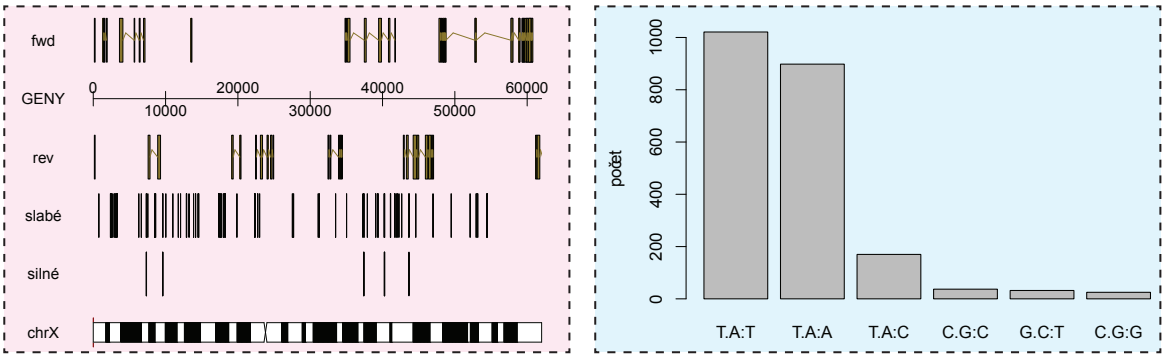
Obrázek 2: Schéma výsledného balíčku. Existující třídy z *Bioconductoru* jsou zvýrazněny červeně.

3 DEMONSTRAČNÍ PŘÍKLAD

Levá část obrázku 3 dává do souvislosti pozice genů s výskytem potenciálního triplexu na chromozomu X organismu *Caenorhabditis elegans*. Pomocí balíku `GenomeGraph` jsou vykresleny stopy

genů na dopředném i reverzním vlákně. Stopa pro triplexy s dosaženým skóre v rozsahu 17–20 bodů je označena jako *slabé*. Triplexy se skóre větším než 20 bodů jsou znázorněny ve stopě *silné*.

Pravá část obrázku 3 zobrazuje výstup analýzy četnosti výskytu tripletů tvořících nalezené triplexy na tomtéž chromozomu X organismu *C. elegans*.



Obrázek 3: *Vlevo:* souvislost pozice genů s výskytem triplexů. *Vpravo:* šest nejčastěji se vyskytujících tripletů tvořících triplexy.

4 ZÁVĚR

Triplexy hrají důležitou roli v mechanismech regulace transkripce, rekombinace DNA a mutagenese, proto je nezbytné vytvářet takové aplikace, které je dokáží nejen identifikovat a vizualizovat, ale také dávat do souvislostí s výsledky dalších vědeckých studií.

Transformace dříve samostatných programů [1], [2], [3], [4] do podoby kompaktního balíčku pro prostředí *R/Bioconductor* naplňuje tento cíl a přináší důležitou přidanou hodnotu v podobě začlenění do existujícího softwarového ekosystému *Bioconductoru*. Molekulární biologové tak mohou podstatně jednodušším způsobem provádět další filtrace a analýzy výsledků včetně budování nových anotací a dávání do souvislostí s existujícími.

Balíček byl ve spolupráci s autory původního algoritmu [1] doplněn o podrobnou dokumentaci s uživatelskou příručkou a odeslán k prvotní revizi tvůrcům *Bioconductoru*.

REFERENCE

[1] LEXA, Matej, Tomáš MARTÍNEK, Ivana BURGETOVÁ, Daniel KOPEČEK a Marie BRÁZDOVÁ. A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics.* Oxford (Velká Británie), 2011, roč. 27, č. 18, s. 2510-2517. ISSN 1367-4803.

[2] KOPEČEK, Daniel. *Rozšíření a optimalizace programu pro vyhledávání triplexů v DNA sekvencích.* 2011. Bakalářská práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Matej Lexa.

[3] RAJDL, Kamil. *Funkce pro manipulaci a vizualizaci molekulárních dat v prostředí R.* Brno, 2012. Bakalářská práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Matej Lexa.

[4] ZRŮNA, Michal. *Vyhledávání triplexů v DNA sekvencích.* Brno, 2012. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Tomáš Martínek.

[5] RAJESWARI, Moganty Raja. DNA triplex structures in neurodegenerative disorder, Friedreich's ataxia. *J Biosci.* 2012, roč. 37, č. 3, s. 519-32.