Vizualizace sekundární struktury RNA s využitím existujících struktur

Richard Eliáš, vedúci práce **David Hoksza**

Matematicko-fyzikální fakulta

richard.elias@matfyz.cz

Motivácia a ciele práce

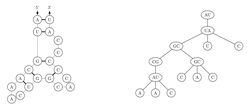
Molekula RNA sa stáva predmetom mnohých štúdií, vďaka čomu rastie dopyt po nástrojoch pomáhajúcich pri jej analýze. Vlastnosti molekuly sú síce ovplyvnené primárou štruktúrou (poradím nukleotidov v reťazci), no viac závisia na ich priestorovom usporiadaní (terciárna štruktúra). My sa v práci zaoberáme trochu zjednodušeným modelom - sekundárnou štruktúrou. Tú reprezentuje zoznam nukleotidov spojených väzbou. Tieto nukleotidy musia byť blízko aj v priestore a tak nám sekundárna štruktúra relatívne dobre aproximuje terciárnu, pre ktorú neexistujú spoľahlivé metódy zisťovania štruktúry už ani pre relatívne malé molekuly.

Prvým krokom pri analýze RNA molekuly je často rozbor obrázka jej sekundárnej štruktúry. Medzi základné kritéria ktoré musia obrázky molekúl spĺňať patrí rovinnosť nakreslenia, kreslenie loopov na kružnice a stemy na priamky. Pri porovnávaní štruktúr sa využíva taktiež kreslenie častí majúcich podobnú funkciu a tvar na rovnaké miesta v obrázkoch, čo pomáha lepšej orientácií v molekule a pomáha pri hľadaní konzervovaných častí. V súčastných nástrojoch (mFold, RNAViz, RNAView, $\cdots$) sa toto posledné kritérium nedodržuje.

Cieľom našej práce je umožniť vizualizovanie molekúl podľa predstáv biológov. Tie bude reprezentovať obrázok vzorovej molekuly podľa ktorej sa budeme snažiť nakresliť cieľovú.

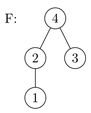
Stromová reprezentácia RNA a použitie *tree-edit-distance* algoritmu

RNA sekundárnu štruktúru reprezentujeme ako usporiadaný zakorenený strom. Vďaka tomu môžeme využiť stromové algoritmy, ako napríklad *tree-edit-distance* algoritmus. Ten rekurzívnym vzorcom spočíta vzdialenosť medzi stromami a následne vie aj transformovať jeden strom na iný. Na príklade ukážeme postup výpočtu a transformáciu medzi stromami.



$$\delta(\emptyset,\emptyset) = 0$$
$$\delta(F,\emptyset) = \delta(F - r_F,\emptyset) + c_{del}(r_F)$$
$$\delta(\emptyset,G) = \delta(\emptyset,G - r_G) + c_{ins}(r_G)$$
$$\delta(F,G) = \begin{cases} \delta(F - r_F, G) + c_{del}(r_F) \\ \delta(F, G - r_G) + c_{ins}(r_G) \\ \delta(F - F_{r_F}, G - G_{r_G}) \\ \quad +\delta(F_{r_F} - r_F, G_{r_G} - r_G) \\ \quad +c_{upd}(r_F, r_G) \end{cases}$$

F:



G:

TreeDistance :

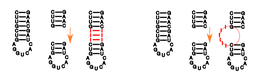
	A	B	C
1	0	0	2
2	1	1	1
3	0	0	2
4	3	3	1

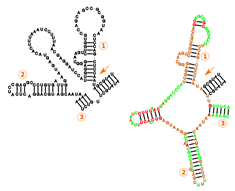
ForestDistance :

	A	B	C
1	0	1	2
2	1	2	1
3	2	1	2
4	3	2	1

Kreslenie chýbajúcich častí v molekule

Po transformácií šablónovej molekuly na cieľovú, získavame čiastočnú vizualizáciu cieľovej RNA a jej zvyšok potrebujeme dopočítať. Po operáciách delete nám v obrázku ostávajú prázdne diery, naopak po insertoch potrebujeme pre dané vrcholy urobiť v obrázku miesto.





Mazanie vrcholov stromu je inverzná operácia ku vkladaniu a tak si ho nebudeme ukazovať. Vkladanie báz do loopu (okrem multibranch) je jednoduché, vytvoríme si iba novú kružnicu na ktorú všetky bázy uložíme. Vkladanie báz do stemu potrebuje pre nich najprv urobiť miesto a tak celú štruktúru posunie smerom od rodičovského vrcholu v strome.

Pre multibranch loopy je to trochu zložitejšie, chceme sa totiž vyhnúť prípadom, kedy ju musíme celú prekresliť, keďže pri tom vznikajú veľké problémy s prekryvmi. Prekresleniu štruktúry sa nevyhneme, ak sa jedná o vkladanie veľkého počtu báz, alebo vkladáme celú novú vetvu RNA.

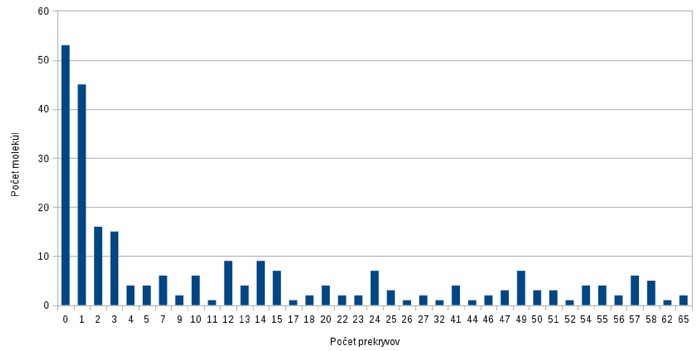
Výsledky experimentov

Program sme testovali na reálnych obrázkoch 16 molekúl malej podjednotky 18S ribozomálnej RNA z CRW databázy. Z testov sme získali 256 výsledných vizualizácií (každý s každým).

Celkové štatistiky prekryvov ukazuje graf. Pre prípady molekúl, ktoré nepotrebovali prekresliť multibranch loop sa štatistika zmenila - do 10 prekryvov bolo 150 nakreslení, nad 10 iba 29.

Počty prekryvov v závislosti na vzdialenosti TED nieje až taká zjavná. Ak vezmeme najbližšie štruktúry, je v priemere asi 5 prekryvov a so zväčšujúcou sa vzdialenosťou rástli výkyvy hlavne kvôli tomu, že niektoré štruktúry boli dobre konzervované a na druhej strane bolo pár extrémov, ktoré sa vizualizovali veľmi ťažko.

V budúcnosti by bolo vhodné upraviť kresliace algoritmy, implementovať otáčanie vetiev RNA stromov v prípade indikovania prekryvov alebo pridať interaktívny nástroj na úpravu obrázkov, čím by užívateľ prekryvy mohol ručne odstrániť.



Nástroj TRAVeLer

V rámci práce sme implementovali nástroj TRAVeLer (Template RnA VisuaLization) schopný vizualizovať aj veľke rRNA molekuly podľa vstupného vzorového obrázka. Implementujeme v ňom *tree-edit-distance* algoritmus, ktorý nám transformuje jednu molekulu na druhú. Následným použitím nášho dokresľovacieho algoritmu vznikne výsledná vizualizácia.

Na príklade ilustrujeme vstupy a výstupy programu. Vstupom sú dve RNA sekundárne štruktúry, cieľová a vzorová, ktorá potrebuje aj obrázok. Výstupom je obrázok cieľovej molekuly, alebo mapovanie medzi štruktúrami (ak si chceme vygenerovať viac typov obrázkov, nemusíme znovu počítať mapovanie).

```
# ./traveler --match-tree mouse.fasta --template-tree human.ps human.fasta \
      --all mouse_to_human
```

Aby bolo hľadanie spoločných a rozdielnych častí jednoduchšie, zaviedli sme farebné kódovanie nukleotidov v obrázku: **červená** - insert, **zelená** - edit, **modrá** - prekreslenie báz, **hnedá** - prekreslenie multibranch loopy.

Príklady vizualizácií: človek (K03432) a žaba (X04025), myš (L24489) a cikáda (U06478), žiabronôžka (X01723) a pásomnička (U27015)