

# 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

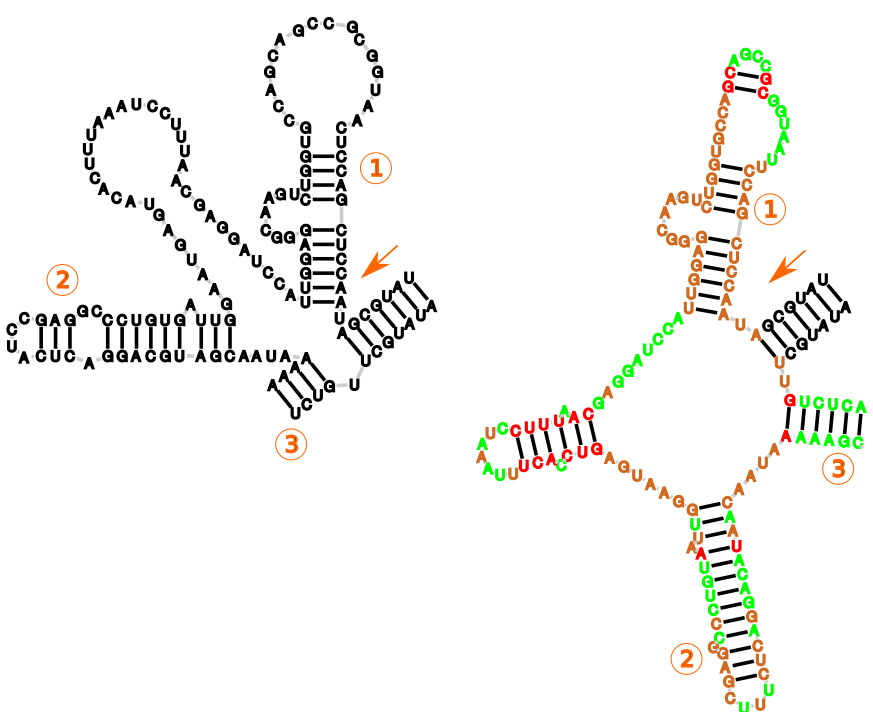
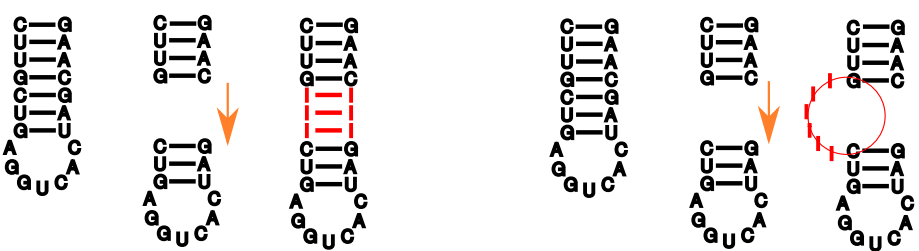
Molekula RNA sa stáva predmetom mnohých štúdií, vďaka čomu rastie dopyt po nástrojoch pomáhajúcich pri jej analýze. Vlastnosti molekuly sú síce ovplyvnené primárnu štruktúrou (poradím nukleotidov v reťazci), no viac závisia na ich priestorovom usporiadaní (terciárna štruktúra). My sa v práci zaoberáme 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 steny na priamkách. Pri porovnávaní štruktúr sa využíva taktiež kreslenie častí majúcich podobnú funkciu a tvar na rovnaké miesta v obrázkoch, čo pomáha lepšej orientácii v molekule a napomáha nájsť konzervované časti v molekulách. V súčasných nástrojoch (mFold, RNAviz, RNAView, ...) sa toto posledné kritérium nedodržuje, čo má za následok ťažké nachádzanie konzervovaných častí v molekulách.

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

## Kreslenie chýbajúcich častí v molekule

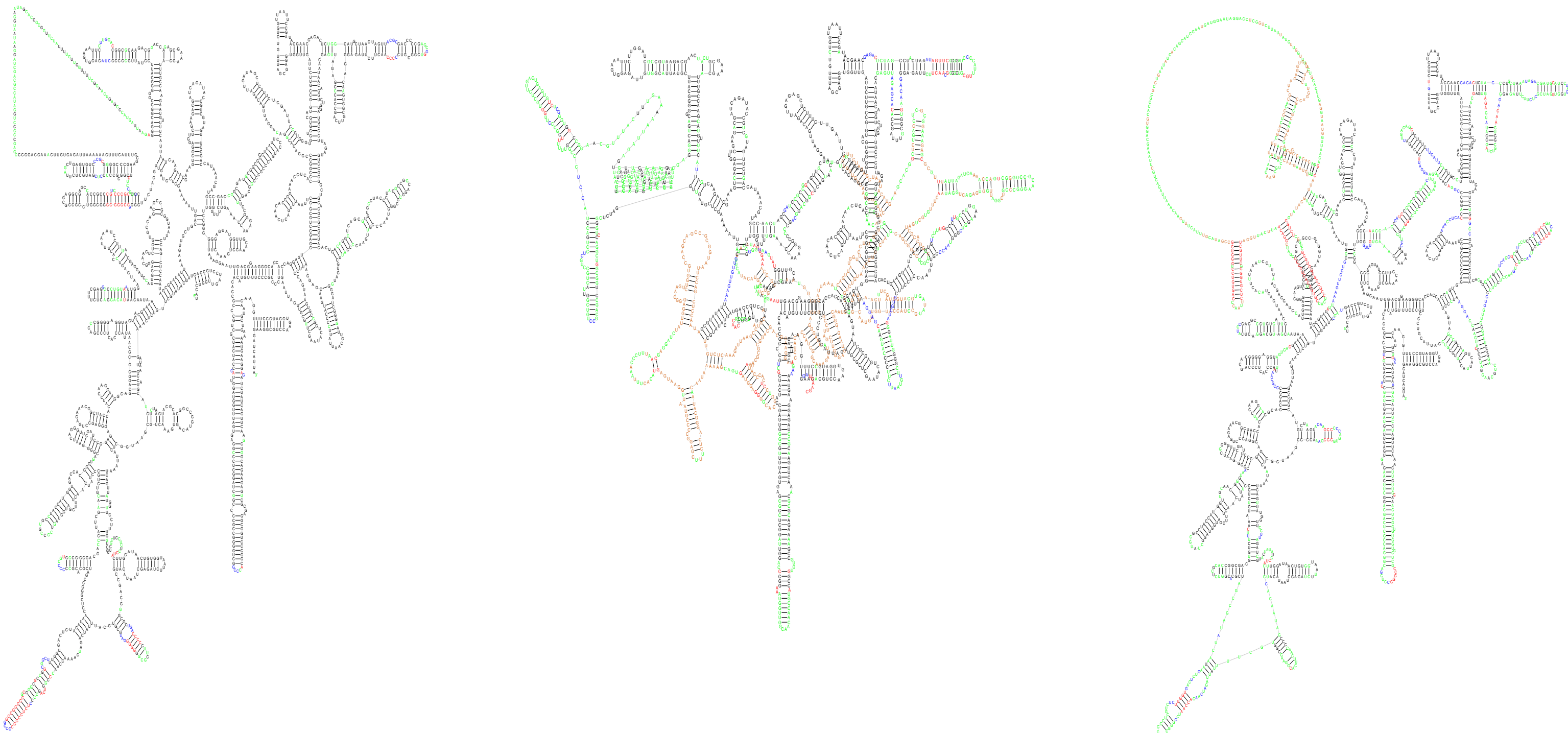
Po transformácii šablónovej molekuly na cieľovú, získavame čiastočnú vizualizáciu cieľovej RNA a jej zvyšok potrebujeme dopočítať. Po operáciách delete nám v obrázku ostávajú prázdne miesta, naopak po insertoch potrebujeme pre dané vrcholy urobiť v obrázku miesto.



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

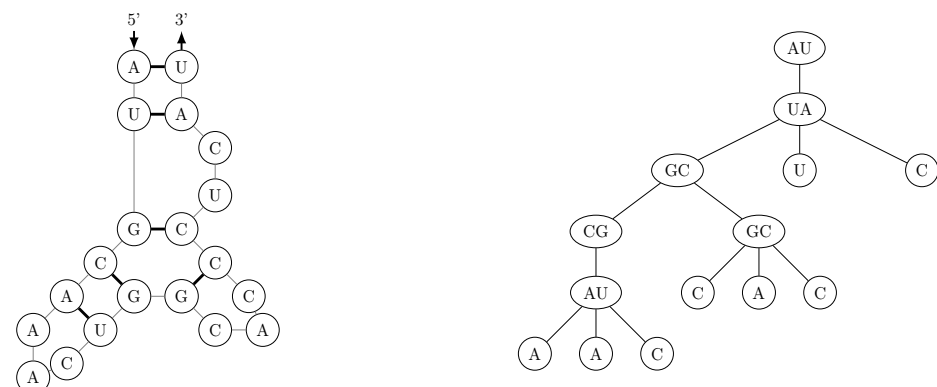
Pre multibranch loopy je to trochu zložitejšie, chceme sa totiž vyhnúť prípadom, kedy ju musíme celú prekresliť, keďže pri tom vznikajú veľké problémy s prekryvmi. Prekresleniu štruktúry sa nevyhneme, ak sa jedná o vkladanie veľkého počtu báz, alebo vkladáme celú novú vetvu RNA.

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

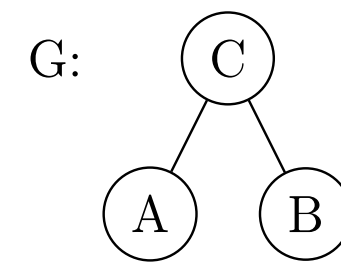
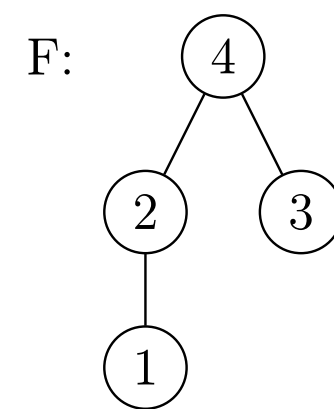


## Stromová reprezentácia RNA a použitie *tree-edit-distance* algoritmu

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.



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



<i>TreeDistance</i> :			
	0	0	2
	1	1	1
	0	0	2
	3	3	1

<i>ForestDistance</i> :			
	0	1	2
	1	2	1
	2	1	2
	3	2	1

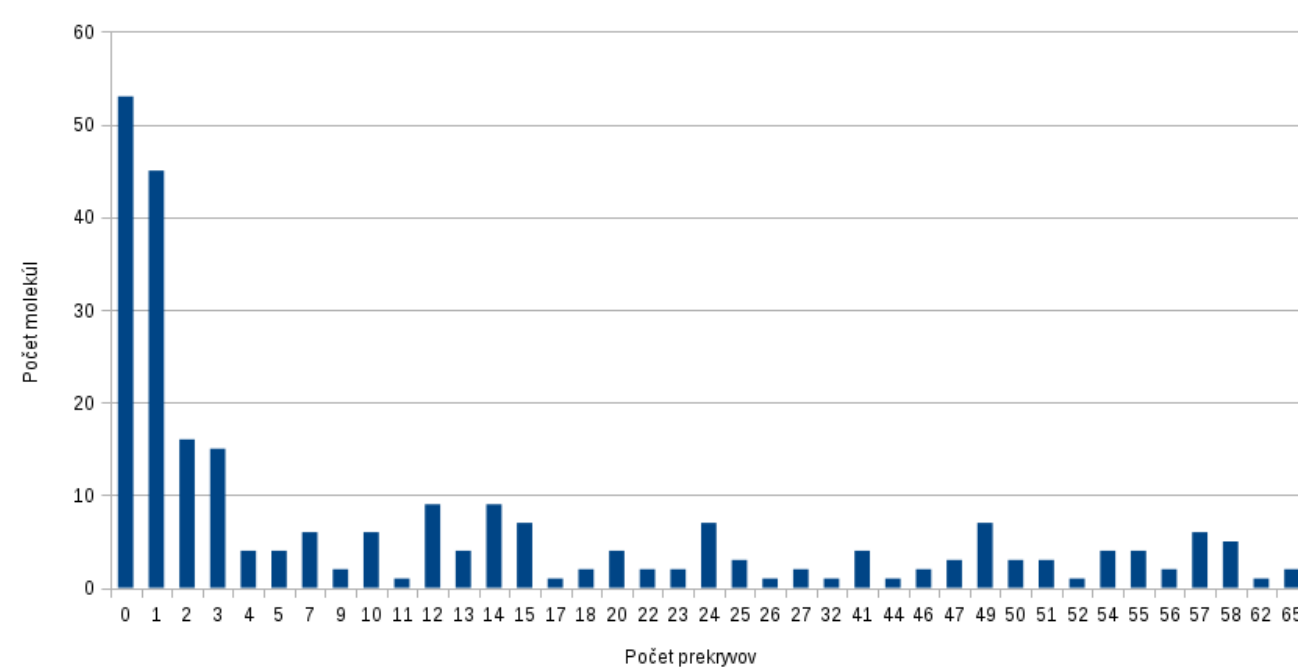
## Nástroj TRAVeLer

V rámci práce sme implementovali nástroj TRAVeLer (Template RnA VisuaLization) schopný vizualizovať aj veľké 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, vzorová potrebuje aj obrázok. Výstupom je obrázok cieľovej molekuly, alebo mapovanie medzi štruktúrami (ak si chceme vygenerovať viac typov obrázkov, nemusíme nanovo počítať mapovanie).

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

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

## 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.