



**MATEMATICKO-FYZIKÁLNÍ  
FAKULTA**  
Univerzita Karlova

## **BAKALÁŘSKÁ PRÁCE**

Michal Hercík

### **Webový plugin pro vizualizaci sady sekundárních struktur RNA**

Katedra softwarového inženýrství

Vedoucí bakalářské práce: doc. RNDr. David Hoksza, Ph.D.

Studijní program: Informatika

Studijní obor: Programování a vývoj software

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne .....

Podpis autora

TODO: podekovani

Název práce: Webový plugin pro vizualizaci sady sekundárních struktur RNA

Autor: Michal Hercík

Katedra: Katedra softwarového inženýrství

Vedoucí bakalářské práce: doc. RNDr. David Hoksza, Ph.D., Katedra softwarového inženýrství

Abstrakt: TODO

Klíčová slova: bioinformatika RNA sekundární struktura web plugin

Title: Web plugin for multiple RNA secondary structure visualization

Author: Michal Hercík

Department: Department of Software Engineering

Supervisor: doc. RNDr. David Hoksza, Ph.D., Department of Software Engineering

Abstract: TODO

Keywords: bioinformatics RNA secondary structure web plugin

# Obsah

<b>Úvod</b>	<b>3</b>
<b>1 Úvod do problematiky</b>	<b>4</b>
1.1 Seznámení s biologickými pojmy . . . . .	4
1.1.1 RNA . . . . .	4
1.1.2 Hairpin loop . . . . .	4
1.1.3 Bulges . . . . .	4
1.1.4 multibranch loop . . . . .	4
1.2 Vizualizace sekundárních RNA struktur . . . . .	4
1.3 Podobné projekty . . . . .	6
1.3.1 VARNA . . . . .	6
1.3.2 RNAStructViz . . . . .	7
1.3.3 Forna . . . . .	8
1.3.4 R-chie . . . . .	9
1.3.5 Shrnutí existujících nástrojů . . . . .	10
1.4 Kreslení grafů na základě šablony . . . . .	10
1.4.1 TRAVeLer . . . . .	11
1.4.2 R2DT . . . . .	11
<b>2 Metody vizualizace a porovnání</b>	<b>12</b>
2.1 Důsledek využívání vzorové struktury . . . . .	13
2.2 Překládání struktur . . . . .	13
2.2.1 Zarovnání . . . . .	14
2.2.2 Průhlednost struktur . . . . .	16
2.2.3 Rozmazání struktur . . . . .	16
2.3 Obarvení struktur . . . . .	16
2.4 Transformace na vzor . . . . .	17
2.5 Mapovací čáry . . . . .	18
2.6 Využití stromu . . . . .	20
2.7 Zvolené metody . . . . .	22
<b>3 Dokumentace</b>	<b>23</b>
3.1 Vstupní data . . . . .	23
3.1.1 Struktura dat . . . . .	23
3.1.2 Získání vstupních dat . . . . .	24
3.2 Objektový návrh . . . . .	25
3.3 Uživatelské seznámení s knihovnou . . . . .	28
3.3.1 Vytvoření objektu RnaVis a vykreslení struktur . . . . .	28
3.3.2 Zoom a posouvání canvasu . . . . .	29
3.3.3 Zarovnání struktur . . . . .	29
3.3.4 Mapovací čáry . . . . .	31
3.3.5 Transformace na vzorovou strukturu . . . . .	31
3.4 Použití canvasu . . . . .	32
3.4.1 Knihovna D3.js . . . . .	32
3.4.2 SVG . . . . .	32

3.4.3 Canvas . . . . .	32
3.5 Demonstrace metod . . . . .	33
<b>Závěr</b>	<b>34</b>
<b>Seznam použité literatury</b>	<b>35</b>
<b>Seznam obrázků</b>	<b>36</b>
<b>Seznam tabulek</b>	<b>38</b>
<b>Seznam použitých zkratek</b>	<b>39</b>
<b>A Přílohy</b>	<b>40</b>
A.1 První příloha . . . . .	40

# Úvod

TODO: uvod

# 1. Úvod do problematiky

## 1.1 Seznámení s biologickými pojmy

### 1.1.1 RNA

**Rozdělit do více kapitol. Jedna z kapitol struktura RNA.**

RNA (zkratka z anglického ribonucleic acid) je biomolekula, která hraje klíčovou roli v procesu přenosu genetické informace u všech živých organismů. RNA se skládá z řetězce nukleotidů, které obsahují cukr ribózu, fosfátovou skupinu a jednu z pěti dusíkatých bází (adenin, guanin, cytosin, uracil nebo inosin). Existují různé typy RNA, jako jsou messenger RNA (mRNA), ribozomální RNA (rRNA) a transfer RNA (tRNA), které mají každý svou specifickou funkci v buňce.

RNA sekundární struktura se týká způsobu, jakým se molekula RNA skládá na sebe díky vzniku bázevých párů mezi komplementárními nukleotidy. Bázevých párování se děje mezi dusíkatými bázemi RNA nukleotidů, přičemž adenin (A) se páruje s uracilem (U) a guanin (G) se páruje s cytosinem (C).

RNA sekundární struktura je důležitá, protože může ovlivnit to, jak RNA molekula funguje. Například stem-loop struktura v mRNA molekule může ovlivnit přístupnost mRNA k ribozomům, což je buněčný mechanismus zodpovědný za překládání mRNA na proteiny.

### 1.1.2 Hairpin loop

### 1.1.3 Bulges

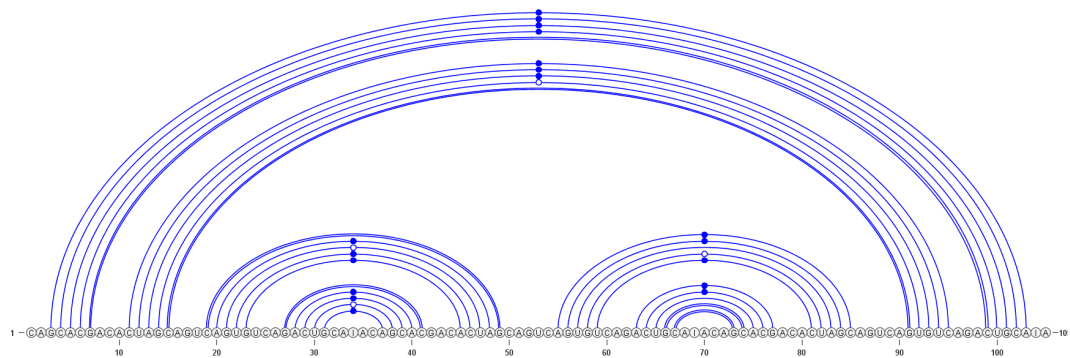
### 1.1.4 multibranch loop

## 1.2 Vizualizace sekundárních RNA struktur

Pro reprezentaci sekundární RNA struktury se používají jak textové, tak grafické způsoby. Pro nás jsou nejzajímavější ty grafické, ze kterých v této části představíme tři nejpoužívanější - arc diagram, circular diagram a radiate diagram. Obrázky ukázek diagramu v této části jsou získané za pomoci nástroje VARNA[2].

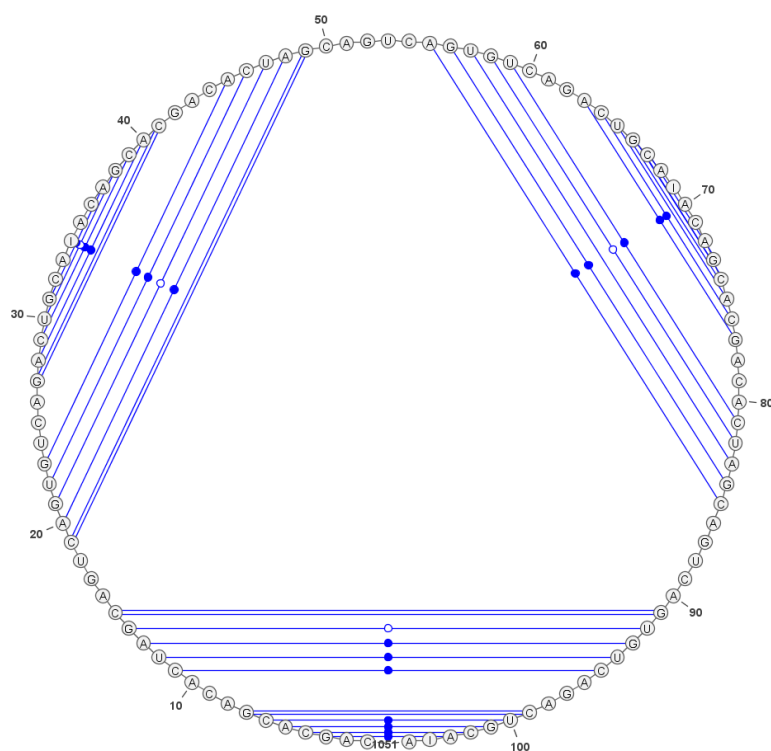
V arc diagramu jsou nukleotidy zobrazeny na rovné čáře ve stejném pořadí jako v sekvenci a bázevých páry nukleotidů jsou spojeny obloukem.





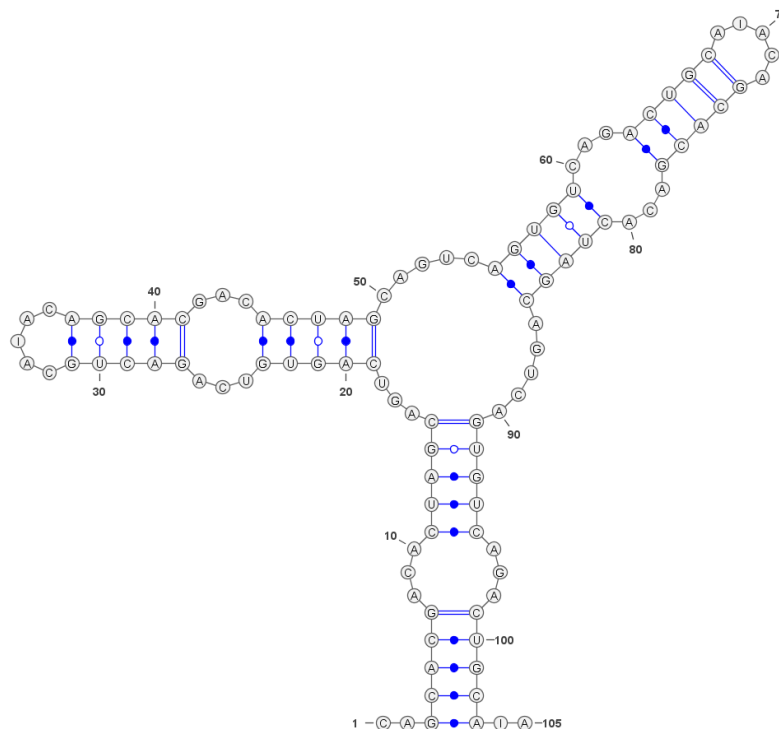
Obrázek 1.1: Ukázka arc diagramu

Circular diagram je velmi podobný. Nukleotidy neleží na rovné čáře, ale po obvodu kruhu. Bázové páry jsou spojeny buď čarou nebo obloukem.



Obrázek 1.2: Ukázka circular diagramu

V radiate diagramu jsou pozice nukleotidů voleny tak, aby bylo možné rozeznat motivy sekundární struktury, jako jsou hairpins, bulges nebo vícevětvené smyčky.



Obrázek 1.3: Ukázka radiate diagramu

Právě schopnost zachytit zmíněné motivy obě předešlé reprezentace postrádají, a proto se radiate diagram používá tam, kde je potřeba detailní vizuální analýza motivů sekundární RNA struktury a její interakce.

## 1.3 Podobné projekty

chybi motivacni uvod, který by řekl, co chceme delat, aby bylo mozne pochopit podobne k cem

Rádi bychom čtenáře seznámili s některými nástroji, které jsou používány pro vizualizaci sekundárních RNA struktur. Většina z nich jsou programy s uživatelským rozhraním a mohlo by se proto zdát zbytečné je zmiňovat nebo porovnávat s naší knihovnou. Nicméně u níže zmíněných programů není důležité řešení samotného uživatelského rozhraní, jako především druh zvolených metod pro vizualizaci a následné porovnávání.

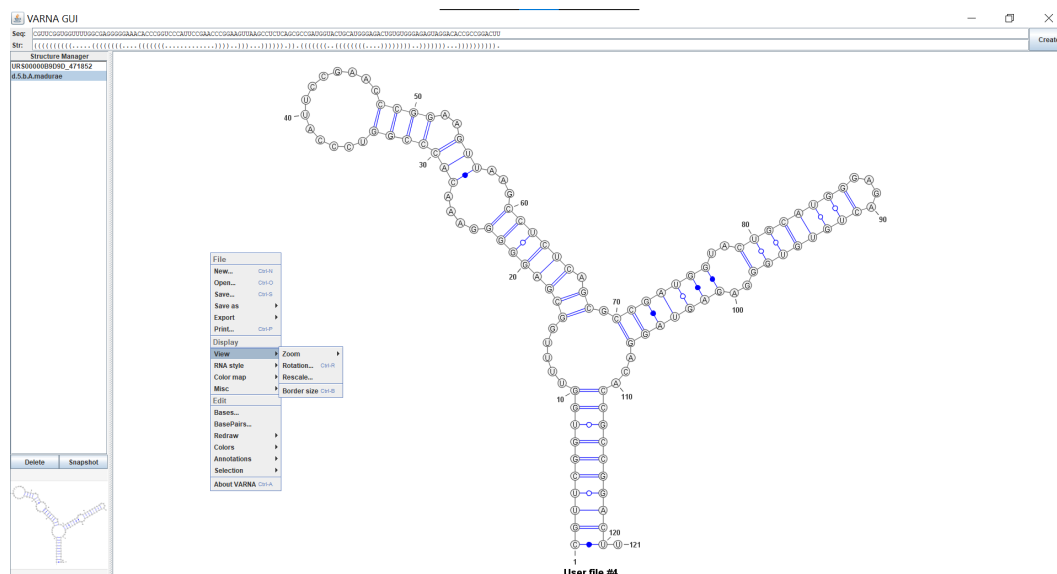
Z velkého množství existujících nástrojů, byla snaha vybrat takové, které mají rozdílné přístupy a nabízí nejširší paletu funkcí.

### 1.3.1 VARNA

VARNA (Visualization Applet for RNA) je nástroj pro automatické kreslení, vizualizaci a anotaci sekundárních RNA struktur, navržený jako doprovodný software pro webové servery a databáze.

VARNA implementuje algoritmy pro vykreslení všech tří výše zmíněných diagramů, podporuje různé textové formáty pro vstup i výstup a je schopen exportovat kresbu do rastrových nebo vektorových formátů. Umožňuje ruční úpravy

a strukturální anotace výsledku kresby a je považován za standard pro práci se sekundárními strukturami RNA.

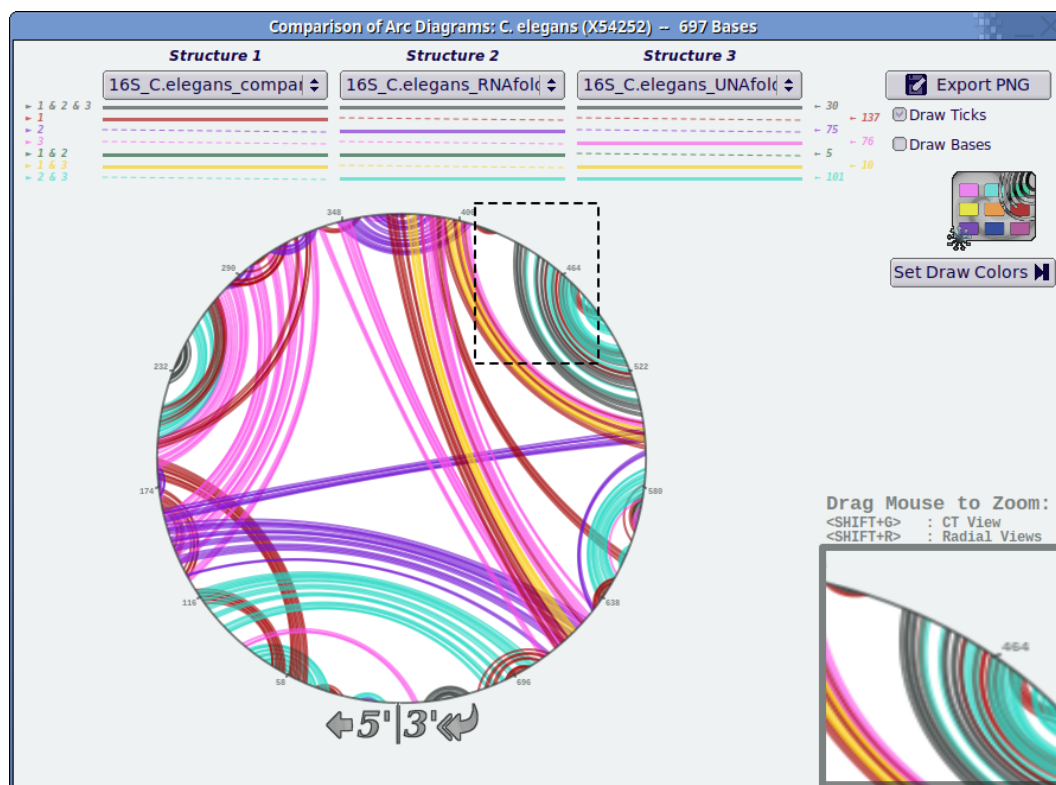


Obrázek 1.4: Snímek nástroje Varna. Zobrazená struktura je d.5.b.A.madurae.

### 1.3.2 RNAStructViz

RNAStructViz[1] je grafický nástroj pro analýzu sekundárních RNA struktur. Jeho předností je vizuální porovnání tří konfigurací v circular arc diagramu. Doplněné zabudovaným prohlížečem CT-style<sup>1</sup> souboru a prohlížečem radial diagramu podstruktury, která je přímo propojená s arc diagram oknem skrze nástroj pro výběr zoom. Mezi další funkce patří vypočítání číselných informací a možnost exportu obrázků a dat pro pozdější použití.

<sup>1</sup> CT formát souboru slouží k ukládání informace o sekvenci a bázeových pářích.

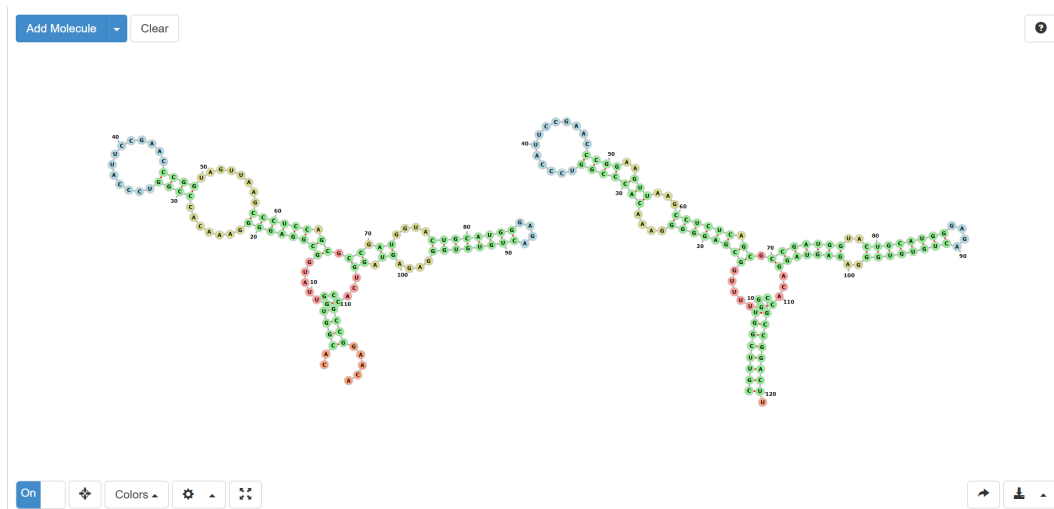


Obrázek 1.5: Snímek nástroje rnaStructViz, zobrazující tři struktury RNA.

### 1.3.3 Forna

Forna[4] (force-directed rna) nabízí webové rozhraní a server, který umožňuje uživateli vložit sekundární RNA strukturu ve formátu dot-bracket a zobrazí ji jako force-directed graf<sup>2</sup>. Uživatel může následně upravit pozice přetažením myši a lze i upravovat přímo strukturu.

<sup>2</sup> <https://cs.brown.edu/people/rtamassi/gdhandbook/chapters/force-directed.pdf>



Obrázek 1.6: Snímek nástroje Forna. Nalevo odvozená sekundární RNA struktura URS00000B9D9D\_471852 od struktury d.5.b.A.maduræ napravo.

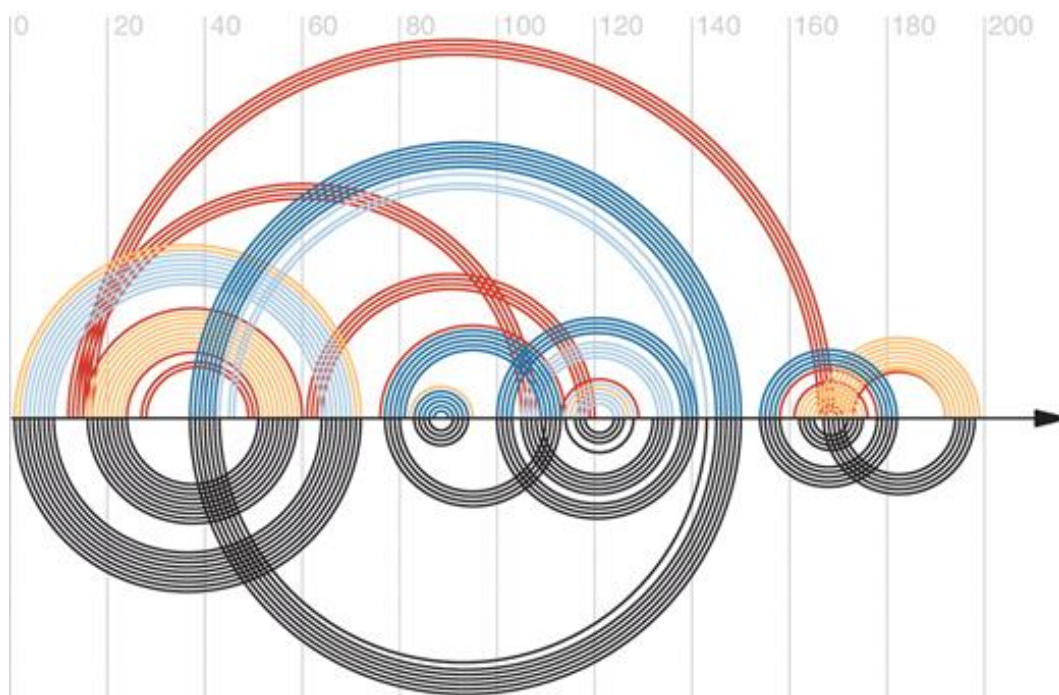
### 1.3.4 R-chie

R-chie [5] je web server, který umí vygenerovat šest různých typů arc diagramu. Vývoj tohoto nástroje byl se zaměřením především na složitější struktury, které nelze hezky nakreslit v radial diagramu. R-chie umí vygenerovat diagram pro porovnávání dvou sekundárních RNA struktur. Důležitým cílem byla možnost generovat diagramy pro velké množství dat, proto také nenabízí grafické rozhraní a s ním spojenou interakci se strukturami.

Projekt také nabízí balíček napsaný v jazyce R<sup>3</sup> zvaný R4RNA, který umožňuje spuštění programu lokálně a napříč operačním systémem.

---

<sup>3</sup> <https://www.r-project.org/>



Obrázek 1.7: Výsledný arc diagram nástroje R-chie, zobrazující dvě struktury. První struktura je nad horizontální čarou a druhá pod ní.

### 1.3.5 Shrnutí existujících nástrojů

Nástroje představené v této kapitole se soustředí především na práci s circular diagramem nebo arc diagramem, a právě pouze pro tyto diagramy nabízí nějaké metody pro porovnávání omezeného množství sekundárních struktur RNA. Forna podporuje pouze radial diagram, ale porovnávání dvou struktur, které sice jdou zobrazit vedle sebe, už nijak neusnadňuje.

Varna Podporuje všechny tři zmíněné diagramy, ale nelze ani zobrazit dvě sekundární rna struktury vedle sebe. Velkou výhodou nástroje VARNA by byla možnost použití na webu, ale k tomu používá Java Applets <sup>4</sup>, které jsou od roku 2017 považované za zastaralé <sup>5</sup>.

Ze zmíněných projektů je nejpodobnější tomu našemu R-chie, který se snaží usnadnit porovnávání sekundárních RNA struktur a nabízí i knihovnu napsanou v jazyce R. Liší se pak v samotném přístupu, protože jejich rozhraní generuje pouze statické circular nebo arc diagramy.

## 1.4 Kreslení grafů na základě šablony

Níže jsou zmíněné dva projekty, které úzce souvisí s naší knihovnou, protože produkují data ve formátu, se kterým pracuje naše knihovna a metody použité ke generování takových dat jsou klíčové pro naši knihovnu.

<sup>4</sup> <https://docs.oracle.com/javase/tutorial/deployment/applet/index.html>

<sup>5</sup> <https://www.oracle.com/java/technologies/javase/9-deprecated-features.html>

### 1.4.1 TRAVeLer

Traveler[3] je nástroj pro vizualizaci cílové sekundární struktury, využívající existující rozložení dostatečně podobné RNA struktury jako vzor. Traveler je založený na algoritmu, který konvertuje cílovou a vzorovou strukturu do odpovídající stromové reprezentace a využije stromovou editační vzdálenost společně s modifikací rozložení k přetvoření vzorové struktury do cílové. Traveler přijme na vstupu sekundární strukturu a vzor rozložení a na výstupu dá rozložení cílové struktury. Je to tedy command-line open source nástroj schopný rychle generovat rozložení i pro největší RNA struktury za poskytnutí dostatečně podobného rozložení.

Do vzniku Traveleru neexistoval žádný nástroj, který by dokázal velké struktury vizualizovat ve standardní notaci, se kterou jsou biologové naučení pracovat a porovnávat struktury napříč druhům.

### 1.4.2 R2DT

R2DT[6] je metoda pro predikci a vizualizaci široké škály sekundárních RNA struktur ve radial diagramu. R2DT je postaveno na knihovně se 3 647 vzory reprezentujícími většinu známých RNA struktur. R2DT se používá na ncRNA<sup>6</sup> (non-coding RNA) sekvencích z RNACentral<sup>7</sup> databáze a vytvořila více než 27 miliónů diagramů<sup>8</sup>, čímž tvoří největší světovou sadu dat s 2D RNA strukturami. Pro vizualizaci neboli 2D rozložení používá R2DT právě výše zmíněný nástroj Traveler.

---

<sup>6</sup> RNA, která se nepřekládá do proteinů

<sup>7</sup> <https://rnacentral.org/>

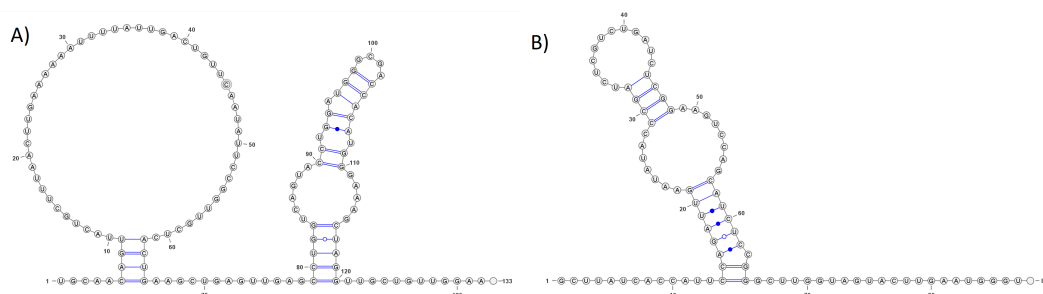
<sup>8</sup> Číslo je aktuální k datu 11.4. 2023

## 2. Metody vizualizace a porovnání

Cílem této knihovny není pouze vizualizovat sekundární strukturu RNA, ale také usnadnit analýzu rozdílů a podobností mezi více strukturami RNA. Proto jsme se zaměřili na práci s radial diagramy, které nejlépe zobrazují motivy struktury a zároveň jsou nejpřirozenější reprezentací.

Kromě toho jsme viděli potenciál generování rozložení na základě vzorové struktury, jak to dělá nástroj Traveler. Výstupem Traveleru je soubor ve formátu JSON, který obsahuje informace o vzoru každého nukleotidu a provedených úpravách - přidání, odebrání, přejmenování a přesunutí nukleotidu.

Například pokud bychom použili nástroj Varna na zobrazení následujících dvou struktur musíme všechny podobnosti vypořizovat sami.

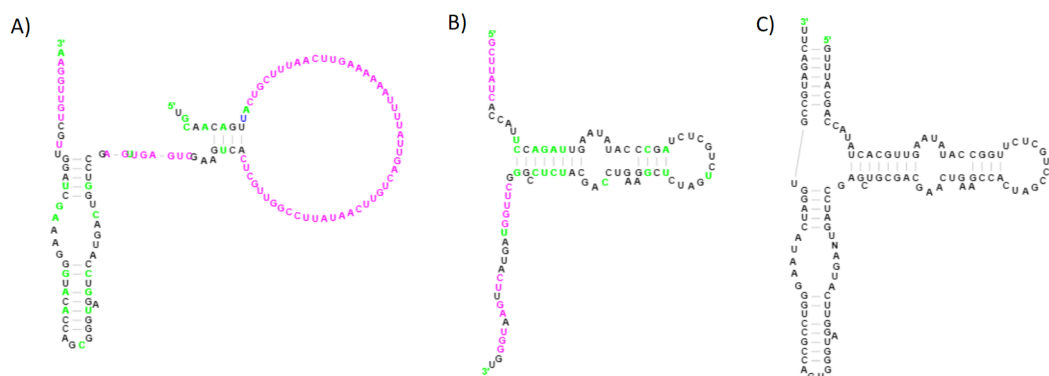


Obrázek 2.1: Dvě sekundární RNA struktury s RNACentral ID A) URS00006E712C, B) URS0000AB09C9.

Sice není těžké vypořizovat některé podobné motivy, ale není na první pohled jasná podobnost sekvence.

Na generování obou struktur používá nástroj Traveler stejnou vzorovou strukturu, která je v tomto konkrétním případě podobná obou odvozeným strukturám. Podobnost odvozených struktur ke vzorové je až na výjimky běžná, a proto jsme se rozhodli tuto podobnost předpokládat. Následující obrázek je vizualizace stejných struktur, za použití nástroje Traveler spolu se vzorovou strukturou.





Obrázek 2.2: Dvě sekundární RNA struktury s RNACentral ID A) URS00006E712C, B) URS0000AB09C9 vygenerované nástrojem Traveler a C) vzorové sekundární RNA struktury d.5.e.S.oshimae.

Přidáním struktury, ze které jsou obě struktury odvozené, je mnohem jasnější, na která místa koukat při hledání podobností a rozdílů. Přesto nemusí být nějaká podobnost hned jasná, a proto jsme se zaměřili na způsoby, jak znázornit mapování nukleotidů na vzorovou strukturu.

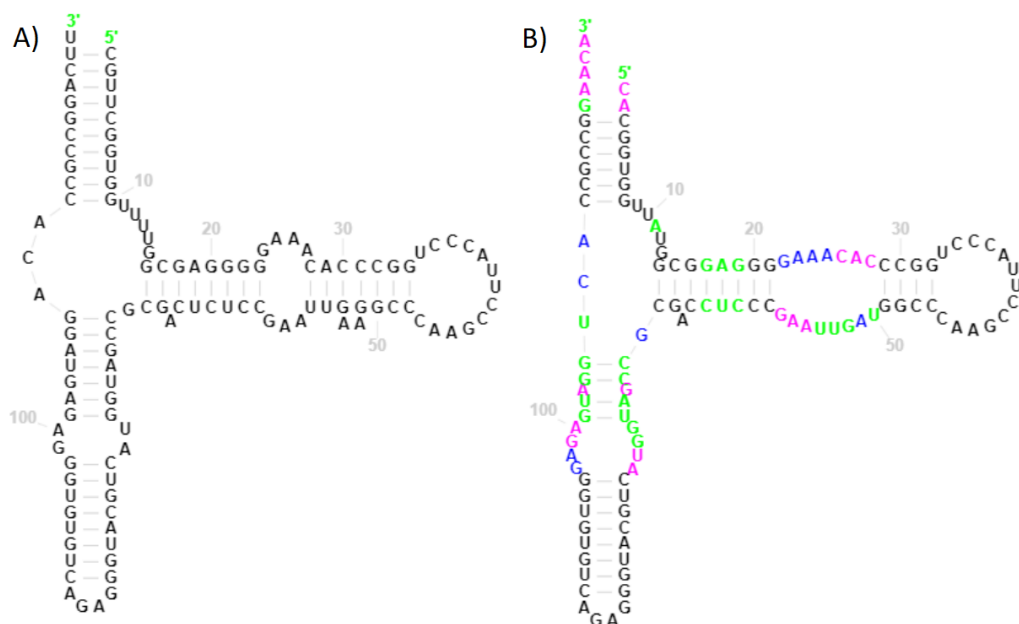
## 2.1 Důsledek využívání vzorové struktury

Využívání vzorové struktury pro analýzu má zřejmý a důležitý důsledek. Náš nástroj usnadňuje analýzu  $N$  struktur, které byly vygenerované ze stejné vzorové struktury.

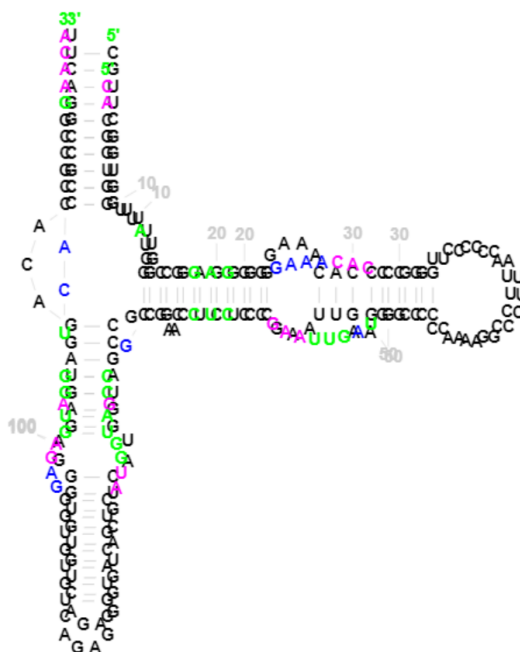
Protože Traveler umožňuje zadat, která vzorová struktura se má použít pro generování je teoreticky možné, porovnávat jakékoliv dvě sekundární RNA struktury, pokud bude jejich rozložení vygenerované na základě stejné vzorové struktury. Pokud se ale pokusíme vygenerovat rozložení struktury pomocí vzorové struktury, která je úplně jiná, nebude existovat žádná podobnost se vzorovou strukturou, a tím pádem naše metody, které spoléhají na podobnost vzorové struktury s vygenerovanou nebudou užitečné.

## 2.2 Překládání struktur

Vygenerované sekundární struktury jsou podobné vzorové struktuře, a tím pádem bývají podobné i ostatním vygenerovaným strukturám ze stejné vzorové struktury. Dává proto smysl se pokusit struktury přes sebe přeložit, aby se spojily společné části a vynikly rozdíly. Pouhým přeložením podobných struktur přes sebe však získáme výsledek, který neposkytuje příliš zajímavé informace a je málo přehledný.



Obrázek 2.3: A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae vedle sebe.



Obrázek 2.4: A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae přeložené přes sebe.

### 2.2.1 Zarovnání

Je potřeba najít způsob, jak řešit problém přesunu a zarovnání struktury, protože manuální manipulace pomocí přetažení myši nebo zadávání pozice může

být zbytečně obtížná, zejména pokud se snažíme dosáhnout přesného zarovnání. Proto je velmi užitečné umožnit zarovnání sekundární RNA struktury na konkrétní nukleotid nebo skupinu nukleotidů ze vzorové struktury.

Pokud je vybrán vzorový nukleotid  $v$  pro nukleotid  $n$  z ostatních struktur, jehož vzor je  $v$ , je celá struktura přesunuta tak, aby se nukleotidy  $v$  a  $n$  překrývaly.

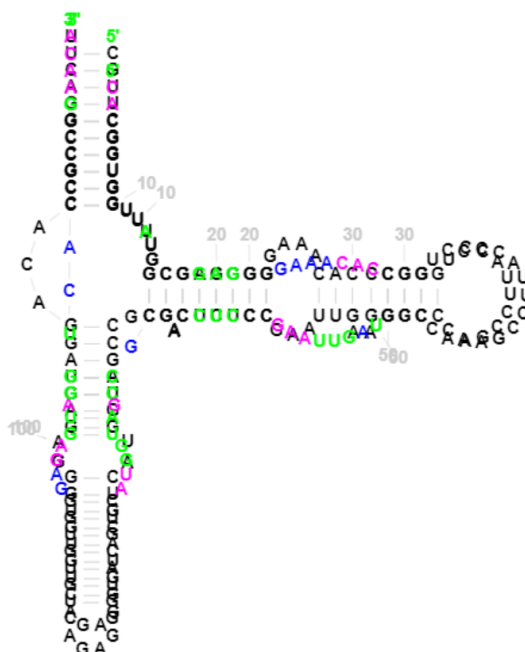
## Skupiny nukleotidů

Pokud se snažíme najít větší skupinu nukleotidů, na které chceme strukturu zarovnat, koukáme se na posunutí pro konkrétní nukleotidy a následně vybereme buď posunutí které zarovná nejvíc nukleotidů nebo to které zarovná skupinu nukleotidů které jsou v části struktury, která nás právě zajímá.

Pro hledání konkrétního posunutí pro zarovnání jsme použili naivní způsob, který nezaručuje nejlepší možné zarovnání, ale zároveň je poměrně přímočarý a dává rozumné výsledky, díky již zmíněné podobnosti struktur.

Postupně se prochází každá struktura. V první iteraci se porovná pozice každého nukleotidu s pozicí ve vzorové struktuře. Na základě pozice se roztřídí nukleotidy. Třídění funguje tak, že do stejné skupiny indexované posunutím se přidávají všechny vzorové nukleotidy, na které se struktura zarovná daným posunutím. Na konci první iterace se skupiny, které jsou menší než  $x$  odeberou.

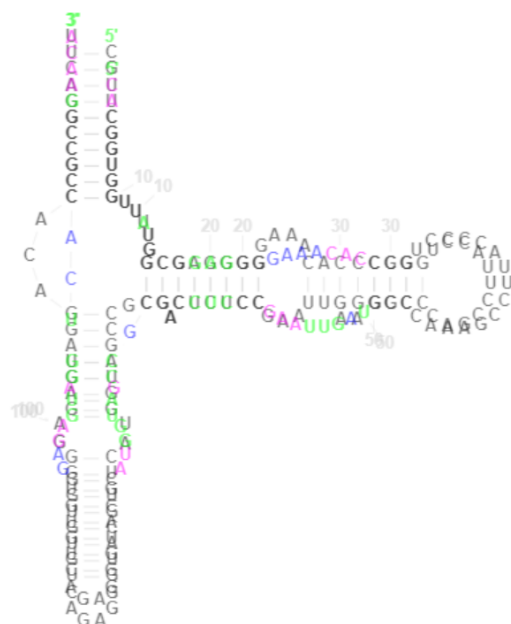
V druhé iteraci se postupuje podobně, ale používají se poslední vytvořené skupiny na filtrování, tím se snažíme docílit, zarovnání na stejné nukleotidy. Filtrování pomocí skupin ingoruje všechny nukleotidy, jejíž vzor není v nějaké skupině, pomocí které filtrujeme. Pokud během iterace nevzniknou žádné skupiny, řeší se iterace daná iterace zvlášť bez filtrování.



Obrázek 2.5: A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae přeložené přes sebe a zarovnané.

## 2.2.2 Průhlednost struktur

V přeložených a zarovnaných strukturách nelze snadno rozeznat, které nukleotidy jsou společné a překrývají se a které nejsou. Přidáním průhlednosti je možné tuto situaci rozlišit, protože překrývající se nukleotidy budou mít sytější barvu než ty, které se nepřekrývají.



Obrázek 2.6: A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae přeložené přes sebe, zarovnané a s průhledností.

## 2.2.3 Rozmazání struktur

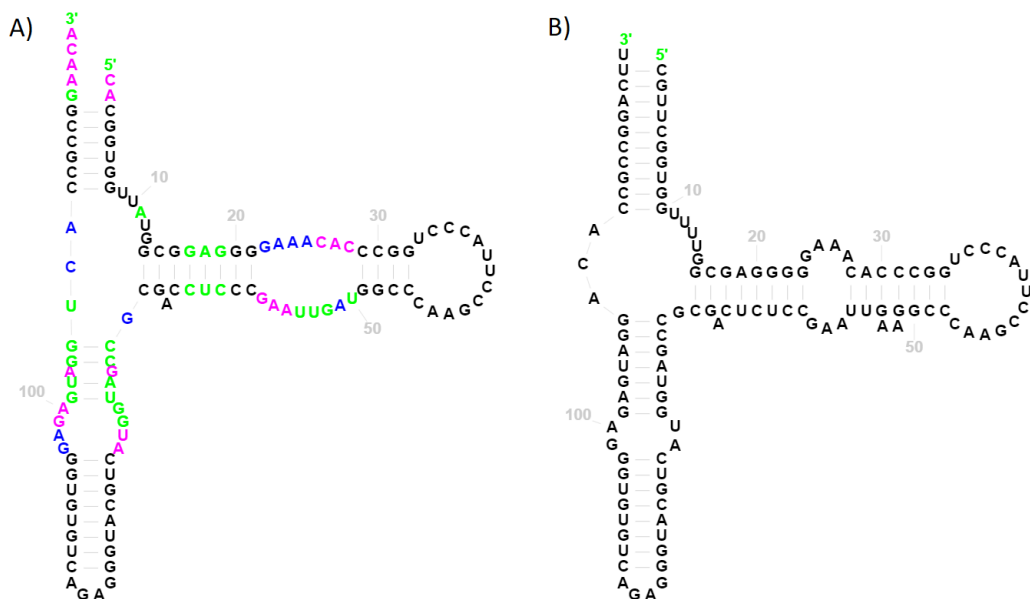
Zarovnávání struktur bohužel neřeší všechny výzvy. Výsledné obrázky se mohou zdát rozmazané. To je způsobeno tím, že ačkoli má nukleotid vzorový nukleotid, který je stejný, jeho pozice se může v rozložení mírně lišit v důsledku metody generování dat. Tento fakt může způsobit, že popisky nukleotidů vypadají rozmazaně.

Jako přímočaré řešení se může zdát posunutí jednotlivých nukleotidů, které jsou blízko sebe, aby dokonale překrývaly jejich vzory. Věříme, že by to vyřešilo zmíněný problém bez významné deformace struktury.

## 2.3 Obarvení struktur

Vstupní data obsahují barevné označení nukleotidů. Slouží k lepšímu zorientování se ve struktuře vzhledem ke vzorové struktuře. Jejich význam je následující.

Černá barva značí, že nukleotid leží na poloze vzorového nukleotidu se stejným názvem. Zelenou barvou jsou označeny ty nukleotidy jejíž vzorový nukleotid bylo třeba přejmenovat. Modrou barvou jsou vyznačeny posunutý nukleotid. A poslední růžovou barvu mají nově přidáné nukleotidy.

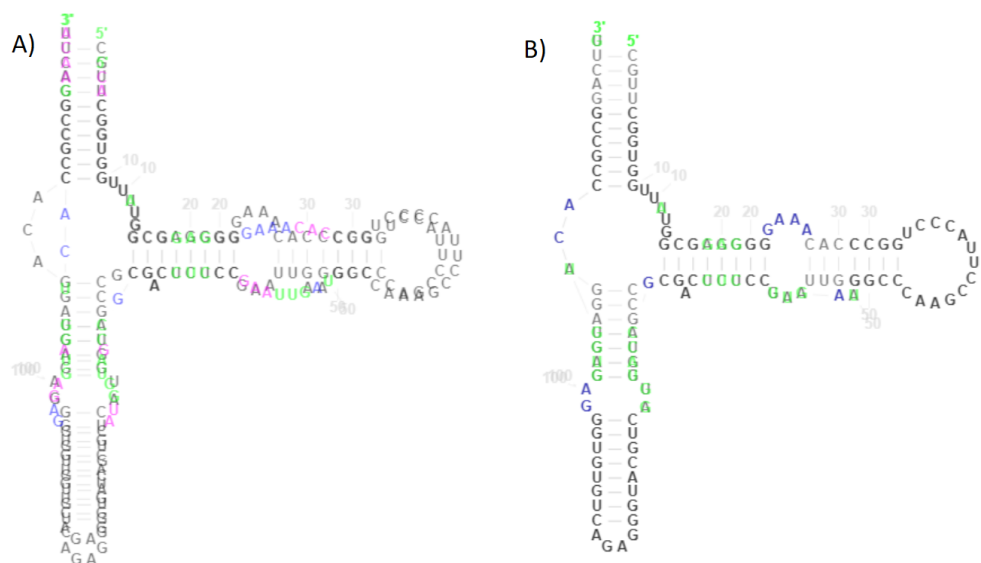


Obrázek 2.7: A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae.

## 2.4 Transformace na vzor

Užitečnou metodou je transformace mezi vzorovou a cílovou strukturou. Každý nukleotid, který má svůj vzorový nukleotid, se přemístí na pozici vzorového nukleotidu a nukleotidy, které ve vzoru nejsou, jsou skryté. Tato metoda je velmi užitečná pro práci s dvěma strukturami, které jsou si podobné, nebo pro získání počátečního přehledu o tom, co je na co namapováno.

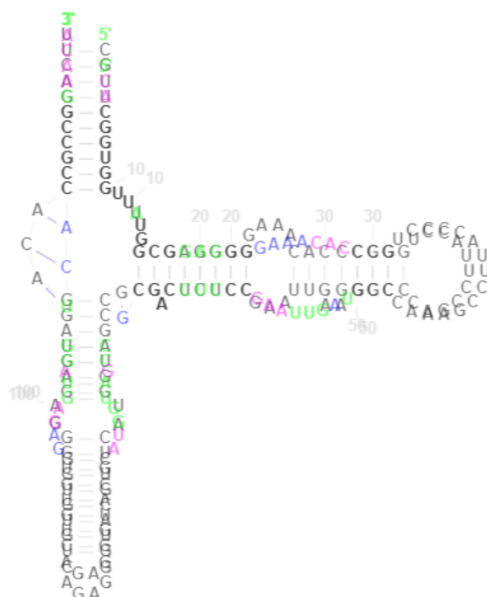
Slabou stránkou této metody je její použití při práci s více než dvěma strukturami nebo strukturami, které jsou velmi odlišné. V takových situacích se na obrazovce děje mnoho věcí a je obtížné se soustředit a vyzorovat něco užitečného.



Obrázek 2.8: Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí vzorové struktury d.5.b.A.madurae přeložené přes sebe A) před transformací a B) po transformaci.

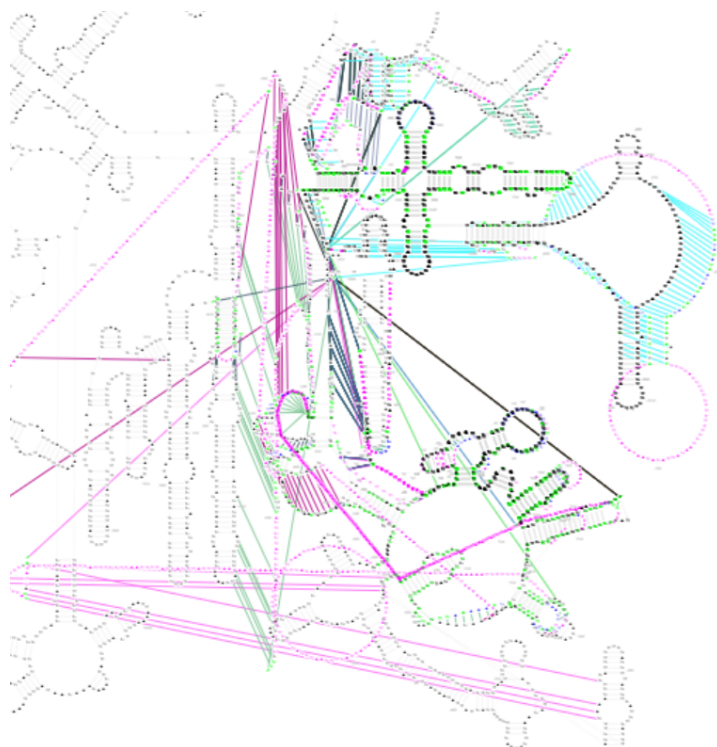
## 2.5 Mapovací čáry

Vědomost o tom, který nukleotid se na co mapuje, může být důležitá pro odhalení rozdílů a podobností mezi strukturami. V našem úsilí zprostředkovat tuto informaci již před transformací na vzorovou strukturu jsme přišli s čarami, které spojují každý nukleotid s jeho vzorovým nukleotidem.



Obrázek 2.9: Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí vzorové struktury d.5.b.A.madurae přeložené přes sebe s mapovacíma čárama.

Bohužel tento způsob se zvětšující se velikostí struktury stává velmi nepřehledným, přesto si myslíme že můžou být užitečné.

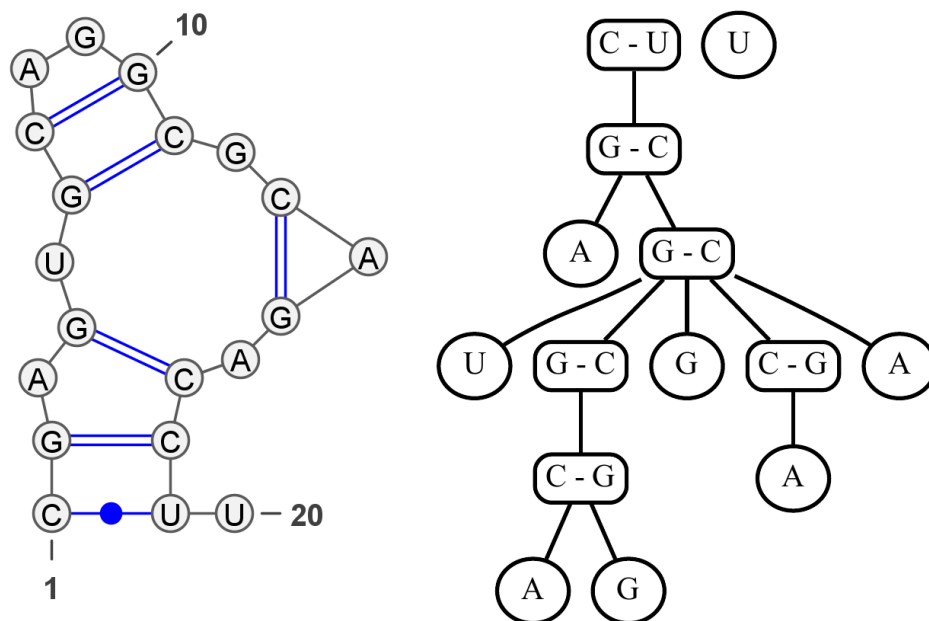


Obrázek 2.10: Výřez z mnoha struktur vygenerovaných nástrojem Traveler pomocí vzorové struktury DD\_28S\_3D přeložených přes sebe s mapovacími čarami. Každá struktura má vlastní barvu mapovacích čar.

## 2.6 Využití stromu

V projektu Traveler byla použita stromová reprezentace pro sekundární RNA strukturu, ve které je vnitřní vrchol, tedy má více jak jednu hranu, básový pár a listem, tedy má jednu nebo žádnou hranu, je nezpárovaný nukleotid. Strom pak lze vytvořit následovně. První nepárové nukleotidy ze začátku a z konce přidáme jako samostatné vrcholy. Potom jdeme postupně po básových párech. První básový pár tvoří kořen. Další básově páry připojujeme a tvoříme strom. Pokud narazíme na nezpárovaný nukleotid přidáme ho jako list k poslednímu přidanému básovému páru. Větvení struktury vyústí ve větvení stromu.





Obrázek 2.11: Vlevo je uměle vytvořená struktura zobrazená nástrojem Varna. Vpravo je jeho stromová reprezentace.

Stromovou strukturu jsme neměli v úmyslu použít na nic konkrétního, ale chtěli jsme prozkoumat možnosti lokálních transformací struktury pro dosažení zarovnání více podobných částí, které jsou třeba jenom posunuté, jako je to například v následující části dvou struktur, které jsou zarovnané dvěma způsoby.



Obrázek 2.12: První způsob zarovnání.



Obrázek 2.13: Druhý způsob zarovnání.

Po zvážení jsme dospěli k názoru, že není možné provést úpravu struktury, která by významně nedeformovala strukturu. Jakákoli úprava části struktury znamená posunutí nějakého zbytku struktury, což může zasáhnout do jiné části. Proto by bylo nezbytné odebrat některé nukleotidy. Ty, které nejvíce překážejí, jsou ty přidané. Pokud bychom odebrali tyto přidané nukleotidy, dostali bychom se na stejný výsledek jako s transformací struktury na vzorovou strukturu.

## 2.7 Zvolené metody

Naše knihovna přímo podporuje překládání a zarovnání struktur na konkrétní nukleotid nebo skupinu nukleotidů bez zarovnání na úrovni jednotlivých nukleotidů pro odstranění rozmazání struktur. Lze transformovat struktury na vzorovou strukturu a upravovat průhlednost struktur.

## 3. Dokumentace

### 3.1 Vstupní data

#### 3.1.1 Struktura dat

Jedná se o data ve formátu JSON, obsahující všechny potřebné informace o rozložení nukleotidů, jejich párování, velikosti popisků, barvách a tloušťkách čar. Kromě informací o rozložení lze z nich také vyčíst informace o potřebných editacích vzorové sekundární struktury.

Každý nukleotid má svůj index a jméno (A, C, G nebo U). Kromě vlastního indexu a jména je vedle uloženo i jméno a index vzorového nukleotidu. Pokud se jedná o přidaný nukleotid, index vzorového nukleotidu je  $-1$ . Za pomoci těchto informací lze jednoduše zjistit, jestli je nukleotid přidaný nebo přejmenovaný a pokud máme k dispozici i data vzorové struktury, lze pomocí indexu porovnávat jejich souřadnice.

V rámci R2DT projektu vzniká JSON schéma<sup>1</sup>, které by mělo popisovat strukturu vstupních dat. Schéma je stále ve vývoji, proto aktuální výstupy Traveleru neodpovídají schématu a je dost možné, že se formát výstupu bude v budoucnu ještě měnit a naše knihovna se jim bude přizpůsobovat.

Samotná struktura dat není složitá, ale popíšeme zde pouze tu část, kterou aktuálně využíváme, kromě toho, že ostatní data pro nás nejsou důležitá, tak jak již bylo zmíněno samotná struktura dat není pevně daná a může se měnit.

Jedná se o objekt, který má dvě položky - `rnaComplexes` a `classes`, což je pole objektů popisující třídy říkající způsob zobrazení struktury, podobně jako to kaskádové styly<sup>2</sup> (CSS) diktují pro webové stránky.

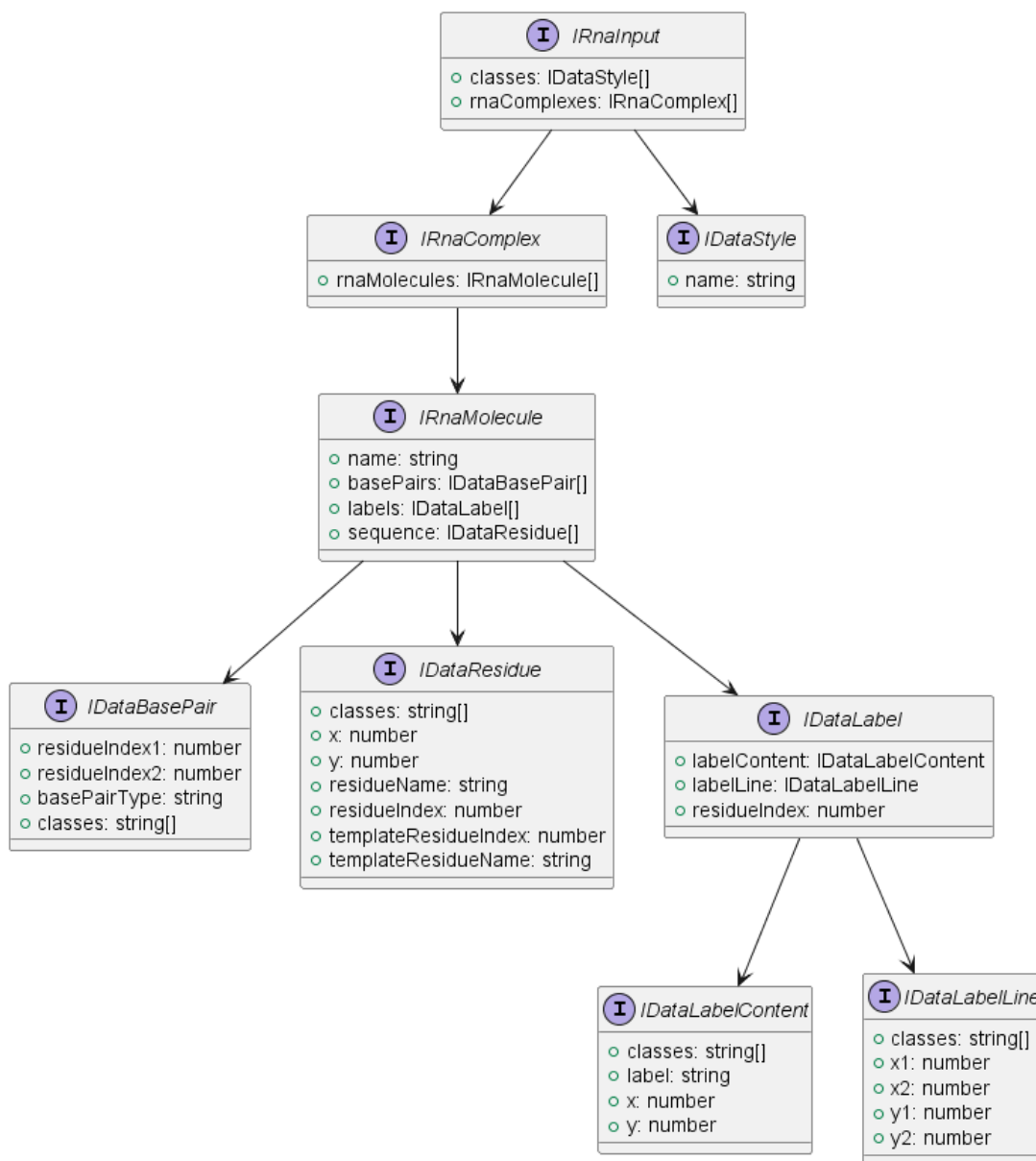
`rnaComplexes` je pole polí sloužící pro popis celých skupin RNA struktur. Naše knihovna pracuje vždy pouze s nultým prvkem. Neviděli jsme důvod to dělat jinak, a pokud by se nějaký důvod našel v budoucnu, neměl by být problém naší knihovnu přizpůsobit situaci (např. rozšířením o novou metodu pro zachování zpětné kompatibility).

V rámci naší knihovny jsme vytvořili interface, který vstupní data musí splňovat. Struktura zbytku dat by měla být jasně viditelná z následujícího diagramu těchto interfaceů.

---

<sup>1</sup> <https://github.com/LDWLab/RNA2D-data-schema>

<sup>2</sup> <https://developer.mozilla.org/en-US/docs/Web/CSS>



Obrázek 3.1: Interface pro vstupní data.

### 3.1.2 Získání vstupních dat

RNACentral je otevřená databáze, která používá R2DT, ke generování rozložení sekundárních RNA struktur. Pro přístup k databázi lze použít jak API<sup>3</sup>, tak přímo posílat SQL dotazy<sup>4</sup>.

Bohužel nezpřístupňuje mapování nukletidů na vzorovou strukturu, přesto je databáze užitečná pro získání vstupních dat, protože obsahuje struktury a ke každé takové struktuře informaci, ze které byla vygenerovaná.

Pro získání testovacích dat jsme využívali veřejný SQL přístup<sup>5</sup>. Vybrali jsme

<sup>3</sup> <https://rnacentral.org/api>

<sup>4</sup> <https://rnacentral.org/help/public-database>

nějaké vzorové struktury různého typu, ke každému jsme vybrali náhodně nejvýše 20 RNA molekul, které použili danou vzorovou strukturu pro generování. Následně jsem použili R2DT, které jsme řekli, jakou vzorovou strukturu použít pro vygenerování vstupních dat.

## 3.2 Objektový návrh

V následující části se pokusíme čtenáře seznámit s objektovým návrhem naší knihovny. V této části dáme obecný pohled na strukturu tříd a později i ukážeme jak se s třídami dá pracovat.

Srdcem celé knihovny je třída **RnaVis**. Má v sobě uložené vrstvy realizované třídou **Layer**, představující jednotlivé struktury. Třída **RnaVis** vykresluje struktury na canvas, nastavuje se přes ní zoom/posouvání a zpřístupňuje většinu funkcionality.

Pro určení vykreslovacích parametrů (např. font, barva, velikost) objektů máme třídu **Styles**. Před každým vykreslením se zeptáme této třídy na vykreslovací parametry pro daný objekt. Důvod, proč jsme zvolili takovéhle řešení problému je, že jsme původně neměli jasnou představu, co a jak bude řešené, a proto nám přišlo nejjednodušší tuto část oddělit a použít stejný způsob, jaký se používá ve vstupních datech.

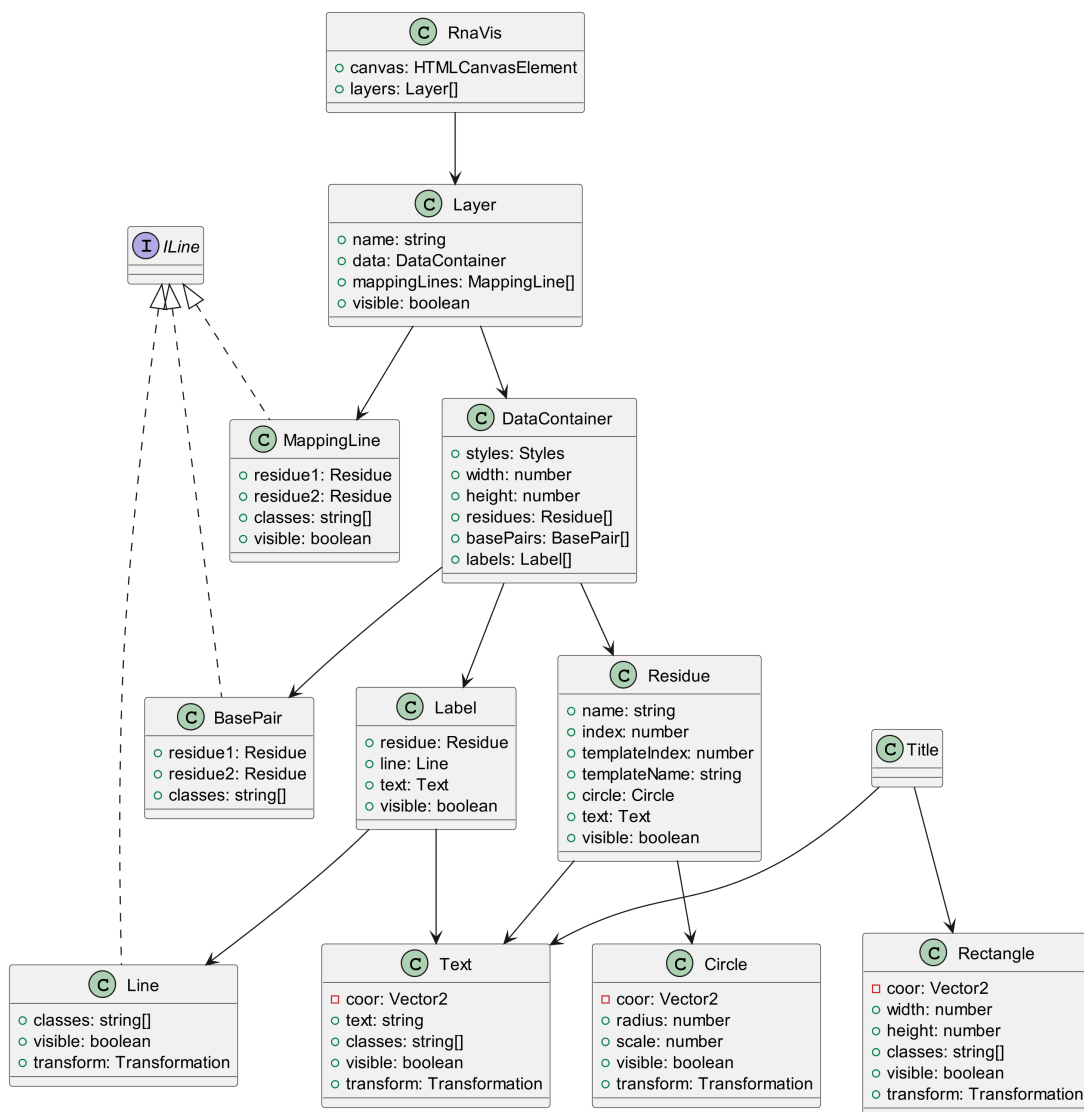
Poslední důležitou velkou částí jsou animace. Tuto funkcionalitu zpřístupňují dvě hlavní třídy - **TranslationAnim**, **VisibilityAnim**. Obě implementují interface **IAnimation**.

Níže jsou tři diagramy tříd, které dohromady obsahují každou třídu. Diagramy se snažíme vyjádřit obecnou strukturu, tím pádem pro přehlednost neobsahují všechny informace - všechny metody, některé privátní vlastnosti a některé závislosti. Pro podrobnější rozbor tříd a metod se doporučujeme podívat do referenční dokumentace<sup>6</sup>.

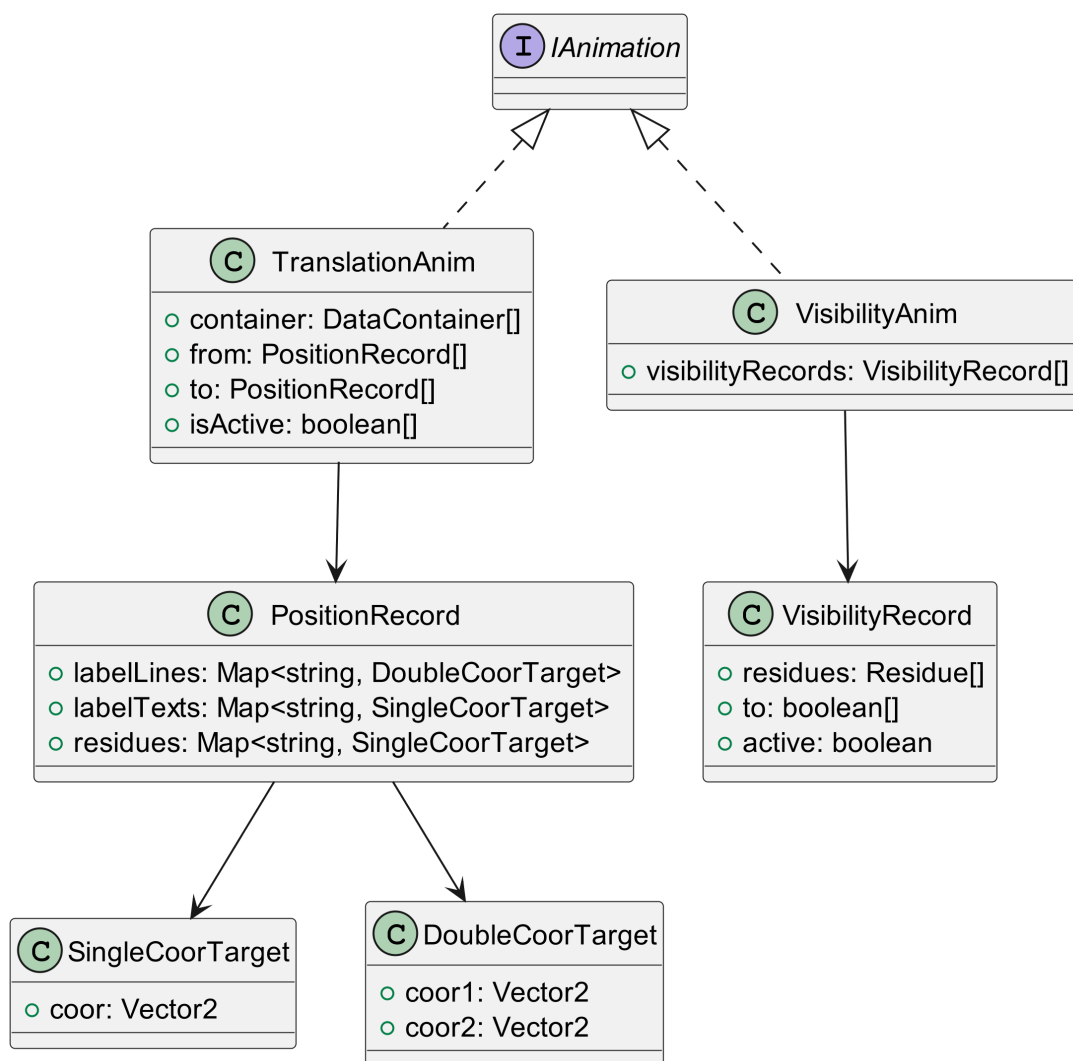
---

<sup>5</sup> Použitý dotaz: <https://gist.github.com/davidhoksza/52df5b0fa5d85fdc4513966d2b50792b>

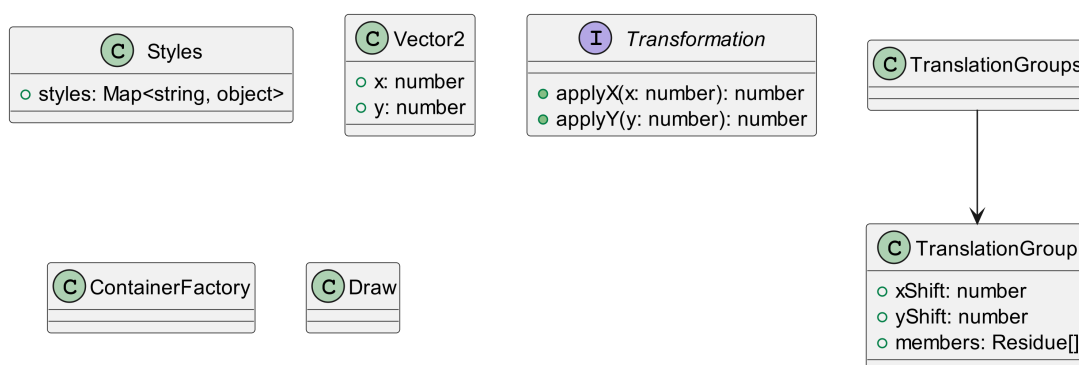
<sup>6</sup> <https://github.com/michalhercik/rna-visualizer/blob/main/lib/docs/README.md>



Obrázek 3.2: Diagram tříd.



Obrázek 3.3: Diagram tříd.



Obrázek 3.4: Diagram tříd.

## 3.3 Uživatelské seznámení s knihovnou

S obecnou strukturou kódu jsme čtenáře seznámili a nyní se pokusíme ukázat, jak lze knihovnu užívat. Bude se jednat o ukázkou implementací základních využití, jako je vykreslení struktur, zarovnání, transformace na vzorovu strukturu a některé další možnosti.

### 3.3.1 Vytvoření objektu RnaVis a vykreslení struktur

V prvním kroce importujeme veškeré potřebné zdroje. Pužité vstupní data `d.5.b.A.madurae.json` a `URS00000B9D9D.json` jsou dostupné na GitHub stránce projektu<sup>7 8</sup>

```
import {
  RnaVis,
  TranslationGroups,
  PositionRecord,
  TranslationAnim,
} from 'rna-visualizer';
import template from './d.5.b.A.madurae.json';
import structure from './URS00000B9D9D.json';
```

Následně vytvoříme `canvas`, na který budeme vykreslovat, nastavíme mu šířku, výšku a přidáme ho do těla dokumentu.

```
const canvas = document.createElement('canvas');
canvas.width = 500;
canvas.height = 500;
document.body.appendChild(canvas);
```

Nyní už můžeme vytvořit samotný objekt `RnaVis`, který nás bude provázet v ostatních ukázkách.

```
const rnaVis = new RnaVis(canvas);
```

Přidáme do něj dvě struktury - `template` a `structure`. Struktury se přidávají pomocí metody `addLayer`, která má tři parametry - přidávanou strukturu, libovolný název a poslední třetí parametr je volitelný a slouží k určení, jestli chceme, aby se struktura vykreslovala a jeho defaultní hodnota je `true`, tedy defaultně se struktura vykresluje. Knihovna předpokládá, že první přidaná struktura, je vzorová struktura, ze které jsou všechny ostatní vygenerované.

```
rnaVis
  .addLayer(template, 'd.5.b.A.madurae')
  .addLayer(structure, 'URS00000B9D9D');
```

<sup>7</sup> [https://github.com/michalhercik/rna-visualizer/blob/main/demo/example\\_data/madurae/URS00000B9D9D\\_471852-d.5.b.A.madurae.json](https://github.com/michalhercik/rna-visualizer/blob/main/demo/example_data/madurae/URS00000B9D9D_471852-d.5.b.A.madurae.json)

<sup>8</sup> [https://github.com/michalhercik/rna-visualizer/blob/main/demo/example\\_data/madurae/d.5.b.A.madurae\\_fromn\\_d.5.b.A.madurae.colored.json](https://github.com/michalhercik/rna-visualizer/blob/main/demo/example_data/madurae/d.5.b.A.madurae_fromn_d.5.b.A.madurae.colored.json)



Struktury lze potom vykreslit pomocí metody `draw`.

```
rnaVis.draw();
```

Celý kód lze zapsat následovně.

```
import {
  RnaVis,
  TranslationGroups,
  PositionRecord,
  TranslationAnim,
} from 'rna-visualizer';
import template from './d.5.b.A.madurae.json';
import structure from './URS00000B9D9D.json';

const canvas = document.createElement('canvas');
canvas.width = 500;
canvas.height = 500;
document.body.appendChild(canvas);

new RnaVis(canvas)
  .addLayer(template, 'd.5.b.A.madurae')
  .addLayer(structure, 'URS00000B9D9D')
  .draw();
```

V dalších ukázkách budeme předpokládat předchozí kroky a budeme na ně navazovat.

### 3.3.2 Zoom a posouvání canvasu

Pro zoom a posouvání používáme `d3.js` knihovnu. Není ale třeba používat jejich rozhraní, protože třída `RnaVis` obsahuje metodu `addZoom`, která přidá jak zoom tak posouvání canvasu.

```
rnaVis.addZoom();
```

### 3.3.3 Zarovnání struktur

Zarovnání struktur docílíme tak, že zavoláme na instanci třídy `RnaVis` metodu `align`, která vrátí pole objektů typu `Vector2`, které reprezentují posunutí pro každou strukturu. Následně zavoláme na instanci třídy `RnaVis` metodu `translate`, které předáme pole obsahující posunutí a metoda posune každou strukturu o příslušné posunutí. Potom už stačí pouze vykreslit struktury metodou `draw`.

```
const translations = rnaVis.align();
rnaVis
  .translate(translations);
  .draw();
```

## Zarovnání struktur na konkrétní skupinu

Metoda `align` má volitelný parametr typu `number`, kterým můžeme vybrat, na kterou skupinu nukleotidů chceme dostat zarovnání.

Při první iteraci se porovnává první přidaná struktura (vzorová) s druhou přidanou strukturou a roztrídí se nukleotidy podle posunutí, které je zarovná. Číselný parametr potom určuje, která skupina se použije pro filtrování v dalších iteracích.

Pokud je parametr prázdný nebo menší než nula, vybere se největší skupina. Pokud je parametr větší, než počet vytvořených skupin metoda vyhodí vyjímku se zprávou `groupIndex >= groups.length`.

V následující ukázce vytvoříme skupiny, které vzniknout při první iteraci a vybereme nejmenší, kterou pak použijeme pro zarovnání struktur. Skupiny vytvoříme statickou metodou `TranslationGroups.create`. První parametr je vzorová struktura, druhý parametr je struktura, kterou chceme zarovnat. Třetí parametr jsou skupiny pomocí, kterých chceme filtrovat a je volitelný a poslední parametr je minimální velikost výstupních skupin, který je taky volitelný a jeho defaultní hodnota je 5.

```
const containers = rnaVis.getDataContainers();
const groupSizes =
  TranslationGroups.create(containers[0], containers[1], null, 20)
    .map(group => group.size());
const groupIndex = groupSizes.indexOf(Math.min(...groupSizes));
const translationsToGroup = rnaVis.align(groupIndex);
rnaVis
  .translate(translationsToGroup)
  .draw();
```

## Zarovnání na nukleotid

Zarovnání na nukleotid obsahuje podobné kroky. Nejdřív si zvolíme nukleotid ze vzorové struktury, na který chceme struktury zarovnávat. Potom na instanci třídy `RnaVis` pomocí metody `getAlignmentToTempResidue` získáme posunutí pro každou strukturu a nakonec struktury zarovnáme a vykreslíme.

```
const templateResidue = rnaVis.layers[0].data.residues[42];
const translationsToResidue =
  rnaVis.getAlignmentToTempResidue(templateResidue);
rnaVis
  .translate(translationsToResidue)
  .draw();
```

## Animace zarovnání

Jakýkoliv přechod je pro oko vždy příjemnější animovat a to knihovna umožňuje pomocí jednoduchého rozhraní pro animace. V následující ukázce ukážeme jak animovat zarovnání struktur.

V prvním kroce vytvoříme pro každou strukturu `PositionRecord`, který obsahuje pozice nukleotidů a popisků. V tomto případě bude obsahovat cílovou pozici

pro každý nukleotid a vytvoříme ho pomocí statické metody `fromTranslation`, která bere jako parametr strukturu a posunutí struktury.

Následně vytvoříme objekt typu `TranslationAnim`, který umožňuje mimo jiné animovat transformaci struktury na daný `PositionRecord`. Potom už jenom zavoláme metodu `animate`, kterou začne celá animace. Jako parametry bere instanci třídy `RnaVis`, kterou používá pro vykreslování každého snímku animace, trvání animace v milisekundách a poslední volitelný parametr je funkce, která se zavolá na konci animace. V ukázce na konci animace jí otočíme a animujeme zpátky do původního stavu.

```
const containers = rnaVis.getDataContainers();
const animationTargets = rnaVis
  .align()
  .map((translation, i)
    => PositionRecord.fromTranslation(containers[i], translation)
  );

const anim = new TranslationAnim(containers, animationTargets);
anim.animate(rnaVis, 1500, () => {
  anim.reverse();
  anim.animate(rnaVis, 1500);
});
```

### 3.3.4 Mapovací čáry

Zobrazení čar, které ukazují mapování nukleotidů, je velmi jednoduché. Mapovací čáry se vytvoří hned při přidání každé vrstvy a jsou defaultně schované, proto je stačí pouze zviditelnit.

```
rnaVis.layers.forEach(
  l => l.mappingLines.forEach(
    m => m.setVisible(true)
  )
);
rnaVis.draw();
```

### 3.3.5 Transformace na vzorovou strukturu

Poslední ukázkou je transformace na vzorovou strukturu. V ukázce nejdříve schováme nukleotidy, které nemají vzorový nukleotid. Potom vytvoříme opět `PositionRecord`, tentokrát pomocí statické metody `fromTemplate`, která každému nukleotidu dá pozici jeho vzorového nukleotidu. Potom už jenom animujeme přechod.

```
const containers = rnaVis.getDataContainers().slice(1);
const targetStructure = rnaVis.layers[0].data;

containers.forEach(c => {
  c.getUnmappableResidues().forEach(
```

```

    r => r.setVisible(false)
  );
});

const targets = containers.map(
  cont => PositionRecord.fromTemplate(cont, targetStructure)
);
new TranslationAnim(containers, targets)
  .animate(rnaVis, 1500);

```

## 3.4 Použití canvasu

### 3.4.1 Knihovna D3.js

D3.js<sup>9</sup> je knihovna v jazyce Javascript pomáhající přivést data k životu podporující především SVG<sup>10</sup>. SVG je webový standard, který lze skvěle kombinovat s ostatními standardy jako je CSS, DOM<sup>11</sup>, Javascript.

D3.js knihovna klade důraz na webové standardy a tím dává uživateli možnost využívat moderní prohlížeče naplno bez dalších frameworků. S knihovnou není nejjednodušší se naučit pracovat, ale její velkou předností je rychlost a široká paleta funkcí pro vizualizace dat.

### 3.4.2 SVG

V počátcích jsme chtěli k zobrazování používat SVG, které v kombinaci s D3.js knihovnou je velmi silný a flexibilní nástroj.

Bohužel SVG je poměrně pomalé a je to poznat už při vykreslování tisícovek objektů. Takového počtu objektů můžeme dosáhnout pouze s jednou sekundární RNA strukturou, my navíc chceme zobrazit více takových struktur a ještě s nimi dynamicky pracovat. Tím se pro nás stalo SVG nepoužitelné.

### 3.4.3 Canvas

Další možností bylo využití canvasu<sup>12</sup>, který slibuje výrazně lepší výkon a lze ho stále jednoduše používat.

Některé pro nás klíčové funkce D3.js knihovny lze využít i pro práci s canvasem. Konkrétně se jedná o zoom, posouvání a animace.

U canvasu jsme nakonec i zůstali, přestože při práci s desítkami velkých struktur vykreslování není plynulé.

---

<sup>9</sup> <https://d3js.org/>

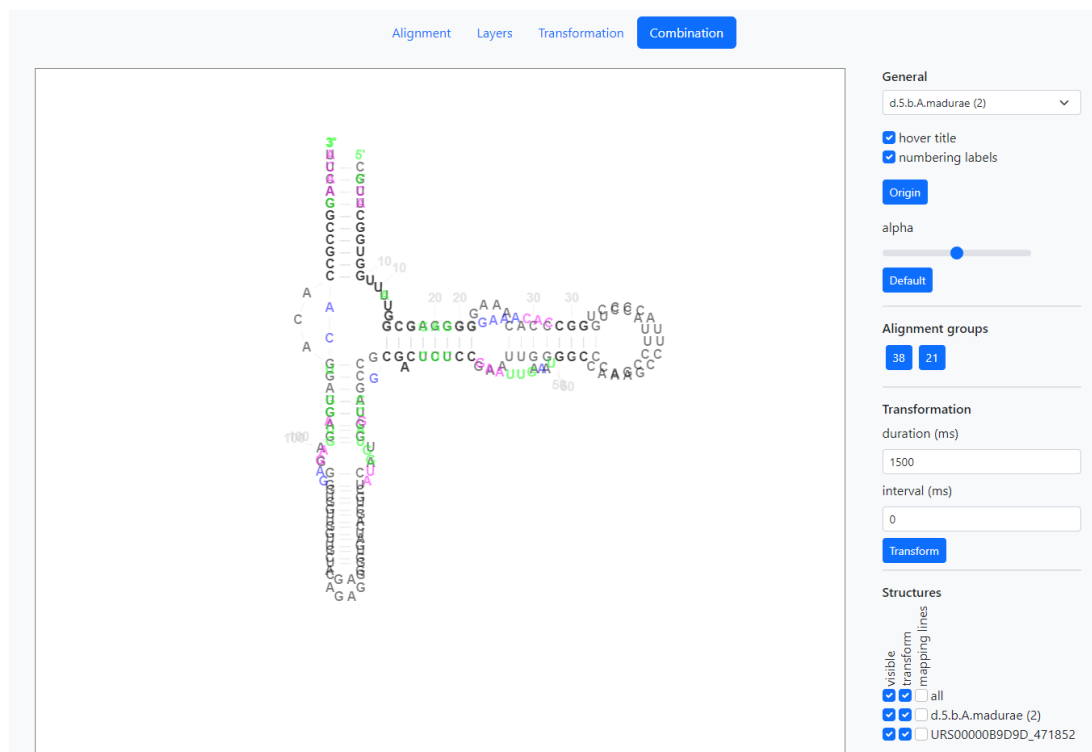
<sup>10</sup> <https://developer.mozilla.org/en-US/docs/Web/SVG>

<sup>11</sup> [https://developer.mozilla.org/en-US/docs/Web/API/Document\\_Object\\_Model](https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model)

<sup>12</sup> [https://developer.mozilla.org/en-US/docs/Web/API/Canvas\\_API](https://developer.mozilla.org/en-US/docs/Web/API/Canvas_API)

## 3.5 Demonstrace metod

V rámci naší knihovny jsme vytvořili také webovou aplikaci<sup>13</sup>, která demonstruje možnosti naší knihovny. Tato aplikace umožňuje uživatelům pracovat s jednotlivými metodami nebo se všemi metodami najednou pro usnadnění porovnání - překládání se zarovnáním, transformace na vzorovou strukturu, mapovací čáry a další možnosti pro zpříjemnění práce.



Obrázek 3.5: Screenshot aplikace vytvořené pro demonstraci funkcionality knihovny.

<sup>13</sup> <https://michalhercik.github.io/rna-visualizer/>

# Závěr

# Seznam použité literatury

- [1] CHENNEY, S., HEITSCH, C., MIZE, C., SWENSON, S., SCHMIDT, M. D., KIRKPATRICK, A. a YOON, I. (2019). Rnastructviz. URL <https://github.com/gtDMMB/RNAStructViz/wiki>. Accessed on March 29, 2023.
- [2] DARTY, K., DENISE, A. a PONTY, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**(15), 1974–1975.
- [3] ELIAS, R. a HOKSZA, D. (2017). Traveler: a tool for template-based rna secondary structure visualization. *BMC Bioinformatics*, **18**(1), 487. ISSN 1471-2105. doi: 10.1186/s12859-017-1885-4. URL <https://doi.org/10.1186/s12859-017-1885-4>.
- [4] KERPEDJIEV, P., HAMMER, S. a HOFACKER, I. L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**(20), 3377–3379. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv372. URL <https://doi.org/10.1093/bioinformatics/btv372>.
- [5] LAI, D., PROCTOR, J. R., ZHU, J. Y. A. a MEYER, I. M. (2012). R-chie : a web server and R package for visualizing RNA secondary structures . *Nucleic Acids Research*, **40**(12), e95–e95. ISSN 0305-1048. doi: 10.1093/nar/gks241. URL <https://doi.org/10.1093/nar/gks241>.
- [6] SWEENEY, B. A., HOKSZA, D., NAWROCKI, E. P., RIBAS, C. E., MADEIRA, F., CANNONE, J. J., GUTELL, R., MADDALA, A., MEADE, C. D., WILLIAMS, L. D., PETROV, A. S., CHAN, P. P., LOWE, T. M., FINN, R. D. a PETROV, A. I. (2021). R2dt is a framework for predicting and visualising rna secondary structure using templates. *Nature Communications*, **12**(1), 3494. ISSN 2041-1723. doi: 10.1038/s41467-021-23555-5. URL <https://doi.org/10.1038/s41467-021-23555-5>.

# Seznam obrázků

1.1	Ukázka arc diagramu . . . . .	5
1.2	Ukázka circular diagramu . . . . .	5
1.3	Ukázka radiate diagramu . . . . .	6
1.4	Snímek nástroje Varna. Zobrazená struktura je d.5.b.A.madurae. .	7
1.5	Snímek nástroje rnaStructViz, zobrazující tři struktury RNA. . .	8
1.6	Snímek nástroje Forna. Nalevo odvozená sekundární RNA struktura URS00000B9D9D_471852 od struktury d.5.b.A.madurae napravo. . . . .	9
1.7	Výsledný arc diagram nástroje R-chie, zobrazující dvě struktury. První struktura je nad horizontální čarou a druhá pod ní. . . . .	10
2.1	Dvě sekundární RNA struktury s RNACentral ID A) URS00006E712C, B) URS0000AB09C9. . . . .	12
2.2	Dvě sekundární RNA struktury s RNACentral ID A) URS00006E712C, B) URS0000AB09C9 vygenerované nástrojem Traveler a C) vzorové sekundární RNA struktury d.5.e.S.oshimae. . . . .	13
2.3	A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae vedle sebe. . . . .	14
2.4	A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae přeložené přes sebe. . . . .	14
2.5	A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae přeložené přes sebe a zarovnané. . . . .	15
2.6	A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae přeložené přes sebe, zarovnané a s průhledností. . . . .	16
2.7	A) Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí B) vzorové struktury d.5.b.A.madurae. .	17
2.8	Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí vzorové struktury d.5.b.A.madurae přeložené přes sebe A) před transformací a B) po transformaci. . . .	18
2.9	Struktura s RNACentral ID URS00000B9D9D vygenerovaná nástrojem Traveler pomocí vzorové struktury d.5.b.A.madurae přeložené přes sebe s mapovacími čárami. . . . .	19
2.10	Výřez z mnoha struktur vygenerovaných nástrojem Traveler pomocí vzorové struktury DD_28S_3D přeložených přes sebe s mapovacími čárami. Každá struktura má vlastní barvu mapovacích čar. . . . .	20
2.11	Vlevo je uměle vytvořená struktura zobrazovaná nástrojem Varna. Vpravo je jeho stromová reprezentace. . . . .	21
2.12	První způsob zarovnání. . . . .	21
2.13	Druhý způsob zarovnání. . . . .	22



3.1	Interface pro vstupní data. . . . .	24
3.2	Diagram tříd. . . . .	26
3.3	Diagram tříd. . . . .	27
3.4	Diagram tříd. . . . .	27
3.5	Screenshot aplikace vytvořené pro demonstraci funkcionality knihovny. 33	

# Seznam tabulek

# Seznam použitých zkratek

## A. Přílohy

### A.1 První příloha