



**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

Úvod	2
1 Úvod do problematiky	3
1.1 Seznámení s biologickými pojmy	3
1.1.1 RNA	3
1.2 Datové formáty	3
1.2.1 JSON	3
1.3 Vizualizace sekundárních RNA struktur	4
1.4 Podobné projekty	5
1.4.1 VARNA	5
1.4.2 RNAStructViz	6
1.4.3 Forna	6
1.4.4 R-chie	6
1.4.5 Shrnutí existujících nástrojů	6
1.5 Příbuzné projekty	7
1.5.1 TRAVeLer	7
1.5.2 R2DT	7
2 Metody vizualizace a porovnání	8
3 Programátorská dokumentace	9
3.1 Volba technologií	9
3.2 Vstupní data	9
4 Uživatelská dokumentace	17
Závěr	18
Seznam použité literatury	19
Seznam obrázků	20
Seznam tabulek	21
Seznam použitých zkratk	22
A Přílohy	23
A.1 První příloha	23

Úvod

TODO: uvod

1. Úvod do problematiky

1.1 Seznámení s biologickými pojmy

1.1.1 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.

1.2 Datové formáty

1.2.1 JSON

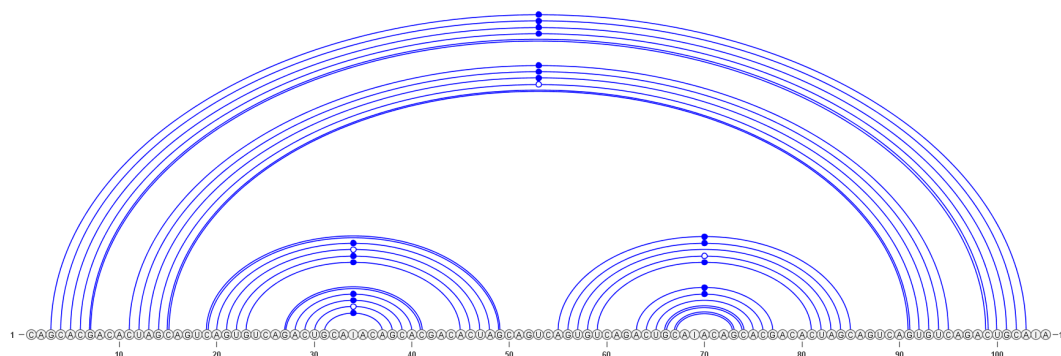
JSON (javascriptový objektový zápis) je datový formát sloužící k ukládání dat organizovaných v polích nebo objektech. Navzdory názvu je na programovacím jazyce nezávislý. Skládá se z dvojice klíč – hodnota. Hodnota je libovolný podporovaný datový typ (např.: boolean, číslo, string, pole, objekt). Níže je ukázka JSON formátu.

```
{
  "basePairs":
  [
    {
      "basePairType": "canonical",
      "classes":
      [
        "bp-line"
      ],
      "residueIndex1": 2,
      "residueIndex2": 118
    }
  ]
}
```

1.3 Vizualizace sekundárních RNA struktur

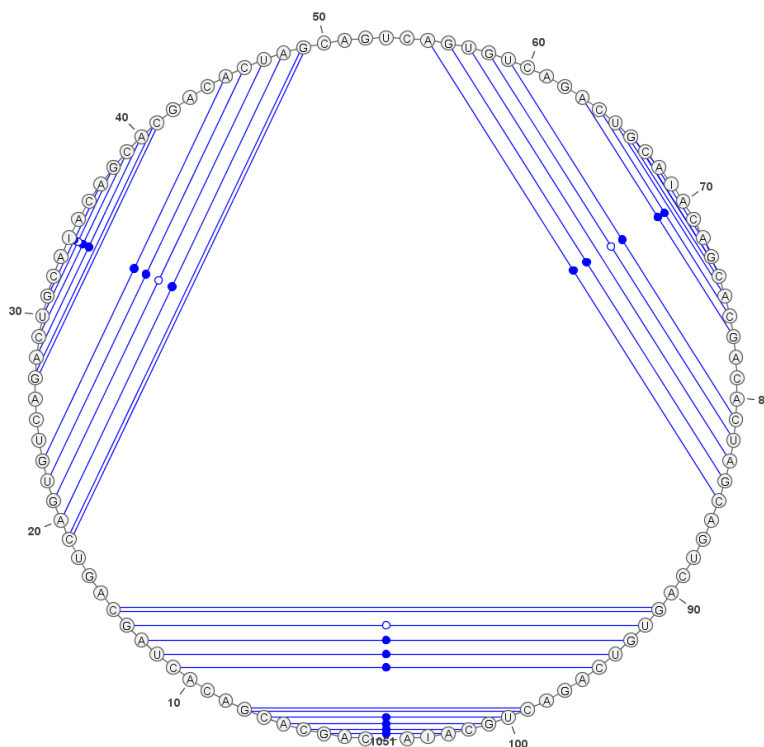
Pro reprezentaci sekundární RNA struktury se používají jak textové, tak grafické způsoby. Pro nás jsou nejzajímavější ty grafické, ze kterých v této části představíme tři nejpoužívanější - linear diagram, circular diagram a radiate diagram. Obrázky ukázek diagramu v této části jsou získané za pomoci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

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



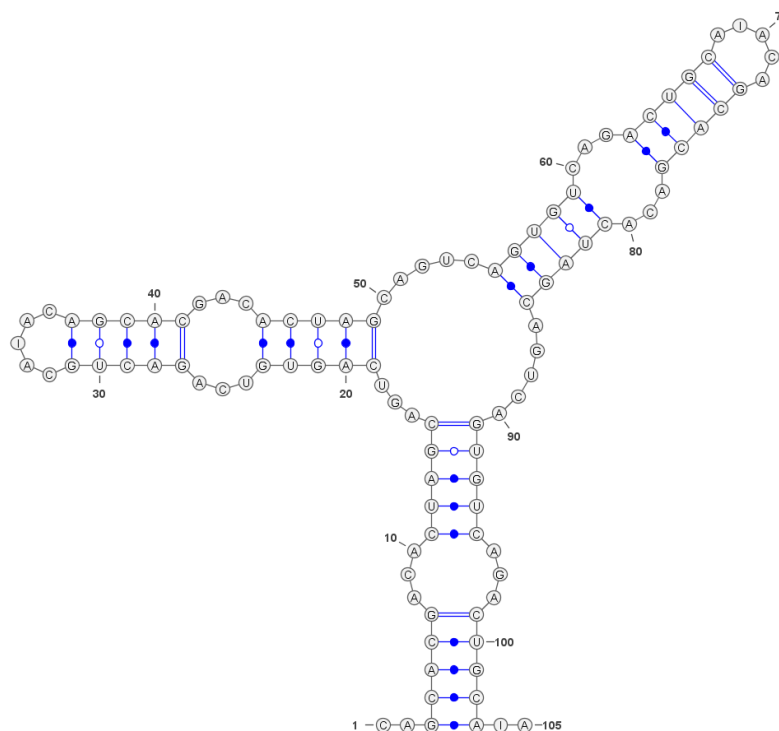
Obrázek 1.1: Ukázka linear diagramu

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

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

1.4 Podobné projekty

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.

1.4.1 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 čtyři kreslicí algoritmy, podporuje různé textové formáty pro vstup i výstup a je schopný 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.

1.4.2 RNAStructViz

RNAStructViz[1] je grafický nástroj pro analýzu sekundárních RNA struktur. Jeho předností je vizuální porovnání tří konfigurací v kompaktním a standardizovaném circular arc diagramu. Doplněné zabudovaným prohlížečem CT-style 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.

1.4.3 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 graf¹. Uživatel může následně upravit pozice přetažením myši a lze i upravovat přímo strukturu.

1.4.4 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 R² zvaný R4RNA, který umožňuje spuštění programu lokálně a napříč operačním systémem.

1.4.5 Shrnutí existujících nástrojů

Nástroje představené v této kapitole se soustředí především na práci s circular diagramem nebo linear 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 Applets³, které jsou od roku 2017 považované za zastaralé⁴.

Za nejpodobnější projekt bychom považovali 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 linear diagramy.

¹<https://cs.brown.edu/people/rtamassi/gdhandbook/chapters/force-directed.pdf>

²<https://www.r-project.org/>

³<https://docs.oracle.com/javase/tutorial/deployment/applet/index.html>

⁴<https://www.oracle.com/java/technologies/javase/9-deprecated-features.html>

1.5 Příbuzné projekty

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.

1.5.1 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.

1.5.2 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 sekvencích z RNACentral⁵ databáze a vytvořila více než 13 miliónů diagramů, čímž tvoří největší světovou sadu dat s 2D RNA strukturami. Pro vizualizaci neboli 2D rozložení používá R2DT právě výše zmíněný nástroj Traveler.

⁵<https://rnacentral.org/>

2. Metody vizualizace a porovnání

Cílem této knihovny je nejen vizualizovat sekundární RNA strukturu, ale především zjednodušit analýzu rozdílů a podobností vícero RNA struktur. Právě proto jsme se zaměřili na práci s radial diagramem.

Zároveň jsme viděli potenciál v generování rozložení na základě vzorové struktury, jako to dělá nástroj Traveler. Výstupem Traveleru je soubor ve formátu JSON, který obsahuje mimo jiné informaci o vzoru každého nukleotidu a i informaci o provedených editacích.

Rozhodli jsme se proto v našich metodách využívat právě výše zmíněného mapování na vzorovou strukturu. Náš nástroj tím pádem je schopný pracovat se strukturama, jejíž rozložení je vygenerované na základě stejné vzorové struktury.

Jednou z metod je transformace z a na vzorovou strukturu. Každý nukleotid, který má vzorový nukleotid se přemístí na pozici vzorového nukleotidu a ty nukleotidy, které vzor nemají jsou schované. Metoda je velmi příjemná pro práci se dvěma strukturama, které si jsou podobné nebo pro počáteční přehled co je na co namapované. Slabá stránka této metody je zjevná při práci s vícero strukturami nebo strukturami, které jsou velmi odlišné. V takových situacích se toho na displeji děje hodně a je složité se soustředit a vypořádat něco užitečného.

Vědět který nukleotid se na co mapuje může být velmi užitečné pro odhalení rozdílů a podobností struktur. Snažili jsme se najít další způsob, jak tuto informaci předat, jediné s čím jsme přišli jsou čáry, které spojují nukleotidy se vzorovým. Bohužel tento způsob se zvětšující se velikostí struktury stává velmi nepřehledným, přesto si myslíme že můžou být užitečné a v naše knihovna je podporuje.

Protože jsou struktury odvozené od stejného vzoru jsou typicky velmi podobné, dává proto smysl mít možnost je přeložit přes sebe, aby splynuli společné části a vynikly ty rozdílné. Manipulací se strukturou ručně ať už přetažením myši nebo zadáním pozice může být zbytečně otravné především kvůli zarovnání. Přijde nám proto velmi užitečné mít možnost zarovnat sekundární RNA strukturu na konkrétní nukleotid nebo skupinu nukleotidů. Obojího lze s naší knihovnou pohodlně dosáhnout, včetně nalezení posunutí, kterým lze zarovnat skupiny nukleotidů

Zarovnávání struktur bohužel nedá vždy na první pohled očekávaný výsledek, protože ačkoli má nukleotid vzorový nukleotid, od kterého se nijak neliší může stále jeho pozice být mírně posunutá. Je to dáno metodou generování dat. Popisky nukleotidů můžou tím pádem vypadat trochu rozmazaně. Jako přímočaré řešení by se mohlo zdát posunout jednotlivé nukleotidy, které jsou blízko, aby dokonale překrývali jejich vzor. Věříme, že by to vyřešilo zmíněný problém, nicméně naše knihovna tuto funkci nijak přímo neimplementuje.

V rámci naší knihovny vznikla i webová aplikace¹, která demonstruje možnosti naší knihovny. Umožňuje pracovat vždy jen s jednou metodou nebo se všemi metodami usnadňující porovnání dvou sekundárních struktur RNA, které jsou zmíněné v této kapitole.

¹<https://michalhercik.github.io/rna-visualizer/>

3. Programátorská dokumentace

3.1 Volba technologií

3.1.1 Programovací jazyk

Volba programovacího jazyka byla poměrně přímočará. Chtěli jsme napsat knihovnu, která se bude používat na webu. Javascript¹ je v tomto případě jasnou volbou, protože to je v podstatě to jediné, co se používá. Přesto Javascript není jazyk, ve kterém je naše knihovna napsaná, protože se jedná o dynamicky typovaný jazyk, což s sebou nese určité výhody pro jednoduché a rychlé psaní kódu, ale u větších projektů se to stává nevýhodou. Naše knihovna je napsaná v Typescriptu², což je nadmnožina Javascriptu, snažící se řešit jeho slabiny a navíc ho lze snadno přeložit do Javascriptu.

3.1.2 SVG vs canvas

3.2 Vstupní data

Jak jsme již zmiňovali, naše knihovna využívá výstupní data nástroje Traveler jako vstupní data. Jedná se o data ve formátu JSON, obsahující všechny potřebné informace o rozložení nukleotidů, jejich párování, velikostech popisků, barvách a tloušťkách čar. Kromě informací o rozložení obsahuje také informace o potřebných editacích vzorové sekundární struktury.

V rámci R2DT projektu vzniká i JSON schéma³, které by mělo popisovat strukturu vstupních dat. Schéma je stále ve vývoji, proto aktuální výstupy R2DT nebo Traveleru neodpovídají schématu a je dost možné, že se jejich výstupy budou v budoucnu měnit a naše knihovna se jim bude přizpůsobovat, TODO: přepsat protože RNAcentral, využívající R2DT, je největší databázi s 2D RNA strukturama.

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 - `classes`, což je pole objektů popisující třídy říkající způsob zobrazení struktury, podobně jako to kaskádové styly⁴ diktují pro webové stránky a `rnaComplexes`.

`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 spl-

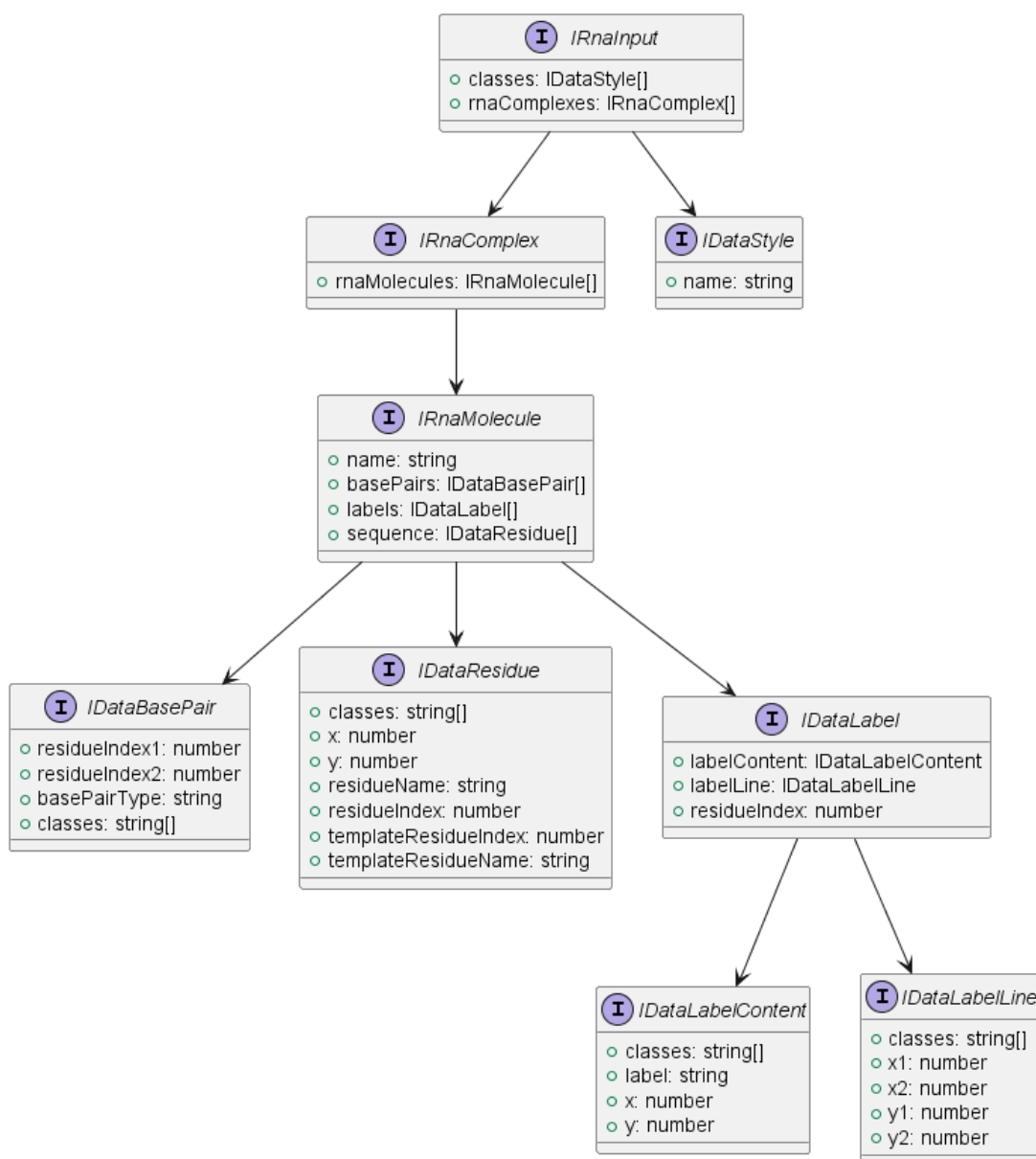
¹<https://developer.mozilla.org/en-US/docs/Web/JavaScript>

²<https://www.typescriptlang.org/>

³<https://github.com/LDWLab/RNA2D-data-schema>

⁴<https://developer.mozilla.org/en-US/docs/Web/CSS>

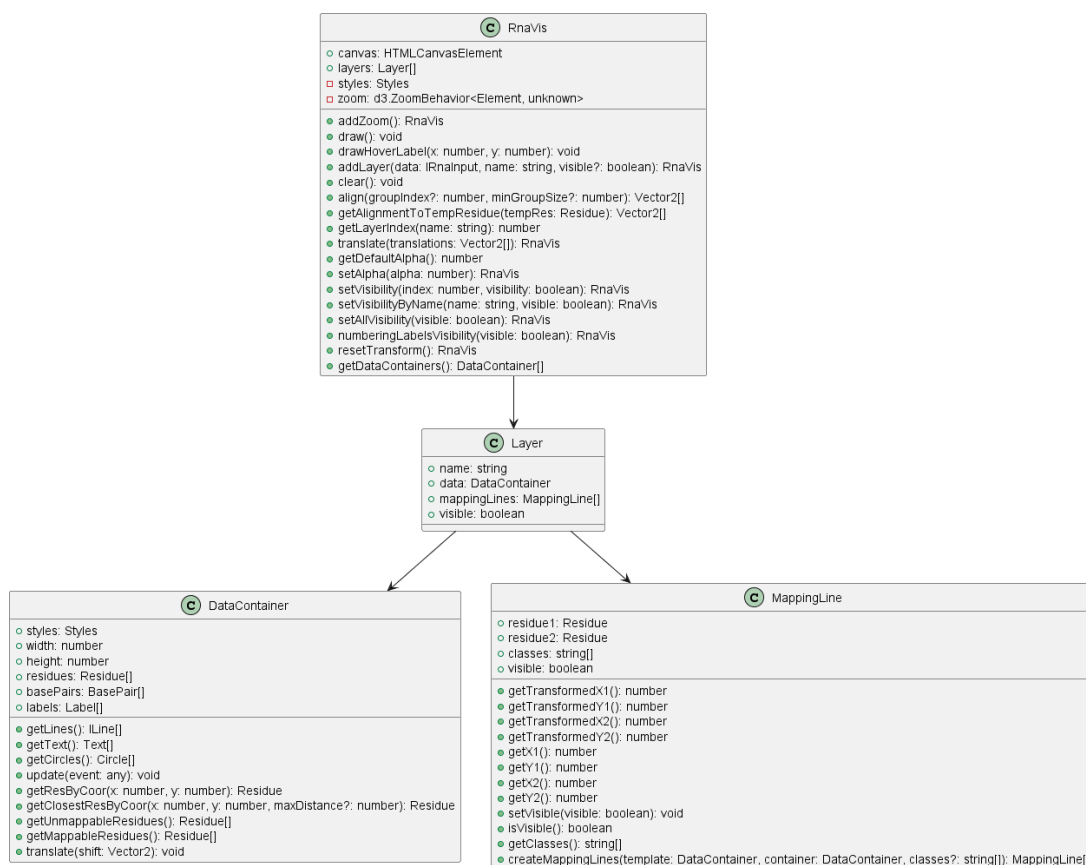
ňovat. Struktura zbytku dat by měla být jasně viditelná z následujícího UML⁵ diagramu těchto interfaceů.



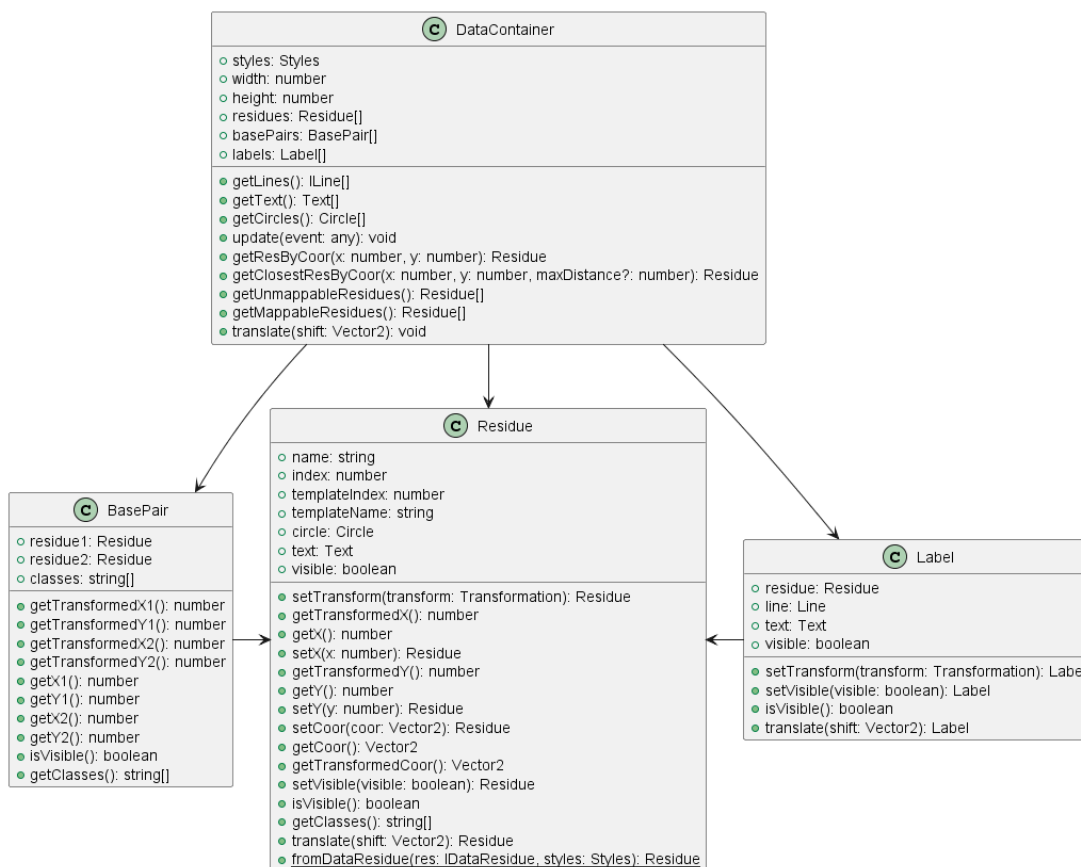
Obrázek 3.1: Interface pro vstupní data

⁵<https://www.uml.org/>

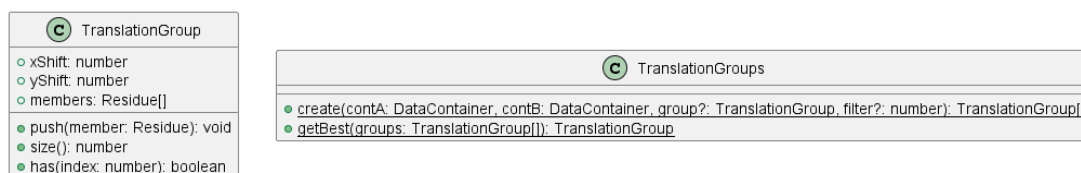
3.3 Objektový návrh



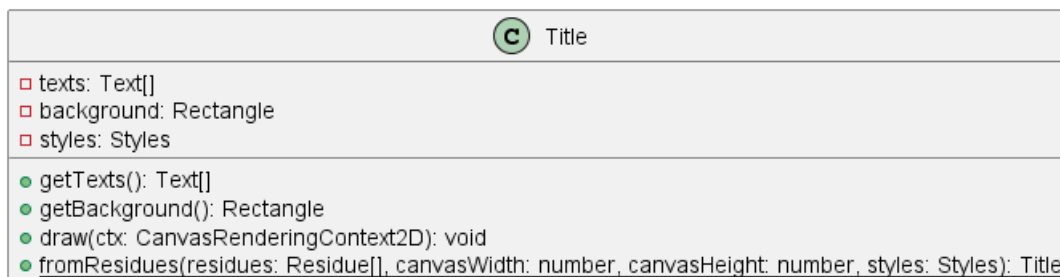
Obrázek 3.2: Interface pro vstupní data




Obrázek 3.3: Interface pro vstupní data



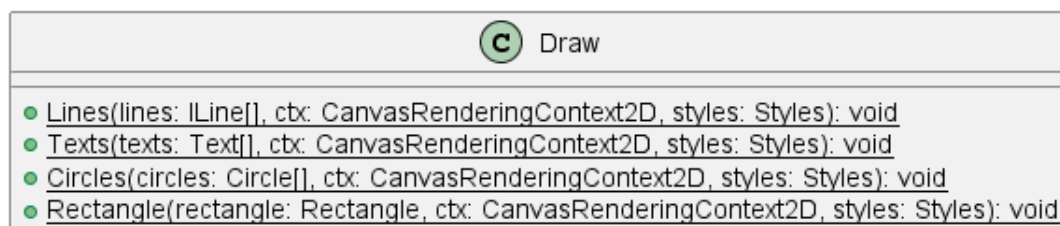
Obrázek 3.4: Interface pro vstupní translationGroups



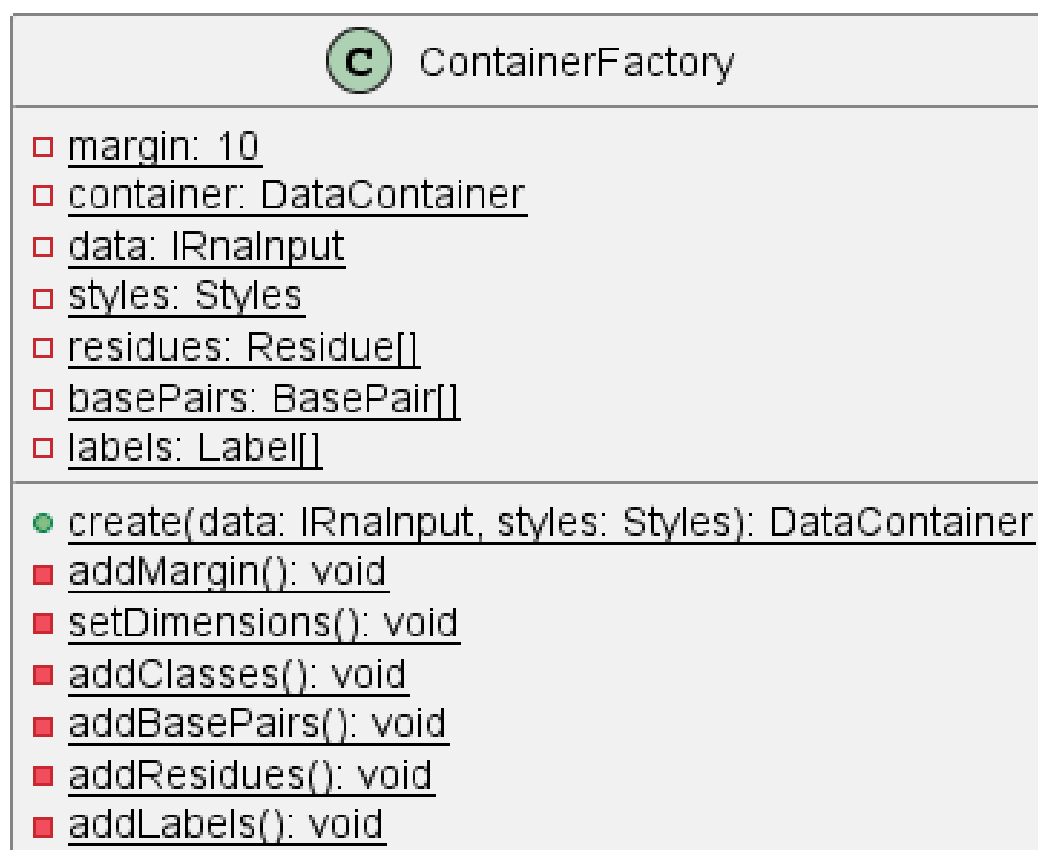
Obrázek 3.5: Interface pro vstupní data

 Styles
<ul style="list-style-type: none"> ❑ default: Map<string, object> ○ styles: Map<string, object> ○ <u>TRANSFORMED CLASS: string</u>
<ul style="list-style-type: none"> ● addFrom(classes: object[]): void ● set(name: string, value: object): void ● get(names: string[]): any ● getProperty(names: string[], property: string): string ● reset(): void ● <u>randomHexColor(): string</u>

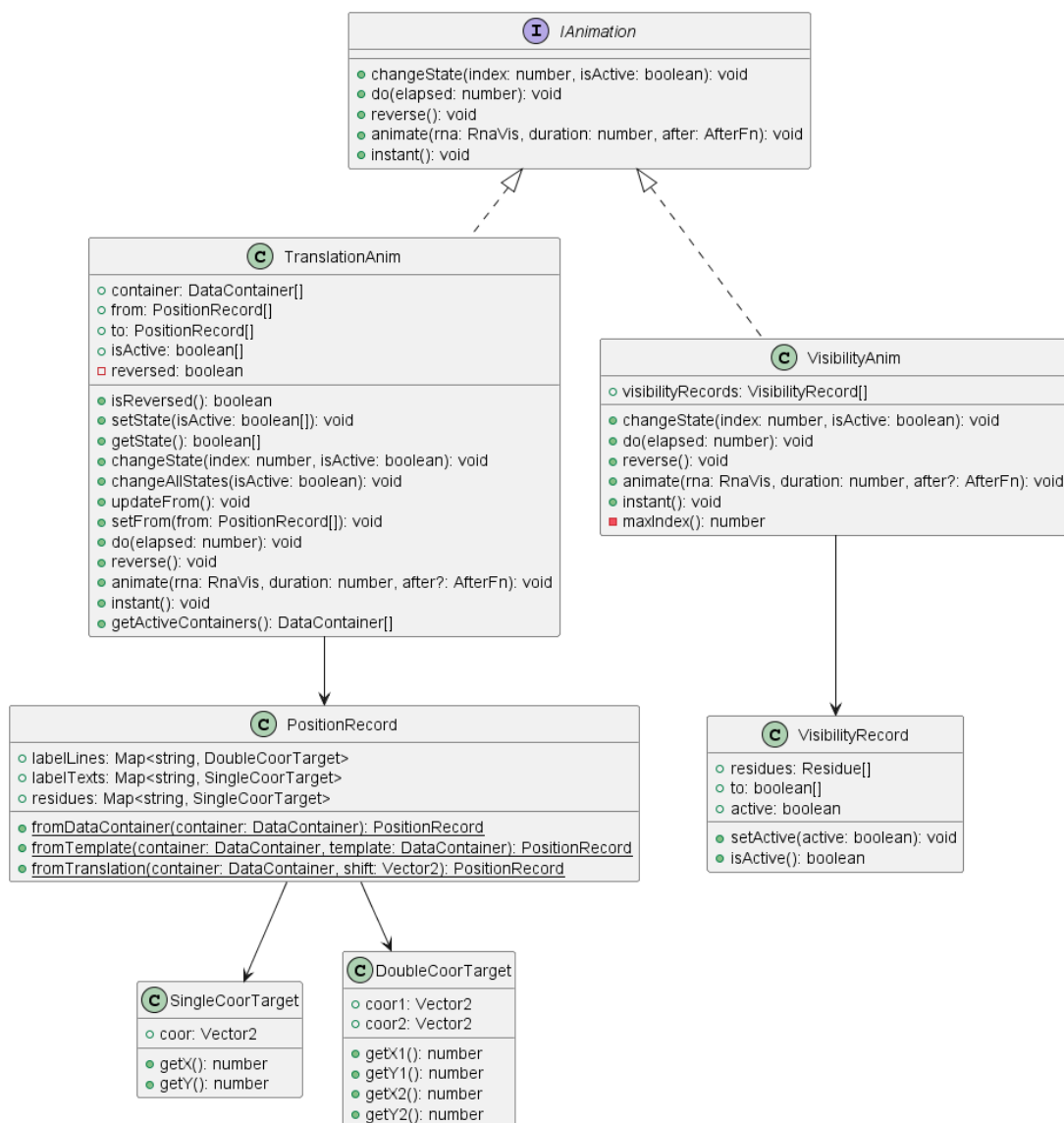
Obrázek 3.6: Interface pro vstupní data



Obrázek 3.9: Interface pro vstupní data



Obrázek 3.10: Interface pro vstupní data



Obrázek 3.11: Interface pro vstupní data

4. Uživatelská dokumentace

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 linear diagramu	4
1.2	Ukázka circular diagramu	4
1.3	Ukázka radiate diagramu	5
3.1	Interface pro vstupní data	10
3.2	Interface pro vstupní data	11
3.3	Interface pro vstupní data	12
3.4	Interface pro vstupní translationGroups	12
3.5	Interface pro vstupní data	12
3.6	Interface pro vstupní data	13
3.7	Interface pro vstupní data	14
3.8	Interface pro vstupní data	14
3.9	Interface pro vstupní data	15
3.10	Interface pro vstupní data	15
3.11	Interface pro vstupní data	16

Seznam tabulek

Seznam použitých zkratek

A. Přílohy

A.1 První příloha