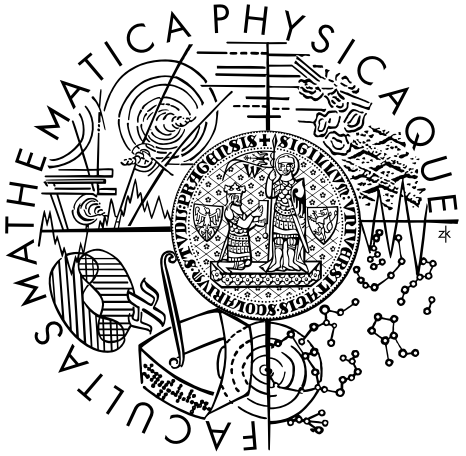
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Rastislav Galvánek

Predikce terciární struktury RNA na základě předlohy

Katedra softwarového inženýrství

Vedoucí bakalářské práce:	RNDr. David Hoksza, Ph.D.
Studijní program:	Informatika
Studijní obor:	Softwarové a datové inženýrství

Praha 2016

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Predikce terciární struktury RNA na základě předlohy

Autor: Rastislav Galvánek

Katedra: Katedra softwarového inženýrství

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

Abstrakt: Práca sa zaoberá návrhom, implementáciou a testovaním nového algoritmu homológnej predikcie terciárnej RNA štruktúry, teda predikcie za pomoci podobnej RNA štruktúry ako vzoru. Zameriava sa na možnosť predikcie dlhých RNA štruktúr v rozumnom čase a dobrej presnosti. Algoritmus je založený na skopírovaní konzervovaných častí štruktúry vzoru a následnom dopredikovaní nekonzervovaných úsekov existujúcim algorimom typu ab initio. Práca je rozdelená na štyri kapitoly. Prvá obsahuje základné informácie o RNA a o jej význame, druhá popisuje spôsoby predikcie RNA a súčasne dostupné metódy predikcie, v tretej je predstavený algoritmus a vo štvrtej sú vyhodnotené výsledky vyvinutého algoritmu.

Klíčová slova: RNA predikcia homológna terciárna

Title: Template-based RNA tertiary structure prediction

Author: Rastislav Galvánek

Department: Department of Software Engineering

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

Abstract: The thesis deals with proposal, implementation and testing of a new algorithm for homologous tertiary RNA structure prediction, which means using a structure similar to the template. It focuses on the possibility of large RNA structures prediction in reasonable time and precision. The algorithm is based on copying conserved parts of template structure into the target structure. The unconserved parts are then predicted by an existing ab initio algorithm. The thesis is divided into four chapters. The first one contains basic information about RNA and its importance. The second one describes different ways of RNA prediction and it contains an overwiev of currently available prediction methods. The third one describes the invented algorithm and the fourth one presents achieved results.

Keywords: RNA prediction homologous tertiary

Chcem sa poďakovať vedúcemu práce RNDr. Davidovi Hokszovi,PhD., za cenné rady, odbornú pomoc a konzultácie, ktoré mi poskytol pri vypracovaní bakalárskej práce. Osobitné poďakovanie patrí mojej priateľke za kontrolu pravopisu a gramatiky v bakalárskej práci.

Výpočtové zdroje boli poskytnuté Ministerstvom školstva, mládeže a športu Českej Republiky v projekte CESNET (Project No. LM2015042) a CERIT-Scientific Cloud (Project No. LM2015085) spadajúce do programu Projects of Large Research, Development and Innovations Infrastructures.

Obsah

Úvod

Ribonukleové kyseliny (RNA) sú makromolekuly skladajúce sa z kombinácií štyroch nukleotidov, pričom dosahujú dĺžku od pár desiatok až po tisíce nukleotidov. RNA molekuly zastávajú viacero dôležitých biologických funkcií. Medzi najdôležitejšie funkcie patrí prenos biologickej informácie z DNA, ako aj ich účať na syntéze bielkovín. Majú zložitú 3D štruktúru, ktorej znalosť sa ukázala ako kľúčová pre porozumenie ich funkcie. Existujú experimentálne metódy na určenie štruktúry, ako napríklad röntgenová kryštalografia, ale tieto metódy sú náročné, drahé a pomalé. Hoci za posledné roky urobili veľký pokrok, rozdiel medzi počtom experimentálne získaých 3D štruktúr a počtom známych RNA molekúl sa tak neustále zväčšuje. Získanie primárnej štruktúry (sekvencie nukleotidov) je dnes už dobre zvládnuté a platí, že 3D štruktúra RNA je určená primárnou sekvenciou RNA. Preto sa skúmajú možnosti, ako 3D RNA štruktúru získať výpočetnými metódami (predikovať) zo znalosti primárnej sekvencie.

Problém predikcie štruktúr RNA sa však tiež ukazuje ako veľmi zložitý, a v súčasnosti stále neexistujú metódy, ktoré by dokázali spoľahlivo a dostatočne presne napredikovať ľubovolnú 3D štruktúru RNA. Existujú rôzne druhy prístupov k predikcii, pričom jedným z nich je homológne modelovanie, čo znamená modelovanie neznámej RNA štruktúry pomocou inej experimentálne získanej RNA štruktúry (vzoru). Tento spôsob sa ukázal ako účinný pri predikovaní bielkovín, ktoré je v súčastnosti lepšie preskúmané, pričom veľa vecí je spoločných s predikciou RNA štruktúr. Po rozvoji experimentálnych metód získavania 3D RNA štruktúry a rozrastajúcej sa databázy experimentálne získaných 3D RNA štruktúr, ktoré môžu slúžiť ako vzory pre predikciu, sa podľa nás tento zatiaľ veľmi slabo preskúmaný smer stáva nádejným pre predikciu dlhých štruktúr RNA.

Cieľom tejto práce je vymyslieť nový algoritmus na predikciu RNA štruktúr, založený na princípe homológneho modelovania. Ten následne naimplementujeme a otestujeme. Algoritmus by mal byť schopný predikovať aj dlhé RNA štruktúry. Dodávame, že väčšina existujúcich algoritmov nedokáže s rozumnou presnosťou predikovať dlhé RNA štruktúry.

1. Úvod do RNA

Pretože sa táto práca venuje problematike vizualizácie RNA v priestore, na začiatok je nutné uviesť základné informáciame o RNA. Najskôr sa oboznámime s funkciami RNA, rôznymi typmi RNA, úrovňami, akými sa dá na štruktúru nahliadnuť, dôvodmi, prečo je štruktúra dôležitá, a nakoniec prečo má predikcia ako spôsob získania štruktúry zmysel.

1.1 Čo je to RNA

Ribonukleová kyselina je popri DNA a bielkovinách jedna z troch hlavných makromolekúl, ktoré sú základom pre všetky známe formy života.

Je tvorená vláknami nukleotidov, typicky sa vyskytuje vo forme jednovláknovej (jednoretiazkovej) molekuly, na rozdiel od dvojvláknovej DNA. Treba dodať, že pri niektorých vírusoch sa môže vyskytovať aj dvojvláknová RNA. RNA je tvorená štyrmi typmi nukleotidov (báz). Sú to adenín (A), guanín (G), cytozín (C) a uracil (U), DNA namiesto uracilu obsahuje tymín (T). RNA má vo svojej cukor-fosfátovej väzbe cukor - ribózu, čím sa líši od DNA, ktorá má deoxyribózu. Väčšia reaktivita ribózy spôsobuje, že RNA môže mať väčšie množstvo priestorových usporiadaní a zastávať viac funkcií než stabilnejšia DNA. Nukleotidy sa spolu viažu vodíkovými väzbami, pričom preferencie vo väzbách sú medzi cytozínom a guanínom, adenínom a uracilom, a medzi guanínom a uracilom. Vlákno RNA sa typicky páruje samo so sebou, čo v kombinácií s tým, že sa može skladať až z niekoľko tisíc nukleotidov často znamená zložitú štruktúru v priestore. Pri DNA platí, že jej dve vlákna sa párujú spolu na princípe komplementarity a vytvárajú špirálovitú štruktúru.

1.2 Typy RNA a ich funkcie

RNA môže zastávať v bunke rôzne funkcie. Známa je hlavne jej funkcia popisovaná v centrálnej dogme molekulárnej biológie. Centrálna dogma hovorí o prenose genetickej informácie medzi biopolymérmi (DNA, RNA, bielkoviny). Obvyklá cesta je replikácia DNA, transkipcia DNA do RNA a translácia z RNA do bielkovín. Nikdy však nenastáva prenos informácie z bielkovín späť do RNA, alebo do DNA. (Crick, 1970)

Prehľad rôznzch typov RNA:

- kódujúca
  - mediátorová RNA (mRNA)
- nekódujúca (ncRNA)
  - ribozomálna RNA (rRNA)
  - transferová RNA (tRNA)
  - funkcionálna RNA (fRNA)
  - mikro RNA (miRNA)

- malá interferujúca (small interfering) RNA (siRNA)
- jadrová (nuclear) RNA (snRNA)
- jadierková (nucleolar) RNA (snoRNA)
- vírusová RNA (vRNA)
- dlhá nekódujúca RNA (lncRNA)
- a iné ...

Kódujúca - mediátorová RNA (mRNA) je prepisovaná podľa sekvencie DNA a neskôr využitá ako vzor pre syntézu bielkovín (proteínov). Proces prepisu informácie z DNA do mRNA sa nazýva transkripcia. Najprv je prepisovaná prekurzorová mRNA (pre-mRNA), ktorá obsahuje aj nekódujúce oblasti - intróny, ktoré sú ďalším spracovaním (alternatívnym splicingom (zostrihom)) z molekuly odstránené a vzniká zrelá (mature) mRNA. mRNA potom putuje do ribozómov, kde prebieha už spomenutá syntéza bielkovín. (Cooper a Hausman, 2004) Sekvencie troch nukleotidov kódujú 20 aminokyselín, z ktorých sú vytvárané bielkoviny. Sekvencia troch nukleotidov (kodón) kóduje práve jednu aminokyselinu.

Kódujúcej RNA sa v bunke nachádzajú približne 2 %. Zvyšných 98 % je označovaných ako nekódujúca RNA. Nekódujúca RNA je taktiež prepísaná z DNA, nekóduje žiaden proteín, zastáva však iné funkcie, ktoré popisujeme nižšie.

Transferová RNA (tRNA) sa väčšinou skladá z približne 80 nukleotidov a vo veľkom počte sa nachádzajú v cytoplazme bunky. Na jednom jej konci sa nachádza antikodón, na opačnom aminokyselina, pričom platí, že príslušnému antikodónu prináleží unikátny typ aminokyseliny. Jej funkcia je naviazať sa na príslušný kodón v ribozóme a pripojiť aminokyselinu na koniec rastúceho polypeptidového reťazca, z ktorého na konci procesu vznikne bielkovina. Tento proces sa nazýva translácia.

Ribozomálna RNA (rRNA) je najčastejšie sa vyskytujúcou molekulou RNA v bunke. Tvorí podstatnú časť ribozómu - veľkú a malú podjednotku, pričom jej veľkosť môže byť viac ako 5000 nukleotidov. (Holzel a kol., 2010) Taktiež napomáha katalitickej aktivite ribozómu. Ribozómy sa viažu na mRNA a vykonávajú syntézu bielkovín. Na jednu mRNA štrukúru môže byť v jednom momente pripojených viacero ribozómov.

Niektoré typy RNA plnia regulačnú funkciu. Napríklad mikro RNA (miRNA) zabraňuje procesu translácie mRNA tým, že sa na ňu naviaže. snRNA sa podieľa na procese splicingu(odstránenie intrónov). snoRNA hrá úlohu pri modifikácii ostatných typov RNA - hlavne rRNA, tRNA a snRNA.

vRNA slúži niektorým vírusom na uchovávanie ich genetickej informácie, môže byť jednovlákonová alebo dvojvláknová.(Patton, 2008)

Poslednou spomenutou kategóriou budú lncRNA. Tieto štruktúry majú veľkosť viac 200 nukleotidov a jeden zo známych zástupcov je gén XIST, ktorý sa uplatňuje pri procese inaktivácie chromozómu X. (Rinn a Chang, 2012)

1.3 Spôsoby náhľadu na štruktúru

Molekula RNA je v skutočnosti súbor atómov usporiadaných v 3D priestore. Existujú však rôzne úrovne pohľadov na molekuly - od zjednodušených, ktoré

zachytávajú menej informácií o molekule ako obsahuje spomínané usporiadanie atómov v priestore, až po také, ktoré presahujú hranice jednej molekuly a zachytávajú vzťahy medzi ňou a inými molekulami. Rozlišujeme nasledujúce úrovne štruktúry RNA

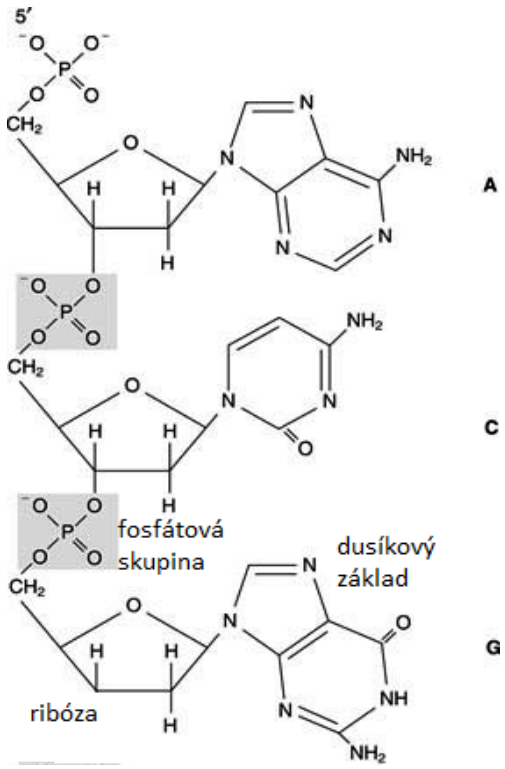
- Primárna
- Sekundárna
- Terciárna
- Kvarciárna

Primárna štruktúra RNA je zoradená lineárna sekvencia nukleotidov, ktoré sú spolu prepojené väzbami.

Nukleotidy obsahujú:

- Dusíkový základ (A, G, C, U)
- 5-uhlíkový cukor - ribózu
- fosfátové skupiny

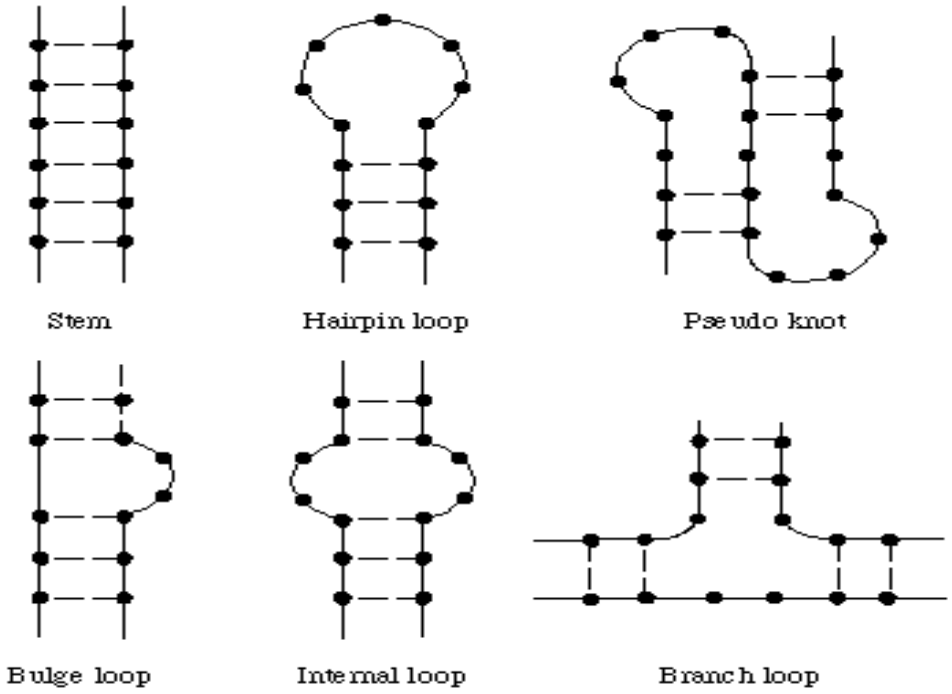
V tejto práci sa nebudeme zaoberať chemicko-biologickými detailami väzieb. Postačí nám znalosť, že k primárnej štruktúre budeme pristupovať ako ku sekvencii písmen, ktoré kódujú jednotlivé nukleotidy. V tejto sekvencii nás bude zaujímať iba poradové číslo a typ nukleotidu.



Obrázek 1.1: Detail primárnej štruktúry sekvencie ACG

Sekundárna štruktúra RNA popisuje vodíkové väzby medzi nukleotidmi. Dva nukleotidy, ktoré spolu tvoria vodíkovú väzbu, sa označujú ako spárované bázy

(base pair). Dá sa povedať, že sekundárna štruktúra je vlastne zoznam spárovaných báz. Tieto spárované bázy vytvárajú špecifické podštruktúry, ako špirála (helix), slučka (loop), pseudouzol (pseudoknot), vlásenka (hairpin loop), vnútorná slučka (internal loop), rozvetvená slučka (branch loop), stopka (stem), bulge loop a iné. Podľa sekundárnej štruktúry je možné zistiť, ako bude bližšie vyzerať terciárna štruktúra, pretože nám ukazuje, ktoré časti sekvencie sa v trojdimenzionálnom priestore budú nachádzať blízko pri sebe. V tejto práci nebudeme ďalej so sekundárnou štruktúrou molekuly pracovať, ani ju nijako využívať.



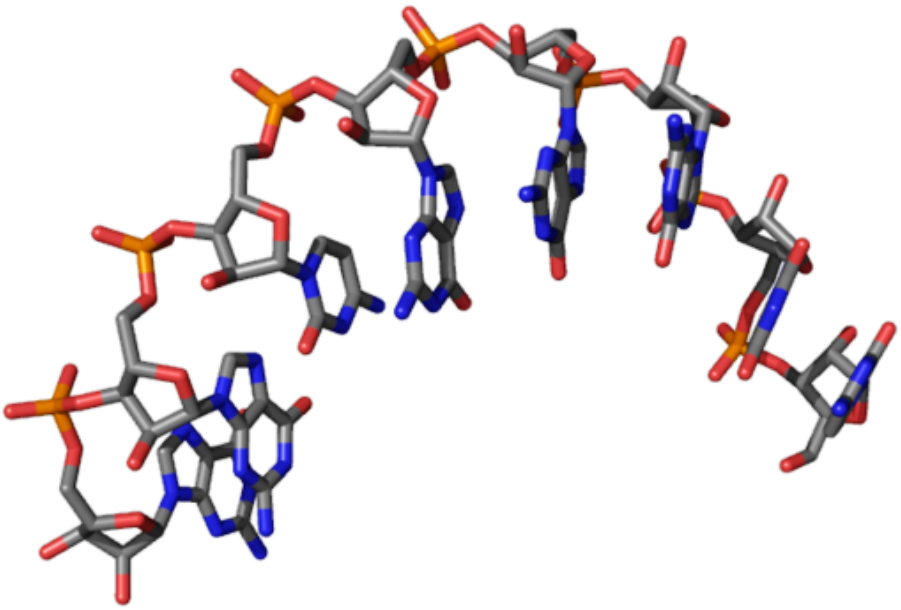
Obrázek 1.2: Vybrané podštruktúry sekundárnej štruktúry

Terciárna štruktúra RNA určuje trojdimenzionálnu štruktúru molekuly v priestore. Je uchovávaná ako množina trojdimenzionálnych súradníc všetkých atómov molekuly. Dáva najpresnejšiu a najužitočnejšiu informáciu o molekule zo všetkých spomínaných úrovní a takisto je najobtiažnejšie ju získať. V tejto práci sa venujeme práve predikcii terciárnej štruktúry, za pomoci inej terciárnej štruktúry, pričom vychádzame z primárnej štruktúry.

Kvarciárna štruktúra RNA popisuje vzťahy medzi celými molekulami RNA - napríklad interakcie medzi jednotlivými molekulami RNA v ribozómoch (Noller, 1984) a taktiež vzťahy medzi RNA a molekulami bielkovín.

1.4 Význam terciárnej štruktúry

Funkcia molekuly do veľkej miery súvisí s jej terciárnou štruktúrou. Zmena terciárnej štruktúry, ktorá vedie ku strate pôvodnej funkcie molekuly, sa nazýva denaturácia. Nasledujúca citácia zdôrazňuje význam 3D štruktútry molekuly: "Štrukturálne vlastnosti RNA sú veľmi dôležité pre jej biologickú funkciu, zahŕňajúc kódovanie (mRNA), prenos genetickej informácie a katalitickú aktivitu



Obrázek 1.3: Detail terciárnej štruktúry

štruktúry. Správne fungovanie RNA vyžaduje vytvorenie zložitej trojdimenzionálnej (3D) štruktúry." (Felden, 2007)

1.5 Získavanie terciárnej štruktúry a zmysel jej predikcie

Získať sekvenciu nukleotidov z molekuly RNA sa stalo možné a relatívne dostupné už v minulom tisícročí. Experimentálne získať 3D štruktúru však ostávalo veľmi ťažké (v roku 2000 boli známe len štruktúry 155 RNA molekúl). Po roku 2000 rapídne stúpa počet experimentálne získaných terciárnych štruktúr (2495 známych štruktúr na začiatku roka 2015 (Westhof, 2015))(vďaka röntgenovej kryštalografii, pokroku v oblasti výpočetných systémov a software) a je možné spracovať aj veľké RNA štruktúry (ako veľké označujeme sekvencie s viac ako 100 nukleotidmi (Holbrook, 2008)). Napriek tomuto pokroku je experimentálna cesta stále veľmi časovo náročná a drahá a množstvo nevyriešených sekvencií naďalej rastie. Preto sa začali skúmať možnosti, ako čo najefektívnejšie štruktúry predikovať.

Spôsoby získavania terciárnej štruktúry:

- Experimentálne metódy
  - X-ray crystallography (Röntgenová kryštalografia)
  - Single-particle cryo-electron microscopy (SPCEM)
  - NMR spectroscopy (NMRS)
- Typy metód výpočetnej predikcie

- Homológne modelovanie
- De novo

Je nutné dodať, že vyššie uvedený zoznam nie je kompletný, a existujú ďalšie metódy, ktoré sú schopné priblížiť rôzne informácie o štruktúre skúmanej molekuly. Metódam výpočetnej predikcie bude venovaná nasledujúca kapitola práce.

Röntgenová kryštalografia bola prvýkrát použitá v roku 1912 na získanie štruktúry anorganických látok. V súčasnosti sa pomocou nej dajú získať štruktúry dlhých sekvencii RNA s rozlíšením lepším ako 2.0 Å. Princíp je nasledovný: prvý a často najzložitejší krok je získanie vhodného kryštálu študovanej molekuly. V druhom kroku je kryštál ožarovaný monochromatickým RTG žiarením, pričom sa na základe zachytenia deformácie lúča určia polohy jednotlivých elektrónov. V poslednom kroku sa zo získaných dát a ďalších chemických vlastností vytvorí presný atómový model štruktúry. Táto metóda je v súčasnosti najviac používaná a dosahujú sa pomocou nej najlepšie experimentálne výsledky. (Smyth a Martin, 2000)

Single-particle cryo-electron microscopy metóda má výhodu v tom, že nie je limitovaná veľkosťou molekuly, na druhej strane rozlíšenie získaných výsledkov sa pohybuje medzi 10 Å a 13 Å.

NMR spectroscopy síce poskytuje presné informácie o polohe jednotlivých atómov, ale je limitovaná dĺžkou sekvencie skúmanej molekuly (do 100 nukleotidov). (Felden, 2007)

2. Existujúce metódy predikcie terciárnej štruktúry

Cieľom predikcie RNA štruktúr je z primárnej sekvencie získať výpočetným spôsobom taký terciárny model štruktúry, ktorý by bol svojou presnosťou porovnateľný s experimentálne získanou štruktúrou. V ďalšom texte budeme pod pojmom štruktúra myslieť terciárnu štruktúru RNA a pod pojmom sekvencia primárnu sekvenciu RNA. V tejto kapitole sa zameriame na odlišné prístupy k predikcii ab initio algoritmami a algoritmami homológneho modelovania.

Na začiatok sa oboznámime s jedným teoretickým poznatkom, ktorý hovorí, že štruktúra je jedinečne určená sekvenciou - teda znalosť unikátnej sekvencie plne postačuje na to, aby bolo možné predikovať unikátnu štruktúru. (Krieger a kol., 2003)

V úvode ešte upozorníme, že časť článkov, na ktoré odkazujeme v tejto kapitole, pojednáva o predikcii proteínov. Z biologického hľadiska však skladanie (folding) proteínov a RNA prebieha veľmi podobným spôsobom, teda poznatky uvedené v takýchto článkoch sú relevantné aj pre RNA. (Moore, 1999)

2.1 Homológne modelovanie

Metóda okrem predpokladu unikátnosti RNA štruktúry uvedeného v úvode zakladá aj na tom, že vďaka povahe evolučného procesu je terciárna štruktúra viac stabilná a mení sa oveľa pomalšie než jej odpovedajúca sekvencia. Rolu tu zohráva fakt, že molekuly, ktroré majú podobnú štruktúru, majú aj podobnú funkciu. (Krieger a kol., 2003)

Princíp homológneho modelovania spočíva v tom, že molekulu nemodelujeme iba z jej primárnej sekvencie, ale použijeme inú molekulu RNA nazývanú „vzor“, ktorej štruktúru poznáme (bola experimentálne získaná). Molekulu, ktorej štruktúru predikujeme, budeme nazývať „cieľ“.

Ďalším krokom je typicky určenie homológnych častí sekvencií (časti, ktoré môžeme preniesť zo vzoru do cieľa). Homológne časti budeme nazývať konzervované úseky a typicky sa určujú zarovnaním (alignment) sekvencií cieľa a vzoru. Čím sú si sekvencie podobnejšie, tým viac konzervovaných úsekov bude existovať. Konzervované úseky je potrebné upraviť, rôzne predikčné metódy na to používajú odlišné stratégie. Prirodzene čím viac konzervovaných úsekov vzíde zo zarovnania, tým jednoduchšie bude vytvoriť kvalitný model štruktúry. Vysoká podobnosť však nie je podmienkou pre použitie tohoto druhu algoritmov a teoreticky je možné napredikovať dobrý model, ak je podobnosť medzi cieľom a vzorom nula. (Rother a kol., 2011)

V ďalšej fáze sa do predikovaného modelu skopírujú konzervované úseky štruktúry z vzoru. Následne je nutné doplniť medzery, ktoré vznikli na miestach, kde sekvencie neboli konzervované. Tu sa prístupy môžu líšiť. Jeden spôsob je použiť databázu častí štruktúr a v nich vyhľadať najlepšie pasujúci chýbajúci úsek. Druhou možnosťou je predikovať medzery metódou de novo.

Týmto je model hotový, môže však ešte prebehnúť konečná optimalizácia modelu.

Výhoda tejto metódy je, že sa ňou (za predpokladu vhodného vzoru) dajú predikovať veľké molekuly, v rozumnom čase s dobrou presnosťou.

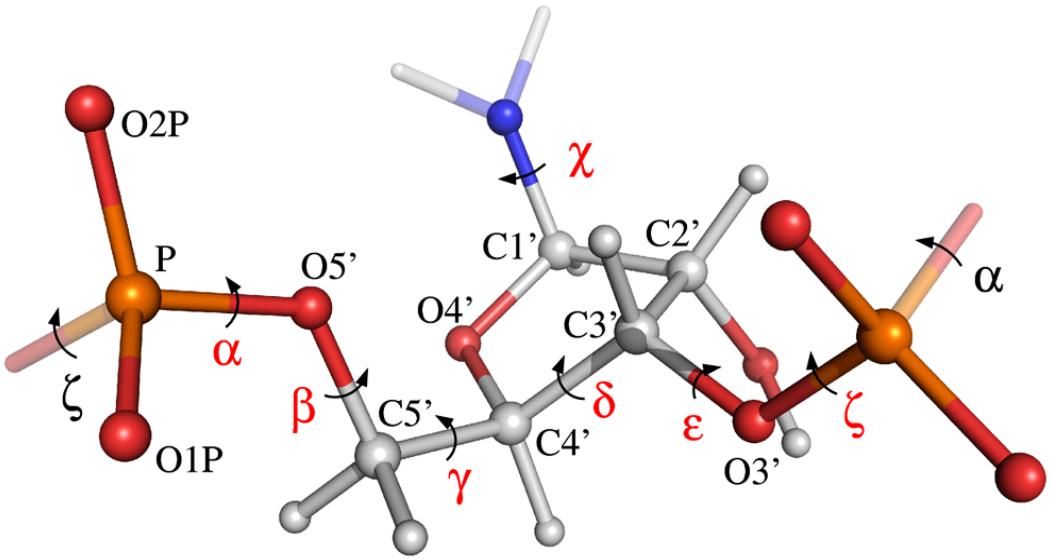
Nevýhodou je, že pri výbere nevhodného vzoru alebo zlom výbere konzervovaných úsekov bude výsledný model veľmi vzdialený skutočnej štruktúre. Ako náročný problém sa taktiež ukazuje doplnenie dlhších nekonzervovaných úsekov.

2.2 De novo predikcia

Algoritmy, ktoré zaraďujeme do skupiny de novo, predikujú terciárnu štruktúru RNA iba z jej primárnej sekvencie nukleotidov. Opierajú sa o fyzikálno-biologické princípy, ktoré riadia skladanie (folding) molekuly v skutočnosti, čo znamená, že voľná energia (množstvo práce, ktorú termodynamický systém dokáže vykonať) vo výslednej molekule je minimálna možná. (Anfinsen, 1973)

Predikcia typicky prebieha tak, že ako prvú potrebujeme energy function (energetická funkcia), ktorá má za úlohu minimalizovať voľnú energiu v predikovanej štruktúre (hľadá globálne minimum riešeného problému). Energy function vedie proces nazývaný conformational sampling, ktorý postupne prechádza rôzne priestorové konfigurácie predikovanej štruktúry. Počas tohoto procesu sa vytvárajú decoys (kandidáti), tí sú ďalej upravovaní a na konci sa z nich vyberie ten s minimom voľnej energie.

De novo metódy sa dajú ďalej rozdeliť na dve podkategórie. Prvá skupina sú „ab initio“ metódy, v ktorých conformational sampling pridáva, modifikuje a odoberá jednotlivé nukleotidy. Druhá skupina sa nazýva „knowledge-based“ metódy, ktoré pracujú s knižnicou krátkych úsekov experimentálne získaných štruktúr, pričom sa snažia nájsť vhodný úsek a pridať ho do predikovaného modelu. Do skupiny knowledge-based patrí aj metóda FARFAR, ktorú v našej práci využívame. (Das a kol., 2010)



Obrázek 2.1: Príklad reprezentácie fragmentu RNA štruktúry (7 uhlov medzi atómami určuje tvar nukleotidu v priestore) (Frellsen a kol., 2009)

V praxi sa hľadajú spôsoby, ako zmenšiť konformačný priestor (možné usporiadania štruktúry) a efektívne v ňom vyhľadávať. To sa rieši napríklad zjednoduše-

nou reprezentáciou štruktúry (výpočet neprebieha so všetkými atómami, nazýva sa to coarse-grained), heuristickými metódami, alebo fixovaním stupňov voľnosti. Ďalším problémom, s ktorým sa de novo metódy stretávajú, je, že optimalizačný problém, ktorý rieši energy function, obsahuje veľa lokálnych miním, ktoré môžu funkciu zmiasť a celý algoritmus skončí predčasne.

Súčasné implementácie de novo algoritmov dokážu dostatočne presne predikovať kratšie sekvencie RNA. Pri dlhých sekvenciách sa stávajú časovo veľmi náročné.

2.3 Prehľad software na predikciu RNA štruktúr

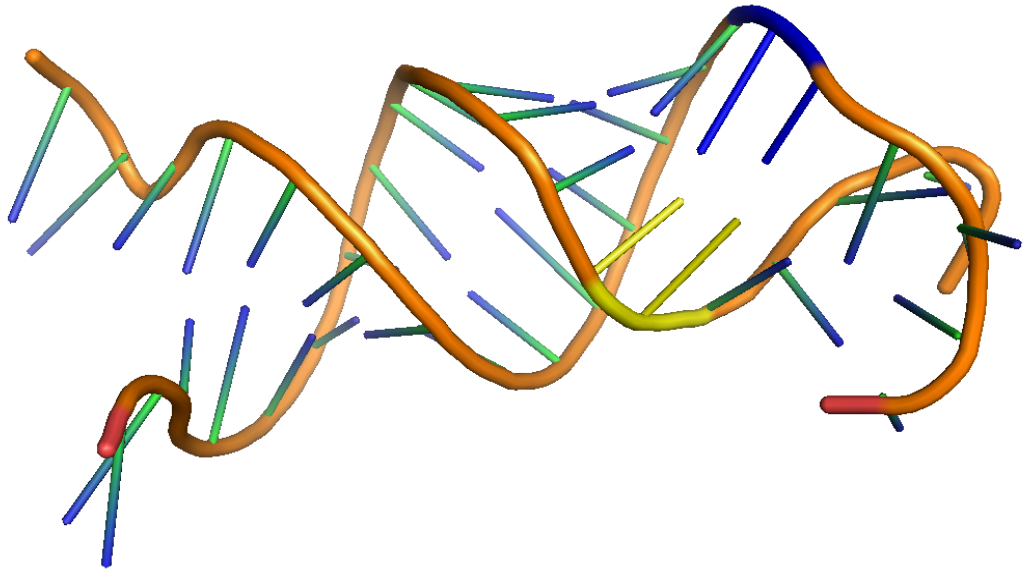
V nasledujúcej tabuľke nájdeme prehľad súčasne dostupného software na predikciu RNA štruktúr. Všetky uvedené metódy sú dostupné ako webserver, alebo zdrojový kód. Zdôrazňujeme, že z nižšie uvedených metód je iba ModeRNA schopný predikovať molekuly RNA, ktoré obsahujú viac ako 500 nukeotidov, je rovnako založený na princípe homológneho modelovania ako náš algoritmus, ale určovanie štruktúry nekonzervovaných úsekov používa knižnicu fragmentov, z ktorých vyberá najlepšie pasujúci. Takisto v súčastnosti nevieme o žiadnych iných automatizovaných metódach, schopných takéto dlhé štruktúry predikovať.

Tabulka 2.1: Prehľad software na predikciu RNA

Názov	Popis	Článok
BARNACLE	Python knižnica pre predikciu RNA štruktúr, zameranie na pravdepodobnostný model konformačného priestoru	(Frellsen a kol., 2009)
FARFAR	automatizovaná knowledge based de novo predikcia RNA, implementované v Rosetta frameworku	(Das a kol., 2010)
MC-Fold MC-Sym Pipeline	MC-Fold najskôr predikuje sekundárnu štruktúru, podľa ktorej potom MC-Sym vytvára štruktúru terciárnu	(Parisien a Major, 2008)
NAST	knowledge-based modelovanie RNA štruktúry, využívajúce coarse-grained model štruktúry	(Jonikas a kol., 2009)
RNAComposer	knowledge-based automatizovaná predikcia štruktúry RNA (do 500 nukleotidov) s využitím sekundárnej štruktúry	(Popenda a kol., 2012)
iFoldRNA	de novo predikcia RNA štruktúry, využívajúca coarse-grained model štruktúry	(Sharma a kol., 2008)
ModeRNA	homológne modelovanie RNA štruktúry	(Rother a kol., 2011)

3. Algoritmus

V tejto kapitole detailne vysvetlíme ako funguje náš algoritmus na predikciu terciárnej štruktúry RNA. Algoritmus môžeme zaradiť do skupiny algoritmov homológneho modelovania, čo znamená, že na základe sekvencie cieľa a štruktúry vzoru predikuje štruktúru cieľovej molekuly. Pripomíname, že tento postup je založený na tom, že podobné sekvencie sa mapujú na podobné terciárne štruktúry, ktoré bývajú konzervované vo väčšej miere ako sekvencie, a teda drobné rozdiely v sekvencii zvyčajne neznamenajú veľké zmeny v štruktúre. Ak však štruktúra nie je v priestore izolovaná, je ovplyvnená interakciami medzi blízkymi reziduami, ktoré pritom v sekvencii môžu byť od seba ľubovolne vzdialené. 3.1



Obrázek 3.1: Príklad vzájomných väzieb nukleotidov, ktoré sú od seba v sekvencii vzdialené. Modrý úsek je tvorený nukleotidmi s identifikačnými číslami 220 a 221, pričom žltý úsek tvoria nukleotidy 156 a 157.

Treba dodať, že primárne sekvencie bývajú uložené v textových súboroch s príponou „fasta“, a terciárna štruktúra v textových súboroch s príponou „pdb“. 3.2 Všetky experimentálne získané makromolekulárne štruktúry je možné nájsť na webových stránkach „The Protein data bank“ (Berman a kol., 2000, http://www.rcsb.org/pdb/), odkiaľ pochádzajú všetky sekvencie a štruktúry použité v tejto práci.

3.1 Stručný popis dôležitých častí algoritmu

V tejto sekcii stručne vysvetlíme všetky dôležité myšlienky algoritmu, detailom a použitému software tretích strán sa budeme venovať v ďalších sekciách.

Na vstup algoritmus dostane primárnu sekvenciu cieľa aj vzoru a terciárnu štruktúru vzoru. Ako prvé prebehne zarovnanie (alignment) sekvencii vzoru a cieľa. Následne je získané zarovnanie modifikované algoritmom posuvného okienka (sliding window) pre lepšiu identifikáciu konzervovaných úsekov. Tento krok má

```
ATOM     43  P     U A 852      59.100 104.889  40.931  1.00 59.77           P
ATOM     44  OP1   U A 852      60.114 104.778  39.857  1.00 95.90           O
ATOM     45  OP2   U A 852      59.307 104.166  42.207  1.00 95.90           O
ATOM     46  O5'   U A 852      58.894 106.433  41.253  1.00 59.77           O
ATOM     47  C5'   U A 852      58.750 107.357  40.191  1.00 59.77           C
ATOM     48  C4'   U A 852      58.554 108.755  40.713  1.00 59.77           C
ATOM     49  O4'   U A 852      57.222 108.946  41.246  1.00 59.77           O
ATOM     50  C3'   U A 852      59.438 109.202  41.853  1.00 59.77           C
ATOM     51  O3'   U A 852      60.750 109.492  41.397  1.00 59.77           O
ATOM     52  C2'   U A 852      58.709 110.457  42.324  1.00 59.77           C
ATOM     53  O2'   U A 852      59.037 111.600  41.562  1.00 59.77           O
ATOM     54  C1'   U A 852      57.236 110.082  42.099  1.00 59.77           C
ATOM     55  N1    U A 852      56.528 109.770  43.352  1.00 95.90           N
ATOM     56  C2    U A 852      55.847 110.808  43.991  1.00 95.90           C
ATOM     57  O2    U A 852      55.758 111.936  43.528  1.00 95.90           O
ATOM     58  N3    U A 852      55.273 110.470  45.190  1.00 95.90           N
ATOM     59  C4    U A 852      55.295 109.232  45.803  1.00 95.90           C
ATOM     60  O4    U A 852      54.787 109.099  46.920  1.00 95.90           O
ATOM     61  C5    U A 852      55.983 108.211  45.068  1.00 95.90           C
ATOM     62  C6    U A 852      56.556 108.505  43.897  1.00 95.90           C
```

Obrázek 3.2: Príklad reprezentácie nukleotidu v pdb súbore, zľava doprava: meno záznamu, poradové číslo atómu, meno atómu, meno rezidua (uracil), identifikátor reťaze (A), sekvenčné číslo rezidua (852), súradnica X, súradnica Y, súradnica Z, zvyšné parametre nepoužívame.

zabezpečiť, aby sa odstránili krátke izolované úseky reziduí v pôvodnom zarovnaní označené ako konzervované. Tabuľka 3.1 ukazuje všetky konfigurácie, ktoré môžu nastať v zarovnaní.

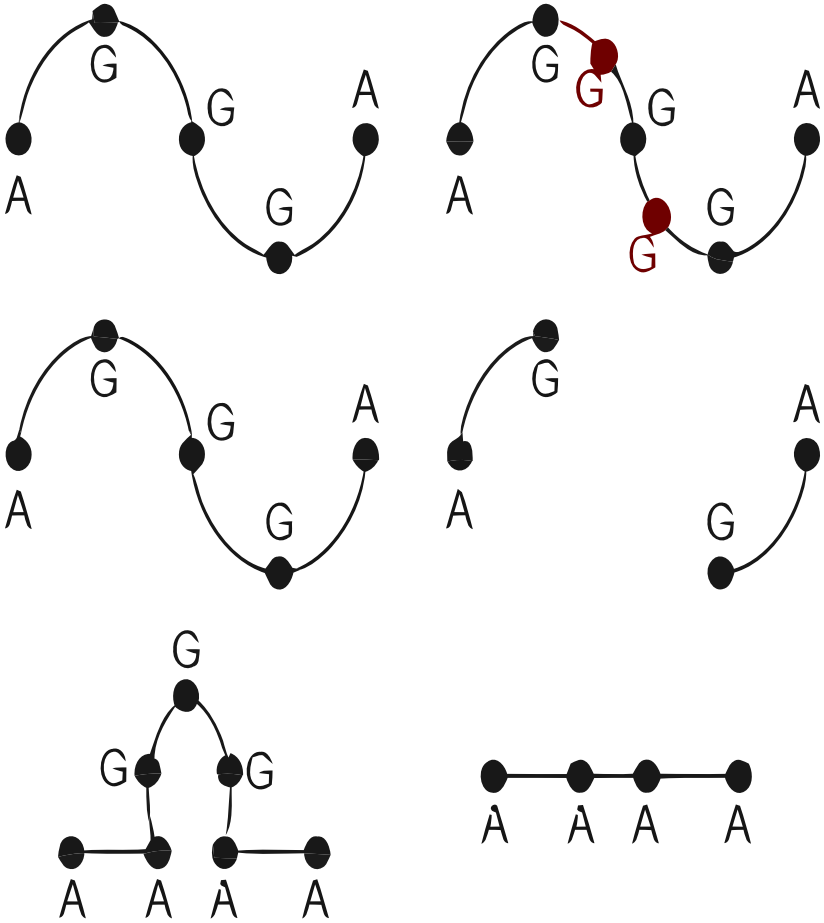
Tabulka 3.1: Príklad rôznych situácii v zarovnaní

Sekvencia	konzervované	nekonzervované	medzera	medzera
cieľ	A	U	-	C
vzor	A	G	A	-

V ďalšom kroku identifikujeme všetky medzery (gaps), ktoré sa v zarovnaní objavili. Tieto môžu spôsobovať problematické situácie pri neskoršej predikcii chýbajúcich úsekov štruktúry. Principiálne môže nastať jedna zo situácií uvedených na obrázku 3.3. Riešime to tak, že nukleotidy z oboch strán medzery označíme ako nekonzervované. Počet nukleotidov takto preznačených je priamo úmerný dĺžke medzery. Problém sa teda snažíme vyriešiť dopredikovaním väčšieho počtu nukleotidov, čo teoreticky dáva predikcii väčší priestor na "spojenie", prípadne „natiahnutie“ určitého úseku štruktúry.

Následne skopírujeme všetky konzervované úseky, ktoré určuje zarovnanie a jeho úpravy. Týmto končí časť predikcie, ktorá sa týkala jednoduchého skopírovania konzervovaných častí štruktúr.

V ďalšej časti pripravujeme vhodný vstup pre algoritmus FARFAR (Das a kol., 2010), ktorý používame na predikovanie chýbajúcich úsekov. Aby bolo možné lepšie porozumieť riešeným problémom, musíme si v krátkosti predstaviť algoritmus FARFAR, ktorý používame. Pomocou FARFAR je možné predikovať štruktúru zo sekvencie, pričom algoritmu môžeme dať na vstup nukleotidy, ktoré už poznáme. V našom prípade sú to konzervované nukleotidy a algoritmus potom predikuje iba chýbajúce úseky.

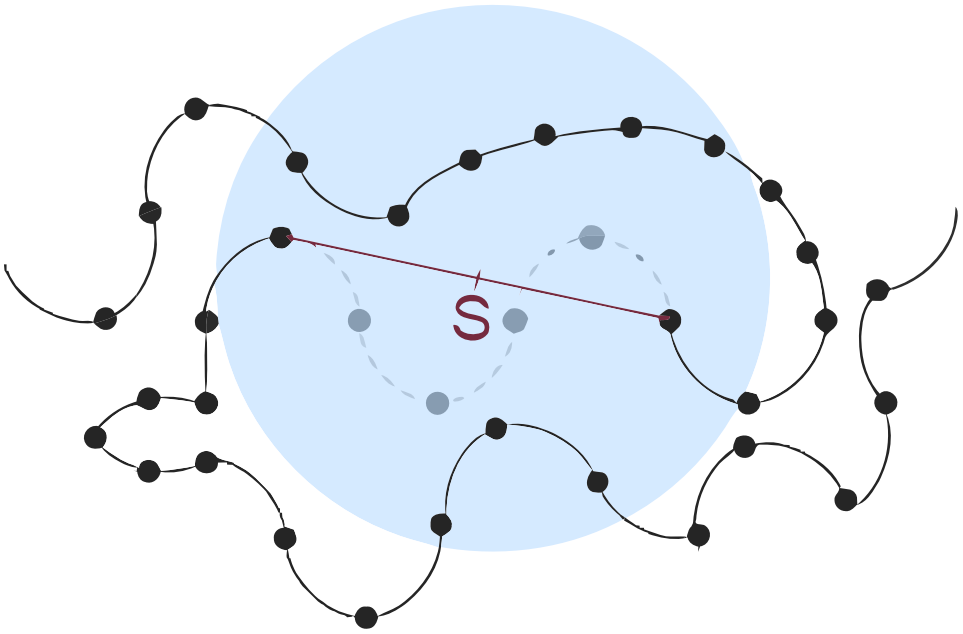


Obrázek 3.3: Problémy s medzerami v zarovnaní znázornené v štruktúre, zhora: prvá situácia zobrazuje vloženie reziduí do súvislej štruktúry, pričom bez toho aby bola modifikovaná okolitá štruktúra nie je možné reziduá pridať. Druhá situácia znázorňuje vymazanie nukleotidov so štruktúry, vzniká medzera medzi nukleotidmi, ktorá nebude ničím nahradená. Tretí prípad ukazuje priaznivý prípad medzery v zarovnaní.

Pri predikcii dlhých štruktúr nemôžeme všetky nekonzervované úseky v predikovanej štruktúre nechať FARFARu predikovať spolu, pretože pri dlhých štruktúrach by to bolo časovo neúnosné (viac v sekcii FARFAR). Navyše vhodným rozdelením úsekov sme schopní nezávisle od seba predikovať jednotlivé časti štruktúry. V prvom rade štruktúru rozdelíme na približne rovnako dlhé úseky. Rozdelenie predikovaných úsekoch na základe sekvencie však nie je úplne vhodné, pretože to, že sú blízko pri sebe dva nukleotidy v sekvencii neznamená, že budú blízko pri sebe aj v štruktúre. Tento problém riešime tak, že dlhšie nekonzervované úseky, z týchto úsekov vyberáme a predikujeme osobitne. Predpokladáme, že správnosť predikcie kratších úsekov problém rozdelenia výraznejšie neovplyvňuje. Krátke nekonzervované úseky sú týmto pripravené na predikciu vo FARFAR.

Vybrané dlhšie nekonzervované úseky budú pripravované tak, že sa na vstup

FARFARu dajú konzervované reziduá, ktoré ležia v guli so stredom ležiacim v strede úsečky, ktorej krajné body tvoria súradnice atómov fosforu krajných (vzhľadom na nekonzervovaný úsek) konzervovaných reziduí. Voľba atómov fosforu je čisto náhodná, a mohla by byť nahradená hocijakým iným atómom nachádzajúcim sa na cukor-fosfátovej chrbtici (backbone) štruktúry. Guľa má taký polomer, aby obsahovala všetky reziduá, ktoré sa môžu dostať do blízkeho kontaktu s ľubovolným reziduom z predikovaného úseku. (obrázok 3.4)



Obrázek 3.4: Znázornenie gule použitej pri príprave predikcie dlhých nekonzervovaných úsekov, slabšie naznačená predikovaná časť štruktúry.

Následne napredikujeme chýbajúce úseky štruktúry. Po skončení predikcie spojíme chýbajúce časti dohromady a získame napredikovaný model štruktúry cieľa.

3.2 Pseudokód

Táto sekcia má za úlohu čo najprehľadnejšie zachytiť poradie vykonávania dôležitých častí algoritmu a pomôcť nám lepšie si vytvoriť obraz o tom, ako algoritmus funguje. Nevenuje sa reprezentácii dát a predstiera, že všetko beží v pamäti počítača, čo síce nie je pravda, ale zvyšuje to prehľadnosť a na princípy algoritmu to nemá vplyv. Premenné začínajú malým písmenom, metódy veľkým. Niekedy je funkcia metódy slovne popísaná. Ak je za premennou „[]“, znamená to, že si pod ňou predstavujeme množinu pdb súborov. Premenné aj metódy sú nazývané tak, aby čo najlepšie popisovali svoju úlohu v algoritme. Informácie o význame parametrov a iných dôležitých detailov sa zas dozvieme v sekcii Detailný popis častí algoritmu, kde sa budeme na metódy z tejto sekcie odkazovať. Na záver uvedieme poznámku k systému názvov. V úsekoch, kde metódy upravujú

zarovnanie (alignment), pojem gap znamená medzeru v zarovnaní, teda pozíciu označenú pomlčkou. Neskôr, po skopírovaní konzervovaných úsekov gap označuje úsek chýbajúcich nukleotidov.

```
Main(fastaTarget, fastaTemplate, pdbTemplate)
{
    CheckTargetMapping
        (fastaTemplate, pdbTemplate)
    alignment := Align
        (fastaTarget, fastaTemplate)
    alignment1 := UseSlidingWindow
        (alignment)
    alignment2 := ProcessGaps
        (alignment1)
    conservedParts := CopyConservedParts
        (alignment2, pdbTemplate)
    mappedConservedParts := MapConservedParts
        (conservedParts, alignment2)
    longParts[] := ProcessLongUnconservedParts
        (mappedConservedParts)
    shortParts[] := ProcessShortUnconservedParts
        (mappedConservedParts)
    predictedParts[] := PredictUnconservedParts
        (longParts[], shortParts[])
    finalModel := ConnectPredictedParts
        (predictedParts)
}

CheckTargetMapping(fasta, pdb)
{
    foreach res in pdb
    {
            if (fasta[res[id]] != res[type])
            ERROR: Input needs manual editing!
            EXIT PROGRAM
    }
}

Align(fastaTarget, fastaTemplate)
{
    string alignment := CallEmbossAln
        (fastaTarget, fastaTemplate)
    return EditAlignmentFormat(alignment)
}

SlidingWindow(alignment, windowLength, minimalLimit)
{
    hL = windowLength DIV 2
    foreach nucleotide in alignment
```

```
46      {
47          check if in [nucleotide[id]-hL, nucleotide[id]+hL]
48          is less than minimalLimit conserved nucleotides
49          if so mark nucleotide as unconserved
50      }
51      return modifiedAln
52  }
53
54  ProcessGaps(alignment, cutoff)
55  {
56      foreach gap in alignment
57      {
58          alignment := mark "cutoff" nucleotides from
59              both sides of the gap as unconserved
60      }
61      return alignment
62  }
63
64  CopyConservedParts(alignment, pdb)
65  {
66      conservedParts = ""
67      foreach nucleotide in pdb
68      {
69          if (IsConserved(nucleotide, alignment))
70              conservedParts += nucleotide
71      }
72      return conservedParts
73  }
74
75  MapConservedParts(conservedParts, alignment)
76  {
77     map nucleotide id's from current state (nc with
78     id x corresponds to x-th nc in template fasta) to
79     state where nc id correctly corresponds to
80     nucleotide in target fasta
81     return modifiedConservedParts
82  }
83
84  ProcessLongUnconservedParts
85      (conservedParts, ncAvarageLength, minLengthGapLimit)
86  {
87      longParts = []
88      foreach gap in conservedParts
89      {
90          if length(gap) <= minLengthGapLimit
91              continue
92          startNc := conserved nucleotide before gap
93          endNc := conserved nucleotide after gap
```

```
94          ph="phosphor"
95          s := FindMiddle(startRes[ph], endRes[ph])
96          p := length(gap) * ncAvarageLength / 2
97          ncsInSphere := all nucleotides inside sphere(p, s)
98          preparedPart := PrepareCorectFormatForFARFAR
99              (ncsInSphere)
100         longParts[] += preparedPart
101     }
102     return longParts[]
103 }
104
105 ProcessShortUnconservedParts
106     (conservedParts, lengthOfSection, maxLengthGapLimit)
107 {
108     createdSections = []
109     createdSections[] = divide "conservedParts" into
110         sections with length of "lengthOfSection"
111     preparedSections = []
112     foreach section in createdSections[]
113     {
114         preparedSections[] += PrepareCorectFormatForFARFAR
115             (section, maxLengthGapLimit)
116     }
117     return preparedSections[]
118 }
119
120 PredictUnconservedParts(lgUnconsPts[], shUnconsPts[])
121 {
122     predictedParts = []
123     allParts := lgUnconsPts[] + shUnconsPts[]
124     foreach input in allParts
125     {
126         predictedParts[] += CallFARFAR(input)
127     }
128     return predictedParts[]
129 }
130
131 ConnectPredictedParts(predictedParts[])
132 {
133     connect all conserved and predicted nucleotides
134     into single file in case of duplicity
135     (they are almost the same) choose only one
136 }
```

3.3 FARFAR

V tejto sekcii si povieme viac o tom, ako pracujeme s FARFAR.

FARFAR dokáže predikovať RNA štruktúru z primárnej sekvencie. Je možné mu na vstup dať okrem sekvencie predikovanej štruktúry aj pdb súbor, ktorý obsahuje ľubovolnú časť nukleotidov danej štruktúry. Tieto nukleotidy pri predikcii nebude modifikovať a dopredikuje iba chýbajúce medzery medzi nimi. Je dokonca možné zvoliť len predikciu určených medzier. Je tu podmienka, že poradie nukleotidov v sekvencii musí odpovedať číslovaniu nukeotidov vo vstupnom pdb súbore.

FARFAR ako výstup generuje kandidátske štruktúry, pričom na konci sa vyberie tá s najnižšou voľnou energiou. Preto platí, že čím väčší počet štruktúr vygenerujeme, tým je pravdepodobnejšie, že sa medzi nimi bude nachádzať štruktúra z nižšou voľnou energiou. FARFAR pracuje nedeterministicky, čo znamená, že pre dva rovnaké vstupy môžu vzniknúť dva rôzne výstupy.

Nakoniec upozorníme, že čím viac nukleotidov je zahrnutých v predikcii (platí aj o tých vo vstupnom pdb), tým dlhšie trvá vygenerovanie jednej kandidátskej štruktúry. Takisto čím dlhší je predikovaný úsek (súvislý úsek z ktorého sa žiadne nukleotidy nenachádzajú vo vstupnom pdb), tým je ho zložitejšie napredikovať, a je väčšia šanca, že bude napredikovaný nesprávne. Toto považujeme za jeden z najväčších problémov nášho algoritmu, pretože presnosť výsledkov FARFAR s rastúcou dĺžkou predikovanej sekvencie klesá. Podľa testov publikovaných autormi algoritmu (Das a kol., 2010) totiž platí, že pri predikcii štrnástich štruktúr dlhých 6 až 13 nukleotidov je priemer RMSD do 2 Å, ale pri predikovaní jedenástich štruktúr dlhých 13 - 23 nukleotidov je priemer RMSD 6.5 Å.

Pre lepšiu predstavu o tom, ako môžu prebytočné nukleotidy vo vstupnom pdb súbore ovplyvniť dĺžku behu FARFAR, uvádzame nasledujúci príklad. Predikovali sme 9 nukleotidov, pričom sme zvolili rozdielne vstupné pdb súbory. Prvý obsahoval všetky konzervované nukleotidy predikovanej štruktúry, druhý obsahoval iba nukleotidy vybrané v metóde ProcessLongUnconservedParts (riadok 84). V prvom prípade, kedy vstupný pdb súbor obsahoval všetkých 129 konzervovaných nukleotidov, trvalo vygenerovanie 10 kandidátskych štruktúr 133 minút. V druhom prípade obsahoval pdb súbor 41 vybraných konzervovaných nukleotidov a vygenerovanie rovnakého množstva kandidátskych štruktúr trvalo 57 minút.

3.4 Detailný popis častí algoritmu

V tejto časti sa venujeme detailom, ktoré neboli spomenuté v prvej sekcii tejto kapitoly. Jedná sa hlavne o parametrizované časti, kde spomenieme aký efekt na priebeh predikcie pri zmene parametrov očakávame. Zmienime sa taktiež o programe Emboss (Rice a kol., 2000) a algoritme FARFAR (Das a kol., 2010) a o dôvodoch ich výberu.

Začneme metódou SlidingWindow (riadok 42), ktorá závisí od dvoch parametrov. Sú to windowLength a minimalLimit. Najprv zadefinujeme pojem konzervovaný nukleotid (tiež konzervovaná pozícia). Je to nukleotid cieľa, ktorý je zarovnaný na rovnaký typ nukleotidu vzoru, a ak ho nevyradíme v ďalšom spracovaní (pojem označiť nukleotid za nekonzervovaný), bude jeho štruktúra skopírovaná zo vzoru do modelu cieľa. Metóda pre každý nukleotid skontroluje, či sa v jeho okolí, ktoré je určené parametrom windowLength, nachádza viac ako minimalLimit konzervovaných nukleotidov. Ak nie, nukleotid označí ako nekonzervovaný(táto zmena neovplyvní parametre výpočtu pre ešte nerozhodnuté nukleotidy). Apliká-

cia algoritmu má teda za následok odstránenie niektorých konzervovaných nukleotidov, ktoré budú neskôr musieť byť dopredikované algoritmom FARFAR. To má zabrániť, aby veľmi krátke izolované konzervované úseky ovplyvnili výsledok predikcie. V prípade, že by bol nukleotid chybne označený ako zarovnaný a nachádzal sa medzi dvomi dlhými nekonzervovanými úsekmi štruktúry, tak by FARFAR predikovanie obochoch nekonzervovaných úsekov bolo ovplyvnené jedným chybne konzervovaným nukleotidom. Keďže FARFAR konzervovanými nukleotidmi nepohybuje, snažil by sa na chybne konzervovaný nukleotid napojiť predikované časti. 3.5

```
#=======================================
#
# Aligned_sequences: 2
# 1: TARGET
# 2: TEMPLATE
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 175
# Identity:     148/175 (84.6%)
# Similarity:   148/175 (84.6%)
# Gaps:          15/175 ( 8.6%)
# Score: 667.5
#
#
#=======================================

TARGET             1 GGCCGACGGAGGCGCGCCCGAGAUGAGUAGGCUGUCCCAUCAGGGGAGGA     50
                     .|||||||||||||||||||||||||||||||||||||||||||||||||
TEMPLATE           1 ---GGACGGAGGCGCGCCCGAGAUGAGUAGGCUGUCCCAUCAGGGGAGGA     47

TARGET            51 AUCGGGGACGGCUGAAAGGCGAGGGCGCCGAAGGGUGCAGAGUUCCUCCC    100
                     ||||||||||||||||||||||||||||||||.|.|||||||||||||||
TEMPLATE          48 AUCGGGGACGGCUGAAAGGCGAGGGCGCCGAAGCGAGCAGAGUUCCUCCC     97

TARGET           101 GCUCUGCAUGCCUGGGGGUAUGGGGAAUACCCAUACCACUGUCACGGAGG    150
                     |||||||.||.||||||||..||||||||||.||||||||||.|||.|
TEMPLATE          98 GCUCUGCUUGGCUGGGGGUGAGGGGAAUACCCUUACCACUGUCGCGAA--    145

TARGET           151 UCUCUCCGUGGAGAGCCGUCGGUC-    174
                         .|.|||||||||||   |
TEMPLATE         146 ------AGCGGAGAGCCGUC---CA    161
```

Obrázek 3.5: Ukážka zarovnania dvoch štruktúr, zelenou sú zvýraznené údaje, ktoré hovoria o miere podobnosti štruktúr. Červenou farbou je zvýraznený po zarovnaní konzervovaný nukleotid, ktorý bude metódou SlidingWindow (veľkosť okienka 10, počet konzervovaných nukleotidov v okienku 5) označený ako nekonzervovaný. Oranžovou sú vyznačené po zarovnaní konzervované úseky, ktoré budú metódou ProcessGap preznačené na nekonzervované úseky (parametrizované tak, že z každej strany medzery sa preznačí polovica dĺžky medzery).

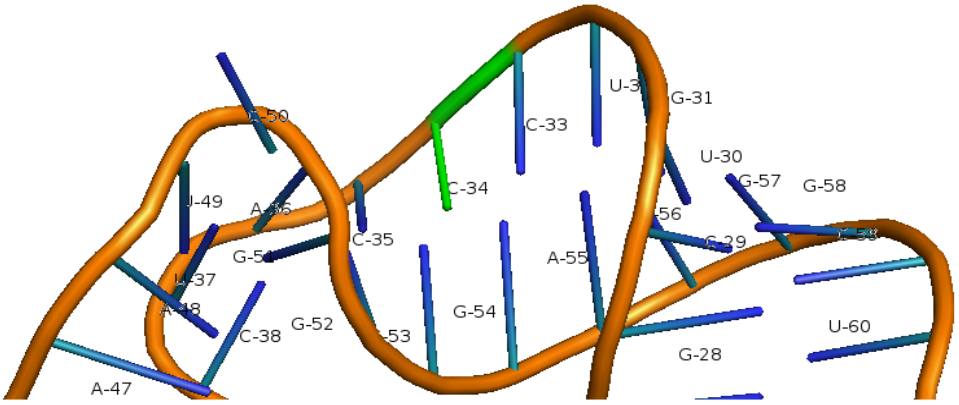
Metóda ProcessGap (riadok 54), ktorá ma parameter cutoff, slúži na upravenie medzier (gaps), ktoré sa nachádzajú v zarovnaní. Pri otázke koľko nukleotidov odstrániť z okolia medzery tak, aby mal FARFAR šancu dotknutý úsek štruktúry čo najlepšie napredikovať, berieme do úvahy dĺžku medzery, ktorú budeme škálovať. Táto metóda môže za istých okolností výsledok predikcie zlepšiť, ale takisto môže vyrobiť zbytočne dlhý nekonzervovaný úsek. Jej účinnosť budeme testovať. Možné situácie sú znázornené na obrázku 3.3.

MapConservedParts (riadok 75) mapuje konzervované nukleotidy na fasta sekvenciu cieľa. Princíp je znázornený na obrázku 3.6.

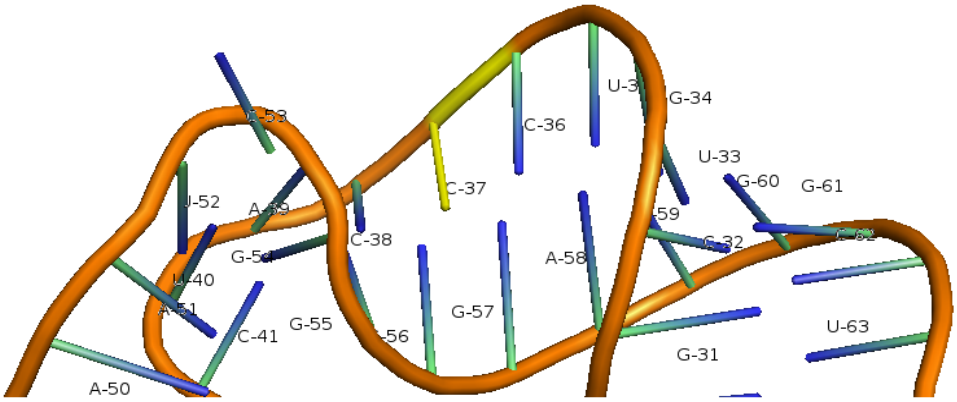
Metóda ProcessLongUnconservedParts (riadok 84) s parametrami minLengthGapLimit a ncAvarageLength najskôr vyhľadá medzery dlhšie ako minLengthGapLimit. Následne pre každý nájdený úsek vytvorí guľu a všetky nukleotidy, ktoré táto guľa obsahuje, skopíruje do vstupného súboru pre FARFAR (obrázok 3.4, pseudokód). Takto zabezpečíme, aby mal FARFAR pri predikovaní chýbajúceho úseku čo najviac informácií o okolitej štruktúre. Znalosť okolitej štruktúry je dôležitá, pretože blízke nukleotidy na seba vzájomne vplývajú (vytváranie chemických väzieb). Takisto sa zmenšia možnosti umiestnenia predikovaného úseku do priestoru, čo môže výrazne zjednodušiť prácu FARFAR predikcií. Tiež sa zabráni možným zrážkam predikovaných úsekov s konzervovanými. Vďaka tomuto kroku môžeme vybraný úsek so všetkými relevantnými informáciami, ktoré nám poskytuje konzervovaná časť štruktúry, predikovať nezávisle od ostatných. Otázku, akej veľkosti guľu potrebujeme, aby obsiahla všetky reziduá s ktorými by predikovaná časť mohla prísť do kontaktu, sme vyriešili nasledovne. Pozorovaním sme zistili, akú vzdialenosť v štruktúre priemerne predstavuje jeden nukleotid. Je to priemerne 2,3 Å + 0,2 Å rozptyl. Táto hodnota sa výnásobí polovicou veľkosti medzery. Dodáme ešte, že zbytočne veľkú guľu nechceme voliť preto, lebo viac nukleotidov vo vstupnom súbore spomaľuje predikciu.

Metóda ProcessShortUnconservedParts (riadok 105) pripravuje krátke nekonzervované úseky pre predikciu FARFAR algoritmom. Hranicu sme volili tak, že od 5 nukleotidov je už úsek dlhý. Pri krátkych úsekoch predpokladáme, že vďaka ich dĺžke nenastávajú situácie, kde by bolo potrebné zaobarať sa ich okolím tak, ako ako to robíme v prípade dlhých medzier. Ak je sekvencia veľmi dlhá, delíme ju na sekcie, ktorých dĺžka je určená parametrom lengthOfSection (rozumný parameter je 300). Tie potom samostatne predikujeme. Prístup použitý pri dlhých medzerách by sa dal bez úpravy použiť aj v prípade krátkych medzier a fungoval by aspoň rovnako dobre. Obmedzenie však v tomto prípade predstavuje organizácia Metacentrum, ktorej výpočetné kapacity využívame na spúšťanie FARFAR. Už pri súčasnom spôsobe samostatného predikovania dlhých úsekov sa v prípade dlhej RNA molekuly vygeneruje niekoľko desiatok predikcií nekonzervovaných úsekov. V prípade, že by sme samostatne predikovali aj krátke úseky, by tento počet mohol narásť niekoľkonásobne, pretože sa v štruktúre typicky vyskytuje viac krátkych nekonzervovaných úsekov ako dlhých. Metacentrum by pri takom množstve vytvorených úloh začal zadržiavať ich spúšťanie, čo by viedlo k nárastu času potrebného pre predikciu. (viac o Metacentre v sekcií Implementácia)

V metóde PredictUnconservedParts (riadok 120) už len spustíme predikciu všetkých úsekov naraz. Tie po skončení spojíme do jednej štruktúry v metóde ConnectPredictedParts (riadok 131). Pri spájaní narazíme na opakovaný výskyt konzervovaných nukleotidov v spájaných súboroch, ktoré vznikli v metóde Pro-




```
TARGET·····1·GGCCGACGGAGGCGCGCCCGAGAUGAGUAGGCUGUCCCAUCAGGGGAGGA·····50¶
·····.||||||||||||||||||||||||||||||||||||||||||||||||||||||||||¶
TEMPLATE····1·---GGACGGAGGCGCGCCCGAGAUGAGUAGGCUGUCCCAUCAGGGGAGGA·····47¶
¶
TARGET·····51·AUCGGGGACGGCUGAAAGGCGAGGGCGCCGAAGGGUGCAGAGUUCCUCCC····100¶
·····|||||||||||||||||||||||||||||||||.|.|||||||||||||||||||||¶
TEMPLATE····48·AUCGGGGACGGCUGAAAGGCGAGGGCGCCGAAGCGAGCAGAGUUCCUCCC·····97¶
¶
TARGET····101·GCUCUGCAUGCCUGGGGGUAUGGGGAAUACCCAUACCACUGUCACGGAGG····150¶
·····|||||||.||.|||||||..||||||||||.||||||||||.||.|··¶
TEMPLATE····98·GCUCUGCUUGGCUGGGGGUGAGGGGAAUACCCUUACCACUGUCGCGAA--····145¶
¶
TARGET····151·UCUCUCCGUGGAGAGCCGUCGGUC-····174¶
·····.|.||||||||||···|·¶
TEMPLATE····146·------AGCGGAGAGCCGUC---CA····161¶
```



Obrázek 3.6: Obrázok zhora dole: Prvá je zobrazená časť template štruktúry, každý nukleotid má pri sebe poznámku s jeho typom a identifikátorom. Druhé je zobrazené celé zarovnanie dvoch sekvencií. Ako tretia je zobrazená časť template štruktúry, čiastočne spracovaná našim algoritmom - po metóde MapConservedParts. Zelenou je zvýraznený nukleotid cytozínu s identifikátorom 34. V alignmente je zvýraznený úsek, kde je tento nukleotid zarovnaný s cieľom, a tri medzery vo vzore, kvôli ktorým je nutné mapovanie. Žltou je zvýraznený ten istý nukleotid cytozínu, ale po mapovaní má už identifikátor 37.

cessLongUnconservedParts pri prríprave okoloia predikovaného úseku. Tieto odstránime, pričom nezáleží, ktoré zvolíme, pretože duplicitné nukleotidy pri predikcii FARFAR nemodifikuje, a preto budú v priestore na skoro rovnakých pozíciach. Pozície jednotlivých atómov sa môžu líšiť v maximálne desatinách Å, pretože

FARFAR po úspešne ukončenej predikcii na nej vykoná tzv. „atom refinement“.

3.5 Implementácia

V tejto sekcii sa oboznámime so software, ktorý používame, a s ostatnými implementačnými detailami a problémami.

Náš kód je napísaný v programovacom jazyku Python. Hlavným dôvodom je knižnica BioPython (Cock a kol., 2009). Tá obsahuje metódy uľahčujúce prácu s pdb súbormi, fasta súbormi a so software Emboss Needle, ktorý používame na zarovnanie sekvencii.

Emboss (Rice a kol., 2000) na vstupe berie dve sekvencie vo formáte fasta a zarovná ich. Problém je, že výstupný formát zarovnania BioPython nepodporuje, a preto je ho nutné spracovať pomocou regulárnych výrazov.

Na predikciu nekonzervovaných úsekov používame algoritmus FARFAR, ktorý je naimplementovaný v balíku Rosetta. Rosetta je dodávaná ako balík C++ zrojových súborov, alebo ako ich skompilovaná verzia. V oboch prípadoch je ale nutné použiť unixový systém. Na bežnom PC by taktiež nebolo možné predikciu paralelyzovať z dôvodu nízkeho výkonu. Preto sme sa rozhodli Rosettu nainštalovať na hardware Virtuálnej organizácie Metacentrum (https://metavo.metacentrum.cz/cs/). Táto organizácia spája a poskytuje výpočetné zdroje jej partnerov, ktoré sú potom dostupné akademickým pracovníkom a študentom pomocou SSH a SCP protokolov. Poskytuje nám unixový operačný systém a pre nás dostatočný výpočetný výkon pre paralelizáciu predikcie. Výpočet v Metacentre funguje tak, že si pripravíme jednotlivé úlohy, ktoré zaradíme do fronty úloh. Z nej potom plánovač priraďuje úlohy voľným strojom, ktoré spĺňujú nami špecifikované požiadavky (predpokladaný čas behu, veľkosť RAM, počet CPU). Vo väčšine prípadov je Metacentrum schopné niekoľko desiatok nami zadaných úloh začať počítať do pár hodín. Pri zahltení fronty stovkami úloh však spustenie výpočtu všetkých úloh môže trvať aj viac ako deň - aby neboli obmedzovaní ostatní užívatelia, výber úloh z fronty zohľadňuje množstvo užívateľom používanej výpočetnej kapacity. To, že časť predikcie beží v Metacentre, spôsobuje nemožnosť úplnej automatizácie predikcie. Je nutné ručne presunúť pripravené súbory do Metacentra a spustiť FARFAR. Keďže sa pri predikcii dlhej štruktúry môže spúšťať niekoľko desiatok FARFAR predikcií, máme na to vytvorený shell script.

Na vyhodnocovanie výsledkov používame program PyMol (Schrödinger, LLC, 2015), ktorý dokáže spočítať RMSD (root mean square deviation) experimentálne získanej štruktúry a nášho modelu. Taktiež pomocou neho vizualizujeme výsledky.

Ako dosť zásadný problém sa ukazuje mapovanie sekvencie (fasta) a štruktúry (pdb) vzoru. V pdb súboroch často niektoré nukleotidy chýbajú. Niekedy sa stáva, že poradové čísla nukleotidov v pdb súbore nekorešpondujú s očíslovaním fasta sekvencie (1 - N). Vyskytujú sa súbory, ktoré sú v číslovaní posunuté o konštantu, ale aj také, ktoré majú časť očíslovanú správne a časť nesprávne. Pre druhý prípad (časť očíslovaná správne a časť nesprávne) v kombinácii s chýbajúcimi nukleotidmi je potom veľmi ťažké určiť, ako by malo správne číslovanie vyzerať. Náš algoritmus kontroluje, či sú stiahnuté súbory na seba správne namapované (riadok 25), a v prípade, že nie sú, vyzve užívateľa na manuálnu úpravu súborov.

4. Experimenty

V tejto kapitole predstavíme, aké výsledky pri predikcii rôznych štruktúr dosahuje náš algoritmus. Oboznámime sa s pôvodom a výberom dát, následným priebehom experimnetov, s výsledkami a nakoniec s tým, ako parametre ovplyvňujú výsledky predikcie.

V úvode kapitoly si ešte predstavíme mieru, ktorú využívame pri hodnotení našich výsledkov. Merať budeme rozdiely medzi experimentálne získanou štruktúrou a nami napredikovanou štruktúrou. Obe štruktúry majú rovnakú sekvenciu, na základe čoho sa dajú jasne určiť vzájomne odpovedajúce nukleotidy a následne atómy. RMSD (root-mean-square deviation pozícií atómov) je miera, ktorá vyjadruje priemernú vzdialenosť medzi dvojicami atómov v zarovnaných terciárnych štruktúrach. Väčšinou sa berú do úvahy len atómy ležiace na chrbtici (backbone). My počítame vzdialenosť medzi atómami fosforu. Vypočíta sa ako

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

, pričom δ je vzdialenosť medzi N pármi ekvivalentných atómov. Hodnota je vyjadrená v jednotkách dĺžky, typicky v jednotke Ångström (Å), ktorá je rovná $10^{-10}m$. Na zarovnanie a samotný výpočet RMSD používame funkciu align v programe PyMol (Schrödinger, LLC, 2015). Okrem toho máme naimplementovanú vlastnú funkciu, ktorá vychádza zo zarovnania vytvoreného v PyMol a počíta RMSD štruktúr po odstránení 10% respektíve 20%, respektíve 30% najhorších párov ekvivalentných atómov a naznačuje, či bola výsledná hodnota RMSD ovplyvnená veľmi nepresne napredikovaným úsekom.

4.1 Dáta

Testovacie dáta boli stiahnuté zo stránky ProteinDataBank ako experimentálne získané terciárne štruktúry obsahujúce RNA. Celkovo bolo stiahnutých 1158 pdb súborov, čo sú všetky v súčasnosti dostupné experimentálne získané štruktúry RNA. Tie sme ďalej roztriedili podľa ich identifikátora vlákien (chains), predtože jeden pdb súbor môže obsahovať viacero vlákien a implementácia nášho algoritmu je schopná predikovať len jedno vlákno. Následne sme ich rozdelili do priehradok podľa počtu nukleotidov 4.1.

Spolu máme 2081 štruktúr rozdelených do šiestich priehradok podľa ich veľkostí. Priehradku s počtom nukleotidov 1-50 sme z ďalšieho testovania vylúčili, pretože náš algoritmus je primárne určený na predikovanie dlhých RNA štruktúr. Štruktúry nachádzajúce sa v jednej priehradke považujeme po usporiadaní do dvojíc za potenciálnych kandidátov na vzor a cieľ pre testovanie. Ako ďalšie parametre, podľa ktorých rozdelíme testovacie štruktúry, sme zvolili podobnosť v zarovnaní sekvencií a početnosť medzier v zarovnaní sekvencií, pretože sa ukázalo, že najvýraznejšie ovplyvňujú výsledky predikcie. Tieto dve hodnoty získame tak, že z každej priehradky pre každú dvojicu vypočítame zarovnanie a Emboss nám do výstupného súboru tieto hodnoty uvedie ([A, B] a [B, A] považujeme za dve dvojice). Na základe vypočítaných údajov o zarovnaní štruktúry rozdeľujeme

Tabulka 4.1: Rozdelenie RNA štruktúr podľa počtu nukleotidov.

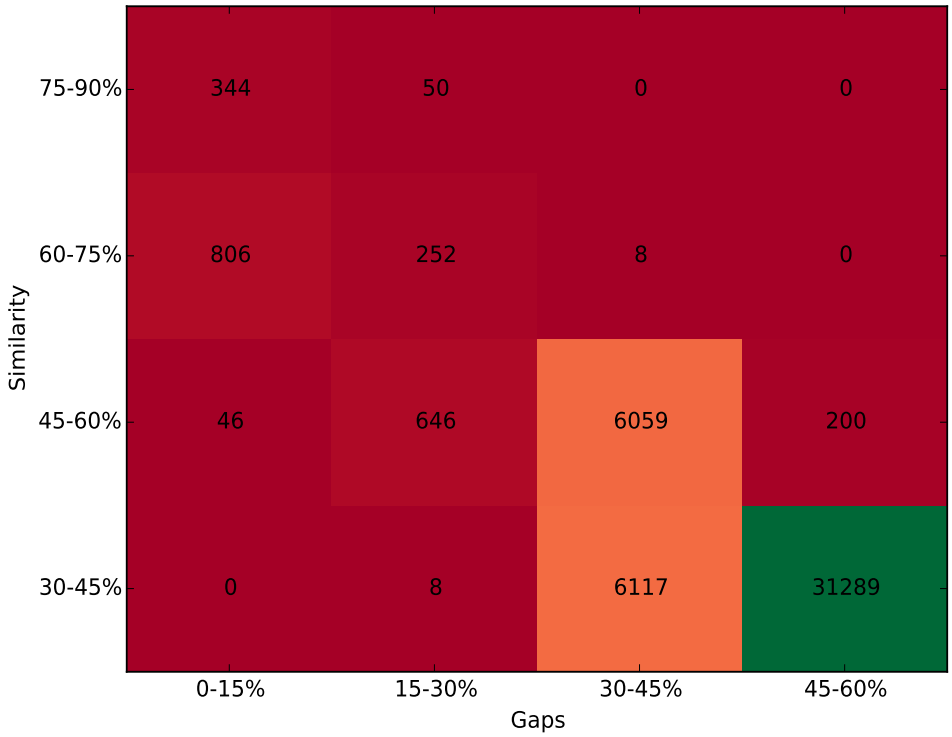
Počet nukleotidov	Počet štruktúr
1-50	1664
51-100	259
101-500	128
501-1000	3
1001-2000	13
2001-10000	14
Súčet	2081

dvojice do nových priehradok, určených podobnosťou zarovnania a množstvom medzier v zarovnaní. Takto vytvorené priehradky budeme nazývať polia. Obe veličiny škálujeme po 15%. Pre podobnosť sme testovali polia v rozsahu od 30% do 90% vzájomnej podobnosti štruktúr a pre početnosť medzier polia v rozsahu od 0% do 60% zastúpenia medzier v zarovnaní. Na nasledujúcich dvoch obrázkoch (heatmapách) je znázornené rozdelenie zarovnaní štruktúr dlhých 51-100 nukleotidov 4.1 a 101-500 nukleotidov 4.2 do polí. Heatmapy ostatných polí neuvádzame pre malý počet štruktúr nachádzajúcich sa v poliach.

Tabulka 4.3: Porovnanie RMSD štruktúr s rovnakou sekvenciou. Ako cieľ slúži štruktúra $3JQ4_A$, ktorá obsauje 2880 nukleotidov. Štruktúry použité v tabuľke ako vzor majú 100% podobnosť sekvencie v zarovnaní medzi sebou a aj s cieľom. Zároveň tvoria v zarovnaní s cieľom cluster(obrázok 4.3).

Vzor	RMSD
$1NJN_0$	1.64
$1NJO_0$	1.63
$1P9X_0$	1.51
$1Z58_2$	1.79
$2O43_A$	1.41
$2O44_A$	1.43
$2O45_A$	1.68
$3FWO_1$	1.56
Priemer	1.58

Zarovnania majú tendenciu zoskupovať sa do clusterov, čo znamená, že viacero rôznych sekvencií má, vzhľadom na jeden cieľ, úplne rovnaké zarovnanie (obrázok 4.3). Štruktúry s rovnakými sekvenciami typicky nie sú identické, avšak vzájomné rozdiely RMSD nebývajú veľké (tabuľka 4.3). Pri použití rôznych vzorov z rovnakého clustera pre jeden cieľ predpokladáme podobné výsledky predikcie, čo nám potvrdzujú testy na dvoch cieľoch a piatich rôznych vzoroch.



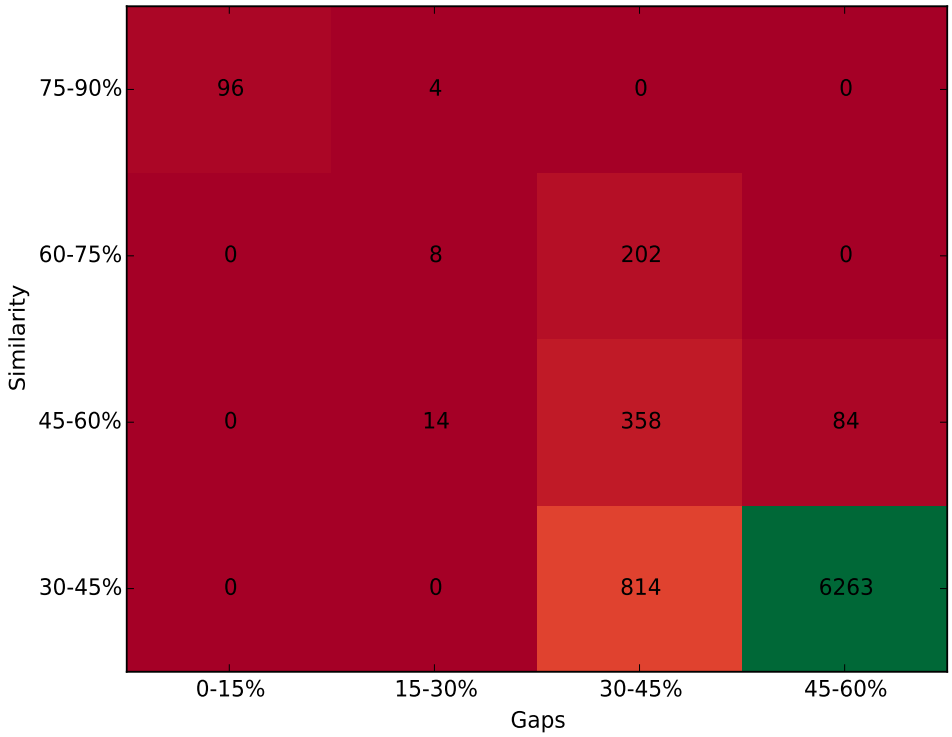
Obrázek 4.1: Rozdelenie zarovnaných dvojíc štruktúr do polí pre veľkosť štruktúr 51 až 100 nukleotidov. Rozdeľuje sa podľa podobnosti a početnosti medzier v zarovnaní. Čísla v poliach znamenajú počet zarovnaní, ktoré vyhovujú danému poľu.

4.2 Experimenty

V tejto sekcii sa budeme venovať popisu priebehu experimentov, výberu testovacích dát a spracovaní výsledkov.

Najskôr skriptom *prepare_structure_similarity.py* vygenerujeme súbory obsahujúce všetky dvojice štruktúr. Takéto súbory vygenerujeme pre každé pole v heatmape (obrázok 4.2). Potom si vyberieme príslušný súbor podľa toho, z ktorého poľa heatmapy chceme predikovať štruktúry. Vybraný súbor dáme ako vstup pre ďalší skript, ktorý má za úlohu vybrať vhodné dvojice a spustiť predikciu. Problém nastáva v tom, že niektoré štruktúry nie sú vhodné ako vzor a najprv potrebujú byť manuálne upravené (viac v poslednom odseku sekcie Implementácia). Preto najprv testujeme, či je štruktúra vhodná ako vzor, a ak nie je, nebudeme ju ako vzor používať. Takto sa vyhneme ručnej úprave štruktúr pri testovaní nášho algoritmu. Ďalšia vec, ktorú sa snažíme zabezpečiť, je, aby boli v testovacích dátach čo najrôznorodejšie štruktúry, takže pridávame obmedzenie, že každá štruktúra sa môže vyskytnúť vo výbere maximálne jedenkrát. Takto vybereme požadovaný počet dvojíc. Vybrané štruktúry sú automaticky spracované a sú vytvorené vstupné súbory pre FARFAR. Z tejto časti experimentov tiež získavame ďalšie informácie o predikcii, ako priemer, medián a smerodajná odchýlka úsekov, ktoré musí FARFAR dopredikovať.

Následne musíme ručne skopírovať vstupné pdb súbory, fasta súbor cieľa (nuk-



Obrázek 4.2: Rozdelenie zarovnaných dvojíc štruktúr do polí pre veľkosť štruktúr 101 až 500 nukleotidov. Rozdeľuje sa podľa podobnosti a početnosti medzier v zarovnaní. Čísla v poliach znamenajú počet zarovnaní, ktoré vyhovujú danému poľu.

```
SEQUENCE '3JQ4_A.fasta':
(100.0%, ['1NJN_0.fasta', '1NJO_0.fasta', '1P9X_0.fasta',
          '1Z58_2.fasta', '2O43_A.fasta', '2O44_A.fasta',
          '2O45_A.fasta', '3FWO_A.fasta']),
(67.9%,  ['1C2W_B.fasta', '3DG0_B.fasta', '3DG2_B.fasta',
          '3DG4_B.fasta', '3DG5_B.fasta'])])
```

Obrázek 4.3: Ukážka výsledkov zarovnania sekvencie $3JQ4_A$ so zvyšnými sekvenciami z priehradky 2001-10000. Názvy chápeme tak, že prvé 4 znaky nesú názov sekvencie (štruktúry) a za podtržítkom je identifikátor vlákna. V zelenom rámiku je zvýraznená cieľová sekvencia, v červených rámikoch sa nachádzajú možné vzory, teda ostatné sekvencie z priehradky. Oranžovými rámikmi sú označené podobnosti zarovnania cieľa s inou sekvenciu z priehradky. Všimneme si, že v tomto prípade sa nám sekvencie rozdelili do dvoch clusterov, pričom sekvencie v hornom červenom rámiku sú úplne rovnaké (štruktúry sa od seba typicky líšia).

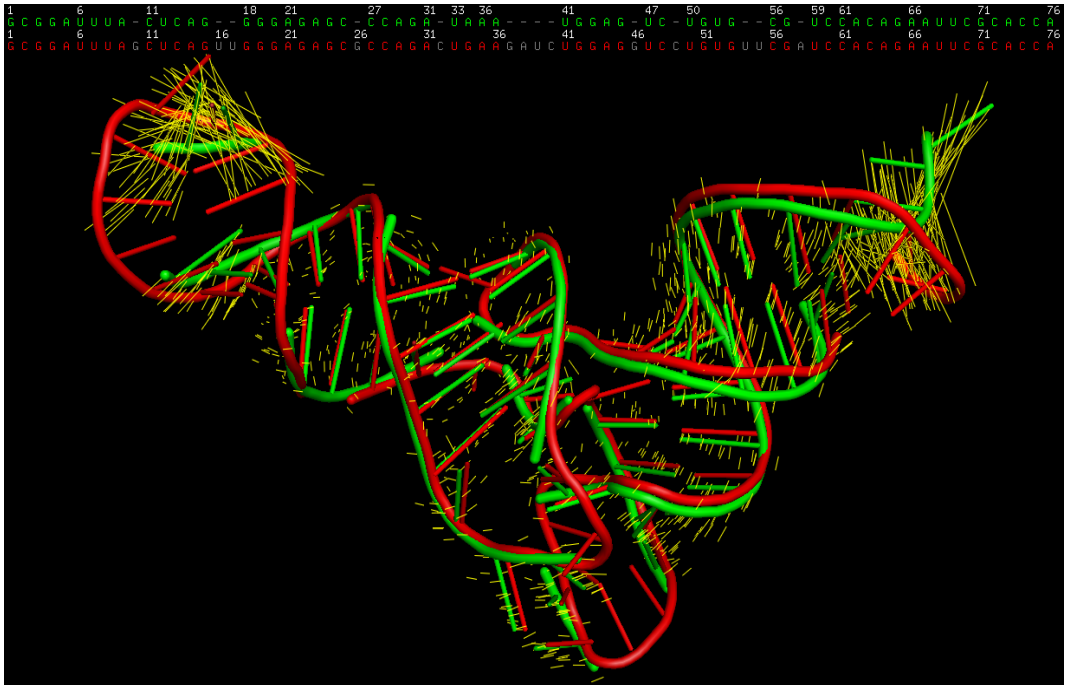
leotidy musia byť kódovane malými písmenami) a súbor s vygenerovanými príkazmi pre FARFAR do Metacentra. Potom spustíme skript, ktorý spúšťa predikciu.

Tento postup opakujeme pre každú predikovanú štruktúru zvlášť. V Metacentre je nutné nastaviť ako dlho chceme, aby nami spustené úlohy maximálne bežali. Tento parameter nastavujeme vždy na 24 hodín (okrem špeciálnych testov). Ak sa nestihne vygenerovať požadovaný počet kandidátskych štruktúr, úloha síce skončí chybou, avšak napredikované kandidátske štruktúry ostávajú k dispozícii vo výstupnom súbore. Preto môžeme povedať, že FARFAR v našom algoritme nie je parametrizovaný iba počtom vygenerovaných kandidátskych štruktúr, ale aj časom 24 hodín. Z toho, že predikcia nekonzervovaných úsekov vo FARFAR trvá 24 hodín a ostatné časti algoritmu trvajú najviac pár minút vyplýva, že celá predikcia ľubovoľne dlhej štruktúry zaberie približene 24 hodín.

Následne spustíme (ešte v Metacentere) skript, ktorý vyberie najlepšiu kandidátsku štruktúru z každej FARFAR predikcie a vytvorí z nich pdb súbory. Tie skopírujeme späť na náš PC, kde spustíme skript na spojenie predikovaných a konzervovaných častí cieľovej štruktúry.

Potom v PyMol otvoríme experimentálne získanú cieľovú štruktúru spolu s nami napredikovanou štruktúrou. Z experimentálne získanej štruktúry odporúčame odfiltrovať všetky riadky, ktoré nekódujú pozíciu predikovaného vlákna, pretože v prípade, že sme tak neurobili, sa občas vyskytli problémy so zarovnaním. Pomocou príkazu align zarovnáme na seba štruktúry a vytvoríme objekt zarovnania 4.4, v ktorom sú určené dvojice odpovedajúcich nukleotidov.

Tento objekt uložíme a dáme ako vstup nášmu ďalšiemu skriptu, ktorý vypočíta RMSD po orezaní 0%, 10%, 20% a 30% dvojíc nukleotidov od seba najviac vzdialených.



Obrázek 4.4: Ukážka zarovnania v PyMol medzi experimentálne získanou (zelená) a predikovanou (červená) štruktúrou molekuly $1EHZ_A$. Žlté čiarky znázorňujú jednotlivé vzdialenosti korešpondujúcich atómov medzi oboma štruktúrami. V hornej časti obrázka môžeme vidieť, ktoré nukleotidy boli zarovnané, a ktoré nie (v originálnej štruktúre niektoré chýbajú).

4.3 Výsledky

Najskôr popíšeme, ako boli nastavené parametre pri predikcii testovacích dát. Význam parametrov a ich úloha pri predikcií je popísaná v predchádzajúcej kapitole, konkrétne v sekcii „Detailný popis“. Preto uvedieme iba názvy metód (predstavených v sekvencií „Pseudokód“) s hodnotami parametrov: *SlidingWindow(alignment, windowLength=10, minimalLimit=5), ProcessGaps(alignment, cutoff=0.5*length(gap)), ProcessLongUnconservedParts(conservedParts, ncAvarageLength, minLengthGapLimit=5)*. Posledné parametre ovplyvňujúce predikciu sú dĺžka behu FARFAR, ktorá bola nastavená na 24 hodín (popísané v sekcií „Experimenty“), a počet vygenerovaných kandidátskych štruktúr, nastavených na limit 100.

4.3.1 Predikcia štrkuktúr dlhých 51-500 nukleotidov

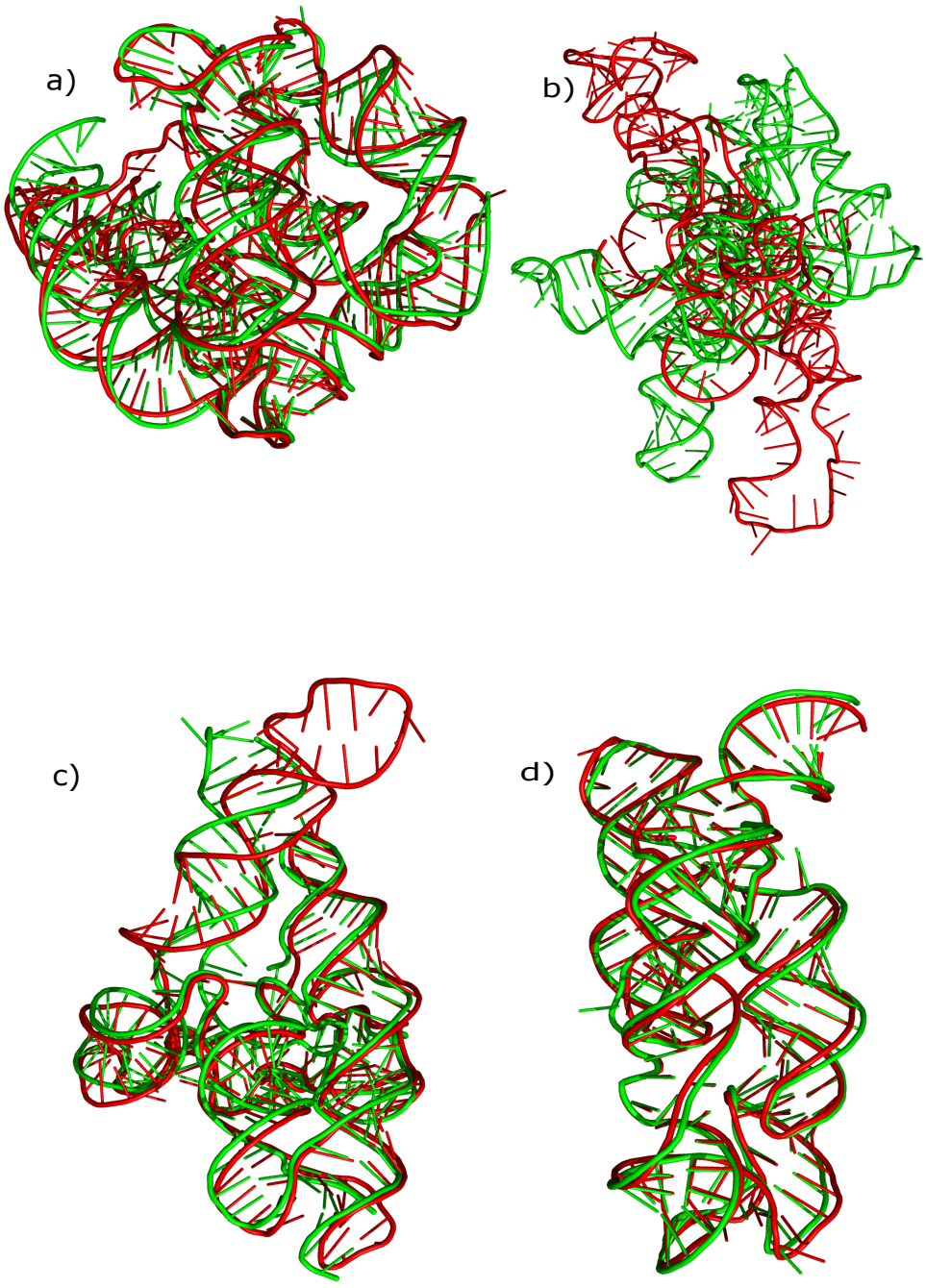
V tejto podsekcii predstavíme aké výsledky dosahuje pri predikovaní štruktúr náš algoritmus (tabuľka 4.5, obrázok 4.5). Testovacie dáta sme vybrali zo štruktúr dlhých 51-500 nukleotidov, ako je to popísané v sekcii Experimenty. Dáta sme vyberali tak, aby bolo každé relevantné pole z heatmáp 4.1, 4.2 zastúpené (pojem relevantné pole vysvetlíme nižšie). Vo výsledkoch nerozlišujeme dĺžku štruktúr, pretože výsledné RMSD sa neukázalo byť závislé od dĺžky predikovanej štruktúry. Pod pojmom relevantné pole myslíme také, ktoré obsahuje aspoň 4 štruktúry, z toho 2 musia byť vhodné ako vzor pre predikciu (aby bolo možné vytvoriť aspoň 2 rozdielne dvojice). Testovali sme len 2 polia s podobnosťou zarovnaní sekvencií dvojíc menšou ako 60%, pretože výsledky sa ukázali byť veľmi nepresné (vysoká RMSD), a teda takéto dvojice (ako cieľ a vzor) sú pre náš algoritmus nevhodné.

Tabulka 4.5: Výsledky testovania nášho algoritmu: dáta pochádzajú z 30 dvojíc vybraných z polí v heatmapách 4.1, 4.2. Tabuľka ukazuje priemernú RMSD a smerodajnú odchýlku pre jednotlivé polia. Na základe týchto výstupných dát môžeme odhadnúť výsledky predikcie podľa toho, do ktorého poľa dvojica štruktúr spadá.

Similarity(%)	Gap(%)	priemer RMSD	std RMSD
30-45	30-45	32.05	6.09
45-60	30-45	32.3	4.78
60-75	0-15	11.88	7.81
60-75	15-30	9.63	2.33
60-75	30-45	8.8	7.2
75-90	0-15	6.02	4.37
75-90	15-30	6.93	4.45

Z výsledkov vyplýva, že s rastúcou podobnosťou (similarity) zarovnania sa znižuje priemerná RMSD predikovanej a experimentálne získanej štruktúry, čo znamená presnejšie napredikované štruktúry. Môžeme si všimnúť, že najväčší rozdiel v presnosti predikcie nastáva medzi podobnosťami zarovnania 45-60% a 60-75%. Smerodajná odchýlka sa vo všetkých prípadoch drží v podobných hodnotách. Tieto hodnoty poukazujú na to, že sa nestáva, aby dvojica s nízkou po-

dobnosťou v zarovnaní mala veľmi nízke RMSD, a naopak, štruktúry s vysokou podobnosťou zarovnania mali RMSD vysokú.



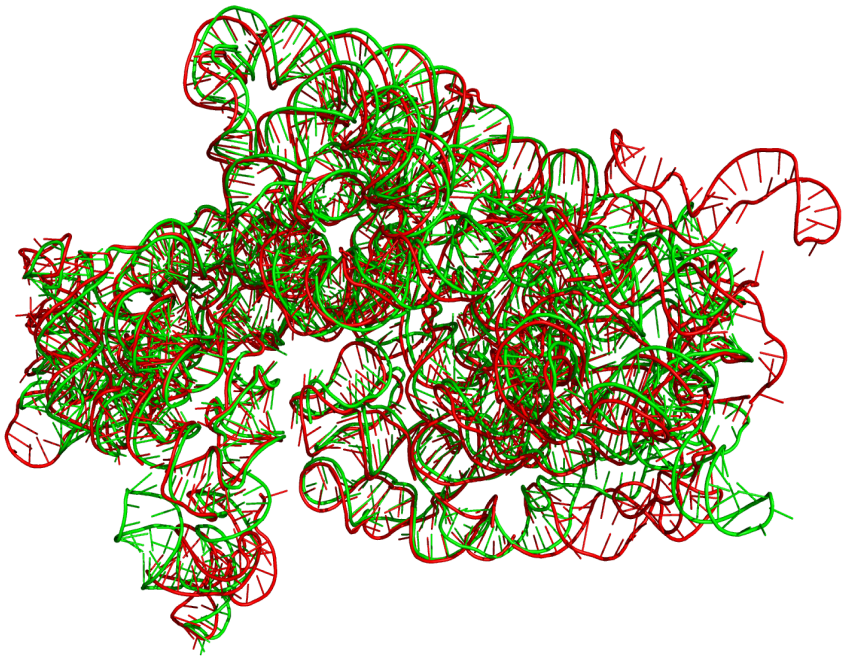
Obrázek 4.5: Príklady zarovnaní nami predikovaných (červenou) a experimentálne získaných (zelenou) štruktúr. Na obrázku sú a) štruktúry 4Y1N s RMSD 5.34, b) 1GRZ s RMSD 35.58, c) 4P95 s RMSD 12.03, d) štruktúry 3D0U s RMSD 1.68.

4.3.2 Predikcia veľmi dlhých štruktúr

V tejto podsekcii prezentujeme výsledky nášho algoritmu pri predikovaní štruktúr dlhých niekoľko tisíc nukleotidov. Parametre nastavujeme rovnako ako pri predikcii testovacích dát, až na pár výnimiek, kedy sme museli nastaviť dĺžku behu FARFAR pri predikovaní veľmi dlhých nekonzervovaných úsekov štruktúry na 48 hodín, pretože za 24 hodín nebol FARFAR schopný vygenerovať ani jednu kandidátsku štruktúru. Výsledky uvádzame v tabuľke4.7. Môžeme si všimnúť, že pri predikcii cieľa 3J2G mal so vzorom 99.4% podobnosť a nekonzervovaných bolo len 9 nukleotidov. RMSD dosiahla 9.6 Å, a preto ak sa sekvencie presne nezhodujú (nie sú v jednom clusteri), tak sa ich štruktúry môžu líšiť výrazne viac, ako prezentujeme v tabuľke 4.3. Konkrétna ukážka sa nachádza na obrázku 4.6.

Tabulka 4.7: Výsledky predikcie dlhých štruktúr.

Cieľ	Dĺžka	Podobnosť(%)	Medzera(%)	RMSD
3DG0	2904	68	15.3	16.5
3DG0	2904	68	15.3	13.62
4JI1	1522	71.6	14.1	14.5
4V6W	1995	68.3	19.9	32.7
3J2G	1533	99.4	0.6	9.6



Obrázek 4.6: Ukážka zarovnania dvoch štruktúr molekuly 4IJ1, ktorá obsahuje 1522 nukleotidov. Červená štruktúra bola predikovaná našim algoritmom, zelená bola experimentálne získana. RMSD zarovnania týchto štruktúr je 14.5.

4.3.3 Vplyv parametrov na výsledky predikcie

V poslednej časti sekcie budeme prezentovať zmeny predikovaných výsledkov po zmene jednotlivých parametrov a vyhodnotíme, či zvolená hodnota parametra v predchádzajúcich testoch predikciu zlepšila, alebo naopak, zhoršila. Ako testovaciu vzorku sme si vybrali 3 dvojice štruktúr 4.9.

Tabulka 4.9: Štruktúry, na ktorých testujeme zmeny parametrov. RMSD je uvedená po predikcii s pôvodnými parametrami.

Cieľ	Dĺžka	Similarity(%)	Gap(%)	RMSD
3DG0	2904	68	15.3	13.62
3D0U	161	84.6	8.6	1.68
3DIO	174	84.6	8.6	12.48

Najskôr sa pozrieme na zmenu dĺžky behu algoritmu FARFAR. Malo by platiť, že čím viac kandidátskych štruktúr bude vygenerovaných, tým bude výsledok predikcie lepší. Na druhej strane, algoritmus je nedeterministický, preto sa zlepšenie nedá zaručiť. Dĺžku behu sme stanovili na 100 hodín a limit vygenerovaných kandidátskych štruktúr na 500. Zaznamenali sme pozitívne aj negatívne výskedky. Percentuálny priemer zlepšenia, prípadne zhoršenia, RMSD bol 14.32% a smerodajná odchýlka bola 5.82. Upozorníme, že v prípade zhoršeného výsledku predikcie, ktorý nastal pri predikcii štruktúry 30DG dlhej 2904 nukleotidov, napriek tomu, že beh FARFAR bol predĺžený, bola pre najdlhší nekonzervovaný úsek vygenerovaná rovnako iba jedna kandidátska štruktúra. Toto považujeme, kvôli nedeterministickosti algoritmu za jednu z možných príčin zhoršenia výsledku.

Pri nastavení parametrov *SlidingWindow(alignment, windowLength=0, minimalLimit=0)*, teda vlastne nepoužitiu algoritmu posuvného okienka sme zistili, že pri dvoch kratších testovacích štruktúrach táto metóda na nekonzervovaných úsekoch nezmenila nič. Pri dlhej štruktúre 30DG už zmeny v nekonzervovaných úsekoch nastali a výsledná predikcia bola o 14.5% horšia. Použitie tejto metódy nám teda prináša buď zlepšenie predikcie, alebo žiadne zmeny pri kratších štruktúrach.

Pri nastavení *ProcessGaps(alignment, cutoff=1)*, čo znamená, že z oboch strán medzery v zarovnaní bude označený ako nekonzervovaný len jeden nukleotid. V tomto prípade sa nám výsledky líšia podľa podobnosti sekvencie štruktúr. V prípade, že podobnosť v zarovnaní bola väčšia ako 60%, sa výsledky vo všetkých vykonaných testoch zlepšili v priemere o 17.07%. Smerodajná odchýlka bola 17.19. Tieto výsledky nás viedli k tomu, vyskúšať s týmito parametrami predikovať aj dvojice so vzájomnou podobnosťou zarovnania 30-45%, pretože tie boli pôvodným rozširovaním nekonzervovaných úsekov dotknuté výraznejšie ako lepšie konzervované štruktúry. V dvoch takýchto pokusoch však bol výsledok predikcie v priemere o 9.48% horší. Pokusy teda poukazujú na to, že by bolo lepšie pri štruktúrach s podobnosťou viac ako 60% zmenšiť parameter cutoff v tejto metóde pre dosiahnutie lepších výsledkov.

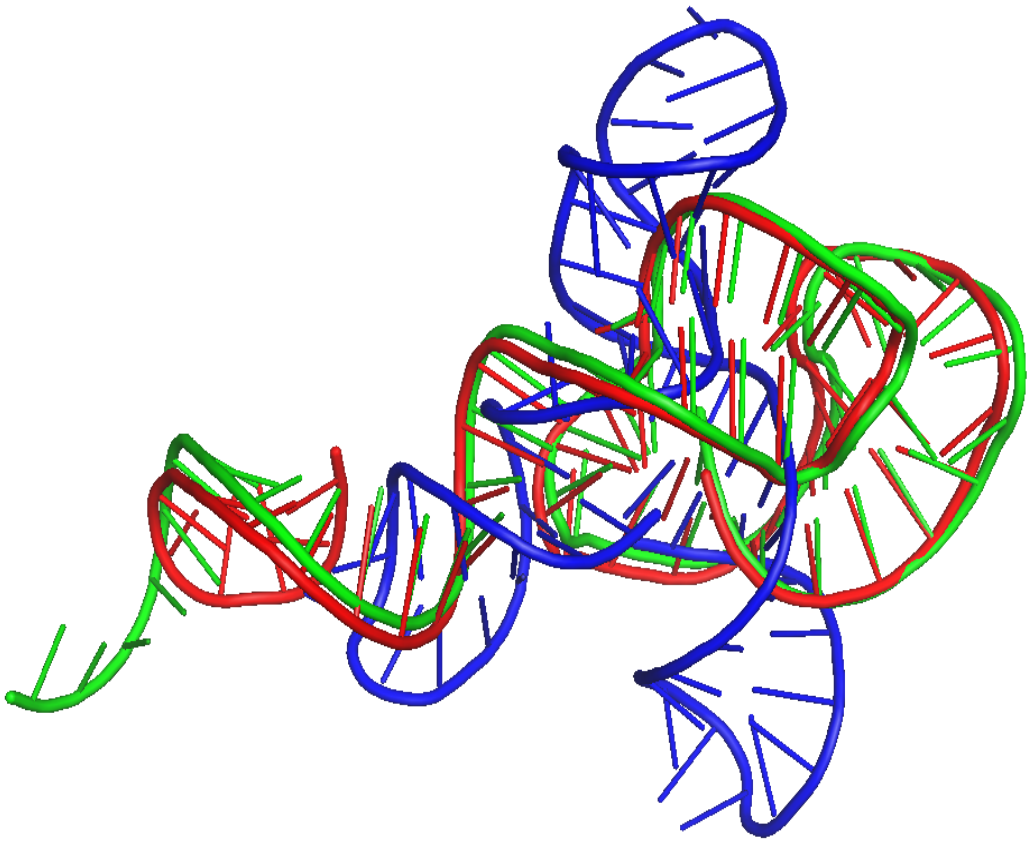
Nakoniec, pri nastavení *ProcessLongUnconservedParts(conservedParts, ncAvarageLength, minLengthGapLimit=∞)*, čo znamená zrušenie samostatného

FARFAR predikovania dlhších nekonzervovaných úsekov, vedeným okolitými nukleotidmi. Táto zmena sa týka hlavne dlhých štruktúr, pretože pri kratších sa všetky nekonzervované úseky budú vo FARFAR predikovať v jednom súbore, teda predikcia bude spomalená, ale FARFAR nepríde o znalosť konzervovaných nukleotidov, ktoré ovplyvňujú predikciu. Predikcia dlhej štruktúry priniesla s takto zmeneným parametrom výsledok horší o 3.97, čo poukazuje na to, že metóda funguje. Pri dvoch kratších štruktúrach nenastali vo výsledkoch výraznejšie zmeny.

4.4 Porovnanie nášho algoritmu s FARFAR

Na začiatok zdôrazňujeme, že FARFAR nie je určený na predikciu dlhých RNA štruktúr. Účelom tejto sekcie je demonšrtovať, že náš algoritmus znamená oproti FARFAR prínos, a že FARFAR nedokáže podobne dobre predikovať dlhšie štruktúry. Nechali sme preto dve štruktúry napredikovať algoritmom FARFAR (24 hodín beh alebo 200 kandidátskych štruktúr) a aj našim algoritmom. Výsledky môžeme vidieť na obrázkoch 4.7 a 4.8. Všimneme si, že okrem toho, že RMSD je v prípade FARFAR niekoľkonásobne väčšia, FARFARom predikované štruktúry sa nepodobajú na tie experimentálne získané, zatiaľ čo našim algoritmom predikované štruktúry áno, a ako sme v úvodnej kapitole spomínali, pre určenie funkcie RNA je dôležitá práve 3D štruktúra molekuly. Nakoniec dodáme, že sme na porovnanie zvolili kratšie štruktúry a pri voľbe dlhších by sa predikcia pomocou FARFAR zhoršovala.

Obrázek 4.7: Výsledné zarovnanie troch štruktúr RNA molekuly 4QK8 s dĺžkou 122 nukleotidov. Na obrázku sa nachádzajú tri štruktúry: experimentálne získana (zelená), predikovaná našim algoritmom pri podobnosti zarovnania sekvencií 61.9%(červená) a predikovaná pomocou FARFAR (modrá). RMSD červenej a zelenej štruktúry je 6.48 a RMSD modrej a zelenej štruktúry je 28.43.



Obrázek 4.8: Výsledné zarovnanie troch štruktúr RNA molekuly 2QUW s dĺžkou 57 nukleotidov. Na obrázku sa nachádzajú tri štruktúry: experimentálne získana (zelená), predikovaná našim algoritmom pri podobnosti zarovnania sekvencií 82.6%(červená) a predikovaná pomocou FARFAR (modrá). RMSD červenej a zelenej štruktúry je 4.95 a RMSD modrej a zelenej štruktúry je 21.59.

Záver

Cieľom práce bolo vytvoriť nový algoritmus predikcie RNA, pracujúci na princípe homológneho modelovania, ktorý bude schopný predikovať aj dlhé RNA štruktúry, čo sa nám podarilo. Vytvorený algoritmus bol následne naimplementovaný a otestovaný. Algoritmus je navrhnutý tak, aby bolo možné predikovať štruktúry ľubovoľnej dĺžky. Aj pri predikovaní dlhej štruktúry dokáže algoritmus dosahovať rozumné výsledky za predpokladu existencie vhodného vzoru. Možnosť predikovať ľubovoľne dlhé štruktúry považujeme za najväčšiu výhodu v porovnaní so súčasne dostupnými metódami, zameranými na predikciu RNA štruktúr, u ktorých je práve dĺžka predikovanej štruktúry limitujúca. Výhoda taktiež je, že vďaka paralelizovaniu predikcie nekonzervovaných úsekov dĺžka štruktúry negatívne neovplyvňuje čas, potrebný pre napredikovanie celej štruktúry.

Výsledky testovania ukázali, že ak existuje dostatočne dobrý vzor pre predikciu, tak je nami navrhnutý algoritmus schopný za 24 hodín napredikovať ľubovoľne dlhú RNA štruktúru pri priemernej RMSD 10 Å.

Našou prácou sme ukázali, že nami použitý spôsob predikovania RNA štruktúr - teda skopírovanie štruktúry konzervovaných úsekov a následné dopredikovanie nekonzervovaných úsekov štruktúry metódou de novo - je jeden z možných spôsobov, ako riešiť problém predikcie štruktúr RNA molekúl.

Nakoniec by sme chceli poukázať na možné vylepšenia algoritmu a možnosti doplnenia ďalších funkcií. Vďaka testovaniu vieme, že niektoré parametre sme nevolili optimálne a ich zmenou môžeme dosiahnuť presnejšie výsledky. Ďalej by sa dalo pridať možnosť automaticky vyhľadať najvhodnejšie vzory pre zadanú sekvenciu cieľa, čo by zrýchlilo prípravu predikcie. Takisto by bolo dobré zlepšiť presnosť predikcie dlhých nekonzervovaných úsekov, čo momentálne považujeme za jeden z najväčších nedostatkov v našom algoritme. Nakoniec by sme chceli proces predikcie plne automatizovať a sprístupniť webserver, ktorý by umožňoval predikciu RNA štruktúr pomocou nášho algoritmu iba zo znalosti primárnej sekvencie RNA molekuly.

Seznam použité literatury

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, N., Weissig, H. a Shindyalov, I. N. (2000). The protein data bank. *Nucleic Acid Reasearch*, **28**, 235–242.

Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. a Hoon, M. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Cooper, G. a Hausman, R. (2004). *The Cell: A Molecular Approach.* 3rd. edition. Sinauer, USA. ISBN 0-87893-214-3.

Crick, F. (1970). Central dogma of molecular biology. *Nature.*

Das, R., Karanicolas, J. a Baker, D. (2010). Atomic accuracy in predicting and designing non-canonical rna structure. *Nat Methods*, **4**, 291–294.

Felden, B. (2007). Current opinion in microbiology. *Current Opinion in Microbiology*, **10**, 286–291.

Frellsen, J., Moltke, I., Thiim, M., Mardia, K. V., Ferkinghoff-Borg, J. a Hamelryck, T. (2009). A probabilistic model of rna conformational space. *Computational Biology*, **5**.

Holbrook, S. (2008). Structural principles from large rnas. *Annual Review of Biochemistry*, **37**, 445–464.

Holzel, M., Orban, M., Hochstatter, J., Rohrmoser, M., Harasim, T., Malamoussi, A., Kremmer, E., Langst, G. a Eick, D. (2010). Defects in 18 s or 28 s rrna processing activate the p53 pathway. *J. Biol. Chem.*, **285**(9), 6364–6370.

Jonikas, M. A., Radmer, R. J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. a Altman, R. B. (2009). Coarse-grained modeling of large rna molecules with knowledge-based potentials and structural filters. *RNA Society*, **15**, 189–199.

Krieger, E., Sander, B. a Vriend, G. (2003). *Structural Bioinformatics.* Wiley-Liss, Inc. ISBN 0-471-20199-5.

Moore, P. B. (1999). *The RNA World.* 2nd. edition. Cold Spring Harbor Laboratory, New Haven, Connecticut USA. ISBN 0-87969-561-7.

Noller, H. (1984). Structure of ribosomal rna. *Annual Review of Biochemistry*, **53**, 119–162.

Parisien, M. a Major, F. (2008). The mc-fold and mc-sym pipeline infers rna structure from sequence data. *Nature*, **452**, 51–55.

Patton, J. (2008). *Segmented Double-stranded RNA Viruses: Structure and Molecular Biology*. Caister Academic Press, USA. ISBN 978-1-904455-21-9.

Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N., Blazewicz, J. a Adamiak, R. W. (2012). Automated 3d structure composition for large rnas. *Nucleic Acids Res.*, **40**, 1–12.

Rice, P., Longden, I. a Bleasby, A. (2000). Emboss: The european molecular biology open software suite (2000). *Trends in Genetics*, **16**, 276–277.

Rinn, J. a Chang, H. (2012). Genome regulation by long noncoding rnas. *Annual Review of Biochemistry*, **81**, 145–166.

Rother, M., Rother, K., Puton, T. a Bujnicki, J. (2011). Moderna: a tool for comparative modeling of rna 3d structure. *Nucleic Acids Research*, **39**, 4007–4022.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.

Sharma, S., Ding, F. a Dokholyan, N. V. (2008). ifoldrna: three-dimensional rna structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.

Smyth, M. S. a Martin, J. H. J. (2000). x ray crystallography. *Journal of Clinical Pathology*, **53**, 8–14.

Westhof, E. (2015). Twenty years of rna crystallography. *RNA Society*, **21**, 486–487.

Seznam obrázků

Seznam tabulek

Přílohy

Priložené CD obsahuje dve zložky. V zložke „metacentrum“ sa nachádzajú skripty, ktoré sa používajú v Metacentre. V zložke „desktop“ sú všetky skripty, ktoré používame na lokálnom počítači.