VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2021 Nikola Polzerová

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY PREDIKCE SEKUNDÁRNÍ STRUKTURY RNA
METHODS FOR PREDICTION OF RNA SECONDARY STRUCTURE

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR
Nikola Polzerová

VEDOUCÍ PRÁCE
SUPERVISOR
Ing. Kateřina Jurečková

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Nikola Polzerová ***ID:*** 193239

Ročník: 3 ***Akademický rok:*** 2020/21

NÁZEV TÉMATU:

Metody predikce sekundární struktury RNA

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma RNA a její struktura, zaměřte se zejména na sekundární struktury RNA. 2) Prostudujte a popište metody pro predikci sekundární struktury RNA. 3) Sestavte dataset RNA sekvencí se známou sekundární strukturou z veřejně dostupných databází. 4) Na základě získaných znalostí implementujte jeden vybraný algoritmus pro predikci sekundární struktury RNA v libovolném programovacím jazyce. Jeho funkčnost ověřte na vytvořeném datasetu. 5) Realizujte alespoň dva další algoritmy pro predikci sekundární struktury RNA. 6) Dosažené výsledky vhodně vyhodnoťte, srovnejte a diskutujte.

DOPORUČENÁ LITERATURA:

[1] JABBARI, Hosna, Ian WARK a Carlo MONTEMAGNO. RNA secondary structure prediction with pseudoknots: Contribution of algorithm versus energy model. PLoS ONE. 2018, 13(4), 1–21. ISSN 19326203. DOI:10.1371/journal.pone.0194583

[2] FALLMANN, Jörg, Sebastian WILL, Jan ENGELHARDT, Björn GRÜNING, Rolf BACKOFEN a Peter F. STADLER. Recent advances in RNA folding. Journal of Biotechnology. 2017, 261(February), 97–104. ISSN 18734863. DOI:10.1016/j.jbiotec.2017.07.007

Termín zadání: 8.2.2021 ***Termín odevzdání:*** 28.5.2021

Vedoucí práce: Ing. Kateřina Jurečková

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

UPOZORNĚNÍ:
Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

Tato bakalářská práce pojednává o sekundární struktuře RNA a konkrétně se zaměřuje na její predikci. Popisuje nejrůznější elementy sekundární struktury a představuje některé metody predikce. V rámci bakalářské práce byly implementovány tři výpočetní metody predikce v programovacím prostředí MATLAB. Konkrétně se jedná o algoritmus Nussinové, Zukerův algoritmus a metodu Crumple. Implementované algoritmy přistupovaly k predikci buď na základě maximalizace bázových párů, nebo minimalizace volné energie. Jejich funkce byla ověřena na vytvořeném datasetu a výsledky byly srovnány se známou sekundární strukturou.

Klíčová slova

RNA, sekundární struktura RNA, maximalizace bázovách párů, minimalizace volné energie algoritmus podle Nussinové, algoritmus podle Zukera, algoritmus pro metodu Crumple

Abstract

This bachelor thesis discuss RNA secondary structure and it's prediction in particular. It describes various secondary structure elements and presents some secondary structure prediction methods. Within the framework of bachelor thesis, three computational methods for secondary structure prediction were implemented in programming and numeric computing platform MATLAB. These methods are Nussinov algorithm, Zuker algorithm and Crumple method. Implemented algorithms approched the prediction in terms of base pair maximalization or free energy minimalization. Their function was verified on the created dataset and the results were compared with known secondary structures.

Keywords

RNA, RNA secondary structure, base pairs maximalization, free energy minimalization, Nussinov algorithm, Zuker algortihm, Crumple method

Bibliografická citace:

POLZEROVÁ, Nikola. *Metody predikce sekundární struktury RNA*. Brno, 2021. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/134377. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce Kateřina Jurečková.

Prohlášení autora o původnosti díla

Jméno a příjmení studenta:	*Nikola Polzerová*
VUT ID studenta:	*193239*
Typ práce:	*Bakalářská práce*
Akademický rok:	*2020/21*
Téma závěrečné práce:	*Metody predikce sekundární struktury RNA*

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

.

V Brně dne: 28. května 2021 ……

podpis autora

Poděkování

Děkuji vedoucí mé bakalářské práce Ing. Kateřině Jurečkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne: 28. května 2021 ……

podpis autora

Obsah

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

RNA patří ke skupině nukleových kyselin a v buňkách je zodpovědná za přenos genetické informace. Jedná se o jednovláknový biopolymer tvořen nukleotidy adeninem, cytosinem, guaninem a uracilem. Spojováním komplementárních bází dochází k vytváření struktur nejrůznějších smyček, výdutí a pseudouzlů. Tyto struktury jsou známé jako sekundární struktura RNA.

Strukturální elementy sekundární struktury RNA mají nespočet funkcí, mezi které patří například modulace epigenetických značek, nebo podpora velkých makromolekulárních komplexů. Také se společně se strukturou terciální účastní posttranskripční regulace genové exprese. Správná predikce sekundární struktury je klíčová pro pochopení funkce a regulace transkriptů RNA.

V posledních desetiletích se dostalo velké pozornosti nejrůznějším experimentálním a výpočetním metodám predikce sekundární struktury RNA. I přes to, že experimentální metody predikce jsou úspěšné, jsou velmi časově náročné, drahé a pro některé sekvence neproveditelné. Právě proto jsou slibnou alternativou bioinformatické výpočetní metody, které pracují s nejrůznějšími hypotézami o tom, jakým způsobem je sekundární struktura RNA tvořena.

K nejznámějším hypotézám patří minimalizace volné energie a maximalizace bázových párů. Minimalizace volné energie popisuje, že nejstabilnější je struktura s nejmenší hodnotou volné energie a proto předpokládá, že RNA vytváří sekundární strukturu tak, aby měla ve výsledku tuto hodnotu co nejmenší. Další možnou hypotézou je maximalizace bázových párů, která předpokládá, že nejstabilnější je ta struktura, která obsahuje co největší počet bázových párů.

Pro bakalářskou práci byly implementovány v programovacím prostředí MATLAB celkem tři výpočetní metody pro predikci sekundární struktury RNA. Všechny implementované metody jsou v textu bakalářské práce postupně představeny a jejich implementace je podrobně popsána. Funkčnost implementovaných algoritmů byla otestována na vytvořeném datasetu RNA sekvencí se známou sekundární strukturou. Pro tvorbu tohoto datasetu bylo čerpáno z volně dostupné databáze RNA sekvencí - RNA STRAND v2.0.

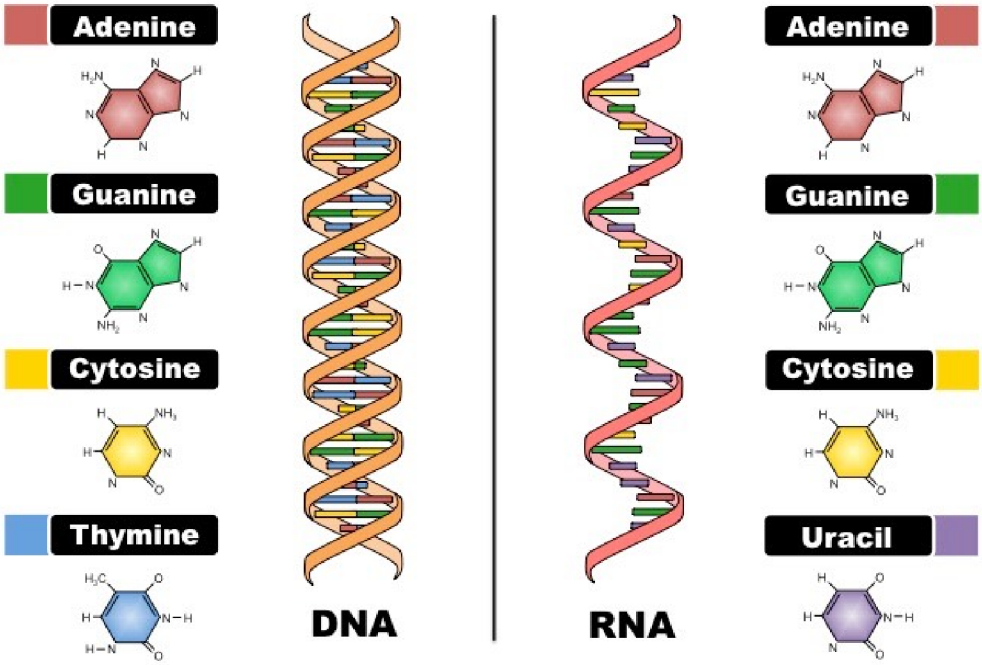
Implementovány byly algoritmus Nussinové, algoritmus Zukera a metoda Crumple. Výsledky byly porovnány se známou sekundární strukturou a mezi sebou. Kvalita implementovaných algoritmů byla ohodnocena pomocí senzitivity, PPV a F-skóre. Predikované sekundární struktury, které se nejvíce odlišovaly od známé sekundární struktury, byly nakonec vykresleny pomocí funkce `rnaplot`, která je volně dostupná v programovacím prostředí MATLAB.

1. RNA A JEJÍ STRUKTURA

1.1 RNA

Ribonukleová kyselina (RNA) patří ke klíčovým molekulám v našich buňkách a je jednou z nukleových kyselin, které jsou zodpovědné za přenos genetické informace z nukleových kyselin do proteinů. Větší pozornosti se RNA začalo dostávat až v šedesátých letech 20. století, kdy Francis Crick představil základní biologické dogma, které definuje tok genetické informace v živých organismech ve směru DNA (deoxyribonukleová kyselina) → RNA → protein. RNA, se obdobně jako DNA, skládá z nukleotidů, které jsou tvořeny vždy jednou ze čtyř bází, monosacharidem ribózou a fosfátovou skupinou. U RNA jsou tyto báze adenin (A), uracil (U), guanin (G) a cytosin (C). Jednotlivé nukleotidy jsou řazeny za sebou a tvoří lineární řetězec. [1]

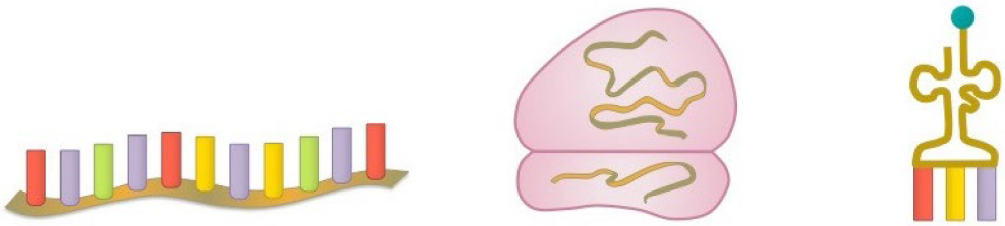
Hlavním rozdílem mezi DNA a RNA je skutečnost, že cukr navázaný v molekule RNA je ribóza. Cukr navázaný v DNA je deoxyribóza [2]. Navázání hydroxylové (-OH) skupiny na 2' uhlíku ribózy v RNA hraje velmi důležitou roli v chemických reakcích, kterých se RNA účastní. Dalším rozdílem je různorodost tvarů, ve kterých RNA nacházíme. Zatímco DNA se nachází nejčastěji ve formě dvoušroubovice, kterou tvoří spojení dvou jednoduchých vláken, viz Obr. **1-1**. Molekuly RNA vytváří nejrůznější vlásenky, smyčky a dvojité nebo trojité šroubovice, vzájemnou interakcí mezi sebou. [1]



Obr. 1-1: Nukleotidy DNA vs. RNA [3]

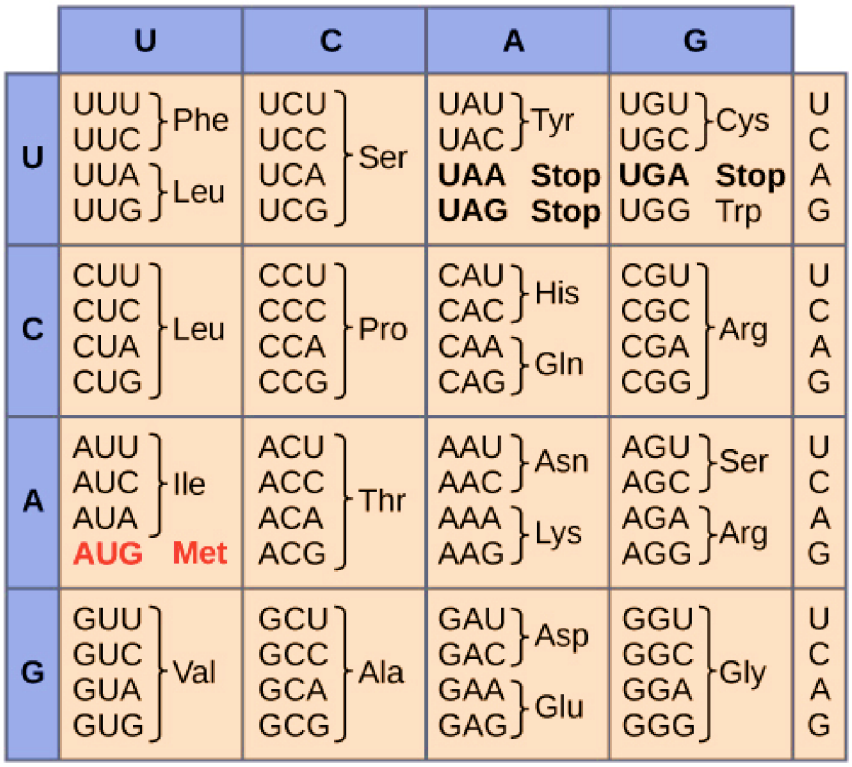
Jednou funkcí RNA je přenos genetické informace mezi DNA a proteinem. [1] Tento přenos má na svědomí mediátorová RNA (mRNA). Ta vzniká v buněčném jádře během procesu transkripce, kdy je DNA sekvence "přepsána" do RNA, následně projde posttranskripčními úpravami, a nakonec je poslána do cytoplazmy, kde slouží jako templát pro syntézu bílkovin během translace. [4]

I přes její nesmírnou důležitost tvoří mRNA jen malé procento z celkové RNA v buňkách. Nejpočetnější skupinou jsou ribozomální RNA (rRNA), která tvoří jádra ribozómů. Ribozómy jsou velké komplexy čtyř rRNA, a několika desítek proteinů. Za pomoci transferové RNA (tRNA) rozluští informaci uloženou a nesenou pomocí mRNA a následně vyrobí konkrétní bílkovinu. [1] Jednotlivé typy RNA jsou zobrazeny na Obr. **1-2**.



Obr. 1-2: Typy RNA, zleva mRNA, rRNA a tRNA [5]

Buňky dekódují informaci uloženou v mRNA čtením nukleotidů ve skupině po třech. Tato trojčlenná skupina nukleotidů se nazývá kodon. Většina kodonů zastupuje konkrétní aminokyselinu. Zbytek slouží jako START a STOP kodony, které značí, kde začíná a končí "překládaný" protein. Jednotlivé kodony jsou čteny z tabulky kodonů, viz Obr. **1-3**, která je čtena pomocí levého a pravého sloupce a horního řádku. Levý sloupec obsahuje první písmeno kodonu, horní řádek obsahuje druhé písmeno kodonu a pravý sloupec obsahuje čtveřici nukleotidů, které se nachází na poslední pozici. [6]



Obr. 1-3: Tabulka jednotlivých kodonů [7]

1.2 Struktura RNA

RNA je typicky jednovláknový biopolymer. Ovšem přítomnost autokomplementárních sekvencí má za následek párování bází a skládání ribonukleotidového řetězce do složitých strukturních forem, jako jsou smyčky a šroubovice. [8][9] Molekuly RNA mají tendenci se párovat do spirálovité struktury všude tam, kde se nacházejí dvě doplňující se části sekvence. Zpravidla tohle párování probíhá na základě Watson-Crickovského teorému za vzniku vodíkových můstků mezi bázemi C-G a A-U. V RNA ovšem dochází k fenoménu a sice k párování mezi méně stabilními páry G-U. [8] Pár vzniklý nedodržením teorému Watson-Crickovského párování se nazývá kolísavý pár bází. [10] Rozlišujeme báze purinové (adenin a guanin) a pyrimidinové (cytosin a uracil). U pyrimidinových i u purinových bází vzniká esterová vazba mezi fosfátovou skupinou a deoxyribózou. [11]

Molekuly RNA vykazují extrémně komplexní a rozmanité struktury, které se účastní několika biologicky významných regulačních procesů, jako jsou molekulární rozpoznání, konformační změna a katalýza. Řetězec RNA dosahuje stability díky tvoření párů mezi komplementárními bázemi, za vzniku nekovalentních vazeb. [2]

Nekovalentní vazby jsou mnohem slabší než vazby kovalentní či iontové. Vznikají všude, kde se k sobě přibližují dvě, nebo více molekul, a významně ovlivňují chování, vlastnosti a reaktivitu vzniklých struktur. Pro vazby mezi jednotlivými nukleotidy v RNA

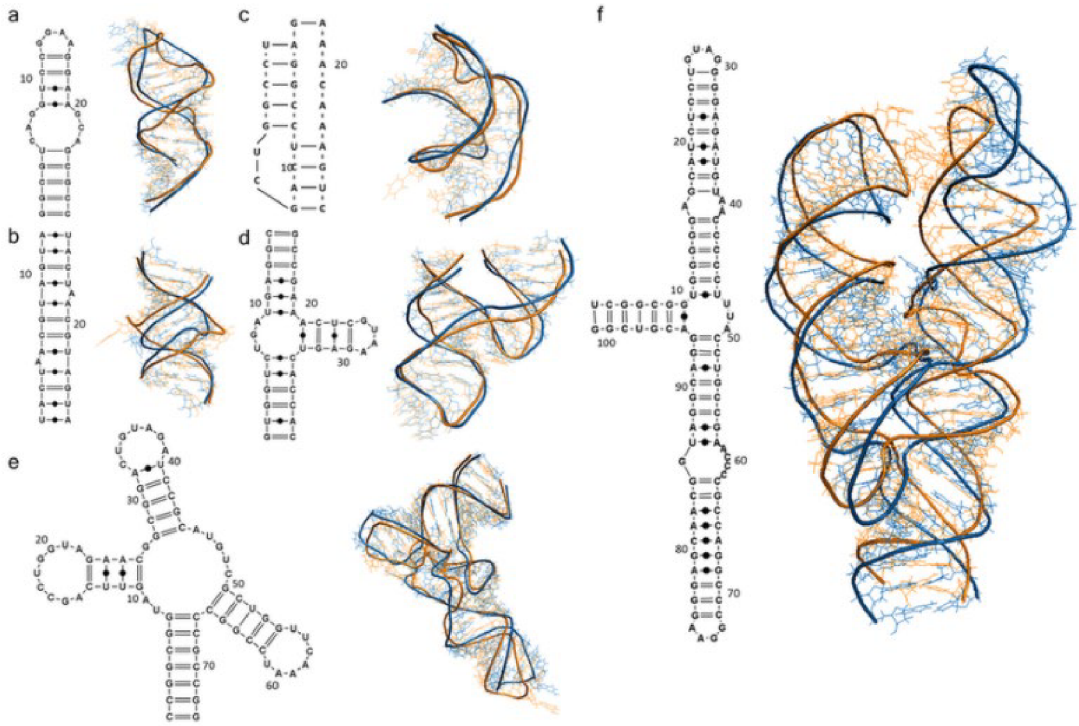
je nutné, aby byly stálé, a zároveň aby bylo možné tyto vazby rozvolňovat a spojovat. Nejsilnější známou nekovalentní vazbou je vazba vodíková (vodíkové můstky). [12] Formace vodíkových můstků mezi komplementárními nukleotidy je doprovázena termodynamickými zákony, zejména vznikem vazebné volné energie. Ta má pro stabilní strukturu RNA vždy zápornou hodnotu a je uváděna v jednotkách kcal/mol. K rozdělení vazby je potřeba vazebnou energii překonat, a tedy struktuře energii dodat, například ve formě tepla. [13]

RNA je schopna tvořit struktury primární, sekundární, terciální i kvarterní. Primární struktura RNA popisuje pouze nukleotidovou sekvenci, vzájemně propojenou pomocí fosfodiesterové vazby, tj. popisuje, jak jsou jednotlivé nukleotidy uspořádány v řadě za sebou, směrem od 5‘ k 3‘ konci. Příklad primární struktury je vyobrazen na Obr. **1-4**. [14]

Obr. 1-4: Příklad primární struktury RNA [15]

Sekundární strukturu RNA lze zjednodušeně popsat jako seznam spárovaných bází přítomných v nukleotidové sekvenci RNA. Existuje celkem 16 možností, jak párovat báze, ovšem jen 6 párů je natolik stabilních, že jsou schopny tvořit stabilní pár. Podrobněji je sekundární struktura RNA popsána v kapitole **1.2.1**. [14]

Známé terciální struktury obsahují převážně jen malé množství motivů, obsahují jen několik málo šroubovic a zpravidla vychází ze sekundární struktury. Některé ze známých motivů terciální struktury jsou uvedeny na Obr. **1-5**, společně se strukturou sekundární, ze které vychází. [16] Terciální struktura může být modelována pomocí molekulární dynamiky, která byla představena v roce 1998. Jedná se o metodu, zkoumající distribuci kovových iontů kolem molekuly RNA. [17] Tato metoda je však příliš pomalá pro simulaci procesu skládání molekul. Efektivnější metoda modelování terciální struktury hledá její možné prvky, které jsou v souladu se strukturou sekundární. Tuto metodu představili v roce 1999 I. Tinoco a C. Bustamante. [16]



Obr. 1-5: Předpovídané sekundární a terciální struktury RNA [18]

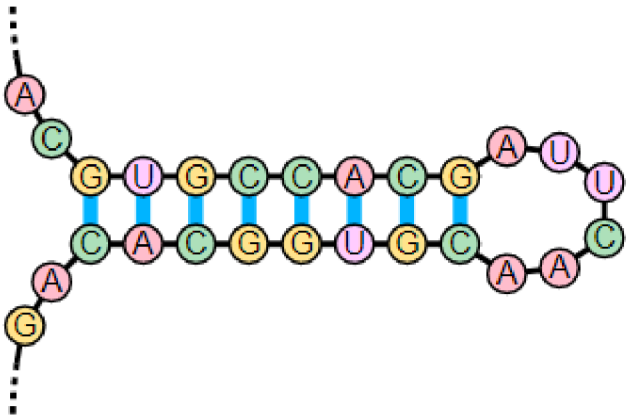
(a) vlásenková smyčka s vnitřní smyčkou, (b) dvojitá smyčka se dvěma výdutěmi, (c) pseudouzel 1KPZ, (d) Hammerhead ribozym RNA 1NYI, (e) tRNA 1J1U, (f) SRP RNA 1Z43

1.2.1 Sekundární struktura RNA

Studium sekundární struktury RNA je klíčové pro pochopení funkce a regulace transkriptů RNA. Jako primární transkript je označována RNA ihned po jejím vzniku. Strukturální elementy sekundární struktury RNA mají nespočet funkcí, mezi které patří například modulace epigenetických značek, nebo podpora velkých makromolekulárních komplexů. Společně s terciální strukturou se také účastní posttranskričpní regulace genové exprese. [19]

Sekundární struktura RNA udává, které komplementární báze v sekvenci vytvořily spojením pomocí vodíkových můstků pár. Rozlišujeme dvě skupiny bází a sice báze purinové (A, G) a pyrimidinové (C, U). Každá báze musí být spojena s bází s ní komplementární. Komplementární báze se nachází v opačných skupinách a stabilní páry tedy vznikají mezi bázemi A-U, C-G a G-U. Kolísavá pár bází G-U porušuje Watson-Crikovský teorém, ale zároveň přispívá k nápadné strukturální formaci. G-C páry mohou vytvářet až tři vodíkové můstky mezi svými Watson-Cricovskými konci, zatímco A-U páry mohou tvořit pouze dva. [20] Vlákno RNA může na základě své primární struktury vytvářet několik sekundárních struktur. Tyto struktury obvykle dělíme na stopky, smyčky, výdutě a pseudouzly, které vznikají nejrůznějším proplétáním RNA vlákna. [21]

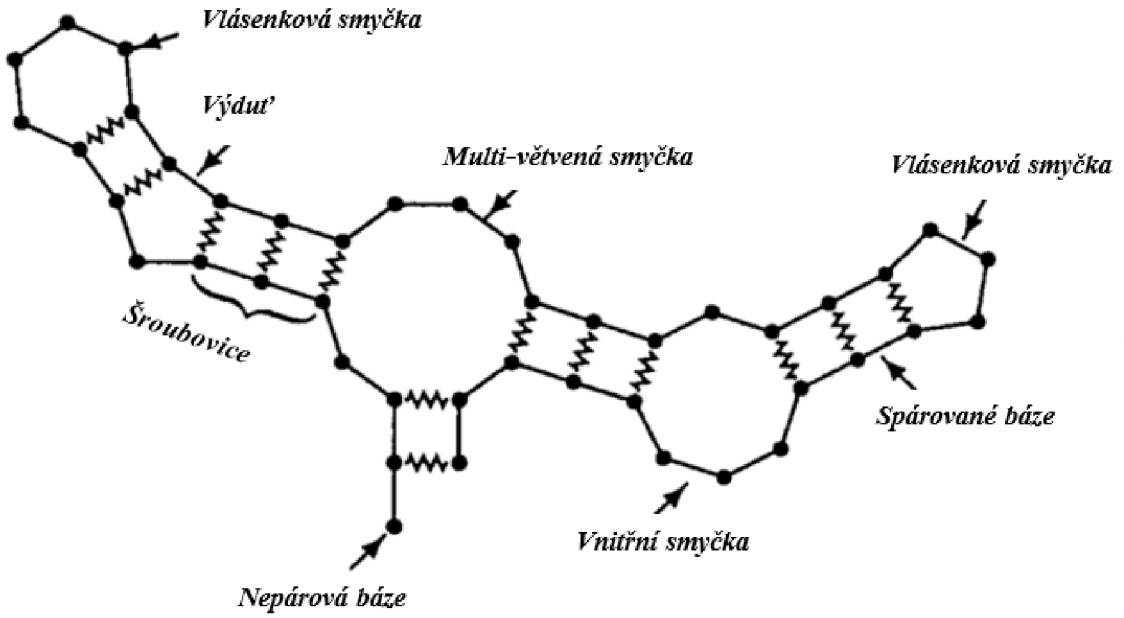
Nejznámější sekundární strukturou je vlásenková smyčka (*hairpin loop*), která vzniká překlopením RNA úseku sama na sebe. Překlopeny jsou k sobě nukleotidy, které se mohou párovat a vzniká tzv. šroubovice, na jejímž konci vzniká jednovláknová smyčka stávající se z nukleotidů, které jsou nepárové. [21] Obě tyto struktury jsou vyobrazeny na Obr. **1-6**. Speciálním případem vlásenkové smyčky je líbající vlásenková smyčka (*kissing hairpin loop*), která vzniká spárováním bází ve dvou oddělených vlásenkových smyčkách. [22]



Obr. 1-6: Vlásenková smyčka a šroubovice (*hairpin loop a helix*) RNA [23]

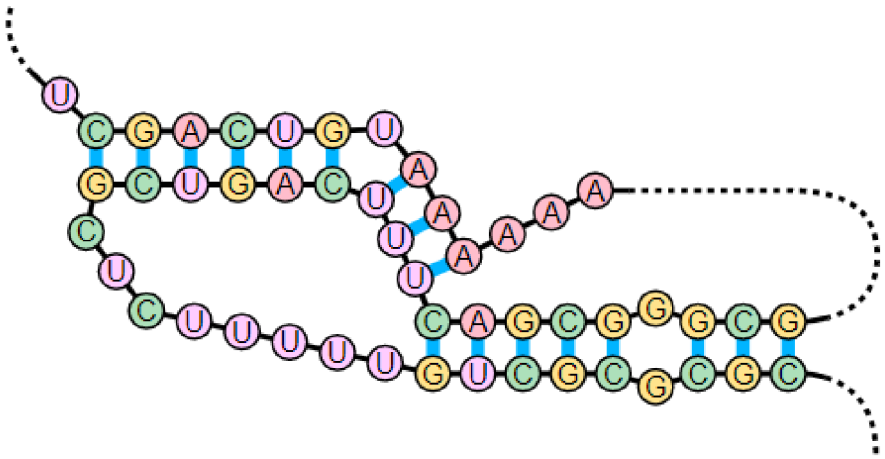
Dále rozlišujeme smyčky vnitřní, vnější a multismyčky. Vnitřní smyčky (*interior loops*) vznikají tam, kde nedochází k párování jednotlivých bází. Vnitřní smyčky se nachází uvnitř sekundární struktury RNA na rozdíl od smyček vláseknových, které se nacházejí na jejím konci. Rozeznáváme vnitřní smyčky symetrické a vnitřní smyčky asymetrické. Speciální případy vnitřních asymetrických smyček jsou nazývány výdutě (*bulges*), které mohou být tvořeny jen jedním nukleotidem. [24] Vnější smyčka (*exterior loop*) spojuje oba, 5‘ a 3‘, konce RNA a alespoň jednou spárovanou bázi. Poslední ze smyček, která je tvořena sekundární strukturou RNA, je multismyčka. Multismyčka (*multi-loop*) je místo v RNA, kde se spojují 3 nebo více spárovaných bází, kdy tyto báze mohou, ale nemusí být odděleny nepárovými bázemi. [21] Speciálním případem multismyčky je multi-větvená smyčka (*multi-branch loop*), která vzniká tam, kde dochází ke spojení několika kmenových smyček do jedné. [25]

Žádná z výše zmíněných struktur se obvykle nevyskytuje samostatně. Je zvykem, že na jedné sekvenci vzniká několik různých strukturálních elementů, které na sebe navazují a jeden v druhý přecházejí. Příklad sekvence, která tvoří několik strukturálních elementů je na Obr. **1-7**.



Obr. 1-7: Rozložení RNA sekundární sktruktury na strukturální elementy, upraveno [26]

Nejsložitějším strukturálním elementem pro sekundární strukturu RNA je struktura pseudouzlu. Tato struktura vzniká vzájemným proplétáním a křížením vazeb, které vznikají mezi komplementárními bázemi. [8] Podstrukturou pseudouzlu je H-typ pseudouzel, který vzniká utvořením bázového páru mezi vlásenkovou smyčkou a jednovláknovou oblastí vlásenky. H-typ pseudouzl tedy obsahuje dvě šroubovicové struktury a dvě smyčky. Příklad pseudouzlu v RNA je vyobrazen na Obr. **1-8**. [27]



Obr. 1-8: Pseudouzel RNA [28]

1.2.2 Vykreslení sekundární struktury RNA

Vyobrazení sekundární struktury RNA je možné pomocí několika grafických přístupů. Nejběžnější a nejjednodušší je závorková-tečková notace, která obsahuje pouze 3 znaky pro nepseudouzlové struktury. [29] Příklad závorkové-točkové notace je uveden na Obr. **1-9**(c).

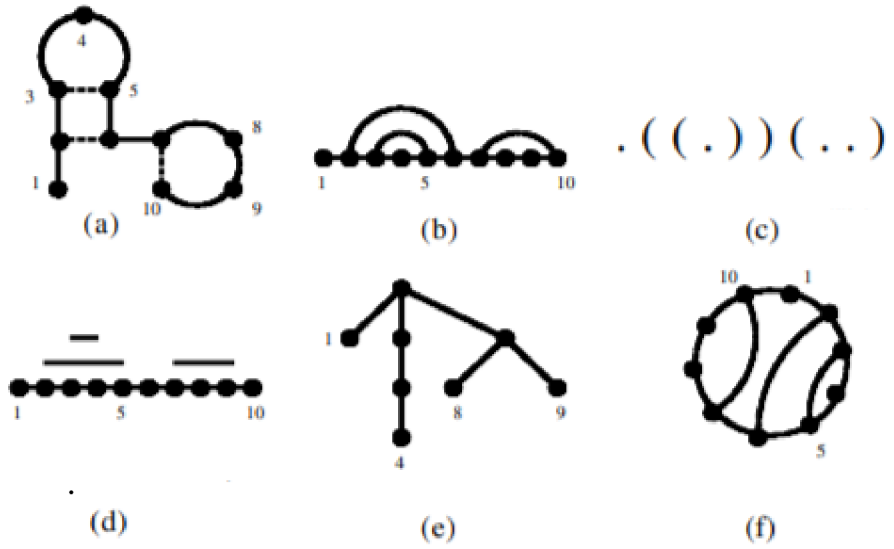
Závorková-tečková notace je jednou z nejběžnějších a nejjednodušších reprezentací sekundární struktury RNA. Tyto znaky znázorňují otevřený a uzavřený pár bází a také místo v sekvenci, kde se nachází smyčka, nebo výduť. Otevřený pár bází vzniká v okamžiku, kdy nukleotid vytvoří pár s nukleotidem nacházejícím se před ním. Uzavřený pár bází vzniká, když nukleotid utvoří pár s nukleotidem, který se nachází za ním. [29]

Znaky reprezentující závorkovou-tečkovou notaci pro nepseudouzlové struktury jsou:

- „**.**“, je kódovacím znakem pro místo v sekvenci, kde nedochází k párování bází, jedná se tedy o místo v sekvenci, kde se nachází nějaká smyčka nebo výduť [29],
- „**)**“, kóduje místo v sekvenci, kde se nachází otevřený pár bází [29],
- „**(**“, kóduje místo v sekvenci, kde se nachází uzavřený pár bází [29].

Nejpřehlednější reprezentace je reprezentace prostým diagramem, vyobrazena na Obr. **1-9**(a). Toto zobrazení ukazuje jednotlivé strukturální elementy a to, jak na sebe navazují. Další možností je reprezentace kruhová, nebo lineární. V těchto reprezentacích jsou báze tvořící pár spojeny pomocí obloukových křivek. V místech, kde dochází ke křížení těchto křivek, lze předpokládat výskyt pseudouzlu. [30] Příklad kruhové a lineární reprezentace je na Obr. **1-9**(b) a Obr. **1-9**(f).

Dále rozlišujeme reprezentaci horou a stromem (*mountain a tree representation*). Reprezentace horou vychází ze závorkové-tečkové notace. Každý pár bází je reprezentován horizontální linií nad primární sekvencí. Tyto horizontální linie jsou skládány nad sebe a nachází se tím výše, čím hlouběji v sekvenci se vytvořený pár nachází. Reprezentace pomocí stromu také vychází ze závorkové-tečkové notace s tím rozdílem, že sdružují substruktury pod označené vrcholy. [30] Příklad reprezentace horou a stromem je na Obr. **1-9**(d) a Obr. **1-9**(e).



Obr. 1-9: Příklady reprezentace sekundární struktury RNA, upraveno [30]

(a) prostý diagram, (b) reprezentace lineární, (c) závorková-tečková notace, (d) reprezentace horou, (e) reprezentace stromem, (f) reprezentace kruhová

2. METODY PREDIKCE SEKUNDÁRNÍ STRUKTURY RNA

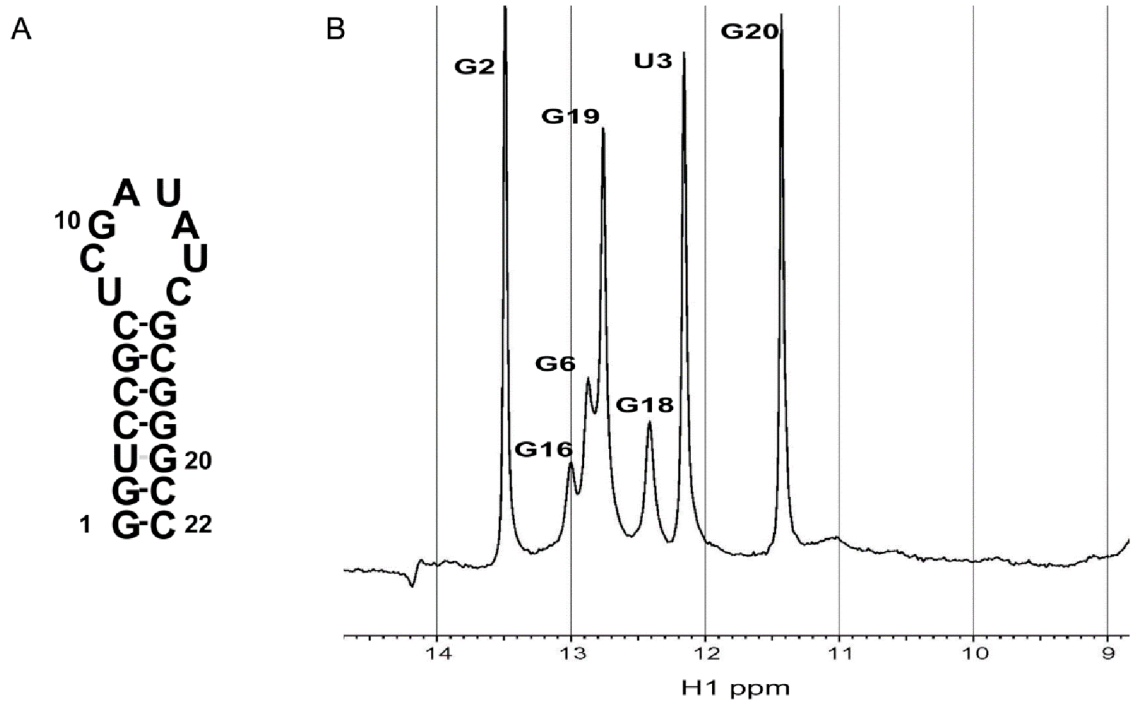
I přes její nesmírnou důležitost se sekundární struktura dočkala větší pozornosti až v posledních desetiletích. Velmi rozšířeným oborem je tedy studium její predikce. Správně predikovaná sekundární sktruktura je klíčová k pochopení všech výše zmíněných procesů, za které je sekundární sktruktura RNA zodpovědná. [31]

Obor metod predikce sekundární struktury můžeme rozdělit na dvě skupiny: metody experimentální a metody výpočetní. Také rozlišujeme metody, které nezahrnují predikce pseudouzlů a metody, které ji zahrnují. Bylo prokázáno, že pseudouzly jsou pro sekundární strukturu RNA funkčně důležité, avšak metody, které jejich predikci zahrnují stále vyžadují mnohá vylepšení. Naopak metody predikující nepseudouzlové struktury jsou velice přesné. [31]

2.1 Experimentální metody predikce

Experimentálních metod predikce sekundární struktury RNA je známých hned několik. Obvykle jsou ale všechny experimentální metody časově i finančně náročné a mnohdy neproveditelné. Mezi nejrozšířenější experimentální metody predikce patří: nukleární magnetická rezonance (NMR), rentgenová (RTG) krystalografie a kryoelektronová mikroskopie. [32]

NMR je šetrná a nedestruktivní metoda, využívající, obdobně jako klasická magnetická rezonance (MR), magnetických vlastností atomových jader. Při vložení zkoumané struktury do vnějšího magnetického pole dojde ke srovnání magnetických momentů jader atomů s magnetickými momenty vnějšího magnetického pole. [33] Orientaci magnetických momentů jader lze měnit, je-li jádrům dodána energie vhodným způsobem. U NMR je tato energie dodána formou radiofrekvenčního záření. Čím vyšší je hodnota magnetického pole a frekvence záření, tím lepší citlivost a rozlišovací schopnost pro zkoumané struktury. NMR dokáže odhalit, jak jsou na sebe molekuly a nukleotidy vázány, jak daleko od sebe se nachází, nebo jaké úhly jejich vazby svírají. Všechny informace o struktuře jsou získány analýzou spekter, která odhalí NMR. Příklad výstupu NMR pro predikci sekundární struktury RNA je na vyobrazen na Obr. **2-1**. [32]



Obr. 2-1: Ukázka predikce sekundární struktury RNA pomocí NMR [34]

(A) predikovaná sekundární struktura, (B) graf analýzy spekter

RTG krystalografie je metoda studující interakce krystalických vzorků s RTG zářením. Dokáže odhalit absolutní strukturu zkoumaného vzorku. Monochromatické RTG světlo prochází zkoumaným vzorkem, přičemž dochází k jeho ohybu (difrakci). Do jaké míry a jakým směrem k difrakci dochází záleží na vnitřní sktruktuře zkoumaného vzorku. Intenzita a směr lomu jsou měřeny a po korekci je určena absolutní struktura vzorku. [35]

Kryoelektronová mikroskopie zkoumá buněčné struktury v atomovém rozlišení. Její princip spočívá ve velmi rychlém ochlazení zkoumaného, zavodněného vzorku na kryogenní teplotu, která se pohybuje okolo -200 ˚C, a následné ponoření do kapalného etanu. V ledu se nestihne vytvořit krystalová mřížka, a vzorek je obalen amorfním ledem připomínajícím sklo, který přesně kopíruje strukturu zkoumaného vzorku. [36]

2.2 Výpočetní metody predikce

Experimentální metody predikce jsou velice přesné a téměř bezchybné. Avšak, jak už bylo řečeno výše, jsou velice časově náročné, drahé a mnohdy proveditelné jen za velice specifických podmínek, nebo nejsou proveditelné vůbec. Proto jsou v této oblasti velice slibnou alternativou výpočetní metody predikce. Ty pracují pouze se vstupní sekvencí nukleotidů a na základě matematických postupů predikují nejoptimálnější strukturu pro zkoumanou sekvenci. [31] Výstupem výpočetních metod predikce je sekundární struktura

nejčastěji ve formě závorkové-tečkové notace, která je podrobněji popsána v předchozí kapitole **1.2.2**.

Výpočetní metody predikce počítají se skutečností, že sekundární struktura není tvořena náhodně, ale že její utváření je závislé na setu několika pravidel. Tyto pravidla jsou celkem 4. První pravidlo přiřazuje jednotlivým bázím v sekvenci čísla od 1 po N a určuje, že párování komplementárních bází může probíhat na pozicích *i* a *j* v případě, že je absolutní hodnota rozdílu mezi *j* a *i* větší nebo rovna 4. Tato podmínka vzdálenosti zajistí, že nedojde k utvoření páru mezi sousedními bázemi. V takovém případě by totiž sekvence RNA nemohla tvořit další struktury. Druhým pravidlem pro utvoření stabilní sekundární struktury je, že páry mohou tvořit pouze komplementární báze. [37] V případě RNA se jedná o báze A-U, C-G a G-U [8]. Třetím pravidlem je ošetřeno, že nukleotid na pozici *i* se vyskytuje v nejvýše jedné spárované bázi. Čtvrté pravidlo říká, že při vzniku dalšího páru bází na pozicích *k* a *l* musí platit $i < k < l < j$. [37]

Výpočetní metody pro predikci sekundární struktury RNA jsou realizovány pomocí algoritmů využívajících dynamického programování. Podstata dynamického programování spočívá ve zjednodušení komplikovaného problému na několik dílčích, menších podproblémů, které se vzájemně překrývají. Konečné řešení původního, komplikovaného problému je pak kombinací všech dílčích řešení. Hlavní výhodou dynamického programování je skutečnost, že není tak časově náročné a zvyšuje efektivitu implementovaného algoritmu. To je způsobeno tím, že předchozí řešení dílčích problémů jsou ukládány a v případě potřeby vytahovány z paměti. Tímto nejsou s každou dílčí částí počítány znovu. [38]

Implementovaný algoritmus predikující sekundární strukturu RNA je nutné nějakým způsobem ohodnotit. Hodnocení korektnosti algoritmu je možné stanovit, například výpočtem senzitivity, pozitivní prediktivní hodnoty (PPV), F-skóre, nebo stanovením složitosti algoritmu. [39]

Senzitivita algoritmu neboli jeho citlivost, stanovuje schopnost algoritmu správně určit páry bází v sekvenci a nabývá hodnot od 0 do 1. [39] Výpočet je dán jako poměr všech správně predikovaných párů bází ku všem bázovým párům v predikované sekvenci. Matematický předpis je zobrazen rovnicí 2.1. [20]

$$Senzitivita = \frac{Počet\ všech\ správně\ predikovaných\ bázových\ párů}{Počet\ všech\ bázových\ párů\ v\ sekvenci} \tag{2.1}$$

PPV hodnota určuje, s jakou pravděpodobností je predikovaný bázový pár skutečně párem a nabývá hodnot od 0 do 1. [39] Výpočet je dán jako poměr správně predikovaných bázových párů v sekvenci ku všem predikovaným bázovým párům. Matematický předpis je zobrazen rovnicí 2.2. [20]

$$PPV = \frac{Počet\ správně\ predikovaných\ bázových\ párů\ v\ sekvenci}{Počet\ všech\ predikovaných\ bázových\ párů} \tag{2.2}$$

F-skóre vyjadřuje míru přesnosti implementovaného algoritmu a nabývá hodnot od 0 do 1. [39] Výpočet je dán jako poměr součinu dvakrát senzitivity a PPV ku součtu senzitivity a PPV. Matematický předpis je zobrazen rovnicí 2.3. [20]

$$F - skóre = \frac{2 * senzitivita * PPV}{senzitivita + PPV} \tag{2.3}$$

Poslední zmiňovanou možností, jak ohodnotit implementovaný algoritmus, je stanovením jeho složitosti. Celkovou složitost implementovaného algoritmu posuzujeme na základě 2 parametrů – čas a paměť. [41] Jako nejlepší algoritmus řešící zkoumanou problematiku je vždy ten, který má co nejmenší jak časovou, tak paměťovou náročnost. Nejvíce popisovanou složitostí je složitost asymptotická, která je nástrojem k popsání a porovnání efektivity algoritmu. [40] Je vyjádřením trendu, jakým roste, ať už časové, nebo prostorové náročnosti, s rostoucím množstvím vstupních dat. [41]

Samotnou asymptotickou složitost je možné vyjádřit pomocí několika notací. První možnou notací je notace Omega (Ω), která vyjadřuje „nejlepší případ“ a to tedy, že náročnost algoritmu bude s přibývajícím množstvím dat růst stejně, nebo více. Druhou možnou a používanou notací, je notace Omikron (O) vyjadřující „nejhorší případ“. Tedy že složitost algoritmu roste stejně, nebo méně. Poslední možnou notací je notace Theta (Θ) popisující složitost algoritmů v „průměrném případě“. Tato notace je vyjádřením složitosti algoritmu, jehož složitost roste s množstvím vstupních dat vždy stejně. [40]

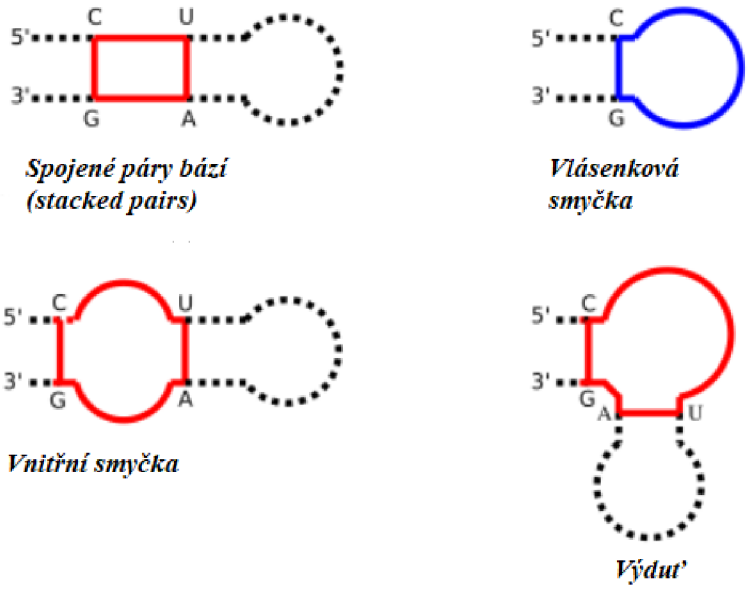
K samotné predikci je možné přistupovat na základě několika teorií. Mezi nejznámější patří predikce na základě maximalizace bázových párů a predikce na základě minimalizace volné energie. Oba tyto přístupy hledají co nejstabilnější sekundární strukturu RNA. Maximalizace bázových párů pracuje s myšlenkou, že nejstabilnější je ta struktura, která obsahuje největší množství párů bází, a tedy minimální množství smyček. Přístup pomocí minimalizace volné energie považuje za nejstabilnější tu strukturu, která má co nejmenší hodnotu volné energie. [42]

Přístup k predikci pomocí volné energie uvažuje, že vytváření a případný rozpad sekundární struktury RNA doprovází termodynamické zákony. Zejména Gibbsova volná energie je veličinou uvozující sekundární strukturu RNA. [31] Termodynamika vzniku a zániku sekundárních struktur vychází obecně z principů přeměny energie a celková termodynamika molekuly RNA odpovídá tedy sumě energií, potřebné k spojení, nebo rozpojení párů bází. [43] V molekulách probíhají veškeré přeměnné procesy za stálého tlaku a stálé teploty, proto je k popisu energických reakcí použita změna Gibbsovi volné energie a změna entalpie, která odpovídá změně celkové vnitřní energie. [44] Matematický přepis pro výpočet Gibbsovy volné energie je zobrazen rovnicí 2.4.

$$\Delta G = \Delta H - T\Delta S\ (kcal/mol), \tag{2.4}$$

kde ΔH znamená změnu entalpie, ΔS změnu entropie a T termodynamickou teplotu (310,15 K = 37 ˚C). [45] Pro výpočet volné energie sekundárních struktur se používají energetické modely, které obsahují experimentálně získané hodnoty pro nejrůznější sekundární struktury RNA.

Energetické modely popisují, jaký je příspěvek volné energie jednotlivých sekundárních struktur k celkové struktuře RNA v jednotkách kcal/mol. [31] Příspěvky jednotlivých struktur byly experimentálně naměřeny podle Freierových pravidel při teplotě 37 ˚C v roztoku NaCl o koncentraci 1 mol/l. Teplota 37 ˚C představuje fyziologickou teplotu a koncentrace 1 mol/l NaCl zajišťuje neutrální pH prostředí. [46] Struktury, pro které byly tyto hodnoty experimentálně určeny jsou čtyři a jsou zobrazeny na Obr. **2-2**. Patří k nim vlásenková a vnitřní smyčka, výduť a stopka (spojené páry bází).



Obr. 2-2: Struktury zohledňované v energetických modelech, upraveno [47]

Nejznámějším a nejhojněji používaným energetickým modelem je model Turnerův. První Turnerův model byl představen v roce 1999, který ovšem uvažoval pouze volnou energii spojených struktur. Jeho obnova z roku 2004 počítá navíc se změnou entalpie strukturálních elementů. Pro RNA uvažuje tento model změny entalpie pro Watson-Crickovské páry, kolísavé páry, vlásenkové smyčky, vnitřní smyčky a výdutě. [48] Hodnoty pro model Turner 2004 jsou uvedeny v Tabulce **2-1** a Tabulce **2-2**.

Model Turnera vychází z termodynamických parametrů nejbližšího souseda (*nearest-neighbour thermodynamics parametrs*). Ovšem pro RNA existuje samostatná sada pravidel, která zahrnuje stabilitu nejrůznějších smyček a výdutí, které jsou tvořené nejrůznějším spojováním komplementárních bází. [46]

Tabulka 2-1: Hodnoty volné energie pro spojené páry (*stacked pairs*) podle modelu Turnera 2004 [kcal/mol] [49]

	A-U	**C-G**	**G-C**	**U-A**	**G-U**	**U-G**
A-U	-0,9	-1,8	-2,3	-1,1	-1,1	-0,8
C-G	-1,7	-2,9	-3,4	-2,3	-2,1	-1,4
G-C	-2,1	-2,0	-2,9	-1,8	-1,9	-1,2
U-A	-0,9	-1,7	-2,1	-0,9	-1,0	-0,5
G-U	-0,5	-1,2	-1,4	-0,8	-0,4	-0,2
U-G	-1,0	-1,9	-2,1	-1,1	-1,5	-0,4

Tabulka 2-2: Hodnoty volné energie pro smyčky a výdutě podle modelu Turnera 2004 [kcal/mol] [49]

Počet bází ve smyčce (výduti)	**Vlásenková smyčka**	**Vnitřní smyčka**	**Výduť**
1	∞	∞	3,8
2	∞	∞	2,8
3	5,4	∞	3,2
4	5,6	1,1	3,6
5	5,7	2,0	4,0
10	6,5	2,5	4,9
15	6,9	2,9	5,4
30	7,7	3,7	6,1

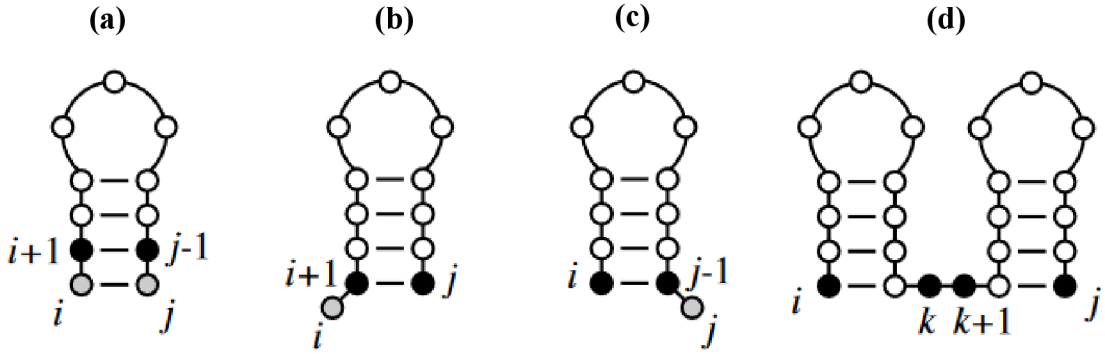
2.2.1 Algoritmus podle Nussinové

Metoda predikce pomocí algoritmu Nussinové byla představena v roce 1978 doktorkou biologie Ruth Nussinovovou, jako jedna z prvních metod pro predikci sekundárních struktur RNA. Jedná se o jednodušší ale velmi efektivní techniku predikce sekundární struktury využívající principy dynamického programování, která by měla, v nejhorším případě, pracovat v $\Theta(n^2)$ čase a $\Theta(n^3)$ paměti. Algoritmus Nussinové predikuje na základě přístupu o maximalizaci párů bází a neuvažuje pseudouzlové struktury. [50]

Algoritmus Nussinové uvažuje set několika pravidel, na jejichž základě stanovuje výslednou predikci sekundární struktury RNA. Tato pravidla mimo jiné určují, které báze v sekvenci mohou tvořit pár. V případě RNA uvažujeme komplementární báze podle Watson-Crikovského párování i kolísavý pár bází. [51] V první řadě je třeba si uvědomit,

že algoritmus Nussinové uvažuje shodnou stabilitu pro všechny možné páry. Dále, je nutné ignorovat příspěvky volné energie, ke kterým dojde vlivem párování, a také destabilizační efekt, který mají na strukturu vzniklé smyčky. V případě, že dochází k predikci na velmi dlouhých sekvencích, dochází před samotnou predikcí k rozdělení sekvence na několik kratších podsekvencí a predikce optimální sekundární struktury probíhá na každé podsekvenci zvlášť. [52]

Algoritmus uvažuje celkem čtyři možnosti párování bází. První možností je, že se nepárová báze na místě *i* naváže na bázi na místě *i+1*, která vytvořila pár s bází na místě *j*. Tato možnost je vyobrazena na Obr. **2-3**(b). Druhou možností je, že se nepárová báze na místě *j* naváže na bázi na místě *j-1*, která vytvořila pár s bází na místě *i*. Tato možnost je vyobrazena na Obr. **2-3**(c). Třetí možnost je navázání spárovaných nukleotidů na místech *i* a *j*, ke spárovaným nukleotidům na místě *i+1* a *j-1*. Tohle spojení je nazýváno spojenými páry bází (*stacked pairs*) a je základem pro vytvoření struktury stopky. Tato možnost je vyobrazena na Obr. **2-3**(a). Poslední možností, jak se mohou párovat komplementární báze, je kombinace dvou optimálních substruktur na pozicích *i,k* a *k+1, j*. Tato struktura je vyobrazena na Obr. **2-3**(d) a je známá jako rozdvojení struktury (*burification*) a je základem pro tvorbu multismyček. [52]



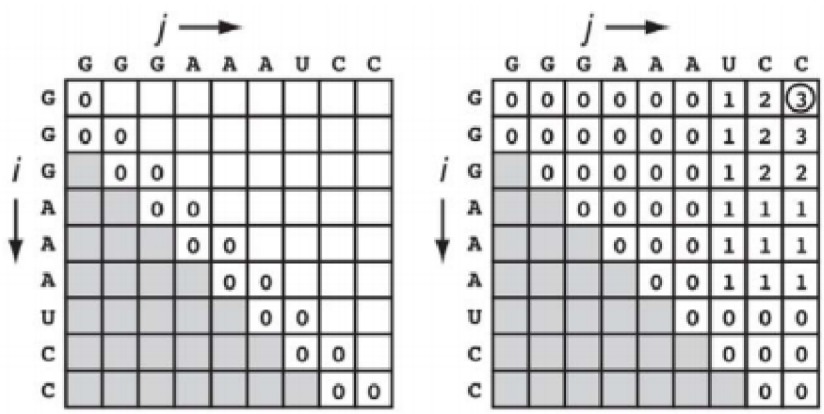
Obr. 2-3: Možnosti predikce struktur pomocí algoritmu Nussinové [53]

Predikce sekundárních struktur pomocí algoritmu Nussinové lze rozdělit na tři dílčí části: vyplnění skórovací matice M, nalezení zpětné cesty a ohodnocení zpětné cesty znaky závorkové-tečkové notace.

V prvním kroku je třeba vyplnit skórovací matici M o rozměrech *n* x *n,* kde *n* je délka zkoumané sekvence. Hlavní diagonála a diagonála pod ní je inicializována nulami. [51] Matice je vyplňována postupně po diagonálách až do pravého horního rohu. Kontroluje se komplementarita jednotlivých bází na pozicích *i, j* a matice je plněna hodnotami podle rovnice 2.5. Poslední řádek rovnice odkazuje na vznik výše zmíněného rozdvojení struktury. [52]

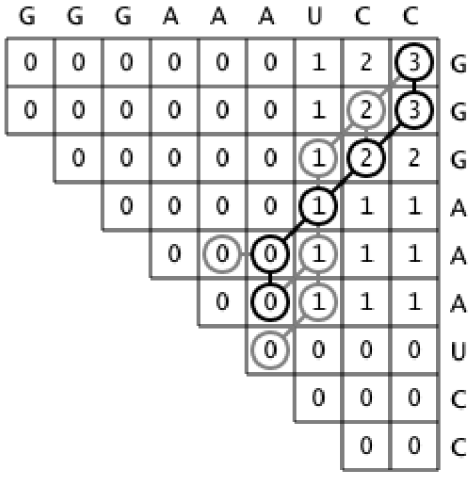
$$M(i,j) = max \begin{Bmatrix} M(i+1, j-1) + \omega(i,j) \\ M(i, j-1) \\ M(i+1, j) \\ max_{i<k<j}[\omega(i,k) + \omega(k+1, j)] \end{Bmatrix} \quad (2.5)$$

Hodnota ω(*i,j*) je rovna 1 v případě, že jsou báze na pozicích *i, j* komplementární. V případě, že báze komplementární nejsou, nabývá ω(*i,j*) hodnoty 0. [52] Příklad vyplněné tabulky pro algoritmus Nussinové je na Obr. **2-4**.



Obr. 2-4: Příklad inicializované a následně vyplněné skórovací matice M [53]

Po vyplnění skórovací matice je nutné nalézt zpětnou cestu. Hledání zpětné cesty začíná v pravém horním rohu, kde se nachází maximální hodnota matice M. Zpětné trasování po maximech probíhá až do chvíle, kdy je nalezeno místo v matici, kde se nenachází žádná hodnota. [55] Směry zpětné cesty jsou možné celkem tři, stejně jako prvky závorkové-tečkové notace. Každému směru je tedy přidělen jeden prvek. Pro přehlednost při hledání zpětné cesty se doporučuje tvořit současně se skórovací maticí M i trasovací matici T. Trasovací matice obsahuje celkem 3 hodnoty, které jsou ukazatelem, ze kterého směru bylo vybráno maximum při vyplňování skórovací matice M. Možné zpětné cesty vyplněnou maticí M jsou vidět na Obr. **2-5**.



Obr. 2-5: Možnosti zpětné cesty pro algoritmus podle Nussinové [54]

Jak je patrné na Obr. **2-5**, neexistuje pouze jedna možnost zpětné cesty, ale několik. Už během implementace algoritmu Nussinové záleží na několika ovlivnitelných parametrech. Prvním je, kterou ze 3 možností pro výběr maxima bude upřednostňovat. Dále je nutné řešit, který směr bude upřednostňován při hledání zpětné cesty a jak bude zpětná cesta ohodnocována prvky závorkové-tečkové notace.

2.2.2 Algorimus podle Zukera

Algoritmus podle Zukera je jedním z nejpopulárnějších metod pro predikci sekundární struktury RNA. Počítá s přístupem minimalizace volné energie a je schopný pracovat v nejhorším případě v $\Theta(n^4)$ čase a $\Theta(n^2)$ paměti. [56] Poprvé byl představen v roce 1981 Michaelem Zukerem a dodnes je hojně používán v nejrůznějších volně dostupných softwarech pro predikci sekundární struktury RNA. Nejlepších výsledků dosahuje pro nepseudouzlové a kratší sekvence. Přesnost jeho predikce klesá, jak délka sekvence roste. Jeho základní verze nepočítá s pseudouzlovými sktrukturami, avšak v dnešní době existuje několik obměn, které predikují i tyto struktury. Pro predikci optimální sekundární struktury RNA počítá s experimentálními hodnotami energetických modelů. [51] Příklady hodnot pro energetický model Turner 2004 jsou uvedeny výše v Tabulce **2-1** a Tabulce **2-2**.

Algoritmus podle Zukera pracuje se dvěma skórovacími maticemi: V a W. Matice V*(i,j)* obsahuje nejmenší hodnoty volné energie pro zkoumané části sekvence. Matice W*(i,j)* obsahuje celkovou minimální volnou energii zkoumané části sekvence. [57] Hodnoty minimální volné energie pro matice V*(i,j)* a W*(i,j)* jsou vybírány podle rovnic 2.6 a 2.7.

$$V(i,j) = min\begin{Bmatrix} eh(i,j) \\ es(i,j) + V(i+1,j-1) \\ VBI(i,j) \\ VM(i,j) \end{Bmatrix} \quad (2.6)$$

$$W(i,j) = min\begin{Bmatrix} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \\ min_{i<k<j}[W(i,k) + W(k+1,j)] \end{Bmatrix} \quad (2.7)$$

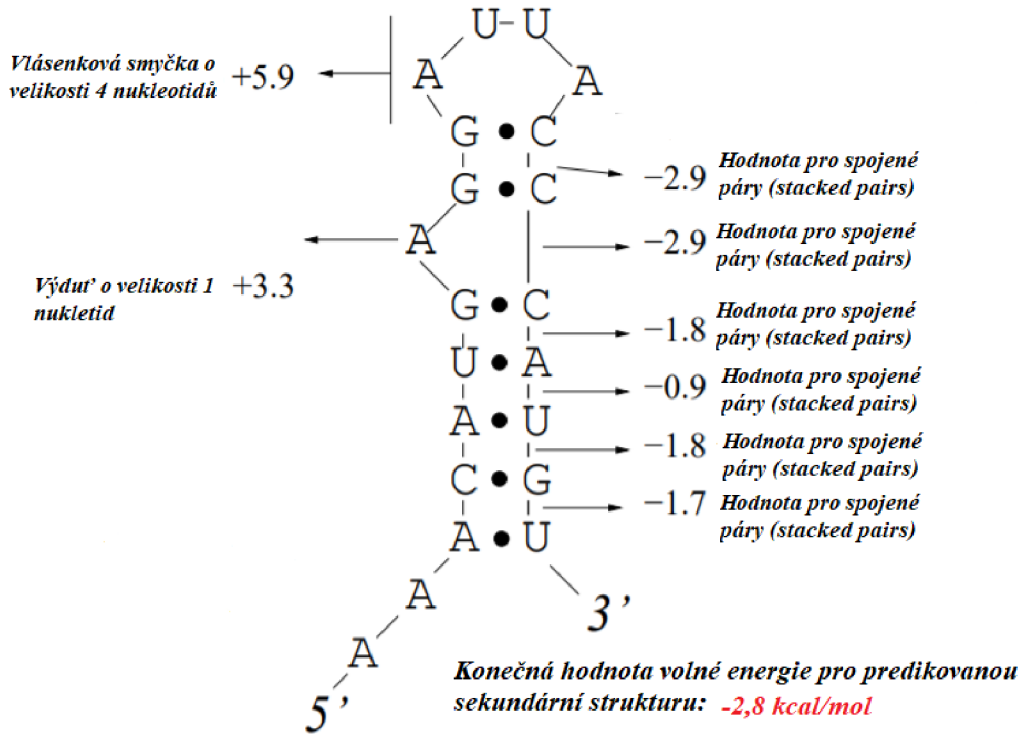
Hodnota *eh* pro matici V značí hodnotu volné energie pro vlásenkovou smyčku, která je uzavřena párem bází na místech *i* a *j*. Hodnota *es* pro matici V zastupuje hodnotu spojených páru bází (*stacked pairs*) na místech *i* a *j* s bázemi na místech *i+1* a *j-1*. *VBI* je volná energie vnitřní smyčky, nebo výdutě, která je uzavřena párem bází na místech *i* a *j*. Hodnoty vnitřní smyčky nebo výdutě jsou brány pro strukturu na pozicích *i'* a *j'*. Tato struktura se nachází uvnitř *i* a *j*. Hodnotu pro výduť uvažujeme v okamžiku, kdy $i' - i > 1$, nebo $j - j' > 1$. Hodnotu pro vnitřní smyčku uvažujeme, pokud $i' - i > 1$ a současně $j - j' > 1$. Konečná hodnota pro *VBI* je vybírána podle vztahu, který je uveden v rovnici 2.8. Hodnota *VM* pro matici V značí místo, kde dochází k rozdvojení struktury (*burification*), a které je základem pro tvorbu multismyček. Konečná hodnota pro *VM* je vybírána podle vztahu, který je uveden v rovnici 2.9, kde konstanta *a* značí konstantní přírůstek volné energie pro vzniklou multismyčku. [51]

$$VBI(i,j) = min\{ebi(i,j,i',j') + V(i',j')\}_{\,i<i'<j'<j} \quad (2.8)$$

$$VM(i,j) = min\{W(i+1,k) + W(k+1,j-1)\}_{i<k<j-1} + a \quad (2.9)$$

Ze skórovací matice W určíme, které báze utvořily pár. Pro lepší vizualizaci je doporučeno tvořit současně s maticí W i trasovací matice T. Získáme tedy vektor čísel značící pozice jednotlivých párů. Tyto pozice jsou pak v konečné závorkové-tečkové struktuře nahrazeny znaky pro otevřené a uzavřené páry bází. [51]

Konečná hodnota volné energie pro strukturu predikovanou algoritmem podle Zukera se nachází v pravém horním rohu matice W. Příklad sekundární struktury a konečné hodnoty volné energie, počítané podle modelu Turner 99, je zobrazen na Obr. **2-6**. [51]

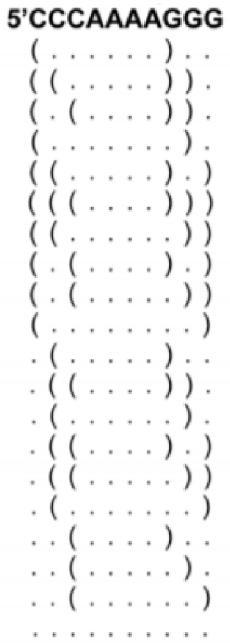


Obr. 2-6: Příklad výpočtu konečné hodnoty volné energie pro sekundární strukturu RNA predikovanou pomocí Zukerova algoritmu, upraveno [51]

2.2.3 Metoda Crumple

Predikce metodou Crumple vytváří všechny možné nepseudouzlové struktury bez ohledu na termodynamiku sekvence a nabízí tedy alternativní přístup k tradičním algoritmům, predikujících sekundární strukturu na principu minimalizace volné energie. Je schopný odhalit nepseudouzlové struktury, které tradičními algoritmy, počítající s minimální volnou energií, nejsou předpovídány, přesto, že přístup minimalizace volné energie je velmi úspěšný. [58]

Obecný přístup predikce pomocí metody Crumple je založený na přístupu, který hledá strukturu ve struktuře. Vstupem do algoritmu, predikující sekundární struktury touto metodou, je tedy nejen sekvence RNA, ale také vstupní sekundární struktura v závorkové-tečkové notaci. Predikce metodou Crumple nepočítá s žádnými maticemi, ale pouze nachází komplementární páry. V okamžiku, kdy je nalezen komplementární pár na pozicích *i*, *j* dochází k uložení této struktury a k rekurzivnímu volání algoritmu, jehož vstupem je stejná sekvence, ale odlišná vstupní sekundární struktura v závorkové-tečkové notaci. Ukázka výstupu algoritmu, který predikuje sekundární struktury RNA pomocí metody Crumple je na Obr. **2-7**. [58]



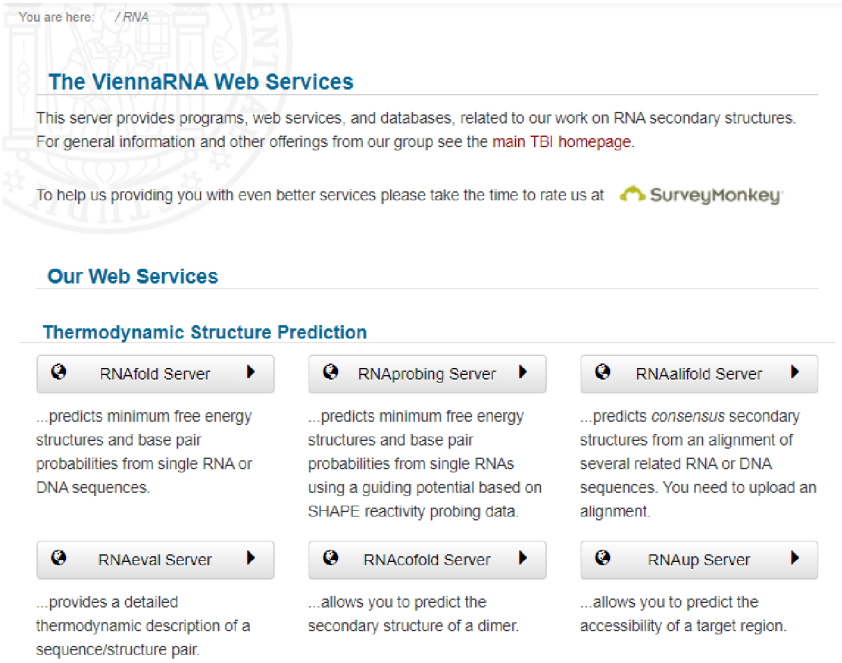
Obr. 2-7: Ukázka výstupu pro algoritmus využívající metodu Crumple k predikci sekundárních struktur RNA [58]

Výstupem algoritmu predikující pomocí metody Crumple je tedy výčet všech možných sekundárních struktur, které je schopna zkoumaná sekvence utvořit. Primárně je ovšem požadováno predikovat jen jednu finální sekundární strukturu. Metoda Crumple je tedy doplňována o přístupy, nebo filtry, pomocí kterých lze získat z výsledného seznamu možných sekundárních struktur pouze jednu nejoptimálnější strukturu. [58]

Finální sekundární strukturu je možné vybrat na základě maximalizace bázových párů a výstupní strukturou bude tedy ta struktura, která obsahuje největší počet párů bází. Další možností, jak vybrat finální sekundární strukturu, je vyčíslení volných energií všech predikovaných struktur. V tomto případě bude finální strukturou ta struktura, která má nejmenší hodnotu volné energie. Více sofistikované algoritmy využívající metodu Crumple k predikci sekundárních struktur bývají doplněné o nejrůznější filtry, které jsou založeny na experimentálních datech. Tyto filtry zahrnují omezení pro: chemicky testovaná dostupná rozpouštědla, enzymatické štěpení párových a nepárových bází, fylogenetické kovariace a minimální délku RNA šroubovice. Minimální délka RNA šroubovice je stanovena pomocí kryoelektronové mikroskopie. Správně určená minimální délka šroubovice má velký vliv na zmenšení prostoru, ve kterém se smotaná RNA nachází. [58]

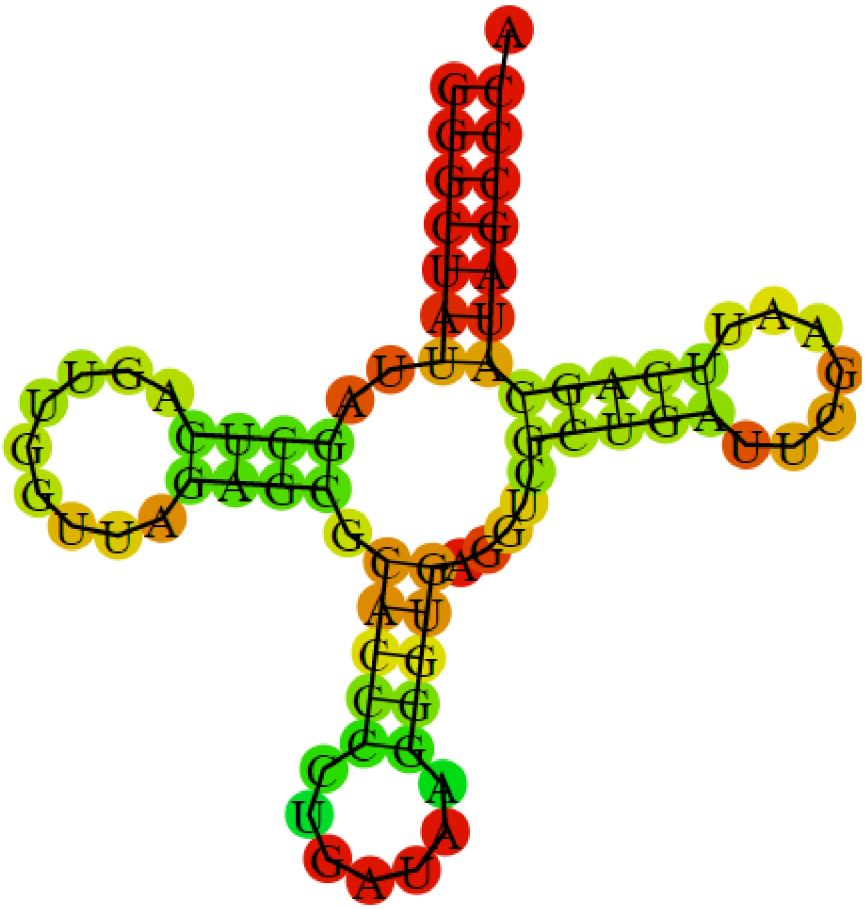
2.2.4 Webové aplikace a bioinformatické funkce predikující sekundární struktury RNA

Protože se predikci sekundárních struktur RNA dostává v posledních letech mnoho pozornosti, existuje několik volně dostupných webových aplikací, které jsou schopné sekundární strukturu predikovat na základě vložené sekvence nukleotidů. [59] Nejznámějším webovou aplikací pro predikci sekundárních struktur RNA je ViennaRNA Web Services. Nejnovější verzí této aplikace je verze 2.0, která využívá parametrů energetického modelu Turner 99 a podporuje čtení formátu FASTA. [60] Náhled hlavní strany této aplikace je na Obr. **2-8**.



Obr. 2-8: Náhled hlavní stránky webu ViennaRNA Web Services

Jak je z náhledu patrné, ViennaRNA Web Services nabízí několik různých funkcí pro predikci sekundárních struktur RNA. První funkcí, která je z webu dostupná, je funkce RNAfold. RNAfold predikuje sekundární struktury na základě přístupu minimalizace volné energie. Další funkcí nacházející se na ViennaRNA Web Services, je funkce RNAalifold, která predikuje shodu sekundárních struktur na základě zarovnání několika podobných sekvencí. ViennaRNA Web Services nabízí výstupní sekundární strukturu ve dvou variantách. Mimo klasické závorkové-tečkové notace, je predikovaná sekundární struktura vyobrazena i pomocí prostého diagramu, příklad tohoto výstupu je na Obr. **2-9**. [60]



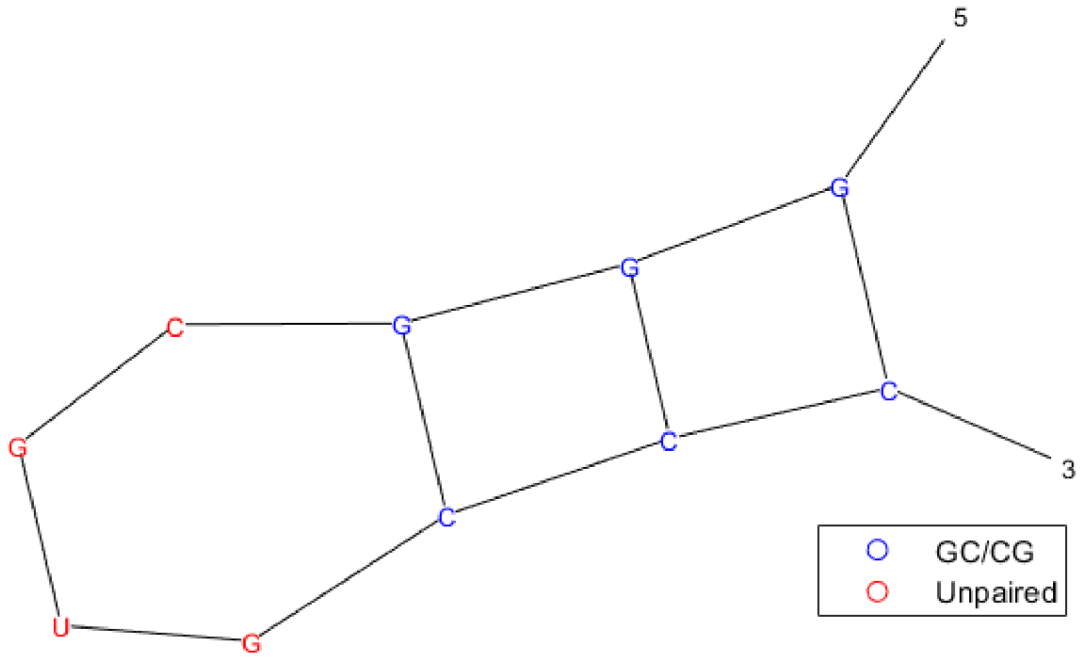
Obr. 2-9: Prostý diagram sekundární struktury RNA pro funkci RNAfold

Mimo volně dostupné webové aplikace existují i nejrůznější bioinformatické funkce predikující sekundární strukturu RNA, které jsou implementovány přímo v konkrétních programovacích prostředích.

Programovací prostředí MATLAB nabízí funkci `rnafold`, která predikuje sekundární strukturu pomocí přístupu minimalizace volné energie. Vstupem do této funkce je RNA sekvence a výstupem je nejčastěji predikovaná sekundární struktura v závorkové-tečkové notaci. Funkce `rnafold` v programovacím prostředí MATLAB umožňuje volbu více než jednoho výstupního parametru. Mimo predikovanou sekundární strukturu v závorkové-tečkové notaci, je možné na výstupu získat také konečnou hodnotu volné energie pro tuto strukturu a skórovací matici, podle které byla finální sktruktura vytvořena. [61]

Další funkcí, kterou programovací prostředí MATLAB nabízí je funkce `rnaplot`. Tato funkce vykresluje sekundární struktury RNA ve formě prostých diagramů, nebo pomocí reprezentace lineární, kruhové, horou nebo stromem. Vstupem pro tuto funkci je tedy sekundární sktruktura v závorkové-tečkové notaci a sekvenci, pro niž byla tato struktura predikována. Dále je na vstupu možné volit a měnit nejrůznější parametry. Jedním z těchto parametrů je parametr `Format`. Tímto parametrem je nastaveno, pomocí

které reprezentace bude sekundární struktura na výstupu vykreslena. Příklad výstupu pro funkci `rnaplot` v programovacím prostředí MATLAB je na Obr. **2-10**. [62]

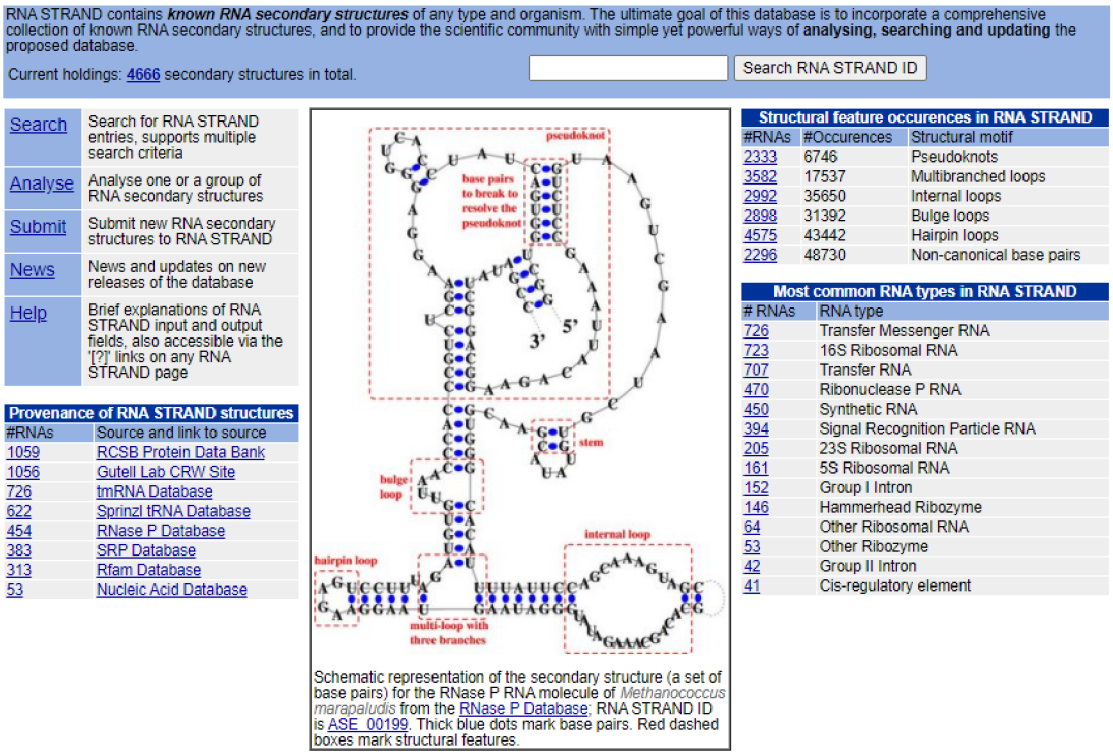


Obr. 2-10: Výstup funkce `rnaplot` v programovacím prostředí MATLAB

3. DATASET RNA SEKVENCÍ SE ZNÁMOU SEKUNDÁRNÍ STRUKTUROU

3.1 Volně dostupná databáze RNA STRAND v2.0

RNA STRAND v2.0 je volně dostupná databáze RNA sekvencí se známou sekundární strukturou. Zahrnuje komplexnější škálu známých sekundárních struktur pro RNA sekvence než předchozí databáze. Náhled úvodní strany pro RNA STRAND v2.0 je na Obr. **3-1**. [63]



Obr. 3-1: Úvodní strana RNA STRAND 2.0 databáze

Řada předchozích databází obsahovala sekundární strukturu pouze pro ty molekuly RNA, pro která byla na základě sekundární struktury predikována struktura terciální. Avšak známé terciální struktury existují jen pro cca 18 % nashromážděných molekul RNA. [63] V současné době obsahuje přesné sekundární struktury pro více než 4666 molekul RNA. Tyto struktury byly získány pomocí experimentálních metod. Nejčastěji používané experimentální metody jsou NMR a RTG krystalografie. [63]

Databáze RNA STRAND v2.0 si klade za cíl poskytnout komplexní informace o všech dostupných strukturálních prvcích pro sekundární strukturu RNA. Tyto informace by mohly být klíčové k pochopení nejrůznějších strukturálních motivů ve specifických typech RNA. Dále by mohly sloužit k odhadu přesnosti výpočetní predikce sekundární

struktury RNA, nebo k vylepšení současných termodynamických modelů pro predikci sekundární struktury. [63]

3.2 Tvorba datasetu RNA sekvencí pro ověření funkčnosti implementovaných algoritmů

Pro potřeby bakalářské práce byl v tabulkovém procesoru Excel vytvořen dataset RNA sekvencí se známou sekundární strukturou. Formát vytvořeného datasetu je **.xlsx*. Na vytvořeném datasetu budou testovány implementované algoritmy pro predikci sekundární struktury RNA. Tyto algoritmy jsou podrobněji popsány v následující kapitole **4**. Všechny sekvence byly získány z volně dostupné databáze RNA STRAND v2.0. Tato databáze je podrobněji popsána v předchozí kapitole 3.1. Vytvořený dataset se známou sekundární strukturou je uveden v Tabulce **3-1**.

Tabulka 3-1: Dataset RNA sekvencí se známou sekundární strukturou

ID RNA STRAND v2.0: *PDB_*	**Sekvence**	**Typ**	**Známá sekundární struktura**
00009	GGUAAUAAGCUCGAGUACC	Intron skupiny I	(((…..((((())))))))
00028	GGGUCUUCGGGUCC	Intron skupiny I	((((((..))))))
00037	GCCACCCUGCAGGGUCGGC	jiná rRNA	(((((((((()))))))).)))
00077	CACGUGCACGUG	mRNA	((((((())))))
00097	GGGCGUGCCC	IRES	(((….)))
00118	GGUUCAGUUGAACC	mRNA	(((((….)))))
00151	GGACUUCGGUCC	mRNA	(((((..)))))
00168	GAGGGUGGAACCGCGCGCGUC CCUC	mRNA	((((……..((….)).))))
00191	GGCUCUCAGUGAGCC	mRNA	(((((…..)))))
00212	CCCCGGGGCCCCGGGG	mRNA	(((((((()))))))))
00226	GACUGGGGCGGUC	tRNA	(((((…)))))
00799	CAUGGGCCGGCCC	mRNA	…((.(()).))
01064	GGUGCGCCCGGUGCGCC	mRNA	(((((((.)))))))
00002	GGCGUAAGGAUUACCUAUGCC	mRNA	(((((.(((….)))))))
Training	GGGAAAUCC		.((..()))

Vytvořený dataset obsahuje celkem 14 RNA sekvencí, které byly získány z databáze RNA STRAND v2.0 a jednu tréninkovou sekvenci, na které byla ověřována funkčnost implementovaných algoritmů před konečnou analýzou. Maximální délka sekvencí, které se v datasetu nachází, je 25 nukleotidů. Všechny sekvence byly vybírány tak, aby obsahovali vlásenkové smyčky, vnitřní smyčky a výdutě nejrůznějších rozměrů.

Sekundární struktury všech RNA sekvencí, které jsou v datasetu použity, byly získány pomocí NMR, experimentální metody pro predikci sekundárních struktur RNA, která je podrobněji popsána v kapitole **2.1**. Vytvořený dataset obsahuje sekvence pro devět molekul syntetických mRNA, jednu molekulu tRNA, jednu molekulu rRNA, dvě molekuly pro introny skupiny I a jednu molekulu IRES.

RNA sekvence a jejich známá sekundární struktura jsou uloženy celkem ve třech sloupcích. První sloupech obsahuje ID z databáze RNA STRAND v2.0, podle kterého je konkrétní RNA sekvence v této databázi dohledatelná. Druhý sloupec obsahuje RNA sekvencí a třetí známou sekundární strukturu.

4. IMPLEMENTOVANÉ ALGORYTMY PRO PREDIKCI SEKUNDÁRNÍ STRUKTURY RNA

Pro predikci sekundární struktury RNA byly implementovány celkem tři algoritmy v programovacím prostředí MATLAB. Implementovány byly tyto algoritmy: algoritmus podle Nussinové, algoritmus podle Zukera a algoritmus pro metodu Crumple. V žádném z implementovaných algoritmů není uvažován vznik rozdvojení struktury (*buriffication*) a multismyček.

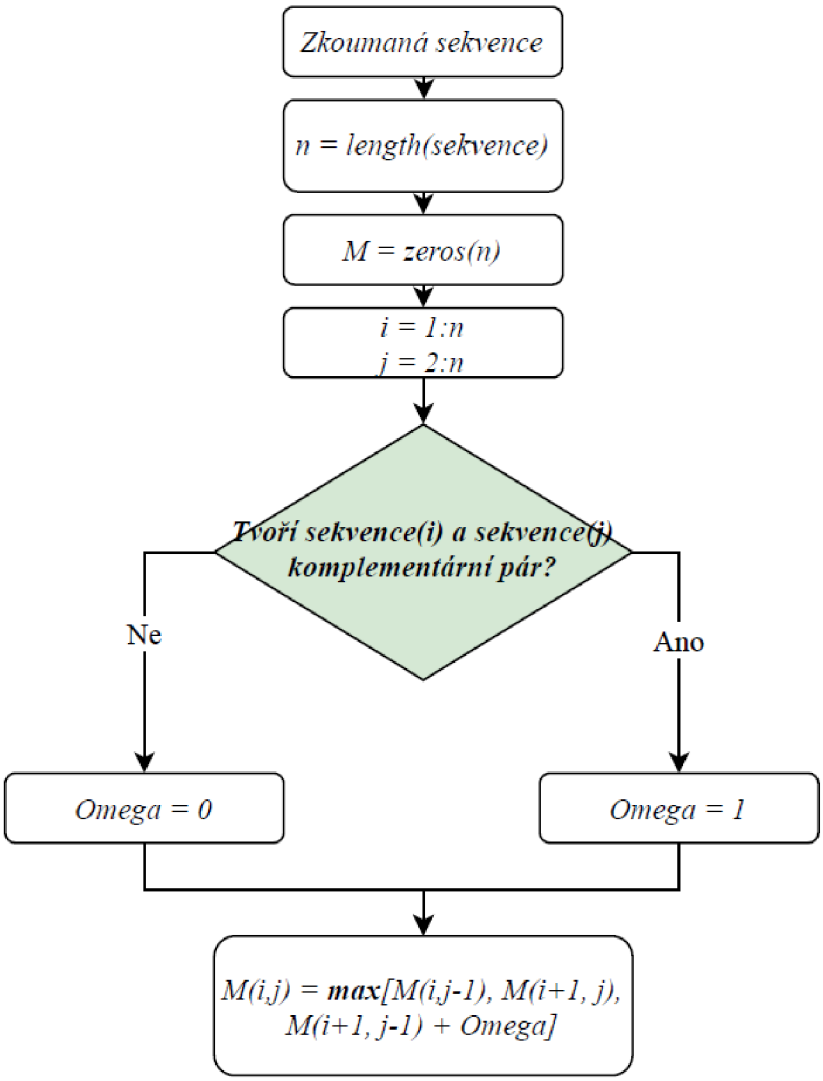
4.1 Algoritmus podle Nussinové

Do implementovaného algoritmu podle Nussinové vstupuje nukleotidová sekvence RNA. Výstupem je predikovaná sekundární struktura. Algoritmus podle Nussinové predikuje sekundární strukturu RNA na základě maximalizace bázových párů. Predikce sekundární struktury RNA algoritmem Nussinové probíhá ve třech krocích: tvorba skórovací matice, tvorba trasovací matice a ohodnocení zpětné cesty elementy závorkové-tečkové notace.

Skórovací matice M má rozměry *n x n*, kde *n* je délka zkoumané sekvence. Hlavní diagonála a diagonála pod hlavní diagonálou jsou inicializovány hodnotami 0. Matice je vyplňována postupně po diagonálách až do pravého horního rohu. Na místo matice M(*i,j*) vybíráme maximum podle rovnice 4.1.

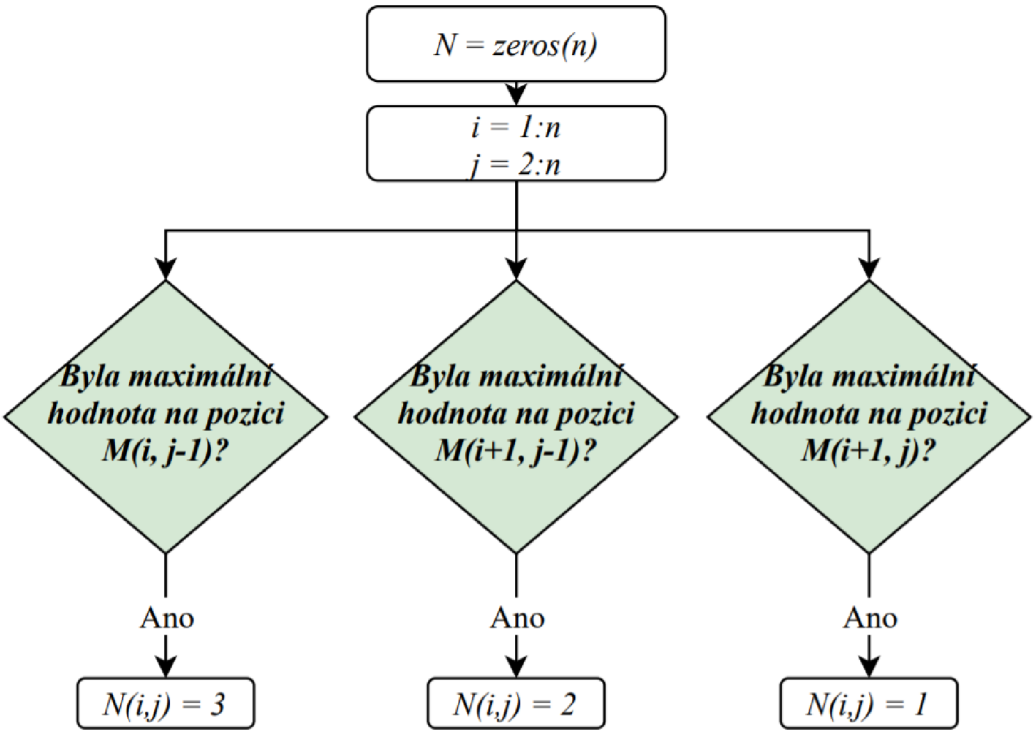
$$M(i,j) = max\begin{Bmatrix} M(i+1,j-1) + \omega(i,j) \\ M(i,j-1) \\ M(i+1,j) \end{Bmatrix} \tag{4.1}$$

Element ω(*i,j*) nabývá hodnoty 1 při komplementaritě bází a hodnoty 0, když báze nejsou komplementární. Celý první krok algoritmu Nussinové je pro lepší vizualizaci graficky znázorněn na vývojovém diagramu na Obr. **4-1**.



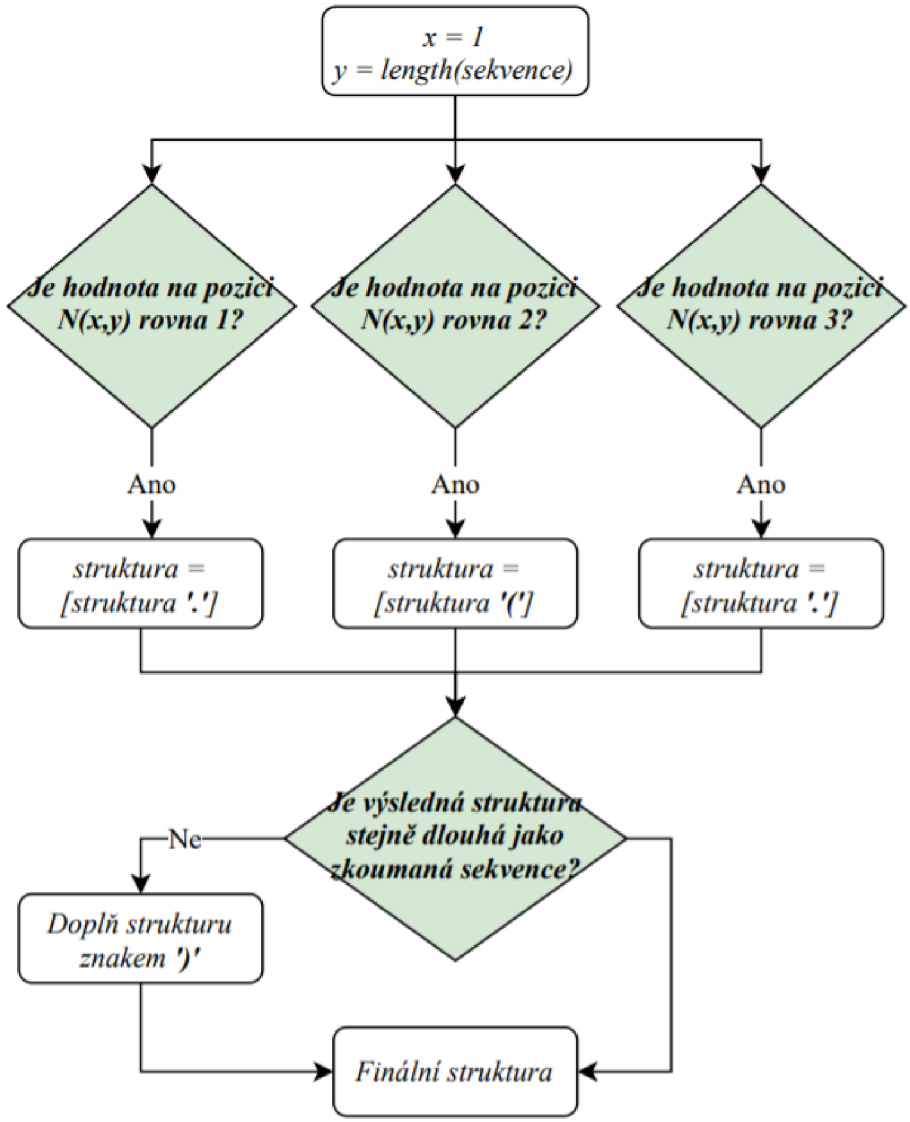
Obr. 4-1: Vývojový diagram pro první krok algoritmu Nussinové

Dalším krokem je tvorba trasovací matice N, která je tvořena současně se skórovací maticí M. Trasovací matice je vyplňována hodnotami 1, 2 a 3. Tyto hodnoty jsou vybírány podle toho, které hodnota z rovnice 4.1 byla vybrána jako maximální. Tento krok je graficky znázorněn společně se skórovací maticí na vývojovém diagramu na Obr. **4-2**.



Obr. 4-2: Vývojový diagram pro druhý krok algoritmu Nussinové

Posledním krokem je nalezení a ohodnocení zpětné cesty elementy závorkové-tečkové notace. Hledání zpětné cesty začíná v pravém horním rohu. Existují tři možné směry zpětné cesty – diagonála, dolů a doleva. Diagonální směr ohodnocujeme znakem závorkové-tečkové notace pro uzavřený pár bází. Směr dolů a doleva je ohodnocen znakem pro smyčku, nebo výduť. Po zpětné cestě putujeme až do okamžiku, kdy narazíme na hodnotu 0, která se nachází na hlavní diagonále, nebo diagonále pod ní. Zbytek sekundární struktury je doplněn znakem závorkové-tečkové notace pro otevřený pár bází. Tento krok je graficky znázorněn na vývojovém diagramu na Obr. **4-3**.



Obr. 4-3: Vývojový diagram pro třetí krok algoritmu Nussinové

Celý algoritmus Nussinové byl implementován v rámci jediné funkce `NussinovAlgorithm`. Jeho výstup bude ohodnocen společně s dalšími implementovanými algoritmy v kapitole **5**. Predikované struktury pouze algoritmem Nussinové jsou znázorněny v Tabulce **4-1**.

Tabulka 4-1: Sekundární struktury predikované algoritmem Nussinové

ID RNA STRAND v2.0: *PDB_*	Známá sekundární struktura	Predikovaná sekundární struktura
00009	(((…..((((.))))))))	((…(..((((.))))))))
00028	((((((..))))))	(((..(((.)))))))
00037	(((((((((.)))))).)))	(((.(((((.)))))))))
00077	((((((.))))))	(((((.)))))))
00097	(((….)))	((((.))))))
00118	((((….))))))	(((((((.)))))))
00151	(((((..)))))	(((..(.)))))
00168	((((……..((….)).))))	..(((.((.(.((((.))))))))))
00191	(((((…..)))))	((..(((.(.)))))))
00212	(((((((.)))))))	(((((((.)))))))
00226	(((((…)))))	((((…(.)))))
00799	…((.((.)).))	…(((((.)))))
01064	(((((((.)))))))	((((((.(.))))))))
00002	(((((.(((….)))))))))	(((((.(((..(.))))))))))
Training	.((..()))	.((..()))

4.2 Algoritmus podle Zukera

Do implementovaného algoritmu podle Zukera vstupuje zkoumaná nukleotidová RNA sekvence. Výstupem je predikovaná sekundární struktura. Algoritmus podle Zukera predikuje sekundární strukturu RNA na základě minimalizace volné energie. V rámci Zukerova algoritmu jsou tvořeny dvě skórovací matice V a W. Obě tyto matice mají rozměry *n x n*, kde *n* je délka zkoumané sekvence. Pomocí matice W jsou nakonec hledány báze, které spolu tvoří pár. Tyto páry jsou pak pomocí závorkové-tečkové notace vyneseny jako finální sekundární struktura.

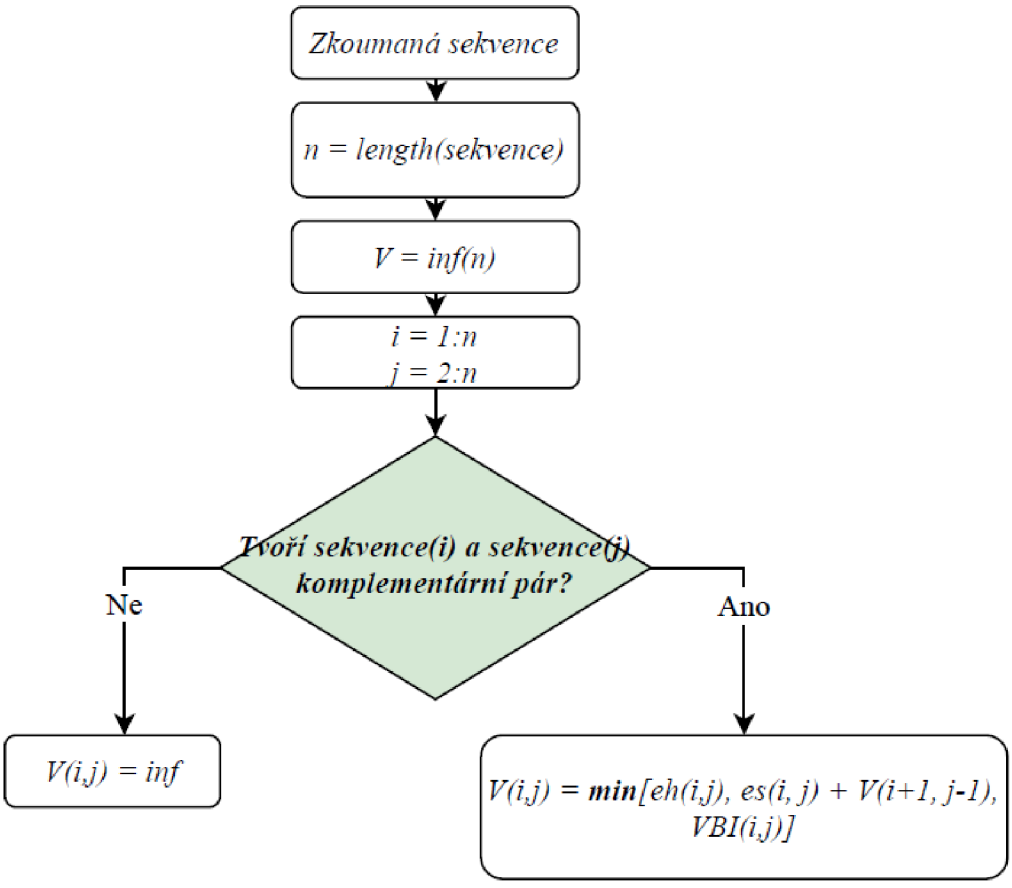
Pro predikci Zukerovým algoritmem byly implementovány celkem tři funkce. Funkce `internalBulgeValue` a funkce `stackedPairsValue`. Tyto funkce počítají příspěvky volné energie pro vnitřní smyčky, výdutě a spojené páry při plnění matice V. Poslední vytvořenou funkcí je funkce `ZukerAlgorithm`, která vybírá minimální hodnoty volné energie pro vlásenkové smyčky, vyplňuje matici W a tvoří konečnou sekundární strukturu. V rámci této funkce jsou volány obě výše zmíněné funkce pro výpočet volné energie.

V prvním kroku Zukerova algoritmu je vyplněna skórovací matice V. Její plnění začíná její inicializací. Hlavní diagonálu a diagonála pod hlavní diagonálou jsou vyplněny hodnotou nekonečno. Na pozici V(*i*,*j*) je pak vybírána minimální hodnota podle rovnice

4.2. Matice V je plněna hodnotami směrem od hlavní diagonály směrem k pravému hornímu rohu.

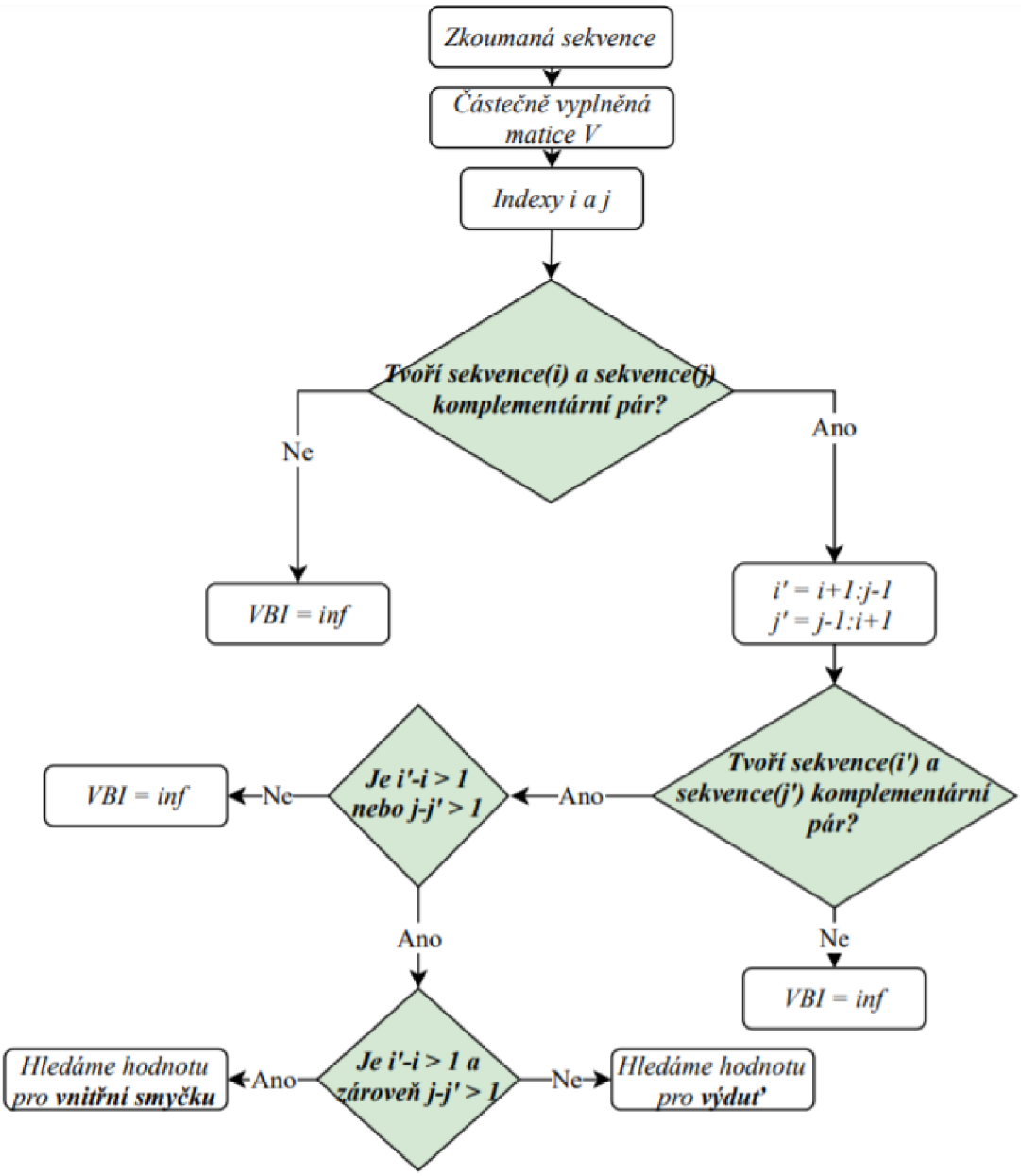
$$V(i,j) = min\begin{Bmatrix} eh(i,j) \\ es(i,j) + V(i+1,j-1) \\ VBI(i,j) \end{Bmatrix} \quad (4.2)$$

Hodnota *eh* značí volnou energii vlásenkové smyčky, která je uzavřena párem bází na pozicích *i* a *j*. Hodnota *VBI* vybírá minimum volné energie pro vnitřní smyčku nebo výduť. Vnitřní smyčka nebo výduť je tvořena uvnitř struktury uzavřené párem na pozicích *i* a *j*. S hodnotou pro vnitřní smyčku je počítáno v případě, že *i' – i >1* a zároveň *j – j' > 1*. S hodnotou pro výduť je počítáno, když *i' – i >1*, nebo *j – j' > 1*, ale ne zároveň. Tabulka hodnot pro vlásenkové smyčky, vnitřní smyčky a výdutě je zpracována v tabulkovém procesoru Excel a uložena jako *Turner04*. Celý první krok Zukerova algoritmu je graficky znázorněn vývojovým diagramem na Obr. **4-4**.



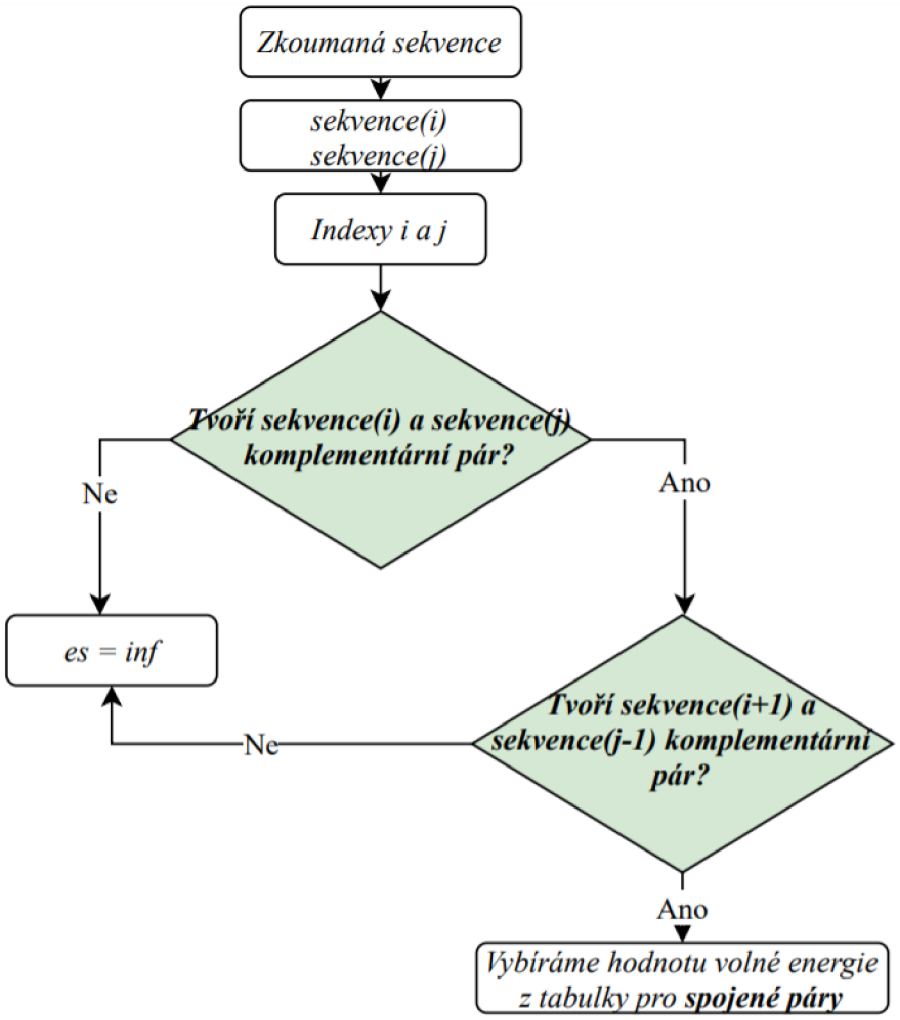
Obr. 4-4: Vývojový diagram pro první krok Zukerova algoritmu

Volné energie pro vlásenkové smyčky jsou vybírány v rámci funkce `ZukerAlgorithm`. Hodnoty volných energií pro vnitřní smyčky a výdutě jsou vybírány v rámci funkce `internalBulgeValue`. Do této funkce vstupuje zkoumaná sekvence, částečně vyplněná matice V a indexy *i* a *j*. Tyto indexy značí pozice nukleotidů, které tvoří pár a uzavírají danou strukturu. V rámci této funkce je kontrolována komplementarita bází na pozicích *i* a *j*, jsou vybírány hodnoty volných energií z tabulky *Turner04* a je vybírána nejmenší hodnota volné energie, která je zároveň výstupem této funkce. Výběr hodnot volných energií pro vnitřní smyčky a výdutě je graficky znázorněn na vývojovém diagramu na Obr. **4-5**.



Obr. 4-5: Vývojový diagram pro výpočet hodnoty *VBI*

Hodnota *es* v rovnici 4.2 zastupuje spojené páry bází, pro které je hodnota volné energie vybírána v rámci funkce `stackedPairsValue`. Vstupem do této funkce jsou zkoumané sekvence, nukleotidy tvořící pár a indexy *i* a *j*. Tyto indexy značí pozice vstupních nukleotidů. Funkce kontroluje, zda jsou komplementární i nukleotidy na pozicích *i+1* a *j-1*. V případě že ano, vzniká v těchto místech spojený pár bází, jehož hodnota je čtena z tabulky, která je zpracována v tabulkovém procesoru Excel a uložena jako *stackedPairs*. Výběr hodnot volných energií pro spojené páry bází je graficky znázorněn na vývojovém diagramu na Obr. **4-6**.

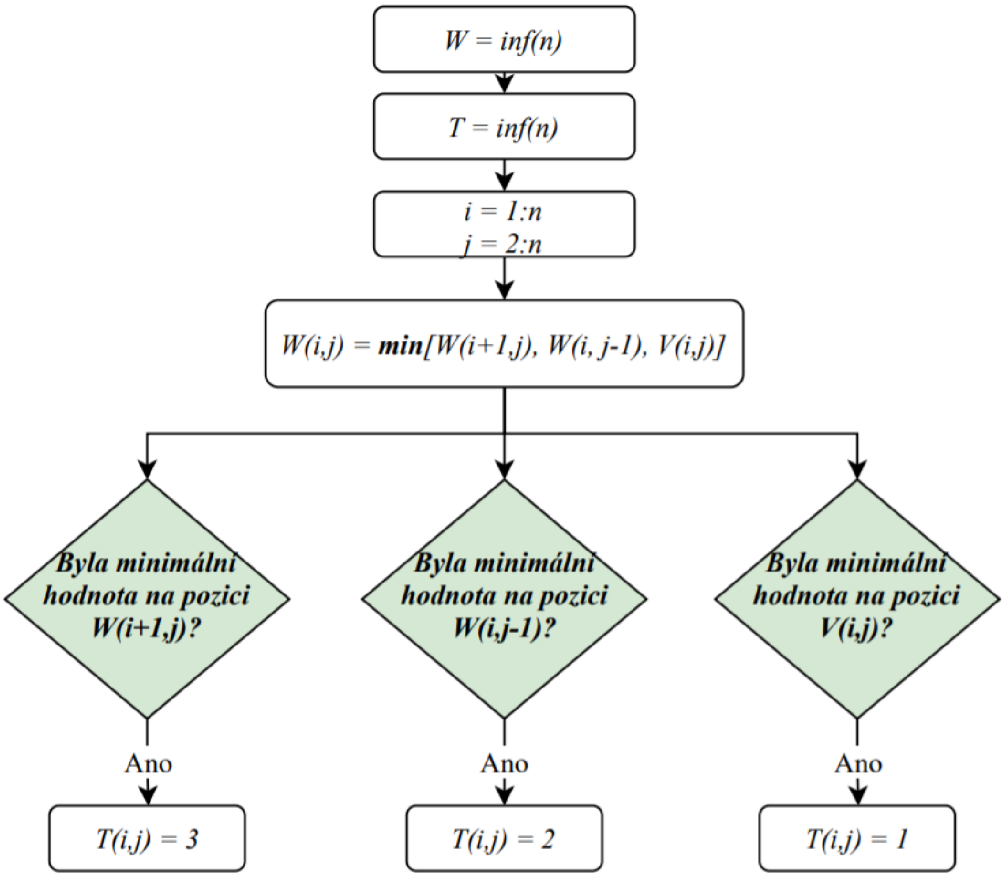


Obr. 4-6: Vývojový diagram pro výpočet hodnoty *es*

V druhém kroku Zukerova algoritmu je vyplněna skórovací matice W a současně i trasovací matice T. Obě matice jsou nejprve vyplněny hodnotami nekonečno na hlavní diagonále a diagonále pod hlavní diagonálou. Na pozici W(*i,j*) je pak vybírána minimální hodnota podle rovnice 4.3. Hodnota celkové volné energie pro strukturu predikovanou Zukerovým algoritmem se nachází v pravém horním rohu matice W.

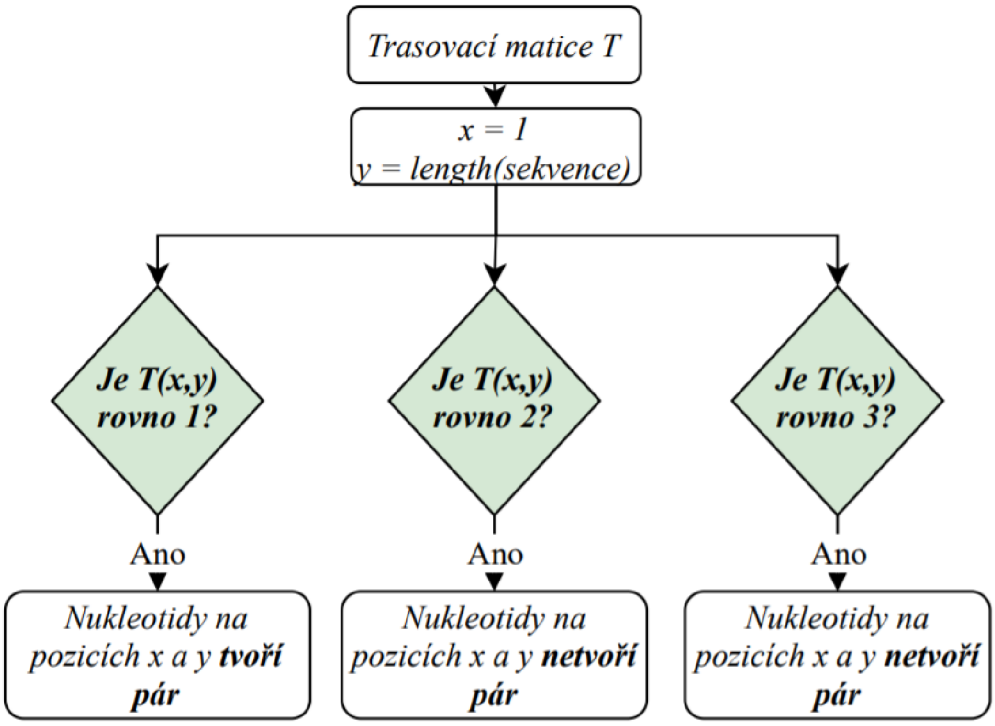
$$W(i,j) = min\begin{Bmatrix} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \end{Bmatrix} \tag{4.3}$$

Matice T je plněna hodnotami zároveň s maticí W a je plněna hodnotami 1, 2 nebo 3 podle toho, která hodnota byla vybrána jako minimální. Tento krok je graficky znázorněn vývojovým diagramem na Obr. **4-7**.



Obr. 4-7: Vývojový diagram pro druhý krok Zukerova algoritmu

Dále je tvořena nulová sekundární struktura, která obsahuje pouze znaky závorkové-tečkové notace pro nepárové nukleotidy. Podle trasovací matice T jsou nalezeny báze, které tvoří páry a tato místa jsou v nulové sekvenci přepsány znaky pro otevřené a uzavřené páry bází. Tento krok je graficky znázorněn vývojovým diagramem na Obr. **4-8**.



Obr. 4-8: Vývojový diagram pro nalezení bází, které tvoří pár

Sekundární struktury predikované Zukerovým algoritmem jsou uvedeny v Tabulce **4-2** společně s hodnotami vypočtené volné energie. Jeho výstup je číselně ohodnocen společně s ostatními implementovanými algoritmy v kapitole **5**, kde je porovnán s výstupy webové aplikace RNAfold a také s výstupy algoritmu pro metodu Crumple s minimalizací volné energie.

Tabulka 4-2: Sekundární struktury a hodnoty volné energie predikované Zukerovým algoritmem

ID RNA STRAND v2.0: *PDB_*	Známá sekundární struktura	Predikované sekundární struktury	Hodnoty volné energie [kcal/mol]
00009	(((…..((((())))))))	((……(.(…))))	1,0
00028	((((((..))))))	(((((…)).)))	-1,3
00037	(((((((((()))))))).)))	((.(((.(…)))).))	-4,1
00077	((((((())))))	((((….))))	-1,3
00097	(((….)))	(((….)))	-0,2
00118	(((((….)))))	(((((….)))))	-1,7
00151	(((((..)))))	(((.(…))))	-1,2
00168	((((……..((….)).))))	(((…(….((….)).))))	-5,1
00191	(((((…..)))))	((..(((…)))))	-3,2
00212	(((((((())))))))	((((.((…))))))	-9,4
00226	(((((…)))))	(((((…)))))	-2,0
00799	…((.(()).))	…(((….)))	-0,2
01064	(((((((.)))))))	(((((((…)))))))	-6,1
00002	(((((.(((….)))))))	(((…(((….))).)))	-4,8
Training	.((..()))	(((…)))	0,6

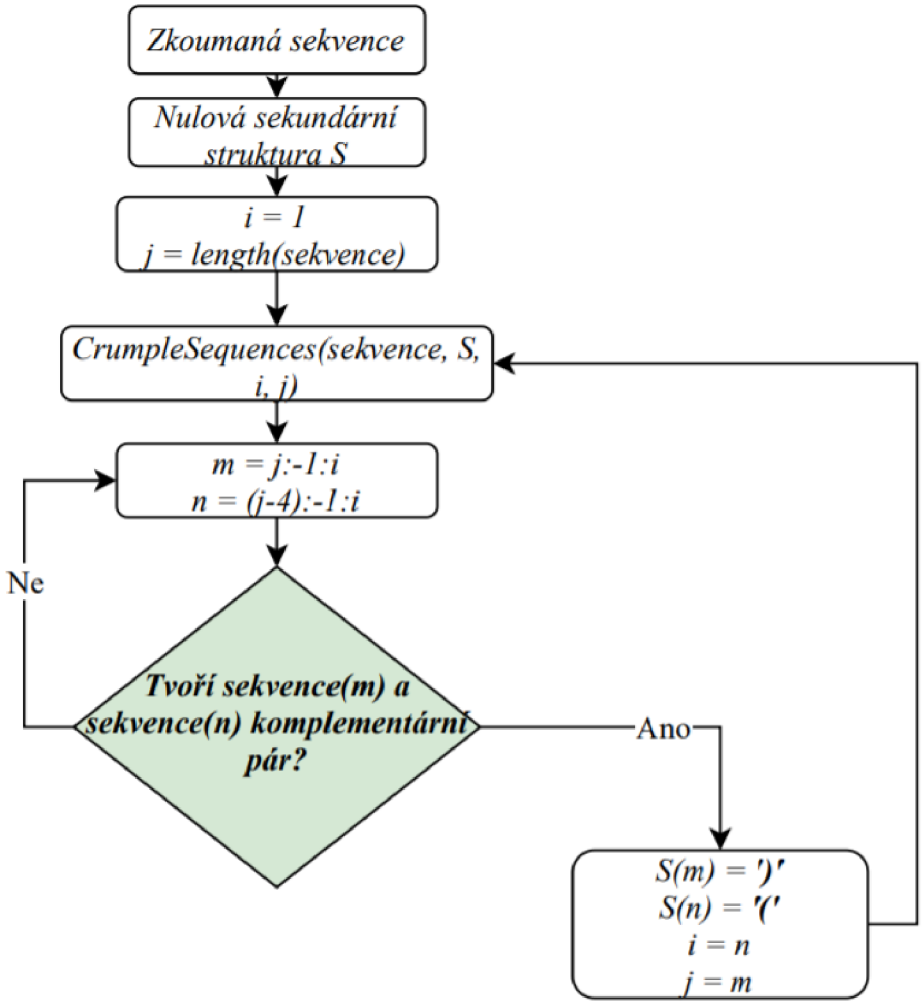
4.3 Algoritmus pro metodu Crumple

Vstupem do implementovaného algoritmu pro metodu Crumple je nukleotidová sekvence. Výstupem z prvotního algoritmu pro metodu Crumple je výčet všech možných sekundárních struktur pro zkoumanou sekvenci. Tento výčet je nakonec eliminován a je z něj vybrána pouze jedna nejoptimálnější sekundární struktura. Tato struktura je vybrána pomocí dvou možných přístupů. Za prvé pomocí přístupu a maximalizaci bázových párů, a za druhé pomocí přístupu o minimalizaci volné energie. Oba přístupy budou číselně ohodnoceny v kapitole **5**, společně s ostatními implementovanými algoritmy.

Prvním krokem implementovaného algoritmu pro metodu Crumple je vytvoření nulové sekundární struktury S. Nulová sekundární struktura obsahuje pouze znak ze závorkové-tečkové notace pro smyčky nebo výdutě. Takto vytvořená nulová sekundární struktura vstupuje do vytvořené funkce `CrumpleSequences` společně se zkoumanou sekvencí a indexy *i* a *j*, značící začátek a konec zkoumané sekvence.

Funkce `CrumpleSequences` hledá všechny možné sekundární struktury pro zkoumanou sekvenci. Nalezené struktury jsou postupně vyplňovány do vstupní nulové struktury. Postupně je kontrolována komplementarita bází na pozicích indexu *j*, který značí konec zkoumané sekvence, a indexu *j-4*. V případě, že dojde k nalezení

komplementárního páru na pozici *j* v nulové sekundární struktuře, je doplněn znak ze závorkové-tečkové notace pro uzavřený pár bází a na pozici *j-4* je doplněn znak ze závorkové-tečkové notace pro otevřený pár bází. Tímto získáme novou nulovou sekundární strukturu, která vstupuje rekurzivně do funkce `CrumpleSequences` společně se zkoumanou sekvencí a novými indexy. V tomto okamžiku hledáme sekundární strukturu v již nalezené sekundární struktuře. Tento krok je graficky znázorněn vývojovým diagramem na Obr. **4-9**.



Obr. 4-9:Vývojový diagram pro funkci `CrumpleSequences`

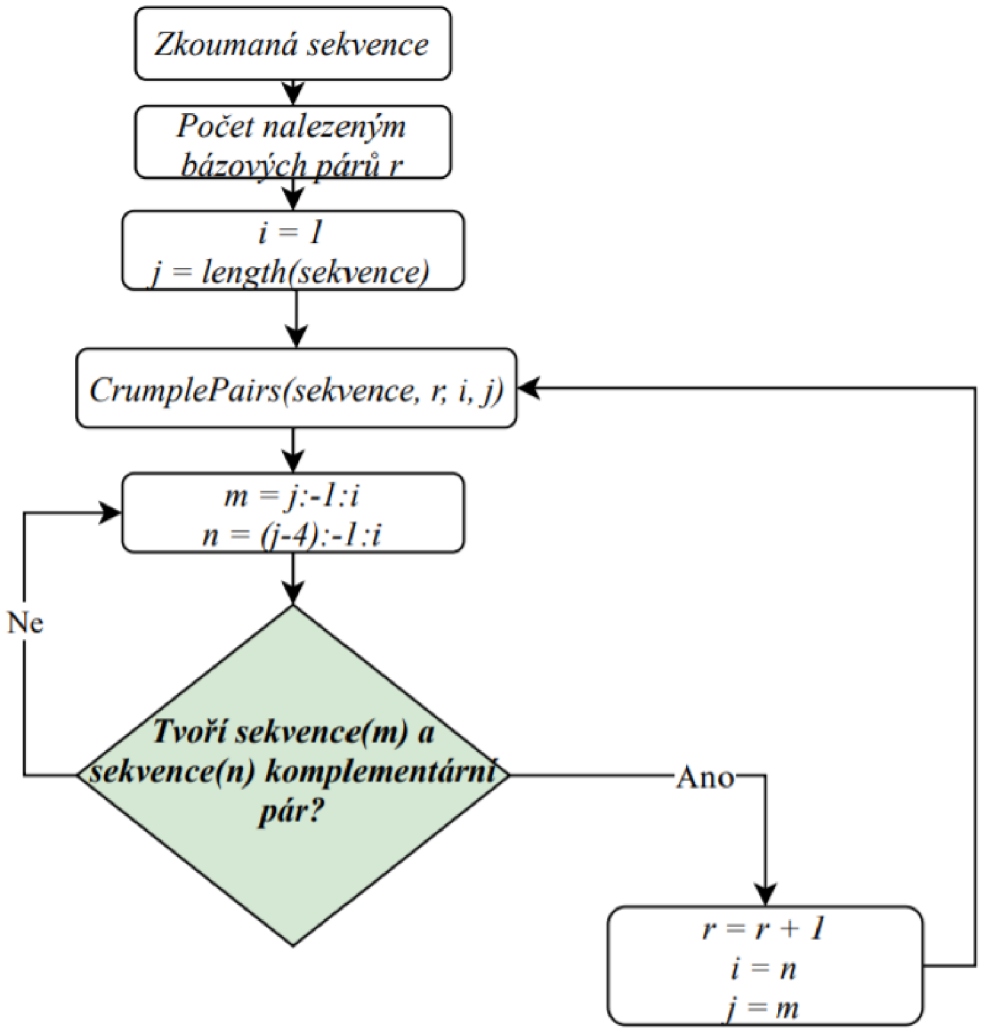
Získaný výčet sekundárních struktur vstupuje do eliminačního procesu, kterým je vybrána pouze jedna nejoptimálnější sekundární struktura.

4.3.1 Maximalizace bázových párů

První možností, jak vybrat nejoptimálnější strukturu z výstupu metody Crumple, je vybrání struktury, která obsahuje největší počet bázových párů. Bázové páry ve všech

predikovaných sekvencích jsou počítány pomocí vytvořené funkce `CrumplePairs`. Do této funkce vstupuje zkoumaná RNA sekvence, proměnná *r* a indexy *i* a *j*.

Proměnná *r* zastupuje počet bázových párů v již nalezené sekvenci a při prvotním vstupu je tedy rovna hodnotě 0. Indexy *i* a *j* značí začátek a konec zkoumané struktury. Při nalezení prvního bázového páru je funkce volána rekurzivně s novou proměnnou *r* a novými indexy *i* a *j*. Tento postup je graficky znázorněn vývojovým diagramem na Obr. **4-10**.



Obr. 4-10: Vývojový diagram pro funkci `CrumplePairs`

Výstupem této funkce je výčet nalezených párů v každé nalezené sekundární struktuře. Pozice počtu bázových párů odpovídá pozicím nalezených sekundárních struktur. V případě shody v počtu vytvořených bázových párů je vybrána ta struktura, která se nachází v seznamu jako první.

Výstup algoritmu pro metodu Crumple bude ohodnocen společně s dalšími implementovanými algoritmy v kapitole **5**. Predikované struktury získané algoritmem

pro metodu Crumple, která vybírá optimální strukturu na základě maximalizace bázových párů, jsou uvedeny v Tabulce **4-3**.

Tabulka 4-3: Sekundární struktury predikované algoritmem pro metodu Crumple s maximalizací bázových párů

ID RNA STRAND v2.0: *PDB_*	**Známá sekundární struktura**	**Predikovaná sekundární struktura**
00009	(((…..((((.))))))))	(((.(((…).…)))))
00028	((((((..))))))	(((.((…)))))
00037	(((((((((.)))))))).)))	((((.(((…))))))).)))
00077	((((((.))))))	((((….))))
00097	(((….)))	(((….)))
00118	(((((….)))))	(((((….)))))
00151	(((((..)))))	(((.(…))))
00168	((((……..((….)).))))	(((((((..(…)).)).)))))
00191	(((((…..)))))	((..(((…)))))
00212	(((((((.)))))))	((((.((…))))))
00226	(((((…)))))	(((((…)))))
00799	…((.((.)).))	…(((….)))
01064	(((((((.)))))))	(((((((…)))))))
00002	(((((.(((….)))))))))	(((((.(((….)))))))))
Training	.((..()))	(((…)))

Algoritmus pro metodu Crumple, ve kterém je vybírána nejoptimálnější struktura pomocí maximalizace bázových párů, je implementován v rámci funkce `CrumpleMaxPairsAlgorithm`. Vstupem do této funkce je zkoumaná nukleotidová RNA sekvence a výstupem nejoptimálnější sekundární struktura. V rámci této funkce je volána i funkce `CrumplePairs`.

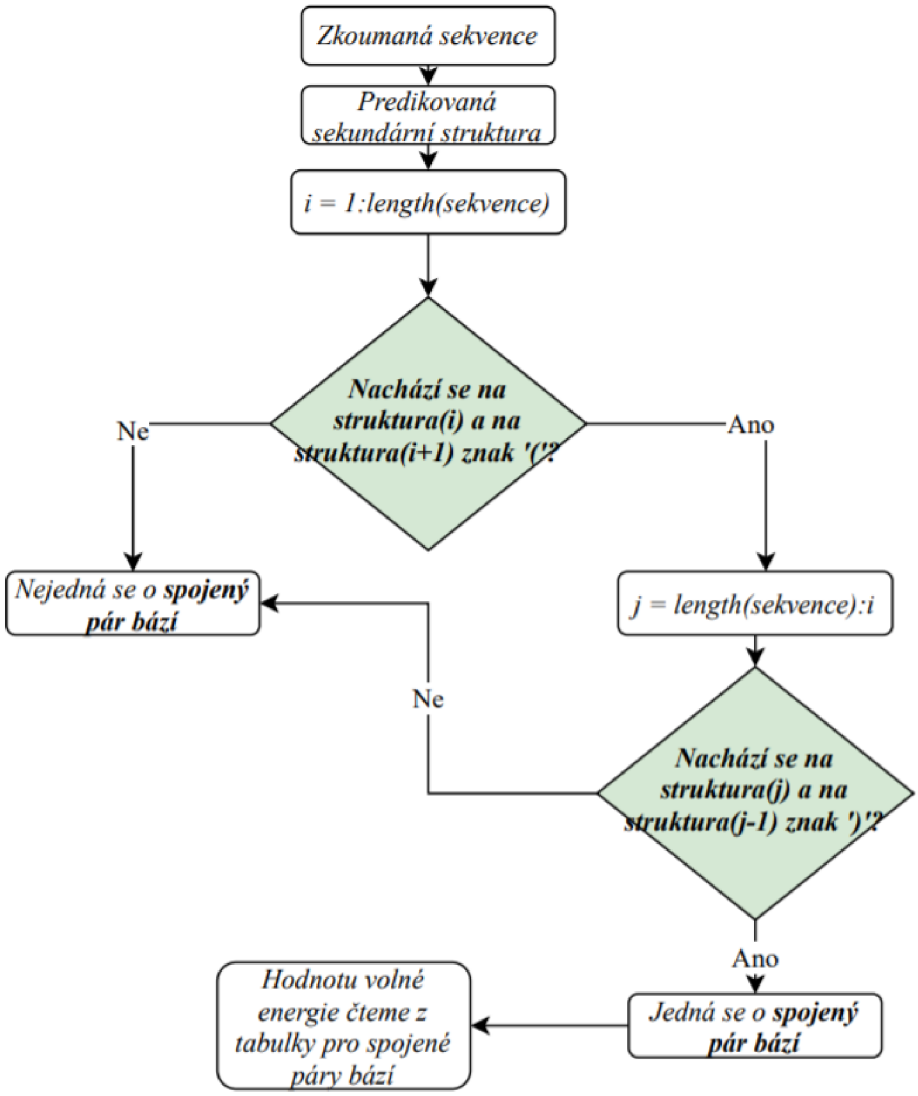
4.3.2 Minimalizace volné energie

Druhou možností, jak vybrat nejoptimálnější strukturu z výstupu metody Crumple je vybrání struktury s nejmenší hodnotou volné energie. Volná energie pro všechny nalezené sekundární struktury je počítána pomocí funkce `structureFreeEnergy`.

Do této funkce vstupuje zkoumaná RNA sekvence a predikovaná sekundární struktura v závorkové-tečkové notaci. Počítá se vznikem celkem čtyř struktur – vlásenková smyčka, vnitřní smyčka, výduť a spojený pár bází. I tato funkce využívá tabulek s volnými energiemi *Turner04* a *stackedPairs*.

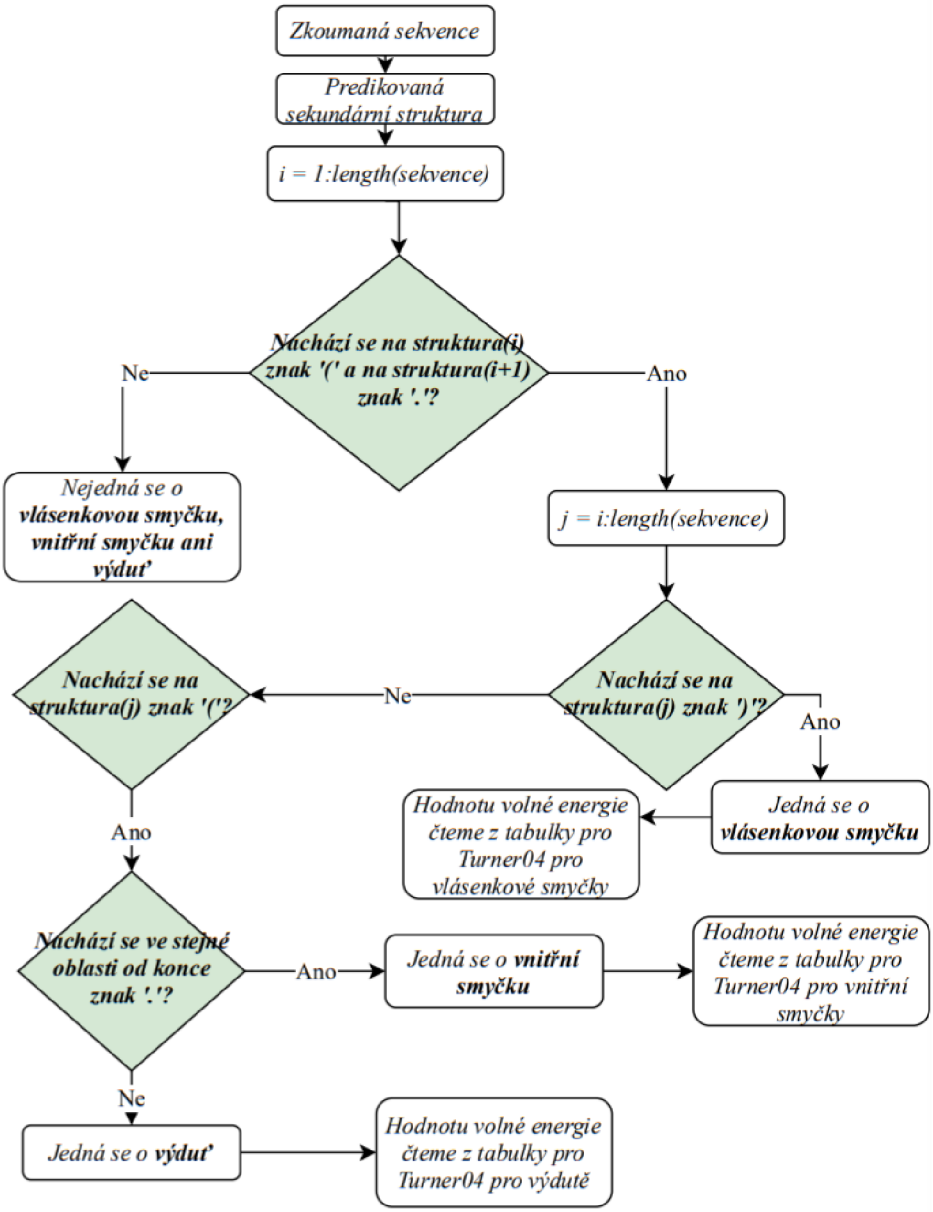
Funkce kontroluje jednotlivé znaky závorkové-tečkové notace a kontroluje jejich pořadí. Znak závorkové-tečkové notace značící uzavřený pár bází následovaný stejným

znakem uvozuje spojený pár bází v případě, že při zpětném procházení nalezené struktury, je jako první naraženo na znak značící otevřený pár bází následovaný stejným znakem. Tento krok je graficky znázorněn vývojovým diagramem na Obr. **4-11**.



Obr. 4-11: Vývojový diagram pro nalezení spojeného páru bází při výpočtu celkové volné energie struktury

V případě, že je znak závorkové-tečkové notace značící uzavřený pár bází následovaný tečkou, hledáme kterým znakem je tečková struktura uzavřena. V případě, že je uzavřena znakem pro otevřený pár bází, je nalezená tečková struktura vlásenková smyčka. V opačném případě je kontrolováno, zda se ve stejné oblasti na druhém konci struktury také nalézá nějaká tečková struktura. Pokud ano, jedná se o vnitřní smyčku, pokud ne, jedná se o výduť. Tento krok je graficky znázorněn vývojovým diagramem na Obr. **4-12**.



Obr. 4-12: Vývojový diagram pro nalezení vnitřních smyček nebo výdutí při výpočtu celkové volné energie struktury

Funkce `structureFreeEnergy` je volána v rámci funkce `CrumpleMinEnergyAlgorithm`. Vstupem do této funkce je zkoumaná nukleotidová sekvence RNA a jejím výstupem je jedna nejoptimálnější sekundární struktura. Sekundární struktury predikované pomocí algoritmu pro metodu Crumple, který vybírá nejoptimálnější sekundární strukturu pomocí minimalizace volné energie, jsou uvedeny spolu s vypočtenými hodnotami volných energií v Tabulce **4-4**.

Tabulka 4-4: Sekundární struktury a hodnoty volné energie predikované algoritmem pro metodu Crumple s minimalizací volné energie

ID RNA STRAND v2.0: *PDB_*	Známá sekundární struktura	Predikované sekundární struktury	Hodnoty volné energie [kcal/mol]
00009	(((…..((((())))))))	((((……..))))	1,0
00028	((((((..))))))	(((((….)))))	-1,3
00037	(((((((((()))))))).)))	(((((((….)))).))	-3,1
00077	((((((())))))	((((….))))	-1,3
00097	(((….)))	(((….)))	-0,2
00118	(((((….)))))	(((((….)))))	-1,7
00151	(((((..)))))	((((….))))	-1,2
00168	((((……..((….)).))))	(((((…...((….)).)))))	-5,9
00191	(((((…..)))))	(((((.….)))))	-3,2
00212	(((((((())))))))	((((((….))))))	-9,4
00226	(((((…)))))	(((((…)))))	-2,0
00799	…((.(()).))	…(((….)))	-0,2
01064	(((((((.)))))))	(((((((…)))))))	-6,1
00002	(((((.(((….)))))))	(((((.(((….))))))))	-3,6
Training	.((..()))	(((…)))	0,6

4.4 Konečný výstup implementovaných algoritmů

Pro bakalářskou práci bylo vytvořeno celkem 11 funkcí, které jsou podrobněji popsány dále v textu. Všechny implementované algoritmy pro predikci jsou volány pomocí jedné vytvořené funkce – `SecondaryStructurePrediction`. Do této funkce vstupuje vytvořený dataset RNA sekvencí se známou sekundární strukturou, jehož celková analýza trvala 5470,71 vteřin. Výstupy této funkce jsou celkem tři.

První v podobě predikovaných sekundárních struktur a hodnot volné energie pro zkoumané sekvence. Tyto hodnoty byly ukládány přímo do struktury vytvořeného datasetu. Výstupní podoba datasetu je datový typ table o rozměrech *15 x 11*, který obsahuje predikované struktury pomocí všech implementovaných algoritmů včetně sekundárních struktur predikovaných pomocí webové aplikace RNAfold, hodnoty volných energií pro algoritmy, které počítali s minimalizací volné energie a původní dataset. Všechny predikované sekundární struktury jsou uvedeny v tabulce v Příloze **1**.

Dále jsou součástí výstupního datasetu hodnoty volné energie pro sekundární struktury predikované implementovaným Zukerovým algoritmem, webovou aplikací RNAfold a algoritmem pro metodu Crumple, který vybírá nejoptimálnější strukturu na základě minimalizace volné energie.

Druhým výstupem funkce `SecondaryStructurePrediction` je ohodnocení kvality predikce všech implementovaných algoritmů, včetně predikce RNAfold. Parametry ohodnocení pro jednotlivé algoritmy jsou senzitivity, PPV a F-skóre. Výpočet těchto parametrů probíhal pro každou predikovanou sekvenci zvlášť a výsledné ohodnocení algoritmů je dáno průměrem těchto jednotlivých ohodnocení.

Třetí výstup jsou vykreslené sekundární struktury s nejnižší senzitivitou pro každou implementovanou metodu. Vykreslení těchto struktur proběhlo pomocí MATLAB funkce `rnaplot`, v rámci funkce `SecondaryStructurePrediction`. Pro vizualizace predikovaných struktur byla zvolena forma prostého diagramu.

Všechny výstupy jsou podrobně okomentovány, graficky doplněny a diskutovány v následující kapitole **5**.

5. VÝSLEDKY IMPLEMENTOVANÝCH ALGORITMŮ

Všechny implementované algoritmy byly otestovány na vytvořeném datasetu RNA sekvencí se známou sekundární strukturou, který je podrobněji popsán v kapitole **3**. Implementované algoritmy byly ohodnoceny pomocí senzitivity, PPV a F-skóre. Pro každý implementovaný algoritmus byla vykreslena predikovaná sekundární struktura, která dosahovala nejmenší hodnoty senzitivity. Vykreslení těchto struktur proběhlo pomocí MATLAB funkce `rnaplot`.

Vizuálním zhodnocením výsledné tabulky predikovaných struktur, která je uvedena v Příloze **1**, lze říct, že predikce pomocí algoritmu Nussinové byla 100% shodná se známou sekundární strukturou ve 3 z 15 případů. Implementovaná Zukerova predikce se shodovala se známou sekundární strukturou ve 3 z 15 případů, predikce RNAfold ve 4 z 15 případů. Predikce metodou Crumple byla shodná se známou sekundární strukturou ve 4 z 15 případů pro eliminaci pomocí maximalizace bázových párů a v 6 z 15 případů pro eliminaci pomocí minimalizace volné energie. Číselně je predikce pomocí jednotlivých algoritmů ohodnocena níže pomocí senzitivity, PPV a F-skóre.

Rozdíly v jednotlivých predikcích jsou patrné. Spolu byly porovnány výstupy těch algoritmů, které predikovali na základě stejné hypotézy. Implementovaný Zukerův algoritmus byl tedy porovnán s výstupy pro RNAfold a výstupy pro metodu Crumple s minimalizací volné energie. Algoritmus Nussinové byl porovnán s implementovanou metodou Crumple s maximalizací bázových párů.

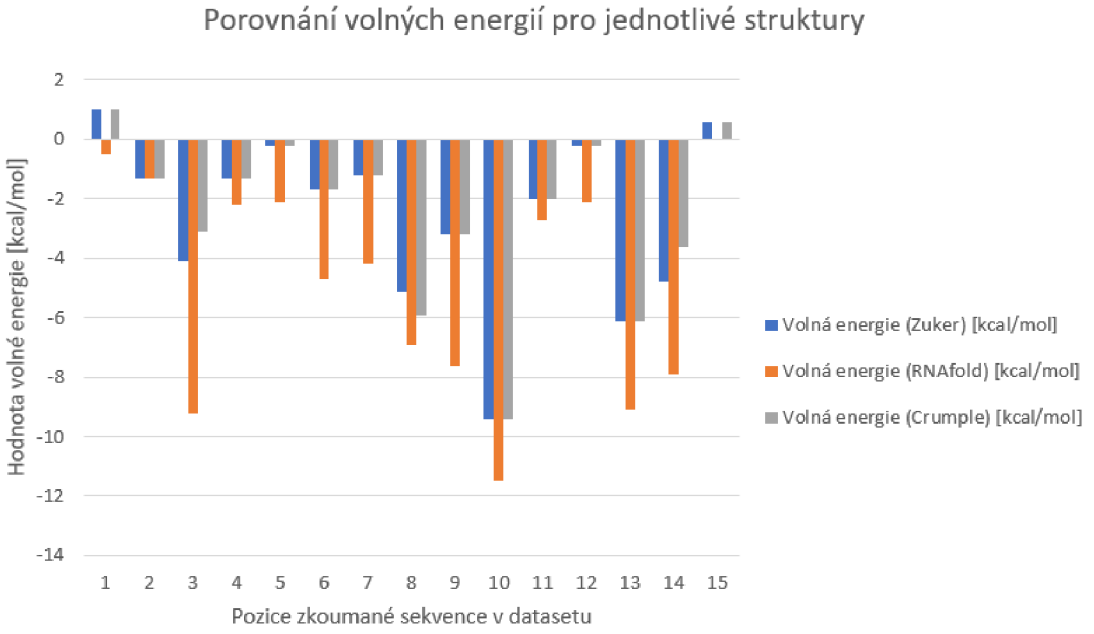
Predikce pomocí implementovaného Zukerova algoritmu byla shodná s predikcí RNAfold v 6 z 15 případů. Implementovaný algoritmus pro metodu Crumple, který vybírá optimální sekundární strukturu pomocí minimalizace volné energie, byl shodný s predikcí pomocí Zukerova algoritmu v 6 z 15 případů a shodný s predikcí pomocí RNAfold v 13 z 15 případů. Algoritmus Nussinové nebyl shodný s predikcí metodou Crumple s maximalizací bázových párů v žádném z případů.

Dále byly hodnoceny získané hodnoty volné energie pro sekundární struktury predikované implementovaným Zukerovým algoritmem, webovou aplikací RNAfold a algoritmem pro metodu Crumple, který vybírá nejoptimálnější strukturu na základě minimalizace volné energie. Hodnoty volných energií jsou uvedeny v Tabulce **5-1**.

Rozdíly ve výpočtech volných energií jsou z tabulky patrné. Velké rozdíly mohou být způsobeny faktem, že volná energie vypočtená webovou aplikací RNAfold počítá s hodnotami pro energetický model Turner 99. Navíc počítá s hodnotami destabilizační energie pro smyčky a výdutě. Výpočet volné energie pro struktury predikované pomocí implementovaných algoritmů počítá s hodnotami energetického modelu Turner 2004 a nepočítá s hodnotami destabilizačních energií. Grafické porovnání volných energií, které byly získány pro jednotlivé struktury, je vyobrazeno na Obr. **5-2**.

Tabulka 5-1: Hodnoty volných energií pro predikované sekundární struktury [kcal/mol]

ID RNA STRAND v2.0: PDB_	Volná energie (Zuker)	Volná energie (RNAfold)	Volná energie (Crumple)
00009	1,0	-0,5	1,0
00028	-1,3	-1,3	-1,3
00037	-4,1	-9,2	-3,1
00077	-1,3	-2,2	-1,3
00097	-0,2	-2,1	-0,2
00118	-1,7	-4,7	-1,7
00151	-1,2	-4,2	-1,2
00168	-5,1	-6,9	-5,6
00191	-3,2	-7,6	-3,2
00212	-9,4	-11,5	-9,4
00226	-2,0	-2,7	-2,0
00799	-0,2	-2,1	-0,2
01064	-6,1	-9,1	-6,1
00002	-4,8	-7,9	-3,6
Training	0,6	0	0,6



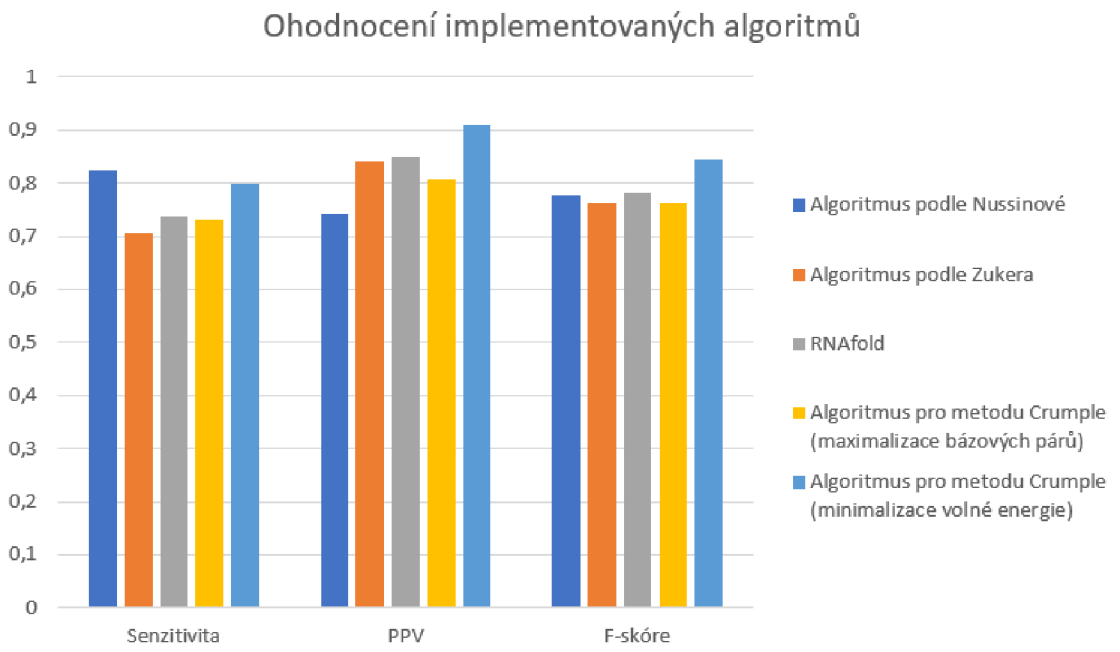
Obr. 5-1: Graf porovnání volných energií pro jednotlivé struktury

Z grafu je patrné, jak se získané hodnoty volné energie liší a shodují. Energie získané pro predikci Zukerovým algoritmem a algoritmem pro metodu Crumple s minimalizací volné energie jsou téměř totožné i přesto, že výsledné struktury se liší. To je s největší pravděpodobností způsobeno tím, že implementovaný algoritmus pro metodu Crumple vybírá v případě více shodným minimálních energií tu strukturu, která je ve výčtu položena nejvýše. Není tedy vyloučeno, že kdyby algoritmus pro metodu Crumple eliminoval shody jinak, byly by predikce těmito algoritmy téměř shodné.

Kvalita každého implementovaného algoritmu byla ohodnocena pomocí senzitivity, PPV a F-skóre. Získané hodnoty senzitivity, PPV a F-skóre jsou uvedeny v Tabulce **5-2** a graficky znázorněny na Obr. **5-2**.

Tabulka 5-2: Ohodnocení implementovaných algoritmů

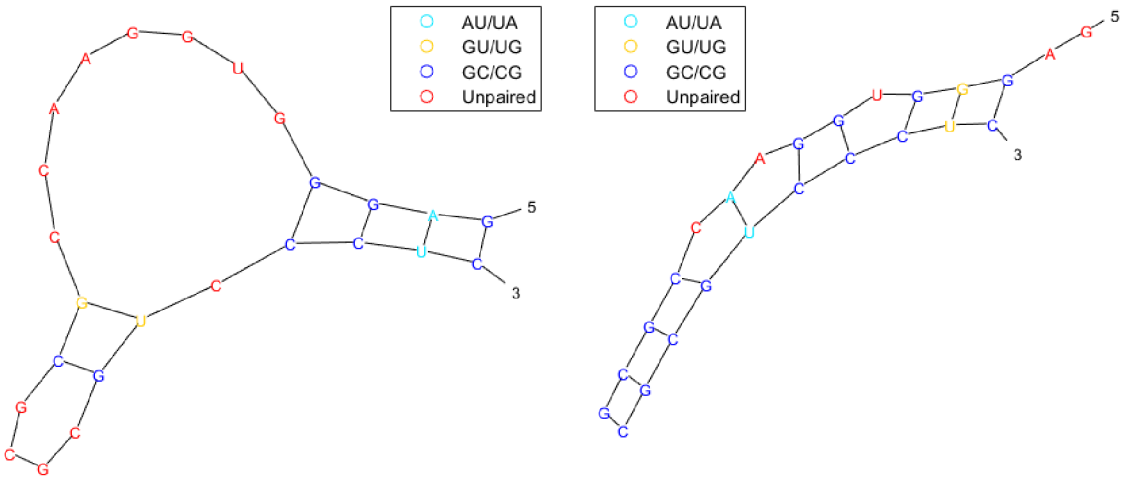
	Senzitivita	**PPV**	**F-skóre**
Predikce Nussinové	0,8245	0,7425	0,7762
Predikce Zukera	0,7067	0,8400	0,7620
Predikce RNAfold	0,7365	0,8492	0,7830
Predikce Crumple (maximalizace párů)	0,7308	0,8079	0,7616
Predikce Crumple (minimalizace energie)	0,7976	0,9103	0,8442



Obr. 5-2: Graf ohodnocení implementovaných algoritmů

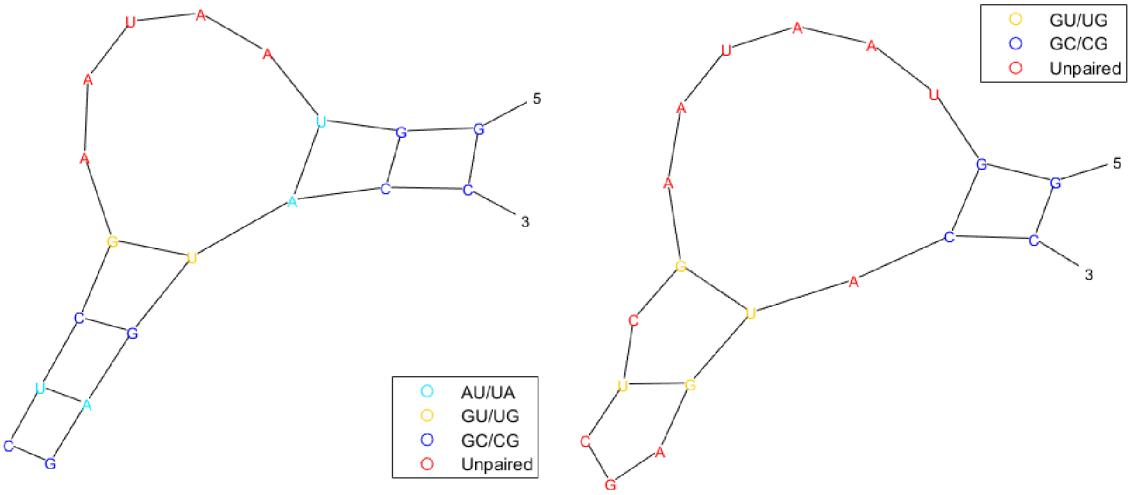
Z tabulky a grafu je patrné, že nejlepší senzitivity dosahuje algoritmus Nussinové a naopak nejhorší algoritmus podle Zukera. Nejvyšší hodnoty PPV a F-skóre dosahuje metoda Crumple s minimalizací volné energie. Implementovaný Zukerův algoritmus a RNAfold dosahují téměř shodných výsledků. Protože RNAfold používá pro predikci právě Zukerův algoritmus, lze říct, že i implementovaný algoritmus podle Zukera byl úspěšný. Celkově dosahuje nejlepších výsledků metoda Crumple s minimalizací volné energie. Potvrzuje tedy myšlenku vyslovenou v kapitole **2.2.3**. Tato myšlenka říká, že metoda Crumple je schopna odhalit nepseudouzlové struktury, které algoritmy využívající minimalizaci volné energie neodhalí.

Sekundární struktury s nejnižší senzitivitou pro každý algoritmus, které byly vykresleny pomocí funkce MATLAB `rnaplot`, jsou zobrazeny společně se známou referenční strukturou na Obr. **5-3**, Obr. **5-4**, Obr. **5-5** a Obr. **5-6**. Rozdíly mezi známou sekundární strukturou a predikovanou sekundární strukturou jsou popsaný pod jednotlivými obrázky.



Obr. 5-3: Porovnání známé sekundární struktury (vlevo) a predikce podle Nussinové (vpravo)

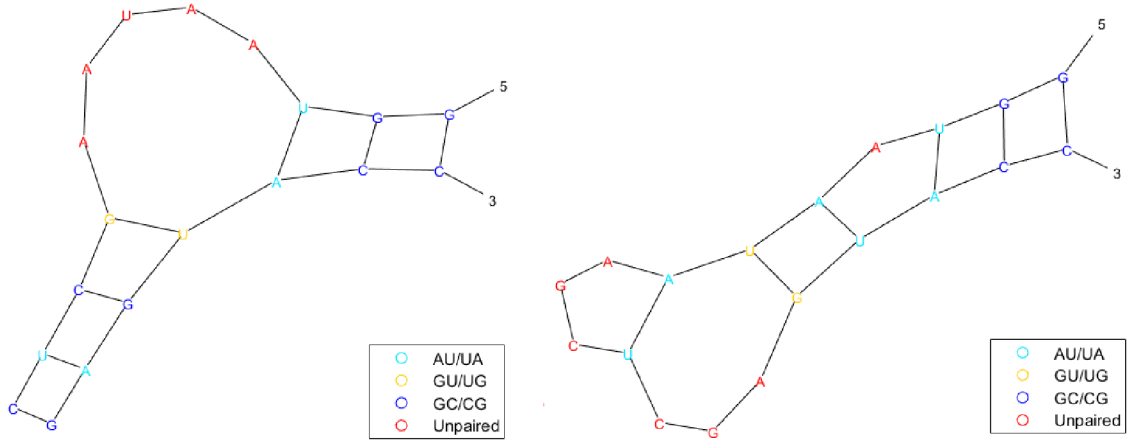
Z Obr. **5-3** je patrné, že algoritmus Nussinové operuje s myšlenkou o maximalizaci bázových párů. Jim předpovězená struktura je tedy spárovaná na všech možných místech a nepredikuje tedy vznik velké vnitřní smyčky. Jedná se o sekundární strukturu pro nukleotidovou sekvenci RNA, která je v datasetu uvedena pod ID PBD_00168.



Obr. 5-4: Porovnání známé sekundární struktury (vlevo) a predikce podle Zukera (vpravo)

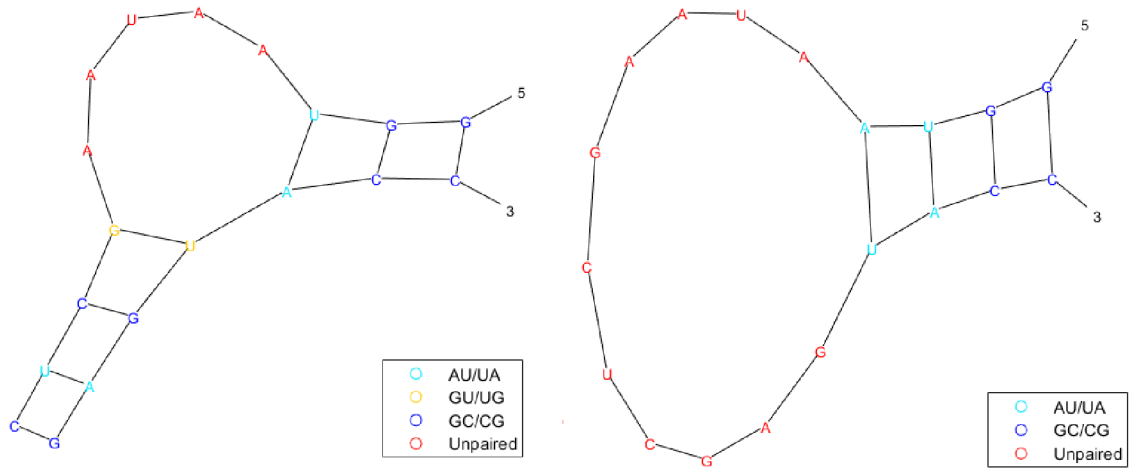
Predikce Zukerovým algoritmem na Obr. **5-4** je téměř shodná se známou sekundární strukturou. Predikovaná struktura obsahuje navíc vlásenkovou a vnitřní smyčku a jednu výduť. Vnitřní smyčka na predikované struktuře vzniká ve stejné oblasti jako výduť na

známé sekundární struktuře. Jedná se o sekundární strukturu pro nukleotidovou sekvenci RNA, která je v datasetu uvedena pod ID PBD_00009.



Obr. 5-5: Porovnání známé sekundární struktury (vlevo) a predikce podle metody Crumple s maximalizací bázových párů (vpravo)

Metoda Crumple s maximalizací bázových párů na Obr. **5-5** se se známou sekundární shoduje v celkem 4 párech bází. Implementované metoda Crumple s maximalizací bázových párů zvolila jako nejoptimálnější strukturu, která obsahuje navíc vlásenkovou smyčku a 2 výdutě. Stejně jako pro predikci Zukerovým algoritmem se jedná o sekundární strukturu pro nukleotidovou sekvenci RNA, která je v datasetu uvedena pod ID PBD_00009.



Obr. 5-6: Porovnání známé sekundární struktury (vlevo) a predikce podle metody Crumple s minimalizací volné energie (vpravo)

Predikce metodou Crumple s minimalizací volné energie na Obr. **5-6** se shoduje se známou sekundární strukturou v prvních třech bázových párech. Na rozdíl od známé sekundární struktury byla navíc predikována vlásenková smyčka o velikosti 11 nukleotidů. I v tomto případě se jedná o sekundární strukturu pro nukleotidovou sekvenci RNA, která je v datasetu uvedena pod ID PBD_00009.

ZÁVĚR

Bakalářské práce byla zaměřena na RNA, konkrétně na její sekundární strukturu a predikci těchto struktur. V teoretické části práce jsou představeny nejrůznější strukturální elementy sekundární struktury a možnosti jejich vizualizace. Dále jsou popsány některé experimentální a výpočetní metody predikce. Vybrané výpočetní metody predikce byly implementovány v programovacím prostředí MATLAB. Kompletní soupis odevzdaných souborů se nachází v Příloze **2.**

V rámci bakalářské práce byly implementovány celkem tři algoritmy predikující sekundární strukturu RNA. Algoritmus podle Nussinové, algoritmus podle Zukera a algoritmus pro metodu Crumple. Metoda Crumple predikuje všechny možné sekundární struktury pro zkoumanou sekvenci, proto je nutné vybrat pouze jednu finální. Finální struktura byla vybírána pomocí dvou přístupů – maximalizace bázových párů a minimalizace volné energie.

Výstupy všech implementovaných algoritmů byly porovnány se známými sekundárními strukturami a mezi sebou. Zukerův algoritmus a algoritmus pro metodu Crumple s minimalizací volné energie byly navíc porovnány s výstupem webové aplikace RNAfold. U těchto dvou algoritmů nebyly pozorovány pouze predikované struktury, ale i výpočty volné energie.

Získané hodnoty volných energií byly téměř totožné i přesto, že predikované struktury byly mírně odlišné. S největší pravděpodobností za to může fakt, že algoritmus pro metodu Crumple s minimalizací volné energie vybírá v případě shody minimálních volných energií tu strukturu, která se nachází ve výčtu možných sekvencí nejvýše. Rozdíly v hodnotách volných energií mezi implementovanými algoritmy a RNAfold byly nejspíše způsobeny tím, že webová aplikace RNAfold využívá pro výpočet volných energií energetický model Turner 99. V bakalářské práci byly pro výpočet volných energie použity parametry energetického modelu Turner 2004.

Kvalita predikce pro všechny implementované metody byla stanovena pomocí senzitivity, PPV a F-skóre. Nejvyšší senzitivity dosahuje algoritmus podle Nussinové, v ostatních parametrech vede metoda Crumple s minimalizací volné energie. Predikce metodou Crumple s minimalizací volné energie také dosáhla nejvíce shod mezi predikovanými a známými sekundárními strukturami. Ty struktury, které dosahovaly pro daný algoritmus nejnižší hodnoty senzitivity byly vykresleny pomocí `rnaplot`, funkce volně dostupné v programovacím prostředí MATLAB.

Celkově se jako úspěšnější jeví přístup minimalizace volné energie oproti přístupu maximalizace bázových párů. Algoritmy, které s tímto přístupem počítaly dosáhly celkově lepších výsledků než algoritmy, které počítaly s přístupem maximalizace bázových párů.

Algoritmy, které počítají s maximalizací bázových párů počítají pouze s teoretickými znalostmi o formaci sekundárních sktruktur. Naopak přístup s minimalizací volné energie

používá pro predikci sekundárních struktur experimentálně získané hodnoty volné energie. Jeho predikce se tedy zakládá na ověřených znalostech, i proto je zřejmě predikce pomocí tohoto přístupu úspěšnější.

CITACE POUŽITÝCH ZDROJŮ

[1] STANĚK, David. *RNA - temná hmota v našich buňkách* [online]. Ústav molekulární genetiky Akademie věd ČR, 2015 [cit. 2020-10-10]. Dostupné z: https://www.img.cas.cz/files/2012/10/RNA-temna_hmota_v_nasich_bukach.pdf

[2] Bhattacharya, S., Mutt, E., & Mitra, A. (2013). RNA Structural Bioinformatics. *Proceedings of Andhra Pradesh Akademi of Sciences*, *15*(January), 101–124.

[3] *Differences between DNA and RNA* [online]. BioNinja [cit. 2020-11-25]. Dostupné z: https://www.ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/dna-versus-rna.html

[4] *The Genetic Code* [online]. OpenStax College [cit. 2020-11-16]. Dostupné z: https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/translation/a/the-genetic-code-discovery-and-properties

[5] *Types of RNA* [online]. BioNinja [cit. 2020-10-10]. Dostupné z: https://www.ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/types-of-rna.html

[6] *The genetic code* [online]. Khan Academy [cit. 2021-01-04]. Dostupné z: https://openstax.org/books/biology/pages/15-1-the-genetic-code

[7] *The Genetic Code* [online]. [cit. 2021-01-04]. Dostupné z: https://openstax.org/books/biology/pages/15-1-the-genetic-code

[8] HIGGS, Paul G. *RNA secondary structure: physical and computational aspects* [online]. University of Manchester, 2000 [cit. 2020-10-16]. Dostupné z: https://www.researchgate.net/publication/12138904_RNA_Secondary_Structure_Physical_and_Computational_Aspects

[9] CHATTERJEE, Kunal. *RNA* [online]. The Ohio State University [cit. 2020-10-16]. Dostupné z: https://www.britannica.com/science/RNA

[10] Campbell, Neil; Reece, Jane B. (2011). *Biology* (9th ed.). Boston: Benjamin Cummings. pp. 339–342. ISBN 978-0321558237.

[11] SCHUDOMA, Christian. *A Fragment Based Approach to RNA Threading* [online]. European Molecular Biology Laboratory [cit. 2020-12-28]. Dostupné z: https://www.researchgate.net/publication/242260995_A_Fragment_Based_Approach_to_RNA_Threading

[12] HOBZA, Pavel a Zdeněk HAVLAS. *Netušená síla slabých vazeb* [online]. Vesmír 89, 2010 [cit. 2020-12-28]. Dostupné z: https://vesmir.cz/cz/casopis/archiv-casopisu/2010/cislo-10/netusena-sila-slabych-vazeb.html

[13] ATKINS, John F, Raymond F GESTELAND a Thomas CECH. RNA worlds: from life's origins to diversity in gene regulation. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press, c2011, xv, 361 p. ISBN 08-796-9946-9.

[14] GOWRI-SHANKAR, Vivrek. *RNA Secondary Structure* [online]. 2003 [cit. 2020-11-16]. Dostupné z: https://www.bioinf.man.ac.uk/resources/phase/manual/node72.html

[15] *RNA and Transcription* [online]. University of Central Florida [cit. 2020-12-28]. Dostupné z: https://www.coursehero.com/file/49949423/10-Transcriptionpptx/

[16] Ahmad, Freed & Mahboob, Shahid & Gulzar, Tahsin & Din, Salah & Hanif, Tanzeela & Ahmad, Hifza & Afzal, Muhammad. (2013). RNA-SSPT: RNA Secondary Structure Prediction Tools. Bioinformation. 9. 873-8. 10.6026/97320630009873.

[17] Yamasaki, S., Amemiya, T., Yabuki, Y. *et al.* ToGo-WF: prediction of RNA tertiary structures and RNA–RNA/protein interactions using the KNIME workflow. *J Comput Aided Mol Des* **33,** 497–507 (2019). https://doi.org/10.1007/s10822-019-00195-y

[18] ZHAO, Yunjie, Yangyu HUANG a Zhou GONG. *Automated and fast building of three-dimensional RNA structures* [online]. 2012 [cit. 2020-10-16]. Dostupné z: https://www.researchgate.net/figure/Predicted-tertiary-structures-of-typical-RNA-molecules-a-a-hairpin-with-internal-loop_fig2_232257514

[19] Vandivier, L. E., Anderson, S. J., Foley, S. W., & Gregory, B. D. (2016). The Conservation and Function of RNA Secondary Structure in Plants. *Annual review of plant biology*, *67*, 463–488. https://doi.org/10.1146/annurev-arplant-043015-111754

[20] FALLMANN, Jörg, Sebastian WILL, Jan ENGELHARDT, Björn GRÜNING, Rolf BACKOFEN a Peter F. STADLER. Recent advances in RNA folding. Journal of Biotechnology. 2017, 261(February), 97–104. ISSN 18734863. DOI:10.1016/j.jbiotec.2017.07.007

[21] *Secondary Structure* [online]. [cit. 2021-01-05]. Dostupné z: http://eternagame.fandom.com/wiki/Secondary_Structural_Motifs_in_RNA

[22] Forsdyke DR (September 1995). "A stem-loop "kissing" model for the initiation of recombination and the origin of introns". *Molecular Biology and Evolution*. **12** (5): 949–58. doi:10.1093/oxfordjournals.molbev.a040273

[23] Svoboda, P., & Cara, A. (2006). Hairpin RNA: A secondary structure of primary importance. Cellular and Molecular Life Sciences, 63(7), 901-908.

[24] Interial Loop. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2021-01-04]. Dostupné z: https://en.wikipedia.org/wiki/Internal_loop

[25] Lua RC, Grosberg Y (2006) Statistics of knots, geometry of conformations, and evolution of proteins. PLoS Comp Biol 2: e45.

[26] Lyngsø, Rune & Pedersen, Christian. (2000). RNA Pseudoknot Prediction in Energy-Based Models. Journal of computational biology : a journal of computational molecular cell biology. 7. 409-27. 10.1089/106652700750050862.

[27] Cao, S., & Chen, S. J. (2009). Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA (New York, N.Y.)*, *15*(4), 696–706. https://doi.org/10.1261/rna.1429009

[28] *Machine learning tool may help us better understand RNA viruses* [online]. Georgia Tech [cit. 2020-10-16]. Dostupné z: https://aihub.org/2020/04/15/machine-learning-tool-may-help-us-better-understand-rna-viruses/

[29] Mattei, E., Ausiello, G., Ferrè, F., & Helmer-Citterich, M. (2014). A novel approach to represent and compare RNA secondary structures. *Nucleic acids research*, *42*(10), 6146–6157. https://doi.org/10.1093/nar/gku283

[30] Moulton, Vincent & Zuker, Michael & Steel, Mike & Pointon, Robin & Penny, David. (2000). Metrics on RNA Secondary Structures. Journal of Computational Biology. 7. 277-292. 10.1089/10665270050081522.

[31] Bhattacharya, S., Mutt, E., & Mitra, A. (2013). RNA Structural Bioinformatics. *Proceedings of Andhra Pradesh Akademi of Sciences*, *15*(January), 101–124.

[32] FIALA, Radovan. *Nukleární magnetická rezonance nepracuje s radioaktivitou* [online]. Brno: Masarykova Univerzita, 2013 [cit. 2020-12-28]. Dostupné z: https://www.em.muni.cz/vite/3560-nuklerani-magneticka-rezonance-nepracuje-s-radioaktivitou

[33] ŽIŽKA, Jan a Vlastimil VÁLEK, et al. *Moderní diagnostické metody. III. díl, Magnetická rezonance*. 1. vydání. Brno : Institut pro další vzdělávání pracovníků ve zdravotnictví, 1996. ISBN 80-7013-225-6.

[34] Zoll, Jan & Hahn, Marc & Gielen, Paul & Heus, Hans & Melchers, Willem & Kuppeveld, Frank. (2011). Unusual Loop-Sequence Flexibility of the Proximal RNA Replication Element in EMCV. PloS one. 6. e24818. 10.1371/journal.pone.0024818.

[35] LOUB, Josef. *Krystalová struktura, symetrie a rentgenová difrakce*. [s.l.]: SPN, 1987.

[36] Kryoelektronová mikroskopie. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2019 [cit. 2020-12-29]. Dostupné z: https://cs.wikipedia.org/wiki/Kryoelektronov%C3%A1_mikroskopie

[37] OLIVER, Carlos G. *Dynamic Programming: Nussinov RNA Folding* [online]. 2017 [cit. 2020-12-29]. Dostupné z: https://cgoliver.com/2017/01/15/Nussinov.html

[38] RICHTER, Karel. *Dynamické programování* [online]. Praha: Katedra počítačů, Fakulta elektrotechnická, České vysoké učení technické v Praze, 2017 [cit. 2020-11-26]. Dostupné z: https://cw.fel.cvut.cz/b172/_media/courses/b6b36dsa/dsa-12-dynamickeprogramovani.pdf

[39] Senzitivita, specificita a prediktivní hodnoty. *Matematická biologie: e-learningová učebnice* [online]. Brno: Institut biostatistiky a analýz Lékařské fakulty Masarykovy univerzity [cit. 2020-11-26]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii--vztah-pravdepodobnosti-statistiky-a-biostatistiky--senzitivita-specificita-a-prediktivni-hodnoty

[40] LEGAULT, Pascale a Karel RICHTA. *Složitost algoritmů* [online]. Praha, 2018 [cit. 2020-11-06]. Dostupné z: https://cw.fel.cvut.cz/b182/_media/courses/b6b36dsa/dsa-3-slozitostalgoritmu.pdf ČVUT, fakulta elektrotechnická.

[41] *Asymptotická složitost* [online]. Vojtěch Hordějčuk, 2008 - 2020 [cit. 2020-11-20]. Dostupné z: https://voho.eu/wiki/asymptoticka-slozitost/

[42] Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. J Mol Biol. 2002 Jun 21;319(5):1059-66. doi: 10.1016/S0022-2836(02)00308-X. PMID: 12079347.

[43] *Termodynamika* [online]. [cit. 2020-12-29]. Dostupné z: https://kof.zcu.cz/vusc/pg/termo09/termodynamics/energy/energy1.htm

[44] Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., & Hofacker, I. L. (2006). Thermodynamics of RNA-RNA binding. Bioinformatics, 22(10), 1177–1182. https://doi.org/10.1093/bioinformatics/btl024

[45] WOLFSHEIMER, S a HARTMANN, AK. *Minimum-Free-Energy Distribution of RNA Secondary Structures: Entropic and Thermodynamic Properties of Rare Events* [online]. 2010 [cit. 2021-5-18]. Dostupné z: https://uol.de/f/5/inst/physik/ag/compphys/download/Alexander/publications/RNAFreeDistr_v4.pdf

[46] Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., & Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(24), 9373–9377. https://doi.org/10.1073/pnas.83.24.9373

[47] *Secondary Structure Loop Decomposition* [online]. [cit. 2021-5-17]. Dostupné z: https://www.tbi.univie.ac.at/RNA/ViennaRNA/doc/html/energy_evaluation.html

[48] Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 2010 Jan;38(Database issue):D280-2. doi: 10.1093/nar/gkp892. Epub 2009 Oct 30. PMID: 19880381; PMCID: PMC2808915.

[49] *Turner 2004 Nearest Neighbors* [online]. 2009 [cit. 2021-5-18]. Dostupné z: https://rna.urmc.rochester.edu/NNDB/turner04/index.html

[50] Nussinov, R., & Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *77*(11), 6309–6313. https://doi.org/10.1073/pnas.77.11.6309

[51] DURBIN, R., S. EDDY, A. KROGH a G. MITCHISON. *Biological sequence analysis* [online]. Cambridge: Cambridge University Press, 1998 [cit. 2021-5-18]. ISBN 0 521 62041 4. Dostupné z: http://www.mcb111.org/w06/durbin_book.pdf

[52] Ahmad, Freed & Mahboob, Shahid & Gulzar, Tahsin & Din, Salah & Hanif, Tanzeela & Ahmad, Hifza & Afzal, Muhammad. (2013). RNA-SSPT: RNA Secondary Structure Prediction Tools. Bioinformation. 9. 873-8. 10.6026/97320630009873.

[53] BARQUIST, Lars. *A brief introduction to computational prediction of RNA secondary structure* [online]. Institute for RNA-based Infection Research [cit. 2020-11-27]. Dostupné z: www.imib-wuerzburg.de/fileadmin/user_upload/3_Education/RNA_workshop/5_RNA_structure_prediction_Barquist.pdf

[54] *An Alternative Traceback Method for Nussinov's RNA Folding Algorithm* [online]. [cit. 2020-12-28]. Dostupné z: http://datamech.com/devan/nussinov-traceback.html

[55] *Nussinov algorithm to predict secondary RNA fold structures* [online]. 2019 [cit. 2020-12-28]. Dostupné z: https://bayesianneuron.com/2019/02/nussinov-predict-2nd-rna-fold-structure-algorithm/

[56] Lei, G., Dou, Y., Wan, W., Xia, F., Li, R., Ma, M., & Zou, D. (2012). CPU-GPU hybrid accelerating the Zuker algorithm for RNA secondary structure prediction applications. *BMC genomics*, *13 Suppl 1*(Suppl 1), S14. https://doi.org/10.1186/1471-2164-13-S1-S14

[57] SHAH, Hardik. *Algorithms For Predicting Secondary Structures Of Human Viruses* [online]. San Jose State University, 2012 [cit. 2020-11-27]. Dostupné z: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1279&context=etd_projects Masters Project's. The Faculty of the Department of Computer Science.

[58] Bleckley, S., Stone, J. W., & Schroeder, S. J. (2012). Crumple: a method for complete enumeration of all possible pseudoknot-free RNA secondary structures. *PloS one*, *7*(12), e52414. https://doi.org/10.1371/journal.pone.0052414

[59] ViennaRNA Package. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2019 [cit. 2020-12-29]. Dostupné z: https://en.wikipedia.org/wiki/ViennaRNA_Package

[60] *ViennaRNA Web Services* [online]. Institute for Theoretical Chemistry [cit. 2020-12-29]. Dostupné z: http://rna.tbi.univie.ac.at/

[61] *Rnafold* [online]. 2007 [cit. 2021-5-18]. Dostupné z: https://www.mathworks.com/help/bioinfo/ref/rnafold.html

[62] *Rnaplot* [online]. 2007 [cit. 2021-5-18]. Dostupné z: https://www.mathworks.com/help/bioinfo/ref/rnaplot.html

[63] Andronescu, M., Bereg, V., Hoos, H. H., & Condon, A. (2008). RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, *9*, 1–10. https://doi.org/10.1186/1471-2105-9-340

SEZNAM SYMBOLŮ A ZKRATEK

Zkratky:

FEKT	…	Fakulta elektrotechniky a komunikačních technologií
VUT	…	Vysoké učení technické v Brně
RNA	…	Ribonukleová kyselina
DNA	…	Deoxyribonukleová kyselina
OH skupina	…	Hydroxylová skupina
mRNA	…	Mediátorová RNA
rRNA	…	Ribozomová RNA
tRNA	…	Transferová RNA
A	…	Adenin
C	…	Cytosin
G	…	Guanin
U	…	Uracil
PPV	…	Pozitivní prediktivní hodnota
NMR	…	Nukleární magnetická rezonance
RTG	…	rentgen/rentgenová
MR	…	Magnetická rezonance
MFE	…	Minimální volná energie

SEZNAM PŘÍLOH

Příloha 1 - Kompletní dataset predikovaných struktur

ID RNA STRAND v2.0: PDB_	Známá sekundární struktura	Predikce podle Nussinové	Predikce podle Zukera	RNAfold	Metoda Crumple (maximalizace párů)	Metoda Crumple (minimalizace energie)
00009	(((…..(((())))))))	((…(..(((())))))))	((……(.(…))))	((((……..))))	(((.(((…)…)))))	((((……..))))
00028	((((((..))))))	(((..(((())))))	(((((…)).)))	(((((….)))))	(((.((…)))))	(((((….)))))
00037	(((((((((()))))))).)))	(((.(((((())))))))))	((.(((.(…)))).))	(((((((….)))).)))	((((.(((…)))))).)))	(((((((….)))).))
00077	((((((()))))))	((((((()))))))	((((….))))	((((….))))	((((….))))	((((….))))
00097	(((….)))	((((())))))	(((….)))	(((…)))	(((….)))	(((….)))
00118	(((((….)))))	(((((((())))))))	(((((….)))))	(((((….)))))	(((((….)))))	(((((….)))))
00151	(((((..)))))	(((..(())))))	(((.(…))))	((((….))))	(((.(…))))	((((….))))
00168	((((……..((….)).))))	..(((.((.(.((((())))))))))	(((…(….((….))..))))	(((((((……..)).)))))	(((((((..(…)).)).)))))	(((((…...((….)).)))))
00191	(((((…..)))))	((..(((.())))))	((..(((…)))))	((((…..)))))	((..(((…)))))	((((.….))))))
00212	(((((((())))))))	(((((((())))))))	((((.((…))))))	((((((….))))))	((((.((…))))))	((((((….))))))
00226	(((((…)))))	((((…())))))	((((…)))))	((((…)))))	((((…)))))	((((…)))))
00799	…((.(()).))	…((((()))))	…(((….)))	…(((….)))	…(((….)))	…(((….)))
01064	(((((((.)))))))	((((((.())))))))	(((((((…)))))))	(((((((…)))))))	(((((((…)))))))	(((((((…)))))))
00002	(((((.(((….)))))))))	(((((.(((..())))))))))	(((…(((….))).)))	((((((.((….))))))))	(((((.(((….)))))))))	(((((.(((….)))))))))
Training	.((..()))	.((..()))	(((…)))	……	(((…)))	(((…)))

Příloha 2 - Soupis elektronických příloh

- Dataset_sekvenci.xlsx – dataset RNA sekvencí se známou sekundární strukturou.
- Turner04.xlsx – energetický model Turner 2004 s hodnotami volné energie pro vnitřní smyčky, vlásenkové smyčky a výdutě.
- stackedPairs.xlsx – hodnoty volné energie pro spojené páry podle energetického modelu Turner 2004.
- SecondaryStructurePrediction.m – funkce, do které vstupuje vytvořený dataset a z níž vystupují kompletní predikce všech implementovaných metod, ohodnocení kvality implementovaných metod a vykreslené nejvíce odlišné predikované sekundární struktury.
- predictionQuality.m – funkce, počítající senzitivitu, PPV a F-skóre pro implementované algoritmy.
- NussinovAlgorithm.m – funkce, predikující sekundární struktury podle algoritmu Nussinové.
- ZukerAlgorithm.m – funkce, predikující sekundárni struktury podle Zukerova algoritmu.
- internalBulgeValue.m – funkce, počítající příspěvky volné energie pro vnitřní smyčky a výdutě v rámci Zukerova algoritmu.
- stackedPairsValue.m – funkce, počítající příspěvky volné energie pro spojené páry bází.
- CrumpleMaxPairsPrediction.m – funkce, predikující sekundární strukturu metodou Crumple s maximalizací bázových párů.
- CrumpleMinEnergyPrediction.m – funkce, predikující sekundární strukturu metodou Crumple s minimalizací volné energie.
- CrumpleSequences.m – funkce, predikující všechny možné sekundární struktury metodou Crumple.
- CrumplePairs.m – funkce, počítající výskyt všech bázových párů ve všech strukturách predikovaných funkcí CrumpleSequences.
- structureFreeEnergy.m – funkce, počítající celkovou volnou energii predikované struktury.
- readme.txt – textový soubor, popisující vstupy, výstupy a návaznost všech odevzdaných elektronických příloh.

Všechny výše uvedené elektronické přílohy jsou podrobně popsány v textu bakalářské práce, kde je jejich popis doplněn i vývojovými diagramy.