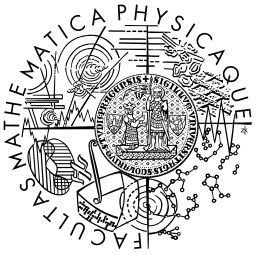


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

**DIPLOMOVÁ PRÁCE**

Bc. Rastislav Galvánek

# Predikce terciární struktury RNA s využitím více vzorů

Katedra teoretické informatiky a matematické logiky

| | |
|---|---|
| Vedoucí diplomové práce: | RNDr. David Hoksza, Ph.D |
| Studijní program: | Informatika |
| Studijní obor: | IUI |

Praha 2019

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

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

V Praze dne ..... Podpis autora

Poděkování.

Název práce: Predikce terciární struktury RNA s využitím více vzorů

Autor: Bc. Rastislav Galvánek

Katedra teoretické informatiky a matematické logiky: Katedra teoretické informatiky a matematické logiky

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

Abstrakt: Abstrakt.

Klíčová slova: klíčová slova

Title: RNA tertiary structure prediction using multiple templates

Author: Bc. Rastislav Galvánek

Department of Theoretical Computer Science and Mathematical Logic: Department of Theoretical Computer Science and Mathematical Logic

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

Abstract: Abstract.

Keywords: key words

# Obsah

# Úvod

Následuje několik ukázkových kapitol, které doporučují, jak by se měla diplomová práce sázet. Primárně popisují použití TEXové šablony, ale obecné rady poslouží dobře i uživatelům jiných systémů.

# 1. RNA štruktúra

V tejto práci sa venujeme automatickej predikcií priestorovej štruktúry ribonukleovej kyseliny, preto sa budeme v prvej kapitole venovať jej významu z biologického hľadiska. Ďalej preberieme možnosti, akým spôsobom a aké podrobné informácie o štruktúre RNA dokáže súčastná veda získať experimentálnym spôsobom a aká je motivácia pre počítačovú predikciu RNA. Nakoniec uvedieme možnosti, ako RNA štruktúru reprezentovať vo formáte textových súborov a aké informácie o štruktúre jednotlive typy súborov uchovávajú.

## 1.1 RNA

Ribonukleová kyselina slúži na prenos alebo uchovávanie genetickej informácie vo všetkých živých organizmoch. Najznámejšia je jej úloha v Centrálnej dogme molekulárnej biológie Crick (1970), kde slúži pri syntéze proteínov z DNA na prenášanie genetickej informácie.

RNA je tiež, ako DNA tvorená štyrmi typmi nukleotidov (báz). Sú to adenín (A), guanín (G), cytozín (C) a uracyl (U). Na rozdiel od RNA sa v DNA namiesto uracylu vyskytuje báza tymín (T). Jednotlivé nukleotidy sú chemicky naviazané na cukre - ribóze, ktorý ich spája do vlákna (v prípade DNA sa jedná o deoxiribózu). Dĺžka vlákna môže byť v závislosti od typu RNA od jednotiek až po tisíce nukleotidov. Pre DNA následne platí, že sa vodíkovými väzbami spájajú dve komplementárne vlákna do špirály, čo čiastočne určuje pravidelný tvar molekuly v priestore. RNA sa však vyskytuje hlavne v jednovláknovej forme, pričom sa vlákno spája vodíkovými väzbami samo so sebou, čo prináša veľkú variabilitu v tvare molekuly. Platí, že vodíkovými väzbami sa navájom viažu nukleotidy cytozínu a guanínu, adenínu a uracylu, guanínom a uracylom.

## 1.2 Druhy a funkcie RNA

Okrem prenosu genetickej informácie pri syntéze proteínov, ktorý pozostáva z replikácie DNA, transkripcie DNA do RNA a nakoniec translácie z RNA do proteínu, pričom sa využívajú rôzne typy RNA, zastáva RNA aj iné funkcie. Slúži napríklad na uchovávanie genetickej informácie niektorých jednoduchých organizmov ako sú vírusy. Tie môžu na uchovanie genetickej informácie používať jednovláknovú RNA, dvojvláknovú RNA a v prípade retrovírusov špeciálny typ RNA, ktorý je schopný prepisovať genetickú informáciu z RNA do DNA procesom reverznej transkriptázy a vložiť tak svoju genetickú informáciu do génu napadnutej bunky. Medzi tieto vírusy patrí napríklad známy vírus HIV. Krupovic a kol. (2018)

Prehľad niektorých typov RNA:

- kódujúca (2%)
  - mediátorová RNA (mRNA)
- nekódujúca (98%)

- ribozomálna RNA (rRNA)
- prenosová RNA (tRNA)
- funkcionálna RNA (fRNA)
- mikro RNA (miRNA)
- malá interferujúca (small interfering) RNA (siRNA)
- jadrová (nuclear) RNA (snRNA)
- jadierková (nucleolar) RNA (snoRNA)
- vírusová RNA (vRNA)
- dlhá nekódujúca RNA (lncRNA)
- ďalšie...

Messenger RNA (mRNA) vzniká pri prepise (transkripcií) DNA v jadre bunky. Najprv je vytvorená pre-mRNA, ktorá obsahuje aj nekódujúce úseky. V ďalšom kroku sa z nej ešte v jadre bunky pomocou procesu nazývaného splicing, odstránia exóny (nekódujúce úseky) za pomoci snRNA, ktorá rozpoznáva sekvenciu báz AGGU označujúcu prechod medzi intrónom a exónom. Následne mRNA putuje von z jadra jeho pórami do cytoplazmy, kde sa naviaže na ribozóm. Alberts B (2002)

Ribozóm obsahuje ribozomálnu RNA, ktorá sa zúčastňuje translácie (prekladu) mRNA na kódovanú bielkovinu. Okrem toho je to typicky najčastejšie sa vyskytujúca RNA v bunke, pričom jej dĺžka môže byť niekoľko tisíc nukleotidov.

Prenosová tRNA sa nachádza v cytoplazme bunky a jej fubkcia spočíva v dopravení správnej aminokyseliny do procesu translácie.

Niektoré typy RNA plnia regulačnú funkciu. Napríklad mikro RNA (miRNA) zabraňuje procesu translácie mRNA tým, že sa na ňu naviaže a zabráni jej spojeniu s ribozómom. snoRNA zas hrá úlohu pri modi

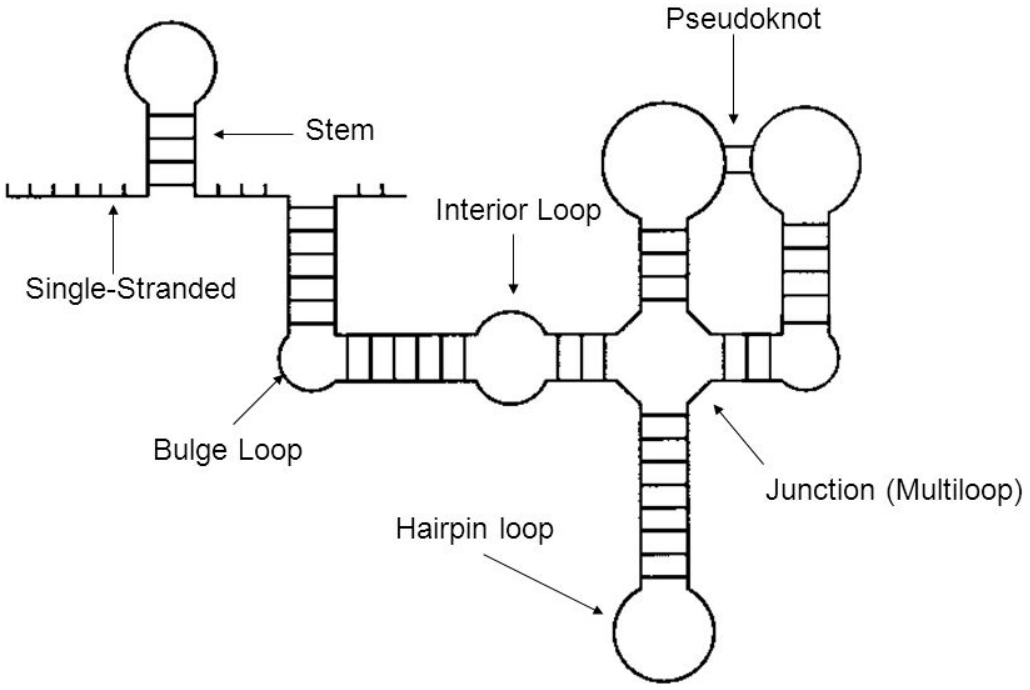
kácii ostatných typov RNA - hlavne rRNA, tRNA a snRNA.

Sekvencie lncRNA mávajú typicky dĺžku okolo 200 nukleotidov a jeden zo známych zástupcov je gén XIST, ktorý sa uplatňuje pri procese inaktivácie chromozómu X. **??**

## 1.3 Reprezentácia a práca s RNA

Fyzicky je RNA v bunke vlastne len mnoho atómov vodíka, kyslíka, uhlíka, fosforu usporiadaných v priestore vďaka chemickým a fyzikálnym väzbám a vlastnostiam atómov. Pre to, aby sme ich mohli spracovávať pomocou počítača potrebujeme vhodnú reprezentáciu štruktúry. Typ reprezentácie závisí od toho, aké informácie o štruktúre chceme mať k dispozícií a takisto aké informácie sme schopní získať. Principiálne môžeme rozdeliť reprezentáciu RNA štruktúr na nasledujúce štyri úrovne:

- Primárna
- Sekundárna
- Terciárna



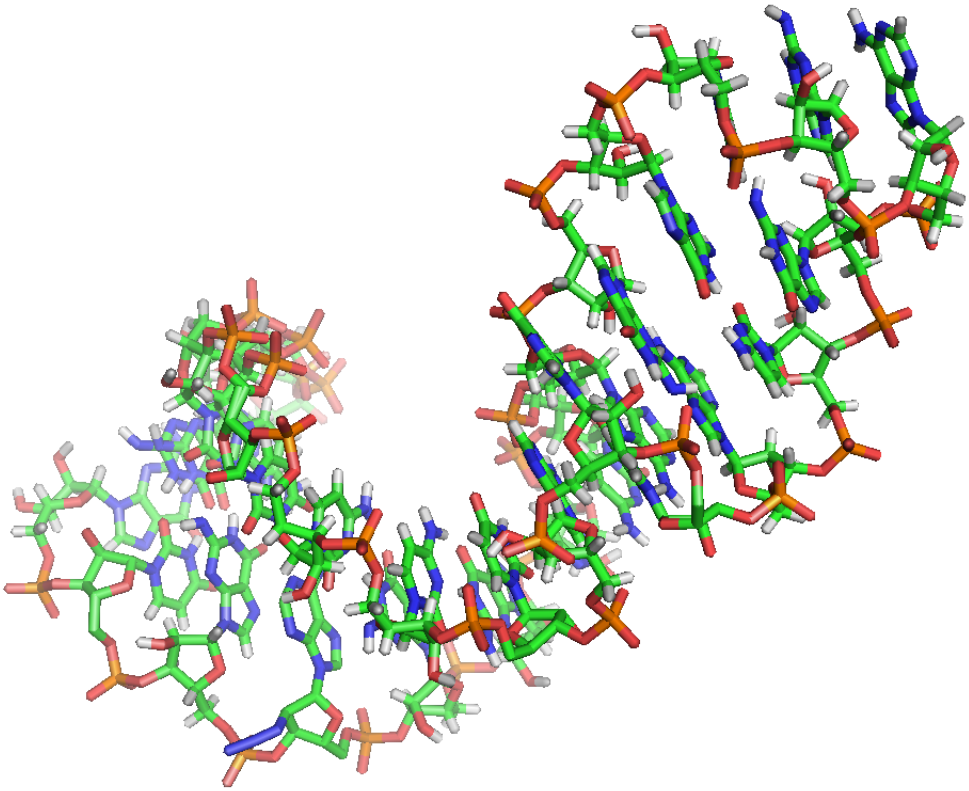

Obrázek 1.1: Príklad nakreslenia sekundárnej štruktúre RNA. Eddy (2004)

- Kvarciárna

Primárna štruktúra je najviac zjednodušená reprezentácia RNA. Určuje len poradie a typ jednotlivých nukleotidov v RNA štruktúre a ďalej ju budeme v práci tiež označovať ako sekvenciu. V počítači ju reprezentujeme ako textový súbor typu fasta, ktorý má na prvom riadku identifikáciu štruktúry (názov, chain) a v ďalších riadkoch sú to len písmena A, G, C, U určujúce presné poradie nukleotidov. V prípade, že typ nukleotidu na niektorej pozícií je neznámy používame písmeno X alebo N. Primárna sekvencia RNA takisto slúži ako jeden zo vstupov pre náš prediktor a definuje sekvenciu štruktúry, ktorú cheme získať.

Sekundárna štruktúra zachytáva vodíkové väzby medzi jednotlivými nukleotidmi vo vlákne RNA. Dva nukleoidy, ktoré sú spojené vodíkovou väzbou označujeme ako base pair. Vďaka spájaniu jednotlivých nukleotidov vieme v sekundárnej štruktúre pozorovať rôzne podštruktúry, ako napríklad helix, loop, pseudoknot, hairpin loop, internal loop, branch loop , stem a ďalšie 1.1. Sekundárnu štruktúru budeme v tejto práci používať na pomoc pri predikcii terciárnej štruktúry, nakoľko nám dáva informáciu o nukleotidoch, ktoré sú spojené vodíkovou väzbou a teda sa nachádzajú blízko pri sebe. Sekundárnu štruktúru molekuly budeme reprezentovať ako textový súbor, kde bodka značí, že na nukleotid danej pozícii nie je spárovaný žiadnou väzbou,base pair spojený vodíkovou väzbou je značený ako valídne uzátvorkovanie jednoduchými zátvorkami a pseudoknot býva reprezentovaný hranatými zátvorkami.

Zmyslom terciárnej štruktúry je popísať presné rozloženie jednotlivých atómov v trojdimenzionálnom priestore za pomoci koordinátov. V našej práci je hlavným cieľom tieto koordináty určiť za predpokladu znalosti primárnej sekvencie a terciárnych štruktúr ďalších RNA molekúl, ktoré sa pokúšame pri predikcii použiť ako vzory.



Obrázek 1.2: Príklad terciárnej štruktúry RNA nachádzajúcej sa v baktérii escherichia coli.

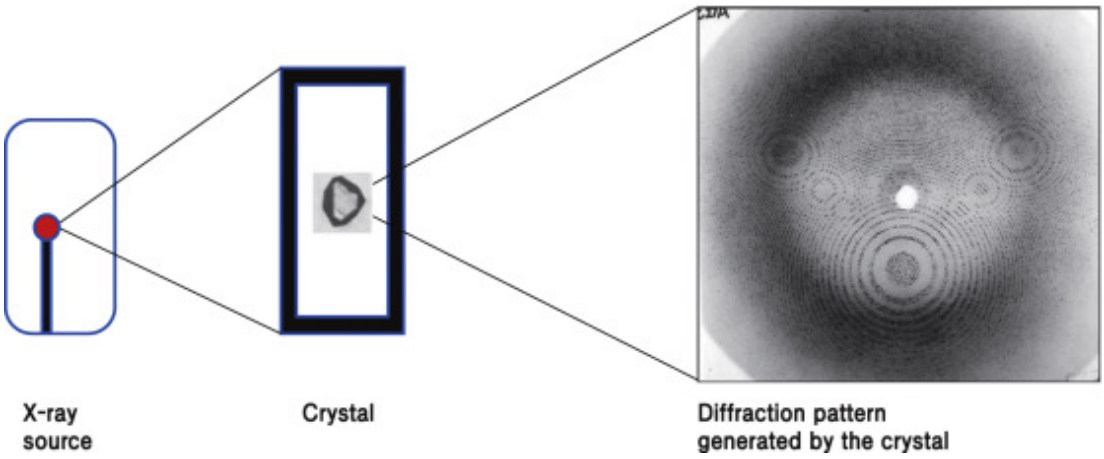
Kvarciárna štruktúra RNA popisuje vzťahy medzi celými molekulami RNA - napríklad interakcie medzi jednotlivými molekulami RNA v ribozómoch Noller (1984) a taktiež vzťahy medzi RNA a molekulami bielkovín.

## 1.4 Význam a získavanie terciárnej štruktúry

Štruktúra RNA priamo súvisí s funkciou, ktorú vykonáva.Tvar štruktúry 1.2 určuje, ktoré enzýmy sa na ňu dokážu pripájanať a prípadne ju modifikovať a s ktorými bielkovinami a nukleovými kyselinami sa dokáže viazať. Bolo preukázané, že väčšina časti RNA štruktúry, ktoré sú schopné sa viazať s inými molekulami nie sú súčasťou žiadneho base pair-u a teda sú v sekundárnej štruktúre označené ako nespárované Schudoma C. (2010). Zmena terciárnej štruktúry, ktorá vedie ku strate pôvodnej funkcie molekuly, sa nazýva denaturácia.

Bolo vyvinutých viacero experimentálnych metód, pomocou ktorých môžeme získať terciárnu štruktúru RNA Felden (2007):

- metódy s vysokou presnosťou
  - X-ray crystallography
  - Cryo-electron microscopy
  - Nuclear Magnetic Resonance (NMR) spectroscopy
- metódy s nižšou presnosťou




Obrázek 1.3: Princíp rentgenovej kryštalografie. Ryu (2017)

- Mass spectrometry
- Chemical probing
- Thermal denaturation
- RNA engineering

X-ray crystallography (rentgenova kryštalografia) funguje principiálne tak, že sa molekula najprv zkryštalyzuje a následne sa nasvieti rentgenovým lúčom. Z kryštálu je lúč odrazený a pritom rozdelený na viacero lúčov. Zmeraním týchto uhlova intenzity odrazených lúčov je následne možné určiť pozície jednotlivých atómov v molekule. Momentálne je to jesna z najpoužívanejších metód na získanie mnohých makromolekulárnych štruktúr. Rozlíšenie získanej štruktúry sa pohybuje okolo 2.0 Å. 1.3

Cryo-electron microscopy metóda využíva zmrazenie molekuly v substancií, ktorá je následne pozorovaná elektrónovým mikroskopom. Princíp metódy je známy približne od roku 1970, ale až donedávna pomocou nej nebolo možné získať tak presné vysledky, ako pomocou rentgenovej kryštalografie. Na druhej strane dĺžka skúmanej štruktúry nie je pri tejto metóde tak limitujúcim faktorom. V roku 2017 bola udelená nobelová cena za chémiu J. Dubochetovi, J. Frankovi a R. Hendersonovi za vyvinutie metódy, ktorou sa dá získať atómová štruktúra molekuly s vysokým rozlíšením.

Metóda Nuclear Magnetic Resonance je založená na pôsobení statického magnetického poľa na jadrá atómov v molekule. Je vhodná hlavne na získavanie kratších štruktúr.

Experimentálne prístupy sa od seba navzájom líšia presnosťou výsledku, dĺžkou štruktúry s ktorou sú schopné pracovať, ale ich hlavnám problémom je, že sú stále časovo náročné a drahé. Pretože získanie primárnej štruktúry RNA a proteinov je oveľa ľahšia úloha, začali byť skúmané aj možnosti, ako predikovať sekundárnu a terciárnu štruktúru za pomoci počítača, čomu sa budeme v našej práci venovať.

# 2. Metódy výpočetnej predikcie

Cieľom výpočetnej predikcie RNA štruktúry je dokázať algoritmicky modelovať terciárnu alebo sekundárnu štruktúru zo znalosti primárnej sekvencie RNA molekuly. Pri takejto predikcií je dôležité, aby sme dostali čo najpresnejší výsledok v porovnaní s experimentálnymi metódami, ale aby výpočové nároky a čas boli výrazne nižšie, ako v prípade experimentálnej rezolúcie štruktúry. Aby malo zmysel sa pokúšať o predikciu štruktúry zo sekvencie, potrebujeme vedieť, že terciárna a teda aj sekundárna štruktúra je do veľkej miery jednoznačne určená štruktúrou primárnou.

Túto otázku môžeme zodpovedať vďaka znalostiam zo skladania (foldingu) proteínov, ktorých výskumu sa venovalo viacej úsilia. Platí, že skladanie bielkovín a RNA prebieha veľmi podobne a preto poznatky o štruktúrach a sekvenciách bielkovín môžeme použiť aj pri RNA. Moore (1999)

Existujú dve hlavné pozorovania, ktoré nám umožnujú štruktúry makromolekúl modelovať Jenny Gu (2009):

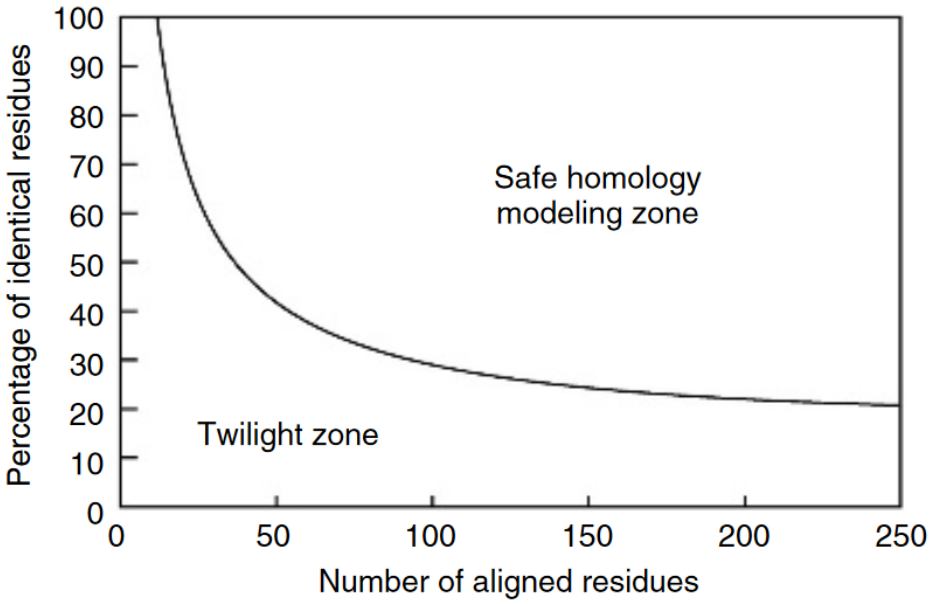
- Štruktúra proteínu je unikátne určená sekvenciou aminokyselín.
- Štruktúra sa zachováva aj pri určitých zmenách v sekvencií a teda platí, že napriek odlišnosti v sekvenciách sú štruktúry veľmi podobné. Je to spôsobené tým, že počas evolúcie štruktúra stále plní podobnú úlohu a preto sa jej tvar ostáva nezmenený aj napriek mutáciam sekvencie. Vďaka rozširujúcej sa databázi makromolekúl (Protein Data Bank) boli získané vzťahy, ako ppodobné si musia byť rovnako dlhé sekvencie, aby sme mohli predpokladať, že aj ich štruktúry sú podobné. Tento vzťah zobrazuje obrázok 2.1.

## 2.1 Ab initio predikcia

Pri ab initio predikcií štruktúry vychádzame iba z primárnej sekvencie a chemicko-fyzikálnych vlastností, vďaka ktorému sa v reálnom svete štruktúra skladá do stabilného tvaru. Algoritmus postupne vytvára kandidátske štruktúry tak, že sa snaží sa minimalizovať funkciu predstavujúcu voľnú energiu (energia, ktorá je ľahko dostupná v systéme). Následne z takto vygenerovaných kandidátov musí vybrať najprirodzenejšiu štruktúru. Najväčším problémom tohoto prístupu je mnoho lokálnych miním vo funkcií predstavujúcej voľnú energiu a preto aj výpočetná zložitosť.

Tieto komplikácie sa dajú čiastočne riešiť viacerými spôsobmi. Jedna cesta je zvýšiť výpočetný výkon - použitie superpočítača, alebo distribuovať výpočet na mnoho výpočetných staníc. Ďalšia je pokus o zmenšenie vyhľadávacieho priestoru a efektívnejšie vyhľadávať kandidátske štruktúry. Jedna metóda je označovaná ako coarse-grained reprezentácie, kde nie sú reprezentované všetky atómy. Využívajú sa taktiež heuristické a pravdepodobnostné metódy na zmenšienie prehľadávaného priestoru.

Stále však platí, že takáto metóda je pre dlhšie štruktúry nepoužiteľná. Napriek tomu, že súšastný state-of-the art umožňuje predikovať štruktúry celkom




Obrázek 2.1: Vzťah dĺžky štruktúr a percentuálneho pomeru identických residuí v sekvenciách určujúce predpoklad, že štruktúry takýchto sekvencií sú podobné. Jenny Gu (2009)

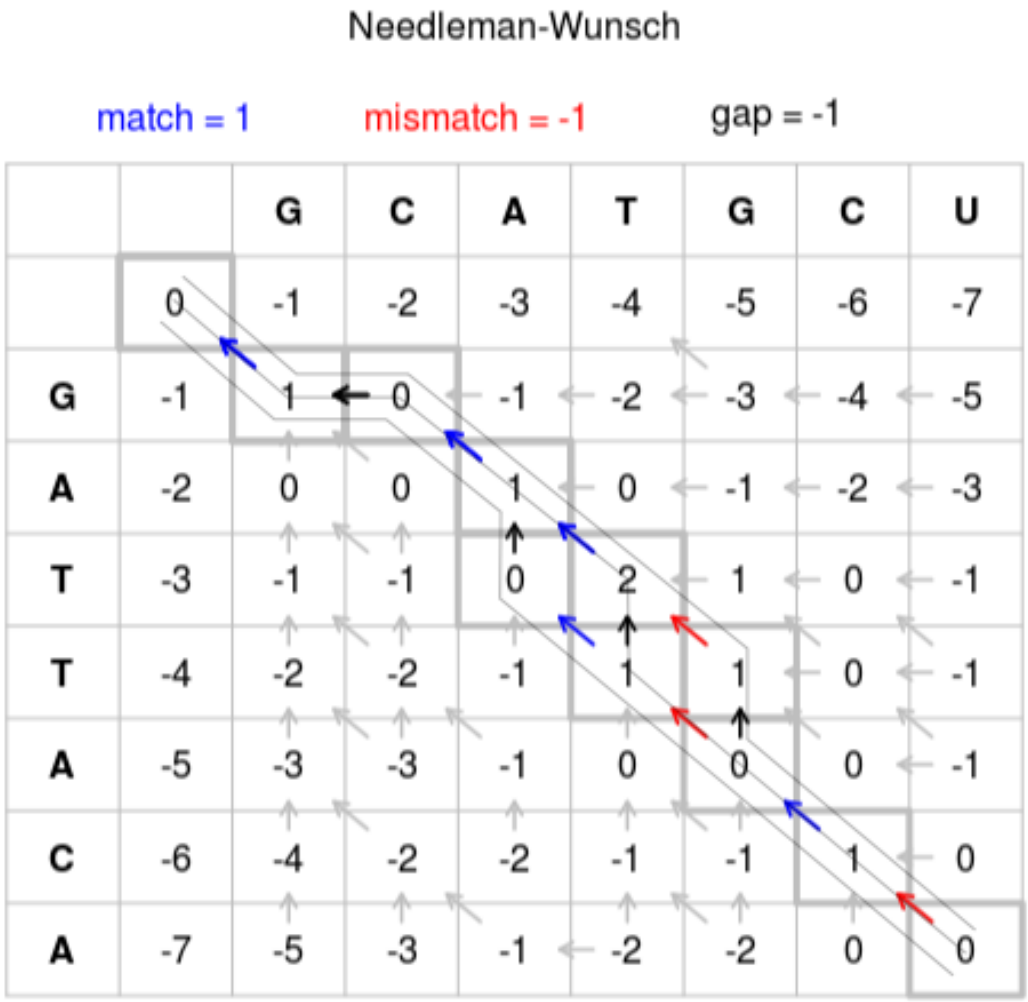
presne s rastúcou dĺžkou sekvencie neúmerne rastie výpočetná náročnosť. Ako príklad uvedieme pokus predikovať štruktúru dlhú 112 nukleotidov, pričom výsledná štruktúra sa líšila od experimentálne získanej len minimálne, výpočet však stál viac ako 100 000 hodín CPU. Qian a kol. (2007)

## 2.2 Knowledge based de novo predikcia

Princíp tohoto typu predikcie je veľmi podobný, ako ten v ab initio metóde ale namiesto samplovania možných usporiadaní atómov používa knižnucu krátkych úsekov štruktúry ( väčšinou dĺžky 2-5 nukleotidov). Algoritmus následne vytvára kandidátske štruktúry tým, že kombinuje jednotlivé krátke úseky štruktúr z knižnice do kandidátskych štruktúr a takisto minimalizuje voľnú energiu modelu. Výhodou je hlavne zrýchlenie generovanie kandidátskych štruktúr oproti ab iniio predikcii. Aj tak je však predikovanie dlhých štruktúr príliš pomalé. Mnohé nástroje preto umožňujú vložiť sekundárnu štruktúru predikovanej sekvencie a tak zmenšiť prehľadávaný priestor.

Ďalší spôsob, ako znížiť prehľadávany priestor je použitie internej reprezentáce štruktúry. V prípade, že atómy reprezentujeme v súradnicami v trojdimenzionálnom priestore, ich síce viem dobre zobraziť, ale takáto reprezentácia má 3*počet _átomov stupňov voľnosti. Výhodnejšie je štruktúru reprezentovať napríklad pomocou reprezentácie uhlov medzi nukleotidmi. 2.2

Nástroj FARFAR Das R. (2010), ktorý používame v našej predikcií patrí tiež medzi knowledge based modelovacie metódy.




Obrázek 2.2: Reprezentácia RNA fragmentu pomocou siedmych uhlov. Frellsen a kol. (2009)

## 2.3 Alignment sekvencií

Zarovnanie dvoch sekvencií slúži na získanie informácie o tom, či su dané sekvencie nejako evolučne, štrukturálne, alebo funkčne príbuzné. Existuje viacero druhov algoritmov zarovnania - napríklad jednoduchý dot plot vhodný na jednoduchú vizualizáciu zarovnania, heuristické metódy ako FASTA a BLAST určené na čo najrýchlejšie porovnanie sekvencie s rozsiahlou databázou ďalších sekvencií, alebo metódy počítajúce najlepšie zarovnanie určené skórovacím systémom za pomoci dynamického programovania.

V tejto práci budeme využívať semiglobálne zarovnanie pomocou algoritmu Needleman–Wunsch Needleman S. B. (1970) implementovaného v programe EMBOSS **?**. Ako vstup algoritmus dostáva dve sekvencie dĺžiek m a n, ktoré chceme zarovnať a hodnoty parametrov gap open (penalizácia v skóre za otvorenie medzery v zarovnaní) a gap extend (penalizácia v skóre za predĺženie medzery v zarovnaní). Algoritmus následne za pomoci dynamického programovania 2.3 vypočíta zarovnanie s najnižším skóre v čase aj priestore O(nm). Výstupom algoritmu sú zarovnané sekvencie a skóre zarovnania. V zarovnaní na určitej pozícii môžu nastať 3 prípady, a to zarovnanie dvor rovnakých reziduí (match), zarovnanie dvoch odlišných reziduí (mismatch) a nakoniec zarovnanie rezidua na medzeru (gap) vloženú do druhej sekvencie. Nami používaná implementácia algoritmu nepenalizuje za medzery v zarovnaní nachádzajúce sa na začiatku, alebo na konci zarovnania, preto je možné ňou zmysluplne zarovnať krátku štruktúru na časť oveľa dlhšej štruktúry.

Okrem globálneho poznáme aj presné lokálne zarovnanie vyriešené algoritmom Smith–Waterman Smith T. F. (1981). Tento algoritmu pracuje taktiež na princípe dynamického programovania a vyhľadáva zarovnanie dvoch subsekvencií s najlepím skóre. Používa sa na nájdenie podobných regiónoch medzi dvomi sekvenciami.

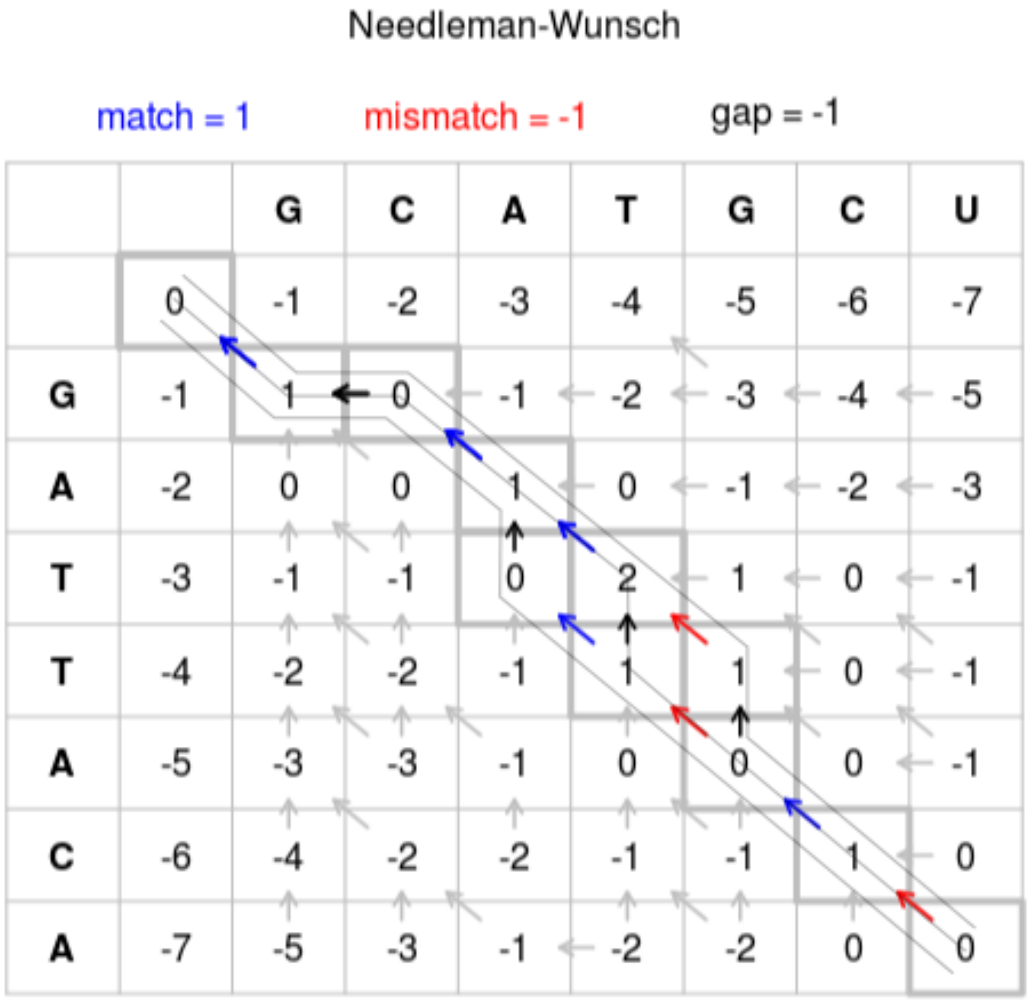
## 2.4 Homológne modelovanie

Tvrdenie zo začiatku kapitoly, ktoré hovorí, že štruktúra si zachováva podobný tvar aj napriek tomu, že jej sekvencia postupne mutuje, umožňuje zmysluplne predikovať štruktúru na základe vzoru.

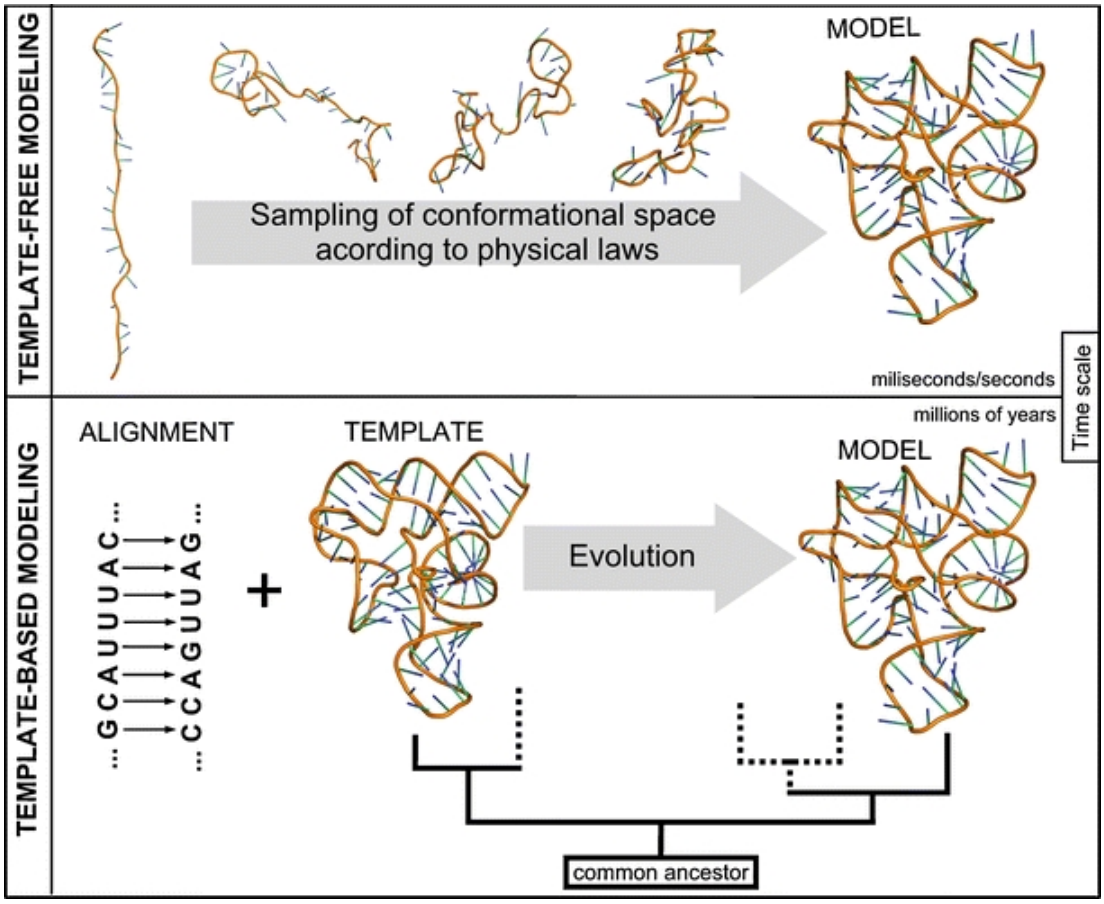
Homológne modelovanie používa na modelovanie neznámej štruktúry zo sekvencie ešte jednu vzorovú sekvenciu (template), ktorej štruktúra je známa, teda získaná za pomoci nejakej experimentálnej metódy. Predikovanú štruktúru zvykneme nazývať cieľ (target).

Prvým krokom je teda určenie vhodnej template štruktúry, pomocou ktorej budeme predikovať target štruktúru. Druhým krokom je globálne zarovnanie oboch sekvencií a získanie konzervovaných úsekov, teda úsekov v ktorých by mali byť obe štruktúry veľmi podobné. Konzervované úseky môžu byť po nejakých úpravách prenesené do cieľovej štruktúry. Z princípu vyplýva, že čím podobnejšie sekvencie budú mať target a template štruktúry, tým viac konzervovaných úsekov bude existovať a tým jednoduchšia a presnejšia by mala predikcia byť.

V treťom kroku musia byť dopredikované nekonzervované (chýbajúce úseky) cieľovej štruktúry. Existujú viacero prístupov. Jedným z nich je knižnica fragmentov, kde sa do chýbajúcej medzery v cieľovej štruktúre snažíme vhodne umiestniť




Obrázek 2.3: Needleman–Wunsch algorithm (2014) Wikipedia dostupné na https://en.wikipedia.org/wiki/Needleman–Wunsch_algorithm 27.05.2019. Príklad jedného z troch najlepších zarovnaní dvoch sekvencií:
GCATG-CU
G-ATTACA




Obrázek 2.4: Porovnanie princípu de novo a template based predikcie Rother K (2011)

fragment štruktúry z knižnice, ďalším je napríklad dopredikovanie medzery ab initio alebo de novo algoritmami.

Takto hotový model sa nakoniec môže optimalizovať použitím algoritmu na minimalizovanie voľnej energie, alebo sa riešia kolízie medzi jednotlivými nukleotidami.

Hlavnou výhodou homológneho modelovania je, že je možné ho použiť na dlhé štruktúry. Problémom môže byť vybrať správnu template štruktúru a dopredikovanie nekonzervovaných úsekov. 2.4

## 2.5 Prehľad existujúcixh nástrojov

Vďaka tomu, že počet dostupných primárnych sekvencií stále rastie rýchlejšie, ako počet experimentálne zistených terciárnych štruktúr, vzniklo mnoho nástrojov na predikciu terciárnej štruktúry RNA. 2.1

## 2.6 ModeRNA

ModeRNA je implementácia algoritmu komparatívneho modelovania RNA s ktorým sme porvnávali nami vyvinutý algoritmus. Je dostupná ako ModeRNA server a ponúka službu kompletnej predikcie submitovanej sekvencie (nájdenie

vhodného template, zarovnanie sekvencií a vytvorenie modelu terciárnej štruktúry). Okrem toho je možné stiahnuť jej zdrojové kódy (Python) a nainštalovať lokálne a používať ju zo scriptu pre automatické hromadné spracovanie.

Ako vstup ModeRNA požaduje zarovnanie template a target sekvencií spolu so súradnicami jednotlivých atómov template štruktúry. Dodané zarovnanie ModeRNA nijako nemodifikuje a od jeho kvality a zvoleného template záleží výsldná presnosť predikcie. Stručný algoritmus moderny vykonáva Rother K (2011):

- Skopírovanie zarovnaných nukleotidov.
- Substitúcia nukleotidov, ktoré boli zarovnané, na iný nukleotid.
- Modelovanie indelov vložením fragmentov štruktúr z knižnice obsahujúcej 131 316 fragmentov dĺžky 2-19 nukleotidov. ModeRNA najprv rýchlym filtrovaním podľa vzdialeností prekrývajúcich sa atómov fragmentu a template-u vyberie 50 najvhodnejších kandidátov, pokúsi sa ich vložiť do medzery a pre každého kandidáta spočíta skóre, pričom vyberie jediného s najlepším skóre a vloží ho do medzery.
- V prípade, že predikovaná štruktúra (backbone) nie je spojitá ModeRNA sa ju pokúsi opäť spojiť.

| Názov | Info | Referencia |
|---|---|---|
| MacroMolecule Builder | komparatívny modeling RNA | Flores a kol. (2011) |
| ModeRNA | komparatívna predikcia s knižnicou databázových fragmentov na predikovanie medzier | Rother a kol. (2011) |
| SimRNA | corase-grained model s Monte Carlo samplingom štruktúr | Boniecki a kol. (2015) |
| FARFAR | knowledge based de novo prediktor | Das R. (2010) |
| RNAComposer | knowledge-based automatizovaná predikcia štruktúry RNA s využitím sekundárnej štruktúry | Biesiada a kol. (2016) |
| iFoldRNA | de novo predikcia RNA založená na corase-grained model | Sharma a kol. (2008) |

Tabulka 2.1: Prehľad niektorých programov určených na predikciu RNA s informáciou o type použitého algoritmu.

# 3. Tabulky, obrázky, programy

Používání tabulek a grafů v odborném textu má některá společná pravidla a některá specifická. Tabulky a grafy neuvádíme přímo do textu, ale umístíme je buď na samostatné stránky nebo na vyhrazené místo v horní nebo dolní části běžných stránek. LaTeX se o umístění plovoucích grafů a tabulek postará automaticky.

Každý graf a tabulku očíslujeme a umístíme pod ně legendu. Legenda má popisovat obsah grafu či tabulky tak podrobně, aby jim čtenář rozuměl bez důkladného studování textu práce.

Na každou tabulku a graf musí být v textu odkaz pomocí jejich čísla. Na příslušném místě textu pak shrneme ty nejdůležitější závěry, které lze z tabulky či grafu učinit. Text by měl být čitelný a srozumitelný i bez prohlížení tabulek a grafů a tabulky a grafy by měly být srozumitelné i bez podrobné četby textu.

Na tabulky a grafy odkazujeme pokud možno nepřímo v průběhu běžného toku textu; místo *„Tabulka 3.1 ukazuje, že muži jsou v průměru o* 9,9 kg *těžší než ženy“* raději napíšeme *„Muži jsou o* 9,9 kg *těžší než ženy (viz Tabulka 3.1)“*.

## 3.1 Tabulky

U **tabulek** se doporučuje dodržovat následující pravidla:

- Vyhýbat se svislým linkám. Silnějšími vodorovnými linkami oddělit tabulku od okolního textu včetně legendy, slabšími vodorovnými linkami oddělovat záhlaví sloupců od těla tabulky a jednotlivé části tabulky mezi sebou. V LaTeXu tuto podobu tabulek implementuje balík `booktabs`. Chceme-li výrazněji oddělit některé sloupce od jiných, vložíme mezi ně větší mezeru.
- Neměnit typ, formát a význam obsahu políček v tomtéž sloupci (není dobré do téhož sloupce zapisovat tu průměr, onde procenta).
- Neopakovat tentýž obsah políček mnohokrát za sebou. Máme-li sloupec *Rozptyl*, který v prvních deseti řádcích obsahuje hodnotu 0,5 a v druhých deseti řádcích hodnotu 1,5, pak tento sloupec raději zrušíme a vyřešíme to jinak. Například můžeme tabulku rozdělit na dvě nebo do ní vložit popisné řádky, které informují o nějaké proměnné hodnotě opakující se v následujícím oddíle tabulky (např. *„Rozptyl* = 0,5 “ a níže *„Rozptyl* = 1,5 “).
- Čísla v tabulce zarovnávat na desetinnou čárku.

| **Efekt** | **Odhad** | **Směrod. chyba**[a] | **P-hodnota** |
|---|---|---|---|
| Abs. člen | −10,01 | 1,01 | — |
| Pohlaví (muž) | 9,89 | 5,98 | 0,098 |
| Výška (cm) | 0,78 | 0,12 | $< 0{,}001$ |

*Pozn:* [a] Směrodatná chyba odhadu metodou Monte Carlo.

Tabulka 3.1: Maximálně věrohodné odhady v modelu M.

- V tabulce je někdy potřebné používat zkratky, které se jinde nevyskytují. Tyto zkratky můžeme vysvětlit v legendě nebo v poznámkách pod tabulkou. Poznámky pod tabulkou můžeme využít i k podrobnějšímu vysvětlení významu některých sloupců nebo hodnot.

## 3.2 Obrázky

Několik rad týkajících se obrázků a grafů.

- Graf by měl být vytvořen ve velikosti, v níž bude použit v práci. Zmenšení příliš velkého grafu vede ke špatné čitelnosti popisků.
- Osy grafu musí být řádně popsány ve stejném jazyce, v jakém je psána práce (absenci diakritiky lze tolerovat). Kreslíme-li graf hmotnosti proti výšce, nenecháme na nich popisky `ht` a `wt`, ale osy popíšeme *Výška [cm]* a *Hmotnost [kg]*. Kreslíme-li graf funkce $h(x)$, popíšeme osy $x$ a $h(x)$. Každá osa musí mít jasně určenou škálu.
- Chceme-li na dvourozměrném grafu vyznačit velké množství bodů, dáme pozor, aby se neslily do jednolité černé tmy. Je-li bodů mnoho, zmenšíme velikost symbolu, kterým je vykreslujeme, anebo vybereme jen malou část bodů, kterou do grafu zaneseme. Grafy, které obsahují tisíce bodů, dělají problémy hlavně v elektronických dokumentech, protože výrazně zvětšují velikost souborů.
- Budeme-li práci tisknout černobíle, vyhneme se používání barev. Čáry rozlišujeme typem (plná, tečkovaná, čerchovaná,...), plochy dostatečně rozdílnými intensitami šedé nebo šrafováním. Význam jednotlivých typů čar a ploch vysvětlíme buď v textové legendě ke grafu anebo v grafické legendě, která je přímo součástí obrázku.
- Vyhýbejte se bitmapovým obrázkům o nízkém rozlišení a zejména JPEGům (zuby a kompresní artefakty nevypadají na papíře pěkně). Lepší je vytvářet obrázky vektorově a vložit do textu jako PDF.

## 3.3 Programy

Algoritmy, výpisy programů a popis interakce s programy je vhodné odlišit od ostatního textu. Jednou z možností je použití LaTeXového balíčku `fancyvrb` (fancy verbatim), pomocí něhož je v souboru `makra.tex` nadefinováno prostředí `code`. Pomocí něho lze vytvořit např. následující ukázky.

```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Menší písmo:

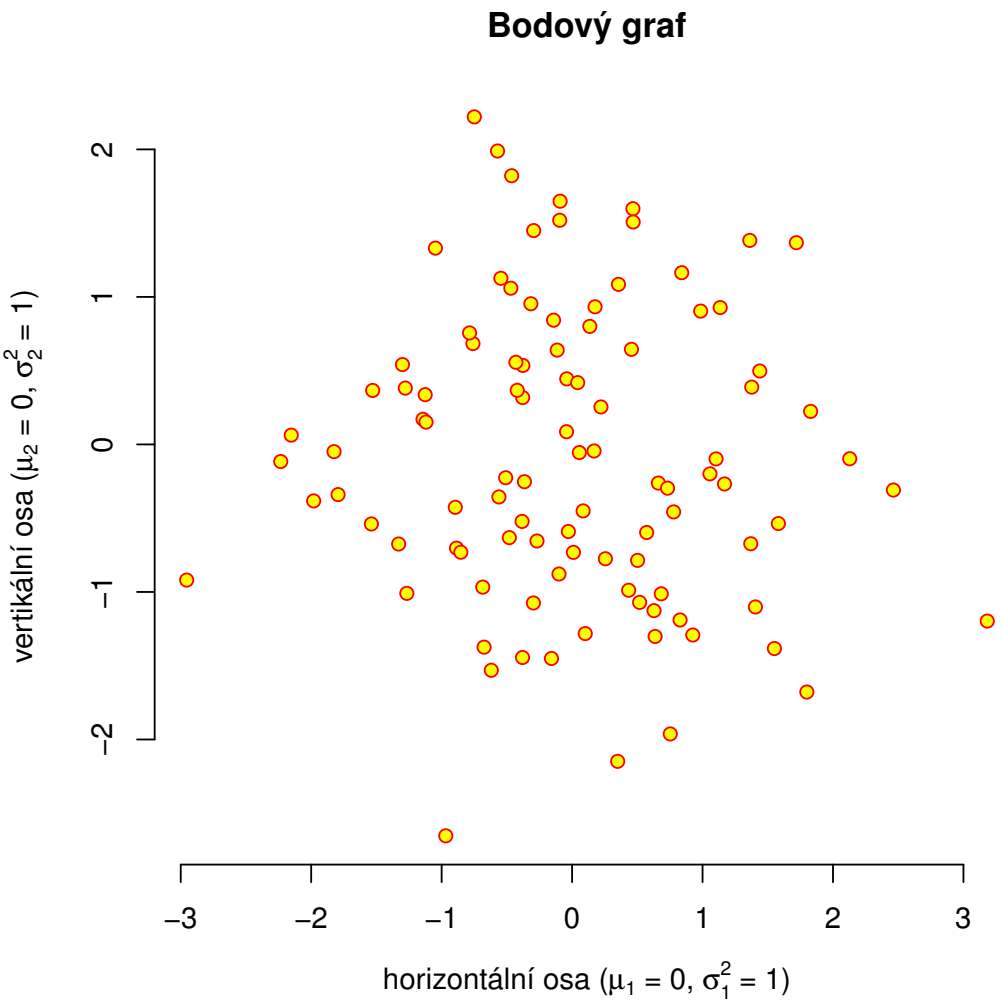
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Bez rámečku:

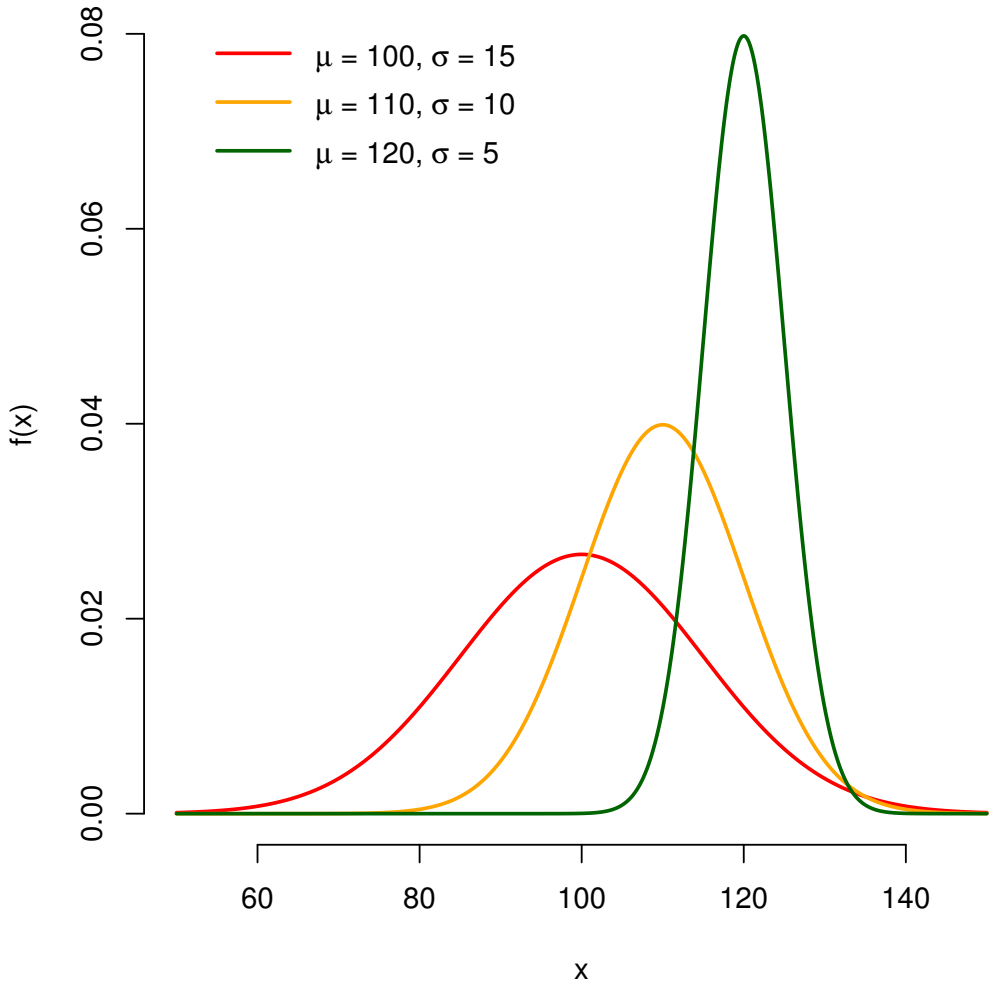
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Užší rámeček:

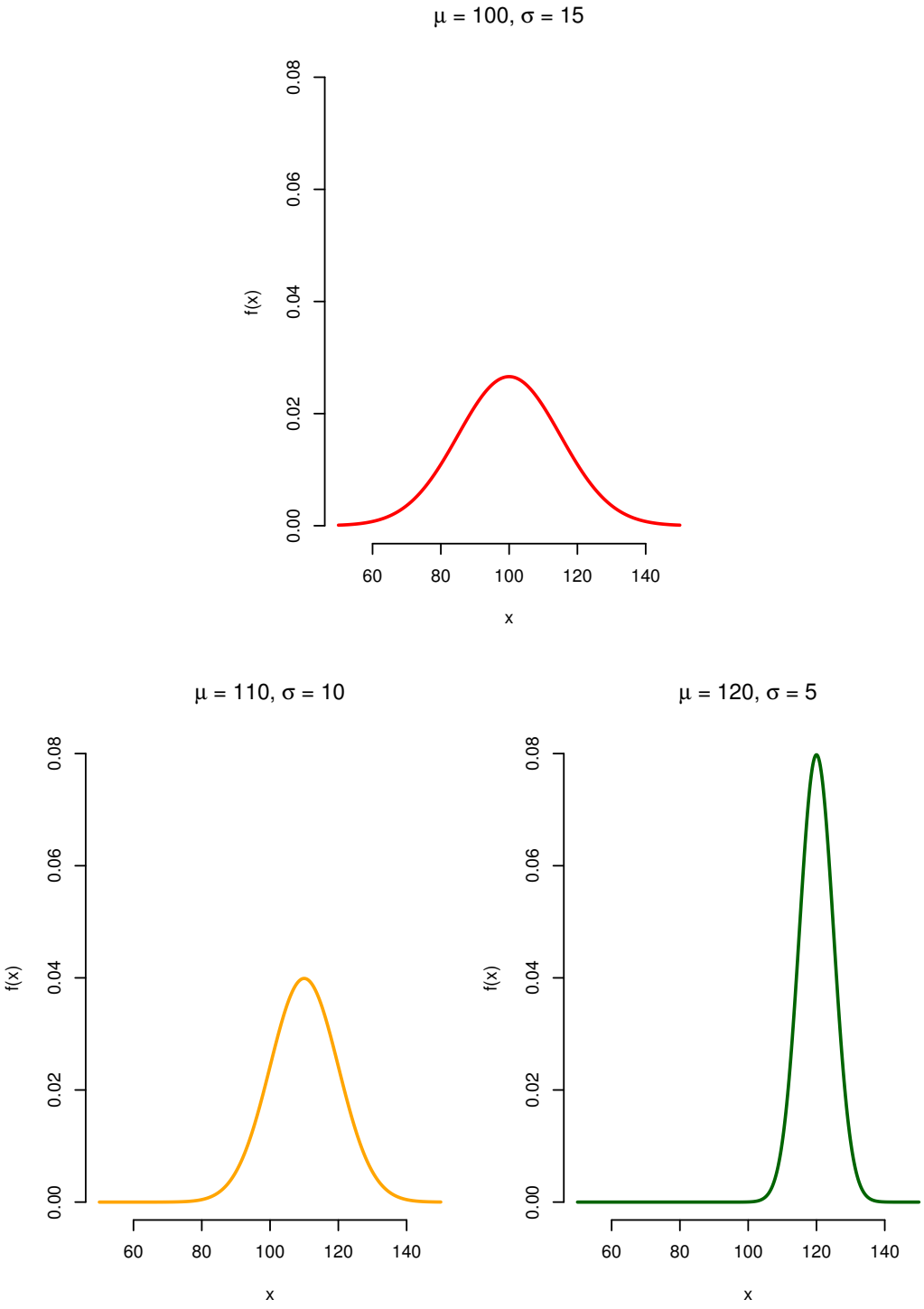
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```




Obrázek 3.1: Náhodný výběr z rozdělení $\mathcal{N}_2(\mathbf{0}, I)$.




Obrázek 3.2: Hustoty několika normálních rozdělení.




Obrázek 3.3: Hustoty několika normálních rozdělení.

# 4. Formát PDF/A

Opatření rektora č. 13/2017 určuje, že elektronická podoba závěrečných prací musí být odevzdávána ve formátu PDF/A úrovně 1a nebo 2u. To jsou profily formátu PDF určující, jaké vlastnosti PDF je povoleno používat, aby byly dokumenty vhodné k dlouhodobé archivaci a dalšímu automatickému zpracování. Dále se budeme zabývat úrovní 2u, kterou sázíme TEXem.

Mezi nejdůležitější požadavky PDF/A-2u patří:

- Všechny fonty musí být zabudovány uvnitř dokumentu. Nejsou přípustné odkazy na externí fonty (ani na „systémové“, jako je Helvetica nebo Times).
- Fonty musí obsahovat tabulku ToUnicode, která definuje převod z kódování znaků použitého uvnitř fontu to Unicode. Díky tomu je možné z dokumentu spolehlivě extrahovat text.
- Dokument musí obsahovat metadata ve formátu XMP a je-li barevný, pak také formální specifikaci barevného prostoru.

Tato šablona používá balíček `pdfx`, který umí LaTeX nastavit tak, aby požadavky PDF/A splňoval. Metadata v XMP se generují automaticky podle informací v souboru `prace.xmpdata` (na vygenerovaný soubor se můžete podívat v `pdfa.xmpi`).

Validitu PDF/A můžete zkontrolovat pomocí nástroje VeraPDF, který je k dispozici na `http://verapdf.org/`.

Pokud soubor nebude validní, mezi obvyklé příčiny patří používání méně obvyklých fontů (které se vkládají pouze v bitmapové podobě a/nebo bez unicodových tabulek) a vkládání obrázků v PDF, které samy o sobě standard PDF/A nesplňují.

Další postřehy o práci s PDF/A najdete na `http://mj.ucw.cz/vyuka/bc/pdfaq.html`.

# Závěr

# Seznam použité literatury


Alberts B, Johnson A, L. J. a. k. (2002). *Molecular Biology of the Cell*. 4th edition. Garland Science, New York. ISBN 0-8153-3218-1.

Biesiada, M., Purzycka, K. J., Szachniuk, M., Blazewicz, J. a Adamiak, R. W. (2016). *Automated RNA 3D Structure Prediction with RNAComposer*, pages 199–215. Springer New York, New York, NY. ISBN 978-1-4939-6433-8. doi: 10.1007/978-1-4939-6433-8_13. URL https://doi.org/10.1007/978-1-4939-6433-8_13.

Boniecki, M. J., Lach, G., Dawson, W. K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K. M. a Bujnicki, J. M. (2015). SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research*, **44**(7), e63–e63. ISSN 0305-1048. doi: 10.1093/nar/gkv1479. URL https://doi.org/10.1093/nar/gkv1479.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(2), 561–563.

Das R., Karanicolas J., B. D. (2010). Atomic accuracy in predicting and designing noncanonical rna structure. *Nature methods*, (7), 291–294.

Eddy, S. (2004). Genome regulation by long noncoding rnas. *Nature Biotechnology*, (22).

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

Flores, S., Sherman, M., M Bruns, C., Eastman, P. a Altman, R. (2011). Fast flexible modeling of rna structure using internal coordinates. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **8**, 1247–57. doi: 10.1109/TCBB.2010.104.

Frellsen, J., Moltke, I., Thiim, M., Mardia, K. V., Ferkinghoff-Borg, J. a Hamelryck, T. (2009). A probabilistic model of rna conformational space. *PLOS Computational Biology*, **5**(6), 1–11. doi: 10.1371/journal.pcbi.1000406. URL https://doi.org/10.1371/journal.pcbi.1000406.

Jenny Gu, P. E. B. (2009). *Structural Bioinformaticsl*. 2th edition. Wiley-Blackwell, New Jerseyl. ISBN 978-0-470-18105-8.

Krupovic, M., Blomberg, J., Coffin, J. M., Dasgupta, I., Fan, H., Geering, A. D., Gifford, R., Harrach, B., Hull, R., Johnson, W., Kreuze, J. F., Lindemann, D., Llorens, C., Lockhart, B., Mayer, J., Muller, E., Olszewski, N. E., Pappu, H. R., Pooggin, M. M., Richert-Pöggeler, K. R., Sabanadzovic, S., Sanfaçon, H., Schoelz, J. E., Seal, S., Stavolone, L., Stoye, J. P., Teycheney, P.-Y., Tristem, M., Koonin, E. V. a Kuhn, J. H. (2018). Ortervirales: New virus order unifying five families of reverse-transcribing viruses. *Journal of Virology*, **92**(12). ISSN 0022-538X. doi: 10.1128/JVI.00515-18. URL https://jvi.asm.org/content/92/12/e00515-18.

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

Needleman S. B., W. C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, (48), 443–453.

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

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. a Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, **450**(7167), 259. doi: 10.1038/nature06249. URL `https://app.dimensions.ai/details/publication/pub.1001896118andhttp://europepmc.org/articles/pmc2504711?pdf=render`. Exported from https://app.dimensions.ai on 2019/05/26.

Rother, M., Rother, K., Puton, T. a Bujnicki, J. M. (2011). RNA tertiary structure prediction with ModeRNA. *Briefings in Bioinformatics*, **12** (6), 601–613. ISSN 1477-4054. doi: 10.1093/bib/bbr050. URL `https://doi.org/10.1093/bib/bbr050`.

Rother K, Rother M, B. M. P. T. B. J. (2011). Rna and protein 3d structure modeling: similarities and differences. *Journal of Molecular Modeling*, (10), 2325–2336.

Ryu, W.-S. (2017). Chapter 2 - virus structure. In Ryu, W.-S., editor, *Molecular Virology of Human Pathogenic Viruses*, pages 21 – 29. Academic Press, Boston. ISBN 978-0-12-800838-6. doi: https://doi.org/10.1016/B978-0-12-800838-6.00002-3. URL `http://www.sciencedirect.com/science/article/pii/B9780128008386000023`.

Schudoma C., May P., N. V. W. D. (2010). Sequence-structure relationships in rna loops: establishing the basis for loop homology modeling. *Nucleic Acids Research*, (38), 970–980.

Sharma, S., Ding, F. a Dokholyan, N. V. (2008). iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**(17), 1951–1952. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn328. URL `https://doi.org/10.1093/bioinformatics/btn328`.

Smith T. F., W. M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, (147), 195–197.

# Seznam obrázků

# Seznam tabulek

# Seznam použitých zkratek

# A. Přílohy

## A.1 První příloha