

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Richard Eliáš

Vizualizace sekundární struktury RNA s využitím existujících struktur

Katedra softwarového inženýrství

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

Studijní program: Informatika

Studijní obor: Obecná informatika

Praha 2016

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

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

V dne

Podpis autora

Název práce: Vizualizace sekundární struktury RNA s využitím existujících struktur

Autor: Richard Eliáš

Katedra: Katedra softwarového inženýrství

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

Abstrakt: Abstrakt .. TODO

Klíčová slova: TODO klíčová slova

Title: RNA secondary structure visualization using existing structures

Author: Richard Eliáš

Department: Department of Software Engineering

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

Abstract: RNA secondary structure data, both experimental and predicted, are becoming increasingly available which is reflected in the increased demand for tools enabling their analysis. The common first step in the analysis of RNA molecules is visual inspection of their secondary structure. In order to correctly lay out an RNA structure, the notion of optimal layout is required. However, optimal layout of RNA structure has never been formalized and is largely habitual. To tackle this problem we propose an algorithm capable of visualizing an RNA structure using a related structure with a well-defined layout. The algorithm first converts both structures into a tree representation and then uses tree-edit distance algorithm to find out the minimum number of tree edit operations to convert one structure into the other. We couple each tree edit operation with a layout modification operation which is then used to gradually transform the known layout into the target one. The optimality of tree edit distance algorithm causes that the common motives are retained and the regions which differ in both the structures are taken care of. Visual inspection and planarity evaluation reveals that the algorithm is able to give good layouts even for relatively distant structures while keeping the layout planar. The new method is well suited for situations when one needs to visualize a structure for with a homologous structure with a good visualization is already available. ii

Keywords: RNA secondary structure, visualization, homology

Poděkování.

Obsah

Úvod	2
1 Úvod do molekulárnej genetiky a bioinformatiky	3
1.1 Co je RNA	3
1.2 Sekundárna štruktúra rRNA + konzervovanosť	3
1.2.1 Motívy	4
1.3 Reprezentácia sekundárnej štruktúry	5
2 Úvod a motivácie	6
3 Tree-edit-distance algoritmus	7
3.1 Hlavná myšlienka TED-u	7
3.2 Značenie	7
3.3 Algoritmy dynamického programovania	8
4 Návod k sazbě	9
4.1 Úprava práce	9
4.2 Jednoduché příklady	9
4.3 Matematické vzorce a výrazy	10
4.4 Definice, věty, důkazy,	11
5 Odkazy na literaturu	13
5.1 Několik ukázek	13
6 Tabulky, obrázky, programy	14
6.1 Tabulky	14
6.2 Obrázky	15
6.3 Programy	15
Závěr	20
Seznam použité literatury	21
Zoznam obrázkov	22
Zoznam tabuliek	23
Seznam použitých zkratk	24
Přílohy	25

Úvod

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

1. Uvod do molekularnej genetiky a bioinformatiky

Na zaciatku prace strucne zoznamime citatela s vybranymi pojмами z bioinformatiky i genetiky.

1.1 Co je RNA

Nositelkami genetickej informacie bunky su molekuly nukleovych kyselin tvorene retazcami nukleotidov, ktore su zakladnymi stavebnymi jednotkami nukleovych kyselin. Vyskytuje sa niekoľko variant nukleotidov (baz). U RNA su to adein (A), guanin (G), cytozin (C), uracyl (U), pri DNA sa namiesto uracylu vyskytuje tymin (T). Medzi jednotlivymi bazami existuju vazby na principe komplementarity. Vodikove vazby existuju medzi bazami A-U a C-G u RNA a podobne A-T a C-G u DNA. Strukturu nukleovych kyselin mozeme chapat podla stupna zjednodusenia

- Primarna struktura - je urcena poradim jednotlivych nukleotidov do polynukleotidoveho retazca
- Sekundarna struktura - je dana 2D priestorovym usporiadanim molekuly
- Terciarna struktura - 3D priestorove usporiadanie molekuly

DNA je dvojlaknova molekula u ktorej spojenie medzi vlaknami sa realizuje na principe komplementarity. Naopak, RNA je iba jednolaknova molekula. V snahe minimalizovat energiu molekuly, RNA komplementaritou vytvara v molekule bazove pary. Tie tvoria sekundarnu strukturu.

V praci budeme primarnou a sekundarnou strukturu mysliet prave primarnu a sekundarnu strukturu RNA, ak nebude povedane inak.

Az donedavna sa myslelo, ze funkcia RNA je iba pri tvorbe bielkovin (mRNA), alebo ako transporter aminokyselin (tRNA). Avsak existuje mnoho dalsich, od relativne malych molekul tvorených desiatkami baz, ktore pomáhajú pri expresii genov (miRNA, siRNA, tmRNA a dalsie), az po velke, tvorene tisickami nukleotidov (rRNA).

Definice 1. *Nech Σ je abeceda $\{A, C, G, U\}$. Potom slovo $W \in \Sigma^n$ nad touto abecedou je sekvencia nukleotidov (baz) RNA.*

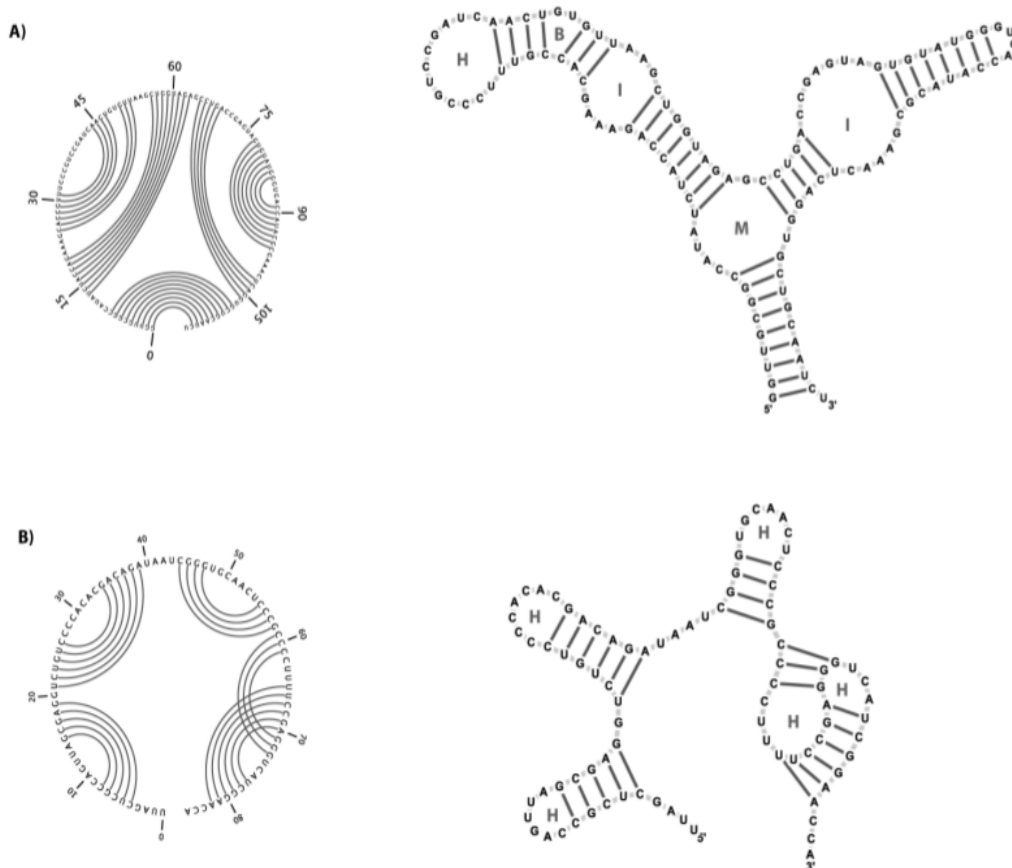
1.2 Sekundarna struktura rRNA + konzervovanost

Ako hlavny objekt zaujmu sme si spomedzi RNA vybrali ribozomalnu. Je to hlavne kvoli jej velkosti a konzervovanosti. Konzervovanostou myslime to, ze napriec celým spektrom organizmov sa sekundarna struktura rRNA velmi nemeni.

Definice 2. *Nech W je sekvencia podľa definície 1 dĺžky n . Sekundárnou štruktúrou označíme množinu \mathbb{S} parov (i, j) takých, že pre dva pary (i, j) a $(k, l) \in \mathbb{S}$ ($BUNO\ i \leq k$) platí jedno z nasledujúcich:*

- $i = k \iff j = l$
- $i < j < k < l$, cize $par(i, j)$ predchadza $par(k, l)$
- $i < k < l < j$, cize $par(i, j)$ obsahuje $par(k, l)$

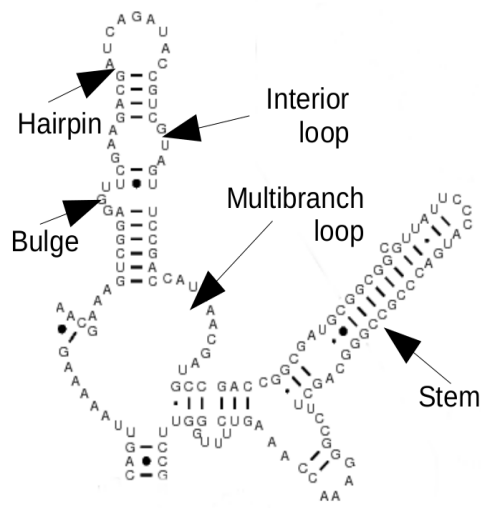
Prva podmienka zabezpečuje, že nukleotid je najviac v jednom bazickom pare, druhá a tretia hovoria o usporiadaní parov, buď sú na sebe nezávislé alebo na seba nadväzujú. Posledná podmienka zakazuje existenciu pseudouzlov (pseudoknots).



Obr. 1.1: Circular Feynman - kruhova reprezentacia sekundarnej struktury

1.2.1 Motivý

Na obrázkoch môžeme pozorovať motívy RNA molekúl, stem/loop, s ďalším možným delením loopov na bulge, interior loop a multibranch loop. V ďalšom rozprávaní nám bude stačiť rozdelenie na stem a loop.



Obr. 1.2: Strukturalne motivy v RNA

1.3 Reprezentacia sekundarnej struktury

Definicia 2 nam ponuka reprezentovat sekundarnu strukturu ako usporiadany les.

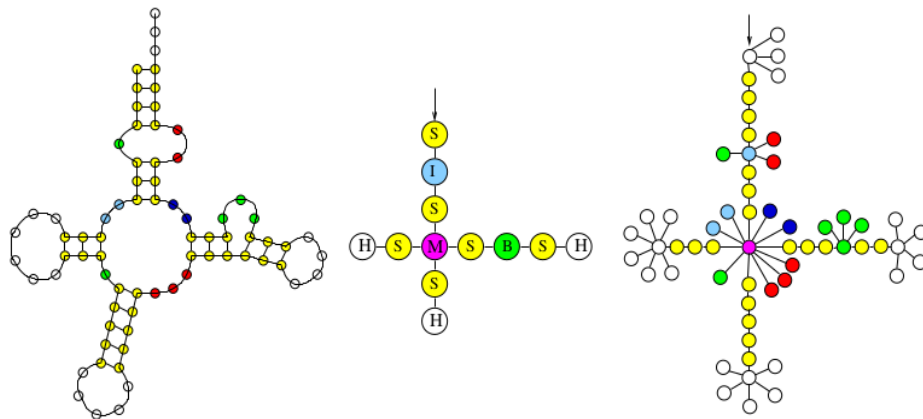


Figure 3: A secondary structure and its tree representations.

Obr. 1.3: Varianty reprezentacie vrcholov

Definice 3. *Usporiadany zakoreneny strom je orientovany graf, v ktorom plati, ze hrany su orientovane vzdy v smere z predka na potomka. Okrem korena ma kazdy vrchol svojho predka. Existuje tu dalej usporiadanie medzi potomkami. Usporiadany les je usporiadana mnozina stromov.*

Kazdy vrchol moze reprezentovat napríklad motiv v strukture RNA, alebo nukleotid, bazovy par, ...

2. Uvod a motivace

V predchadzajucej kapitole sme definovali vybrane casti genetiky, bioinformatiky, teorie grafov.

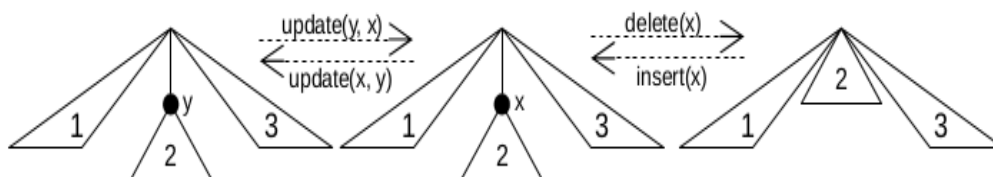
Cielom prace je vytvorit program na vizualizaciu sekundarnej struktury RNA pomocou uz existujucej vizualizacie inej molekuly.

3. Tree-edit-distance algoritmus

Jadro aplikacie lezi v pouziti tree-edit-distance (TED) algoritmu, vďaka ktorému dostaneme mapovanie medzi 2 RNA stromami. Mapovanie nam ukáže spoločne časti oboch RNA stromov. TED algoritmus je obdobou Levenstheinovoho string-edit-distance algoritmu. Problém u retazcov je špeciálnym prípadom TED-u, kedy stromy zdegenerovali na cesty (spojový zoznam).

3.1 Hlavná myšlienka TED-u

Zaklad TED algoritmu je v rekurzivnom vzorci 3.2 z Demaine a kol. (2009) a Pawlik a Augsten (2011). Vzdialenosť medzi lesmi F a G , $\delta(F, G)$ je definovaná ako minimálny počet editačných operácií, ktoré z F urobia G . Používame štandardné editačné operácie - delete, insert, update.



Obr. 3.1: Ukazky TED operácií

Delete, zmazanie vrcholu, znamená pripojiť k predkovi všetkých jeho potomkov so zachovaním poradia medzi nimi. Insert, vloženie vrcholu, je opačná operácia k delete, čo znamená, že vkladáme vrchol medzi rodiča nejakých jeho, po sebe nasledujúcich potomkov. Update iba zmení hodnotu vo vrchole stromu.

3.2 Znamenie

V tejto kapitole sa budeme riadiť znamením Pawlik a Augsten (2011). Teda, používame definíciu stromu a lesa z 3. Ak F je les (strom), N_F označuje množinu jeho vrcholov a E_F množinu jeho hran. Platí ďalej že $E_F \subseteq N_F \times N_F$. \emptyset označuje prázdny strom, resp. prázdny les. Podľa lesa F je graf \tilde{F} s vrcholmi $N_{\tilde{F}} \subseteq N_F$ a hranami $E_{\tilde{F}} \subseteq E_F \cap N_{\tilde{F}} \times N_{\tilde{F}}$. Obdobne to platí aj pre podstrom stromu T . F_v označuje podstrom F zakorenený vo v , t.j. v strome ostávajú iba potomkovia v . $F - v$ budeme znčiť les, ktorý dostaneme zmazaním vrcholu v z F , spolu so všetkými hranami zasahujúcimi do v . Podobne $F - F_v$ budeme znčiť les, ktorý dostaneme zmazaním podstromu F_v z F .

Definícia 4 (Editačná vzdialenosť). *Nech F a G sú dva lesy. Editácia vzdialenosť, tree-edit-distance - $\delta(F, G)$, medzi F a G je rovná minimálnej cene, za ktorú les F transformujeme na G .*

Vo vzorci 3.2 počítame editačnú vzdialenosť $\delta(F, G)$, c_{del} , c_{ins} a c_{upd} sú ceny zmazania, vloženia a editácie vrcholu v strome a r_F a r_G sú korene, buď obidva

najpravejsie alebo najlavejsie (tzn. vyberieme najpravejsi/najlavejsi strom lesa a jeho koren).

$$\begin{aligned}\delta(\emptyset, \emptyset) &= 0 \\ \delta(F, \emptyset) &= \delta(F - r_F, \emptyset) + c_{del}(r_F) \\ \delta(\emptyset, G) &= \delta(\emptyset, G - r_G) + c_{ins}(r_G)\end{aligned}\tag{3.1a}$$

$$\delta(F, G) = \begin{cases} \delta(F - r_F, G) + c_{del}(r_F) \\ \delta(F, G - r_G) + c_{ins}(r_G) \\ \delta(F - r_F, G - r_G) + \\ \delta(r_F - r_F, r_G - r_G) + c_{upd}(r_F, r_G) \end{cases}\tag{3.1b}$$

Obr. 3.2: Rekurzivny vzorec pre vypocet tree-edit-distance

3.3 Algoritmy dynamickeho programovania

Kazdy krok rekurzivie 3.2 je vypocitana v konstantnom case z inych podproblemov, takže doba behu algoritmu zavisí na pocte podproblemov, ktore sa budu pocitat. Rekurgia medzi lesmi F a G vedie k casovej zlozitosti $\mathcal{O}(n^2 \cdot m^2)$ a priestorovej zlozitosti $\mathcal{O}(n \cdot m)$ kde $n = N_F$ a $m = N_G$.

Definice 5 (Dekompozicna strategia).

Zvysovaniu efektivity algoritmu sa venuju Zhang a Shasha (1989), Demaine a kol. (2009) a Pawlik a Augsten (2011). Posledny menovany vyuziva

Definice 6. Celkova dekompozicia lesa (full decomposition) F , $\mathcal{A}(F)$ je mnozina vsetkych podlesov F , ktore dostaneme rekurzivnym odstranenim najlavejsieho alebo najpravejsieho korenoveho vrcholu - $r_R(F)$ a $r_L(F)$ z F a nasledne aj vsetkych jeho podlesov.

$$\begin{aligned}\mathcal{A}(\emptyset) &= \emptyset \\ \mathcal{A}(F) &= F \cup \mathcal{A}(F - r_L(F)) \cup \mathcal{A}(F - r_R(F))\end{aligned}$$

4. Nápořěda k sazřě

4.1 Űprava práce

Vlastní text bakalářské práce je uspořádaný hierarchicky do kapitol a podkapitol, každá kapitola začíná na nové straně. Text je zarovnán do bloku. Nový odstavec se obvykle odděluje malou vertikální mezerou a odsazením prvního řádku. Grafická úprava má být v celém textu jednotná.

Práce se tiskne na bílý papír formátu A4. Okraje musí ponechat dost místa na vazbu: doporučen je horní, dolní a pravý okraj 25 mm, levý okraj 40 mm. Číslojí se všechny strany kromě obálky a informačních stran na začátku práce; první číslovaná strana bývá obvykle ta s obsahem.

Písmo se doporučuje dvanáctibodové (12 pt) se standardní vzdáleností mezi řádky (pokud píšete ve Wordu nebo podobném programu, odpovídá tomu řádkování 1,5; v \TeX u není potřeba nic přepínat). Pro běžný text používejte vzpřímené patkové písmo. Text matematických vět se obvykle tiskne pro zdůraznění skloněným (slanted) písmem, není-li k dispozici, může být zastoupeno kurzívou.

Primárně je doporučován jednostranný tisk (příliš tenkou práci lze obtížně svázat). Delší práce je lepší tisknout oboustranně a přizpůsobit tomu velikosti okrajů: 40 mm má vždy *vnitřní* okraj. Rub titulního listu zůstává nepotištěný.

Zkratky použité v textu musí být vysvětleny vždy u prvního výskytu zkratky (v závorce nebo v poznámce pod čarou, jde-li o složitější vysvětlení pojmu či zkratky). Pokud je zkratek více, připojuje se seznam použitých zkratek, včetně jejich vysvětlení a/nebo odkazů na definici.

Delší převzatý text jiného autora je nutné vymežit uvozovkami nebo jinak vyznačit a řádně citovat.

4.2 Jednoduché příklady

Číslo v českém textu obvykle sázíme v matematickém režimu s desetinnou čárkou: $\pi \doteq 3,141\,592\,653\,589$. V matematických textech se považuje za přípustné používat desetinnou tečku (pro lepší odlišení od čárky v roli oddělovače). Numerické výsledky se uvádějí s přiměřeným počtem desetinných míst.

Mezi číslo a jednotku patří úzká mezera: šířka stránky A4 činí 210 mm, což si pamatuje pouze 5 % autorů. Pokud ale údaj slouží jako přívlastek, mezeru vynecháváme: 25mm okraj, 95% interval spolehlivosti.

Rozlišujeme různé druhy pomlček: červeno-černý (krátká pomlčka), strana 16–22 (střední), 45 – 44 (matematické minus), a toto je — jak se asi dalo čekat — vložená věta ohraňčená dlouhými pomlčkami.

V českém textu se používají „české“ uvozovky, nikoliv “anglické”.

Na některých místech je potřeba zabránit lámání řádku (v \TeX u značíme vlnovkou): u~předložek (neslabičných, nebo obecně jednopísmenných), vrchol~*v*, před *k*~kroky, a~proto, . . . obecně kdekoliv, kde by při rozlomení čtenář „škobrt-nul“.

4.3 Matematické vzorce a výrazy

Proměnné sázíme kurzívou (to \TeX v matematickém módu dělá sám, ale nezapomínejte na to v okolním textu a také si matematický mód zapněte). Názvy funkcí sázíme vzpřímeně. Tedy například: $\text{var}(X) = \text{E } X^2 - (\text{E } X)^2$.

Zlomky uvnitř odstavce (třeba $\frac{5}{7}$ nebo $\frac{x+y}{2}$) mohou být příliš stísněné, takže je lepší sázet jednoduché zlomky s lomítkem: $5/7$, $(x+y)/2$.

Nechť

$$\mathbb{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}.$$

Povšimněme si tečky za maticí. Byť je matematický text vysázen ve specifickém prostředí, stále je gramaticky součástí věty a tudíž je zapotřebí neopomenout patřičná interpunkční znaménka. Výrazy, na které chceme později odkazovat, je vhodné očíslovat:

$$\mathbb{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}. \quad (4.1)$$

Výraz (4.1) definuje matici \mathbb{X} . Pro lepší čitelnost a přehlednost textu je vhodné číslovat pouze ty výrazy, na které se autor někde v další části textu odkazuje. To jest, nečíslujte automaticky všechny výrazy vysázené některým z matematických prostředí.

Zarovnání vzorců do několika sloupečků:

$$\begin{aligned} S(t) &= \text{P}(T > t), & t > 0 & \quad (\text{zprava spojitá}), \\ F(t) &= \text{P}(T \leq t), & t > 0 & \quad (\text{zprava spojitá}). \end{aligned}$$

Dva vzorce se spojovníkem:

$$\left. \begin{aligned} S(t) &= \text{P}(T > t) \\ F(t) &= \text{P}(T \leq t) \end{aligned} \right\} \quad t > 0 \quad (\text{zprava spojité}). \quad (4.2)$$

Dva centrované nečíslované vzorce:

$$\begin{aligned} \mathbf{Y} &= \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ \mathbb{X} &= \begin{pmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{pmatrix}. \end{aligned}$$

Dva centrované číslované vzorce:

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4.3)$$

$$\mathbb{X} = \begin{pmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{pmatrix}. \quad (4.4)$$

Definice rozdělená na dva případy:

$$P_{r-j} = \begin{cases} 0, & \text{je-li } r-j \text{ liché,} \\ r! (-1)^{(r-j)/2}, & \text{je-li } r-j \text{ sudé.} \end{cases}$$

Všimněte si použití interpunkce v této konstrukci. Čárky a tečky se dávají na místa, kam podle jazykových pravidel patří.

$$\begin{aligned} x &= y_1 - y_2 + y_3 - y_5 + y_8 - \cdots = && \text{z (4.3)} \\ &= y' \circ y^* = && \text{podle (4.4)} \\ &= y(0)y' && \text{z Axiomu 1.} \end{aligned} \quad (4.5)$$

Dva zarovnané vzorce nečíslované:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^n f_i(y_i; \boldsymbol{\theta}), \\ \ell(\boldsymbol{\theta}) &= \log\{L(\boldsymbol{\theta})\} = \sum_{i=1}^n \log\{f_i(y_i; \boldsymbol{\theta})\}. \end{aligned}$$

Dva zarovnané vzorce, první číslovaný:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^n f_i(y_i; \boldsymbol{\theta}), && (4.6) \\ \ell(\boldsymbol{\theta}) &= \log\{L(\boldsymbol{\theta})\} = \sum_{i=1}^n \log\{f_i(y_i; \boldsymbol{\theta})\}. \end{aligned}$$

Vzorec na dva řádky, první řádek zarovnaný vlevo, druhý vpravo, nečíslovaný:

$$\begin{aligned} \ell(\mu, \sigma^2) &= \log\{L(\mu, \sigma^2)\} = \sum_{i=1}^n \log\{f_i(y_i; \mu, \sigma^2)\} = \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2. \end{aligned}$$

Vzorec na dva řádky, zarovnaný na =, číslovaný uprostřed:

$$\begin{aligned} \ell(\mu, \sigma^2) &= \log\{L(\mu, \sigma^2)\} = \sum_{i=1}^n \log\{f(y_i; \mu, \sigma^2)\} = \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2. \end{aligned} \quad (4.7)$$

4.4 Definice, věty, důkazy, . . .

Konstrukce typu definice, věta, důkaz, příklad, . . . je vhodné odlišit od okolního textu a případně též číslovat s možností použití křížových odkazů. Pro každý typ těchto konstrukcí je vhodné mít v souboru s makry (`makra.tex`) nadefinované jedno prostředí, které zajistí jak vizuální odlišení od okolního textu, tak automatické číslování s možností křížově odkazovat.

Definice 7. *Nechť náhodné veličiny X_1, \dots, X_n jsou definovány na témž pravděpodobnostním prostoru (Ω, \mathcal{A}, P) . Pak vektor $\mathbf{X} = (X_1, \dots, X_n)^\top$ nazveme náhodným vektorem.*

Definice 8 (náhodný vektor). *Nechť náhodné veličiny X_1, \dots, X_n jsou definovány na témž pravděpodobnostním prostoru (Ω, \mathcal{A}, P) . Pak vektor $\mathbf{X} = (X_1, \dots, X_n)^\top$ nazveme náhodným vektorem.*

Definice 7 ukazuje použití prostředí pro sazbu definice bez titulku, definice 8 ukazuje použití prostředí pro sazbu definice s titulkem.

Věta 1. *Náhodný vektor \mathbf{X} je měřitelné zobrazení prostoru (Ω, \mathcal{A}, P) do $(\mathbb{R}_n, \mathcal{B}_n)$.*

Lemma 2 (Anděl, 2007, str. 29). *Náhodný vektor \mathbf{X} je měřitelné zobrazení prostoru (Ω, \mathcal{A}, P) do $(\mathbb{R}_n, \mathcal{B}_n)$.*

Důkaz. Jednotlivé kroky důkazu jsou podrobně popsány v práci Anděl (2007, str. 29). □

Věta 1 ukazuje použití prostředí pro sazbu matematické věty bez titulku, lemma 2 ukazuje použití prostředí pro sazbu matematické věty s titulkem. Lemmata byla zavedena v hlavním souboru tak, že sdílejí číslování s větami.

5. Odkazy na literaturu

Odkazy na literaturu vytváříme nejlépe pomocí příkazů `\citet`, `\citep` atp. (viz L^AT_EXový balíček `natbib`) a následného použití BibT_EXu. V matematickém textu obvykle odkazujeme stylem „Jméno autora/autorů (rok vydání)“, resp. „Jméno autora/autorů [číslo odkazu]“. V českém/slovenském textu je potřeba se navíc vypořádat s nutností skloňovat jméno autora, respektive přechylovat jméno autorky. Je potřeba mít na paměti, že standardní příkazy `\citet`, `\citep` produkují referenci se jménem autora/autorů v prvním pádě a jména autorek jsou nepřechýlena.

Pokud nepoužíváme bibT_EX, řídíme se normou ISO 690 a zvyklostmi oboru. Jména časopisů lze uvádět zkráceně, ale pouze v kodifikované podobě.

5.1 Několik ukázek

Mezi nejvíce citované statistické články patří práce Kaplana a Meiera a Coxe (Kaplan a Meier, 1958; Cox, 1972). Student (1908) napsal článek o t-testu.

Prof. Anděl je autorem učebnice matematické statistiky (viz Anděl, 1998). Teorii odhadu se věnuje práce Lehmann a Casella (1998). V případě odkazů na specifickou informaci (definice, důkaz, ...) uvedenou v knize bývá užitečné uvést specificky číslo kapitoly, číslo věty atp. obsahující požadovanou informaci, např. viz Anděl (2007, Věta 4.22) nebo (viz Anděl, 2007, Věta 4.22).

Mnoho článků je výsledkem spolupráce celé řady osob. Při odkazování v textu na článek se třemi autory obvykle při prvním výskytu uvedeme plný seznam: Dempster, Laird a Rubin (1977) představili koncept EM algoritmu. Respektive: Koncept EM algoritmu byl představen v práci Dempstera, Lairdové a Rubina (Dempster, Laird a Rubin, 1977). Při každém dalším výskytu již používáme zkrácenou verzi: Dempster a kol. (1977) nabízejí též několik příkladů použití EM algoritmu. Respektive: Několik příkladů použití EM algoritmu lze nalézt též v práci Dempstera a kol. (Dempster a kol., 1977).

U článku s více než třemi autory odkazujeme vždy zkrácenou formou: První výsledky projektu ACCEPT jsou uvedeny v práci Genbergové a kol. (Genberg a kol., 2008). V textu *nenapíšeme*: První výsledky projektu ACCEPT jsou uvedeny v práci Genberg, Kulich, Kawichai, Modiba, Chingono, Kilonzo, Richter, Pettifor, Sweat a Celentano (2008).

6. 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. L^AT_EX se o umístění plovoucích grafů a tabulek postará automaticky.

Každý graf a tabulku očíslovujeme 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 nepřímou v průběhu běžného toku textu; místo „*Tabulka 6.1 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 6.1)*“.

6.1 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 L^AT_EXu 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*“).

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 6.1: Maximálně věrohodné odhady v modelu M.

- Čísla v tabulce zarovnávat na desetinnou čárku.
- 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.

6.2 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 jednolitě černé tmy. Je-li bodů mnoho, zmenšíme velikost symbolu, kterým je vykresluje, 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 intenzitami 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.

6.3 Programy

Algoritmy, výpisy programů a popis interakce s programy je vhodné odlišit od ostatního textu. Jednou z možností je použití \LaTeX ového 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:

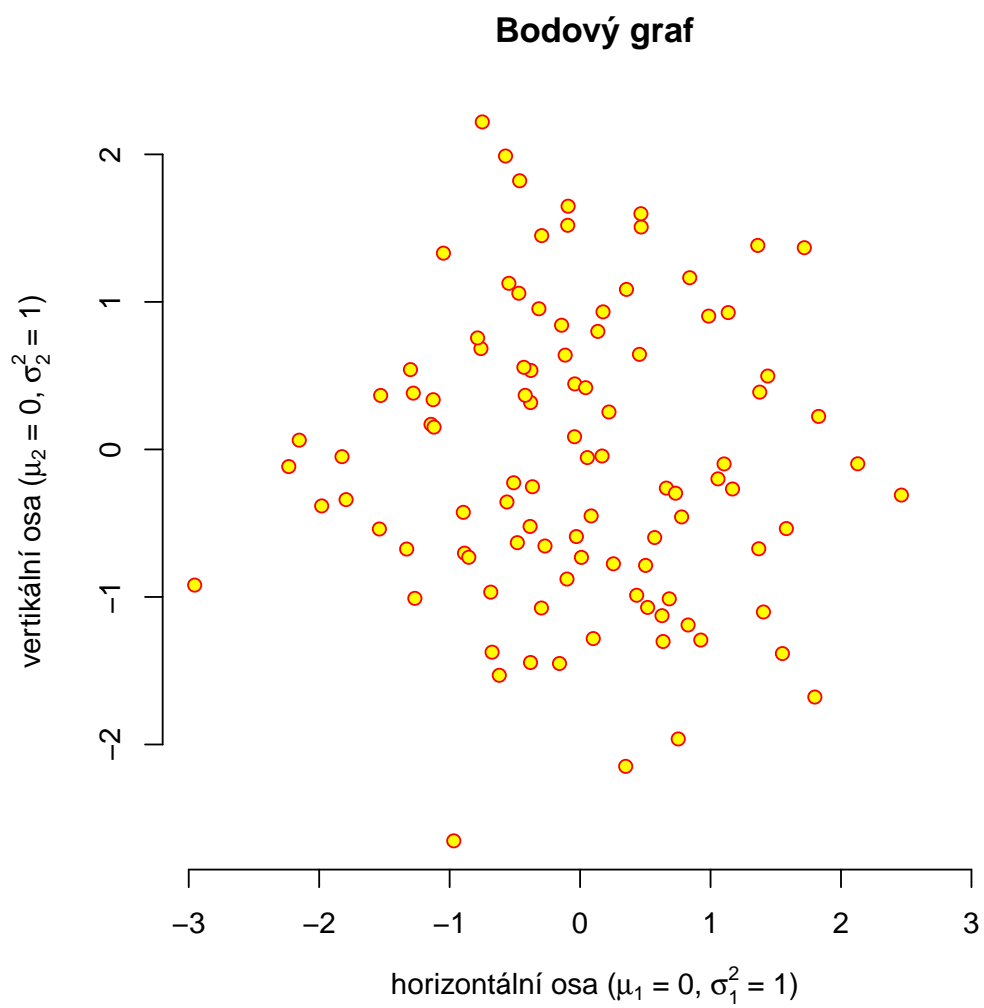
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Bez rámečku:

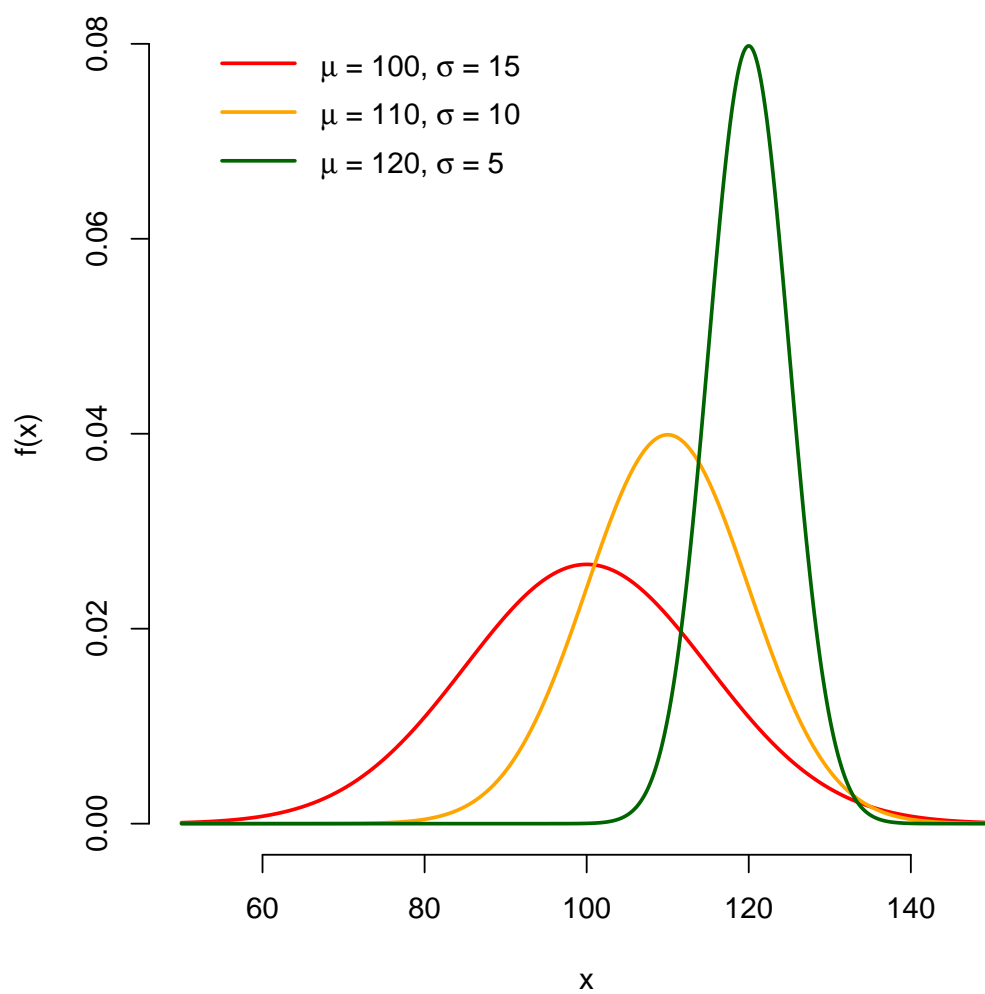
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```

Užší rámeček:

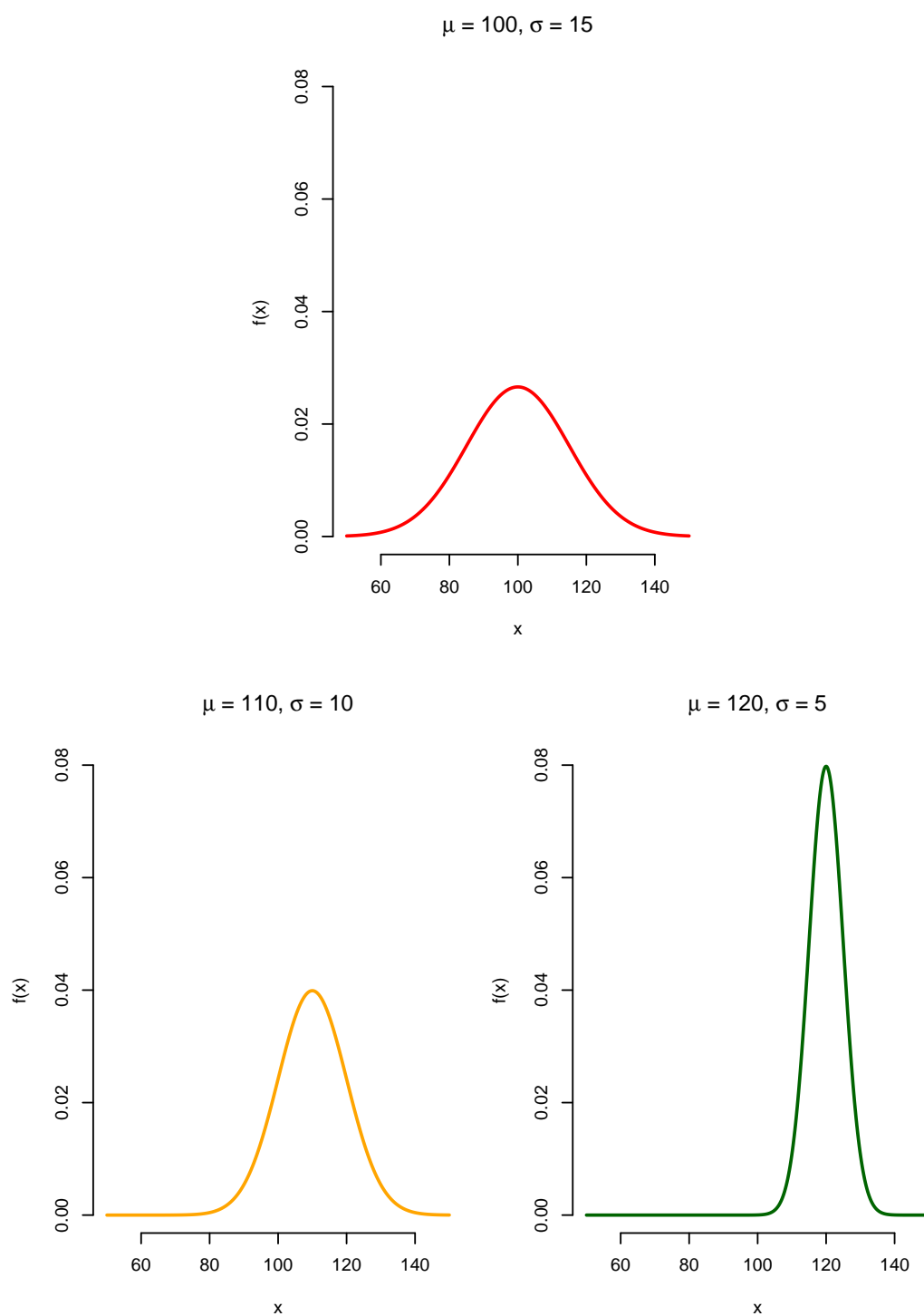
```
> mean(x)
[1] 158.90
> objekt$prumer
[1] 158.90
```



Obr. 6.1: Náhodný výběr z rozdělení $\mathcal{N}_2(\mathbf{0}, I)$.



Obr. 6.2: Hustoty několika normálních rozdělení.



Obr. 6.3: Hustoty několika normálních rozdělení.

Závěr

Seznam použité literatury

- ANDĚL, J. (1998). *Statistické metody*. Druhé přepracované vydání. Matfyzpress, Praha. ISBN 80-85863-27-8.
- ANDĚL, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.
- COX, D. R. (1972). Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, **34**(2), 187–220.
- DEMAINE, E. D., MOZES, S., ROSSMAN, B. a WEIMANN, O. (2009). An optimal decomposition algorithm for tree edit distance. *ACM Transactions on Algorithms*, **6**(1), 2:1–2:19. ISSN 1549-6325. doi: 10.1145/1644015.1644017. URL <http://doi.acm.org/10.1145/1644015.1644017>.
- DEMPSTER, A. P., LAIRD, N. M. a RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38.
- GENBERG, B. L., KULICH, M., KAWICHAJ, S., MODIBA, P., CHINGONO, A., KILONZO, G. P., RICHTER, L., PETTIFOR, A., SWEAT, M. a CELENTANO, D. D. (2008). HIV risk behaviors in sub-Saharan Africa and Northern Thailand: Baseline behavioral data from project Accept. *Journal of Acquired Immune Deficiency Syndrome*, **49**, 309–319.
- KAPLAN, E. L. a MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.
- LEHMANN, E. L. a CASELLA, G. (1998). *Theory of Point Estimation*. Second Edition. Springer-Verlag, New York. ISBN 0-387-98502-6.
- PAWLIK, M. a AUGSTEN, N. (2011). RTED: A robust algorithm for the tree edit distance. *PVLDB*, **5**(4), 334 – 345.
- STUDENT (1908). On the probable error of the mean. *Biometrika*, **6**, 1–25.
- ZHANG, K. a SHASHA, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, **18**(6), 1245 – 1262.

Zoznam obrázkov

1.1	Circular Feynman - kruhova reprezentacia sekundarnej struktury .	4
1.2	Strukturalne motivy v RNA	5
1.3	Varianty reprezentacie vrcholov	5
3.1	Ukazky TED operacii	7
3.2	Rekurzivny vzorec pre vypocet tree-edit-distance	8
6.1	Náhodný výběr z rozdělení $\mathcal{N}_2(\mathbf{0}, I)$	17
6.2	Hustoty několika normálních rozdělení.	18
6.3	Hustoty několika normálních rozdělení.	19

Zoznam tabuliek

6.1	Maximálně věrohodné odhady v modelu M.	14
-----	--	----

Seznam použitých zkratek

Přílohy