Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Identifikácia variantov v dátach nanopórového sekvenovania

Bakalárska práca

2018
Eduard Batmendijn

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Identifikácia variantov v dátach nanopórového sekvenovania

Bakalárska práca

Študijný program:	Informatika
Študijný odbor:	2508 Informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	doc. Mgr. Tomáš Vinař, PhD.

Bratislava, 2018
Eduard Batmendijn

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Eduard Batmendijn
Študijný program: informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Identifikácia variantov v dátach nanopórového sekvenovania
Variant Identification in Nanopore Sequencing Data

Anotácia: Nanopórové sekvenovanie produkuje sekvencie signálov, ktoré je možné porovnať so známou referenčnou DNA sekvenciou. Cieľom práce je navrhnúť a naimplementovať metódy, ktoré by umožnili identifikovať miesta, kde sa prečítané sekvencie signálov odlišujú od referencie. Efektívne metódy zohľadňujúce špecifické vlastnosti dát umožnia spoľahlivú identifikáciu takýchto variantov aj z dát s nízkym pokrytím.

Vedúci: doc. Mgr. Tomáš Vinař, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 31.10.2017

Dátum schválenia: 31.10.2017 doc. RNDr. Daniel Olejár, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: Na tomto mieste by som sa rád poďakoval svojmu školiteľovi Tomášovi Vinařovi za jeho vedenie, ochotu pomôcť a trpezlivosť s mojím prístupom k povinnostiam.

Abstrakt

V niektorých aplikáciách sekvenovania DNA je potrebné zistiť, ako sa sekvenovaná DNA líši od referenčnej DNA postupnosti. V našej práci navrhujeme pravdepodobnostný model, pomocou ktorého na základe referenčnej DNA a signálu zo sekvenátora MinION identifikujeme bázy, ktoré sa v sekvenovanej DNA líšia od referencie. Pri testovaní v ideálnych podmienkach dáva tento model presnejšie výsledky, než postup, ktorý najprv zo signálu štandardným spôsobom určí postupnosť báz a potom túto postupnosť porovná s referenciou.

Kľúčové slová: sekvenovanie DNA, varianty, MinION

Abstract

Some applications of DNA sequencing require us to determine the differences between the sequenced sample and the reference sequence. To address this problem, we propose a new approach based on a probabilistic model that can use the signal from MinION sequencer to identify such differences. Our model outperforms the standard approach of basecalling the signal and comparing the result to the reference, when evaluated under ideal setting.

Keywords: DNA sequencing, variants, MinION

Obsah

Úvod

DNA nesie genetickú informáciu zakódovanú v postupnosti štyroch dusíkatých báz: adenínu, cytozínu, guanínu a tymínu. Od sedemdesiatych rokov dvadsiateho storočia vznikajú stále nové techniky na sekvenovanie DNA, čo znamená pre fyzickú vzorku DNA zistiť jej postupnosť báz.

Nanopórové sekvenovanie je technika sekvenovania DNA, pri ktorej jedno vlákno DNA prechádza veľmi malým otvorom (nanopórom) a ovplyvňuje pritom elektrický prúd prechádzajúci týmto otvorom. Na základe merania zmien tohto prúdu možno určiť, aké bázy prechádzali nanopórom. Nanopórové sekvenovanie má oproti iným technikám sekvenovania niekoľko výhod, umožňuje napríklad pomerne lacno a rýchlo sekvenovať DNA už v malých množstvách. Jeho veľkou nevýhodou je však veľké množstvo chýb, ktorých sa pri sekvenovaní dopúšťa.

V niektorých aplikáciách sekvenujeme DNA, u ktorej je dopredu známa postupnosť jej báz, až na drobné odlišnosti, ktoré nazývame varianty. Cieľom sekvenovania je vtedy nájdenie týchto variantov. Pri použití nanopórového sekvenovania na hľadanie variantov narazíme na problém, že veľa variantov, ktoré nájdeme, sú v skutočnosti iba nepresnosti v sekvenovaní. V tejto práci sa snažíme nájsť spôsob, ako využiť dotatočnú informáciu o sekvenovanej DNA (t. j. to, že sa podobá na známu postupnosť) už vo fáze prevádzania nameraného elektrického prúdu na postupnosť báz a zvýšiť tým presnosť identifikácie variantov.

V kapitole 1 detailnejšie popíšeme proces sekvenovania DNA konkrétnym nanopórovým sekvenátorom MinION. V kapitole 2 presnejšie zadefinujeme cieľ našej práce. V kapitole 3 navrhneme spôsob, ako identifikovať najjednoduchší druh variantov – jednonukleotidové polymorfizmy. V kapitole 4 tento spôsob otestujeme na reálnych dátach z prístroja MinION. Najprv experimentálne nájdeme vhodné parametre pre náš model, a potom ho porovnáme s priamočiarejším prístupom, ktorý najprv určí postupnosť báz zo signálu a takýto výsledok sekvenovania porovná s postupnosťou, na ktorú sa má podobať.

Kapitola 1

Sekvenovanie DNA

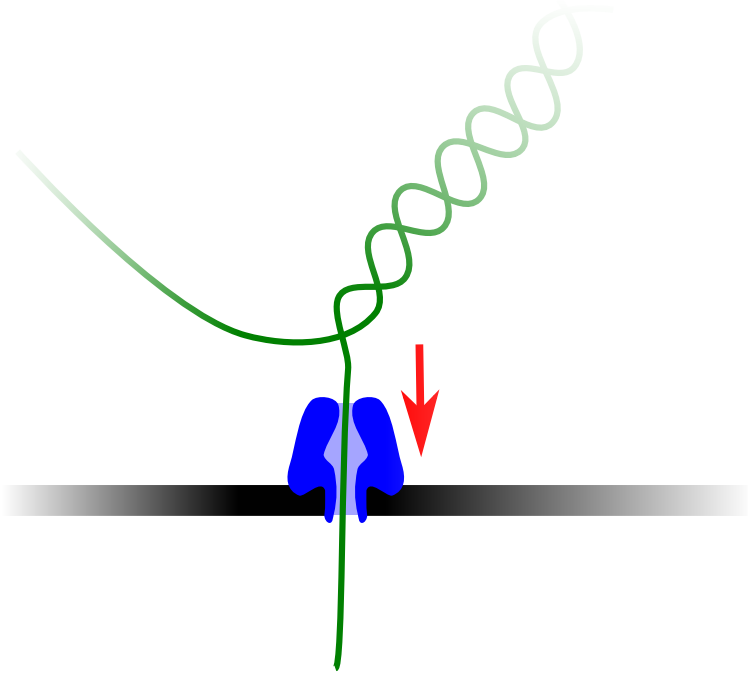
Genetická informácia je v prírode často kódovaná deoxyribonukleovou kyselinou (DNA[1]). DNA je tvorená dvoma vláknami spletenými do tvaru dvojzávitnice. Každé vlákno obsahuje postupnosť dusíkatých báz, ktorá kóduje informáciu. V DNA sa vyskytujú štyri dusíkaté bázy: adenín (`A`), cytozín (`C`), guanín (`G`) a tymín (`T`). Postupnosti báz v jednotlivých vláknach sú komplementárne: na pozíciách, kde má prvé vlákno adenín (resp. cytozín, guanín, tymín) má druhé vlákno tymín (resp. guanín, cytozín, adenín). Na zrekonštruovanie celej informácie nám teda stačí poznať poradie báz v jednom z vlákien.

Proces zisťovania poradia báz v DNA sa nazýva *sekvenovanie* DNA. Techniky sekvenovania DNA boli známe už v sedemdesiatych rokoch minulého storočia a od vtedy sa stále vyvíjajú. Pri sekvenovaní sa určí poradie dusíkatých báz vo fragmentoch DNA, nazývaných *čítania*. Z dostatočného počtu prekrývajúcich sa čítaní sa potom dá zrekonštruovať celá postupnosť báz v DNA. Pre rôzne sekvenačné technológie sú typické rôzne dĺžky čítaní, ktoré produkujú.

1.1 Nanopórové sekvenovanie DNA

Jednou z najnovších sekvenačných technológií je nanopórové sekvenovanie. Vyznačuje sa dlhými čítaniami, nízkou cenou a dostupnosťou prvých dát už počas sekvenovania, ale aj veľkým množstvom chýb v jednotlivých čítaniach [5]. Pri nanopórovom sekvenovaní sa vo vhodne zvolenej membráne vytvorí *nanopór*, t. j. otvor s priemerom rádovo 1nm. Membránou sa oddelia dve komory s elektrolytom, pričom v jednej z komôr sa nachádza aj predpripravená vzorka DNA. Po zavedení elektrického napätia medzi komorami začne nanopórom tiecť iónový prúd. Jedno vlákno DNA sa postupne oddeľuje od druhého a prechádza nanopórom (Obr. 1.1). Časť vlákna, ktorá sa práve nachádza v najužšej časti nanopóru, má vplyv na elektrický prúd tečúci cez nanopór. Rôzne

[1] z anglického ***d****eoxyribo****n****ucleic* ***a****cid*



Obr. 1.1: Prechod jedného vlákna DNA nanopórom.

bázy ovplyvňujú elektrický prúd rôznym spôsobom. Pri sekvenovaní sa meria priebeh elektrického prúdu v čase a na základe jeho zmien sa potom určuje, aké bázy prešli cez nanopór [3].

V našej práci budeme pracovať s dátami zo sekvenátora MinION. Prístroj MinION je nanopórový sekvenátor vyrábaný firmou Oxford Nanopore Technologies. V sekvenátori MinION sa používa polymérová membrána, do ktorej sú zasadené proteínové nanopóry. Sekvenátor obsahuje stovky nanopórov, dokáže teda sekvenovať niekoľko DNA vlákien súčasne. Vo verzii, s ktorou pracujeme, prechádza vlákno DNA cez nanopór rýchlosťou približne 400 báz za sekundu. Hodnota elektrického prúdu sa zaznamenáva 4000-krát za sekundu, teda v priemere zhruba 10- krát na bázu. Namerané hodnoty prúdu sa pre každé čítanie ukladajú do zvlášť súboru vo formáte `.fast5`. Tieto nespracované dáta budeme nazývať *surový signál*.

1.1.1 Normalizácia signálu

Surový signál závisí nielen od úseku DNA nachádzajúceho sa v nanopóre, ale aj od ďalších faktorov, ktoré sa pre rôzne čítania môžu líšiť. Pred ďalším spracovaním je preto potrebné surový signál znormalizovať.

Jednou z metód normalizácie je *mediánová normalizácia*, ktorú navrhujú Stoiber et. al. [11]

Definícia 1. Nech $a_1, a_2, \ldots, a_n \in \mathbb{R}$. Symbolom

$$\text{MEDIAN}_{i=1}^{n}(a_i)$$

budeme značiť medián hodnôt $a_1, a_2, \ldots, a_n$.

Definícia 2. Nech $r_1, r_2, \ldots, r_n$ sú namerané hodnoty surového signálu. Nech

$$M = \text{MEDIAN}_{i=1}^{n}(r_i)$$

a nech

$$D = \text{MEDIAN}_{i=1}^{n}(|r_i - M|).$$

Mediánovo znormalizovaný signál je postupnosť $s_1, s_2, \ldots, s_n$ určená predpisom

$$s_i = \frac{r_i - M}{D}.$$

1.1.2 Určovanie báz

Na základe signálu nameraného zariadením MinION sa určuje, aké dusíkaté bázy prechádzali nanopórom, keď bol tento signál zaznamenaný. Táto úloha je pomerne náročná a v súčasnosti sa stále vyvíjajú lepšie a lepšie riešenia. Programy, ktoré určujú bázy, nazývame *prekladače báz*.

Pri určovaní báz sa využíva fakt, že rôzne bázy pri svojom prechode nanopórom ovplyvňujú signál rôznym charakteristickým spôsobom. V praxi však signál nie je ovplyvnený iba jednou bázou. Pracuje sa preto s predpokladom, že signál je ovplyvnený k po sebe idúcimi bázami, ktoré sú práve najbližšie k nanopóru. Skupinám k po sebe idúcich báz sa hovorí *k-mery*.

Ďalším problémom je, že vlákno DNA cez nanopór neprechádza konštantnou rýchlosťou. Jednotlivým bázam vo výslednej postupnosti preto môžu zodpovedať rôzne dlhé úseky signálu. Prvým krokom pri určovaní báz preto často býva rozdelenie signálu na úseky, v rámci ktorých bola hodnota signálu približne konštantná. Týmto úsekom sa hovorí *udalosti*. Pri ďalšom spracovaní sa predpokladá, že medzi jednotlivými udalosťami sa vlákno DNA väčšinou posunie o jednu bázu. Keďže však rozdelenie signálu na udalosti nemusí presne zodpovedať posunom DNA vlákna v nanopóre, uvažuje sa aj možnosť, že sa vlákno medzi udalosťami neposunulo, prípadne posunulo o viac než jednu bázu.

Niektoré prekladače báz (napr. Nanocall [4]) modelujú prechod DNA vlákna nanopórom ako skrytý Markovovský model. Na základe tohto modelu sa Viterbiho algoritmom vypočíta najpravdepodobnejšia postupnosť báz, ktorá mohla vygenerovať pozorovaný signál.

Iné prekladače báz sú založené na rekurentných neurónových sieťach. Niektoré (napr. DeepNano [1]) pracujú so signálom rozdeleným na udalosti, iné pracujú s nerozdeleným signálom (napr. Chiron [15]).

Najlepšie súčasné prekladače báz majú pre jedno čítanie presnosť okolo 85% až 90%. Ak sa osekvenuje viac kópií rovnakej DNA, skombinovaním dostatočného počtu prekrývajúcich sa čítaní sa dá dosiahnuť presnosť okolo $99,9\%$ [16].

Kapitola 2

Ciele práce

2.1 Varianty v DNA

V prírode sa často vyskytujú dvojice DNA molekúl, ktoré obsahujú veľmi podobnú postupnosť báz, líšiacu sa iba v malých detailoch. Typickým príkladom sú DNA dvoch rôznych jedincov rovnakého druhu. Tieto malé odlišnosti v DNA voláme *varianty*. Najjednoduchšie druhy variantov sú nasledujúce.

Jednonukleotidový polymorfizmus (SNP[1]). Jedna báza z prvej DNA postupnosti sa v druhej postupnosti nahradí inou bázou.

Inzercia. Do postupnosti je vsunutá skupina báz.

Delécia. Z postupnosti vypadne súvislá skupina báz.

2.2 Identifikácia variantov

Pri niektorých využitiach DNA sekvenovania sa sekvenuje vzorka, o ktorej je známe, že by sa mala podobať na inú, už osekvenovanú DNA. Cieľom sekvenovania je potom zistiť, ako sa tieto dve DNA postupnosti líšia. Jedným z takýchto využití je napríklad zisťovanie rezistencie baktérií na antibiotiká [2].

V našej práci sa budeme zaoberať nasledujúcim scenárom. Máme nejakú známu postupnosť dusíkatých báz, ktorú budeme nazývať *referencia*. Ďalej máme vzorku DNA, o ktorej vieme, že sa od referencie líši len veľmi málo. Táto vzorka bola spracovaná prístrojom MinION, máme teda k dispozícii nameraný surový signál z jednotlivých čítaní. Naším cieľom je identifikovať varianty v sekvenovanej vzorke vzhľadom na referenciu. Ideálne by bolo vedieť s dobrou presnosťou určovať varianty už z jedného čítania.

[1] z anglického ***s**ingle **n**ucleotide **p**olymorphism*

Obr. 2.1: Jednonukleotidový polymorfyzmus, inzercia a delécia.

Snažíme sa teda navrhnúť algoritmus s nasledovným vstupom a výstupom (neformálne):

Vstup: referenčná postupnosť dusíkatých báz,

postupnosť nameraných hodnôt surového signálu,

nepovinné: odhad očakávaného množstva variantov

Výstup: popis nájdených variantov (pozícia, typ, skóre)

Náš algoritmus bude nevyhnutne robiť chyby. V niektorých prípadoch nezvládne nájsť variant, ktorý vzorka obsahovala (falošné odmietnutie), v iných prípadoch nájde variant, ktorý neexistuje (falošné prijatie). V niektorých aplikáciách môže byť cena za chyby jedného druhu väčšia, než cena za chyby opačného druhu. Algoritmus preto vráti ku každému z nájdených variantov aj skóre, indikujúce istotu algoritmu, že naozaj ide o variant. Znižovaním minimálneho skóre, ktoré budeme vyžadovať, aby sme nájdený variant považovali za skutočný, bude možné znížit množstvo falošných odmietnutí za cenu zvýšenia množstva falošných prijatí, a obrátene.

Jedným z možných riešení nášho problému je určiť zo signálu bázy pomocou niektorého z existujúcich prekladačov báz a následne už len zisťovať odlišnosti dvoch postupností báz. Problémom tohoto riešenia je nízka presnosť (ak nemáme veľa prekrývajúcich sa čítaní).

Pri tomto prístupe však prekladač báz vôbec nevyužíva fakt, že sekvenovaná postupnosť sa podobá na referenciu. Uvažuje teda podstatne väčší priestor možných výsledných sekvencií, než je nutné. V dôsledku toho upredňostňuje sekvencie, ktoré lepšie vysvetľujú pozorovaný signál, aj keď môžu byť výrazne vzdialené od referencie. Zmenšenie priestoru uvažovaných sekvencií môže navyše znížiť výpočtovú náročnosť určovania báz, prípadne umožniť použitie presnejších techník, ktoré by za normálnych okolností boli príliš časovo náročné.

V našej práci sa preto budeme snažiť nájsť riešenie, ktoré bude pri spustení na jednom čítaní presnejšie, než porovnanie referencie s výstupom z prekladača báz.

Kapitola 3

Identifikácia jednonukleotidových polymorfizmov

V tejto časti navrhneme techniku na odhaľovanie jednonukleotidových polymorfizmov (SNP) v sekvenovaných dátach. Uvažujeme pritom najjednoduchší možný scenár, keď sekvenovaná vzorka neobsahuje žiadne iné varianty a jednotlivé SNPy nie sú príliš blízko seba. Predpokladáme teda, že ak sa sekvenovaná vzorka líši od referencie v nejakej báze, v okolitých bázach sa tieto dve sekvencie zhodujú. Náš postup pri spracovaní jedného čítania sa skladá z troch fáz:

1. Zistenie, ktorej časti referencie toto čítanie zopovedá.
2. Približné zarovnanie signálu k bázam referencie.
3. Odhadnutie pravdepodobnosti SNPu na jednotlivých pozíciách sekvencie.

Na riešenie prvých dvoch fáz používame existujúce nástroje, jadro našej práce je zamerané na tretiu fázu.

3.1 Rámcové zarovnanie čítania k referencii

Keďže čítania, s ktorými pracujeme, nemusia zodpovedať celej referencii (môže ísť o kratšie fragmenty sekvenovanej DNA), v prvej fáze spracovania každého čítania je naším cieľom zistiť, ktorej časti referencie zodpovedá. To robíme v dvoch krokoch:

1. Pomocou prekladača báz preložíme signál v našom čítaní do DNA sekvencie.
2. Nájdeme časť referencie, ktorá sa tejto sekvencii najviac podobá.

Na realizáciu kroku 1 používame prekladač báz Albacore poskytovaný priamo výrobcom sekvenátora MinION. Presnosť tohto prekladača báz sa pohybuje v rozmedzí

85% až 90% [16], výstup teda nebude úplne presne zodpovedať sekvenovanej DNA. V praxi je však takáto presnosť na krok 2 väčšinou dostatočná. Problém, ktorý riešime v kroku 2, je známy ako *zarovnanie sekvencií*. Na zarovnávanie sekvencií v tejto práci používame nástroj BWA-MEM [6].

3.2 Približné zarovnanie signálu

Po dokončení prvej fázy už vieme, ktorej časti referencie by malo zodpovedať spracovávané čítanie. Stále však nevieme, ktoré časti nameraného signálu zodpovedajú jednotlivým bázam v referencii. Aproximáciu tohoto priradenia (zarovnania signálu k referencii) získame pomocou nástroja `resquiggle` z balíčka tombo [14].

Nástroj `resquiggle` surový signál normalizuje použitím mediánovej normalizácie, ktorú vo svojej práci [11] popisuje Stoiber et. al. S takto znormalizovaným signálom budeme pracovať aj vo zvyšku našej práce.

Po znormalizovaní si `resquiggle` rozdelí signál na udalosti, čo sú úseky, kde je hodnota signálu zhruba konštantná. Pre každý úsek vypočíta priemernú hodnotu signálu v ňom. Pomocou dynamického programovania, konkrétne algoritmu z triedy *dynamic time warping*, potom vypočíta najvierohodnejšie priradenie udalostí k bázam [13]. Vierohodnosť daného priradenia je hodnotená pomocou modelu pre očakávané hodnoty signálu pri jednotlivých k-meroch. Tento model pre fixné k (v praxi väčšinou $k = 5$ alebo $k = 6$) obsahuje pre každý zo 4^k možných k-merov informáciu, akú hodnotu signálu očakávame, ak daný k-mer práve prechádza nanopórom.

Pri zarovnávaní `resquiggle` predpokladá, že signál z nášho čítania presne zodpovedá referencii. Keďže sekvenovaná postupnosť obsahuje aj varianty, dá sa očakávať, že v okolí týchto variantov zarovnanie nebude veľmi presné. Dá sa však použiť ako aproximácia, v okolí ktorej budeme hľadať presnejšie zarovnanie.

3.3 Pravdepodobnosť SNPu na jednej pozícii

Pri hľadaní SNPov v sekvenovanej DNA sa pre každú pozíciu v referencii snažíme odhadnúť, či je na tejto pozícii v sekvenovanej postupnosti rovnaká báza ako v referencii alebo je tam SNP. Uvažujeme preto štyri hypotézy. Prvá hypotéza je, že v sekvenovanej postupnosti je na skúmanej pozícii `A` a na okolitých pozíciách rovnaké bázy ako v referencii. Zvyšné tri hypotézy sú analogické, s `C`, `G` a `T`. Jedna z hypotéz teda vždy hovorí, že na danej pozícii nie je variant, zvyšné tri tvrdia, že sa na tejto pozícii nachádza SNP.

Pre každú zo štyroch hypotéz ohodnotíme, ako dobre vysvetľuje signál nameraný v okolí skúmanej pozície. Vezmeme výsek T z referencie obsahujúci skúmanú pozíciu s okolím niekoľkých báz a úsek S z nameraného signálu, ktorý tejto časti referencie

zodpovedá podľa približného zarovnania z kapitoly 3.2. V postupnosti T zmeníme bázu na skúmanej pozícii na tú, ktorá tam má byť podľa našej hypotézy. Takto upravenú postupnosť označme D_B, kde B je báza, ktorú máme v našej hypotéze na skúmanej pozícii. Následne vypočítame podmienenú pravdepodobnosť $P(S \mid D_B)$, že by postupnosť báz D_B pri sekvenovaní vygenerovala signál S.

3.3.1 Pravdepodobnostný model

Podmienené pravdepodobnosti, že by nejaká postupnosť báz D pri sekvenovaní vygenerovala signál S, budú definované jednoduchým pravdepodobnostným modelom.

Model predpokladá, že bázy sekvenovaného vlákna DNA postupne prechádzajú nanopórom, pričom každá báza sa v nanopóre nachádza počas minimálne M po sebe idúcich meraní signálu. Ku každej nameranej hodnote signálu teda môžeme priradiť bázu, ktorá sa v danom čase nachádzala v nanopóre. Tomuto priradeniu budeme hovoriť *zarovnanie signálu.*

Definícia 3. Nech $n, m, M \in \mathbb{Z}^+$, pričom $m \geq nM$. *Zarovnaním signálu dĺžky m k sekvencii o dĺžke n s minimálnou dĺžkou udalosti M* nazývame ľubovoľné zobrazenie $f : \{0, 1, \ldots, m-1\} \to \{0, 1, \ldots, n-1\}$ také, že:

- f je neklesajúce, teda $\forall x, y \in \{0, 1, \ldots, m-1\} : x < y \Rightarrow f(x) \leq f(y)$.
- $\forall i \in \{0, 1, \ldots, n-1\} : |\{x \mid f(x) = i\}| \geq M$.

Množinu všetkých zarovnaní signálu dĺžky m k sekvencii o dĺžke n s minimálnou dĺžkou udalosti M budeme označovať ako $A(n, m, M)$.

Náš model predpokladá, že nameraný signál závisí iba od bázy, ktorá je práve v nanopóre, c_1 báz pred ňou a c_2 báz za ňou. To sa nedá aplikovať na prvých c_1 a posledných c_2 báz sekvenovanej postupnosti. Pri modelovaní preto budeme pre jednoduchosť predpokladať, že DNA postupnosť, ktorej signál modelujeme, išla cez nanopór ako súčasť nejakej dlhšej sekvencie. Na modelovanie signálu, ktorý bol pri tejto postupnosti zaznamenaný, budeme teda potrebovať poznať aj kontext c_1 báz pred začiatkom postupnosti a c_2 báz za jej koncom.

Definícia 4. *DNA sekvenciou dĺžky n s kontextom c_1 a c_2* nazývame ľubovoľné zobrazenie $D : \{-c_1, -c_1 + 1, \ldots, n - 1 + c_2\} \to \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Funkčnú hodnotu D v danom bode i budeme namiesto $D(i)$ značiť ako D_i.

Nech $k = c_1 + 1 + c_2$. Budeme hovoriť, že nejaký k-mer sa nachádza v nanopóre, ak báza, ktorá je v nanopóre, spolu s okolím c_1 báz pred ňou a c_2 báz za ňou tvorí tento k-mer.

Úroveň signálu závisí v našom modeli od k-meru, ktorý sa v čase merania nachádza v nanopóre. Pre každý zo 4^k možných k-merov uvažujeme rozdelenie pravdepodobnosti pre hodnotu signálu. Keďže pracujeme s digitalizovanou hodnotou signálu, existuje iba konečné množstvo hodnôt, ktoré môže signál nadobudnúť. Rozdelenie pravdepodobnosti teda bude pravdepodobnostná funkcia.

Definícia 5. Množinu všetkých možných nameraných hodnôt signálu budeme označovať $\mathbb{S}$.

Čísla M, c_1, c_2 a pravdepodobnostné funkcie pre jednotlivé k-mery sú jediné parametre nášho modelu.

Definícia 6. *k-merový model* je štvorica (M, c_1, c_2, Δ), kde:

- $M \in \mathbb{Z}^+$.
- $c_1, c_2 \in \mathbb{Z}_{\geq 0}$.
- Δ je množina, ktorá pre každé $K \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^{c_1+1+c_2}$ obsahuje rozdelenie pravdepodobnosti δ_K nad množinou $\mathbb{S}$.

V našom modeli budeme jednotlivé merania považovať za nezávislé udalosti.

Definícia 7. Nech (M, c_1, c_2, Δ) je k-merový model, $n, m \in \mathbb{Z}^+$, D je DNA sekvencia dĺžky n s konextom c_1 a c_2, nech $S = (S_i)_{i=0}^{m-1}$ je postupnosť hodnôt z $\mathbb{S}$ a nech $a \in A(n, m, M)$. Pravdepodobnosť vygenerovania signálu S pri prechode sekvencie D nanopórom pri zarovnaní a podľa modelu (M, c_1, c_2, Δ) bude:

$$P(S \mid D, a) = \prod_{i=0}^{m-1} \delta_{(D_{a(i)-c_1}, \ldots, D_{a(i)+c_2})}(S_i).$$

Výslednú podmienenú pravdepodobnosť $P(S \mid D)$ získame marginalizáciou zarovnania a. Pre jednoduchosť budeme predpokladať, že všetky zarovnania z $A(n, m, M)$ sú rovnako pravdepodobné.

Definícia 8. Nech (M, c_1, c_2, Δ) je k-merový model, $n, m \in \mathbb{Z}^+$, D je DNA sekvencia dĺžky n s kontextom c_1 a c_2 a nech $S = (S_i)_{i=0}^{m-1}$ je postupnosť hodnôt z $\mathbb{S}$. Pravdepodobnosť vygenerovania signálu S pri sekvenovaní D podľa modelu (M, c_1, c_2, Δ) je:

$$P(S \mid D) = \frac{\sum_{a \in A(n,m,M)} P(S \mid D, a)}{|A(n, m, M)|}.$$

3.3.2 Výpočet podmienenej pravdepodobnosti

Balík nástrojov tombo obsahuje aj rozdelenia pravdepodobnosti signálu pre jednotlivé k-mery (konkrétne pre $k = 6$ a $c_1 = 2$). Tombo modeluje rozdelenie pre každý k-mer ako normálne rozdelenie. Parametre týchto normálnych rozdelení sa pre jednotlivé k-mery líšia a boli odhadnuté empiricky na základe dát.

Keďže ide o hustotu pravdepodobnosti a nie pravdepodobnostnú funkciu, nemôžeme rozdelenia z tombo priamo použiť ako funkcie δ z nášho modelu. Označme distribúciu z tombo pre k-mer K ako θ_K. Najmenší možný rozdiel, ktorý môžu mať dve hodnoty signálu po digitalizácii, označme r. V našom modeli potom pre k-mer K použijeme pravdepodobnostnú funkciu δ_K definovanú ako:

$$\delta_K(x) = \int_{x-r/2}^{x+r/2} \theta_K(t)\,\mathrm{d}t,$$

ktorú aproximujeme ako:

$$\delta_K(x) = r \cdot \theta_K(x).$$

Podmienenú pravdepodobnosť vygenerovania signálu S dĺžky m pri sekvenovaní postupnosti D dĺžky n s kontextom c_1 a c_2 teda budeme počítať ako:

$$P(S \mid D) = \frac{\sum_{a \in A(n,m,M)} \prod_{i=0}^{m-1} r \cdot \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i)}{|A(n,m,M)|} = \\ = \frac{r^m}{|A(n,m,M)|} \sum_{a \in A(n,m,M)} \prod_{i=0}^{m-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i).$$

Pri vyhodnocovaní hypotéz pre jednu pozíciu pracujeme stále s rovnako dlhou postupnosťou báz a s tým istým úsekom signálu, teda hodnoty r^m a $|A(n,m,M)|$ sú pre všetky štyri hypotézy rovnaké. Keďže pri týchto podmienených pravdepodobnostiach bude dôležitý iba ich pomer, v našej implementácii počítame iba hodnotu

$$\sum_{a \in A(n,m,M)} \prod_{i=0}^{m-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i).$$

Naivný výpočet tejto sumy by bol pomalý, preto ju počítame dynamickým programovaním. Používame algoritmus z rodiny označovanej ako *dynamic time warping*. Algoritmy tohto typu sa používajú, keď chceme zistiť, ktoré časti dvoch postupností si vzájomne zodpovedajú [9].

Pre každé $x \in \{0, 1, \ldots, m\}, y \in \{0, 1, \ldots, n\}$ vypočítame hodnotu

$$\mathrm{DP}[y][x] = \sum_{a \in A(y,x,M)} \prod_{i=0}^{x-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i).$$

Hodnota $\mathrm{DP}[y][x]$ teda zodpovedá podmienenej pravdepodobnosti, že by prvých y báz postupnosti D vygenerovalo prvých x hodnôt signálu. Zjavne platí:

- $\mathrm{DP}[0][0] = 1$
- $\mathrm{DP}[0][x] = 0$ pre $x \geq 1$
- $\mathrm{DP}[y][x] = 0$ pre $x < yM$

Pre $y \geq 1$ a $x \geq yM$ môžeme množinu zarovnaní $A(y, x, M)$ rozdeliť na dve časti:

- $B(y, x, M) = \{a \in A(y, x, M) \mid |\{i \mid a(i) = y - 1\}| = M\}$
- $C(y, x, M) = \{a \in A(y, x, M) \mid |\{i \mid a(i) = y - 1\}| > M\}$

Množina $B(y, x, M)$ obsahuje tie zarovnania, v ktorých je posledná báza priradená presne M hodnotám signálu, množina $C(y, x, M)$ obsahuje zarovnania, kde je posledná báza priradená ostro viac než M hodnotám. Zjavne platí $B(y, x, M) \cap C(y, x, M) = \emptyset$ a vďaka podmienke, že každá báza musí byť priradená aspoň M hodnotám signálu platí aj $B(y, x, M) \cup C(y, x, M) = A(y, x, M)$. Sumu, ktorú počítame, preto môžeme rozdeliť ako:

$$\mathrm{DP}[y][x] = \sum_{a \in B(y,x,M)} \prod_{i=0}^{x-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i) + \sum_{a \in C(y,x,M)} \prod_{i=0}^{x-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i).$$

Skrátením ľubovoľného zarovnania z $B(y, x, M)$ o posledných M hodnôt (teda zúžením jeho definičného oboru na $\{0, \ldots, x - M - 1\}$) dostaneme nejaké zarovnanie z $A(y - 1, x - M, M)$. Toto skrátenie nám vlastne definuje bijekciu medzi $B(y, x, M)$ a $A(y - 1, x - M, M)$. Prvú z našich dvoch súm teda môžeme prepísať ako:

$$\begin{aligned}
&\sum_{a \in B(y,x,M)} \prod_{i=0}^{x-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i) = \\
&= \sum_{a \in A(y-1,x-M,M)} \left(\prod_{i=0}^{x-M-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i) \prod_{i=x-M}^{x-1} \theta_{(D_{y-1-c_1},\ldots,D_{y-1+c_2})}(S_i) \right) = \\
&= \mathrm{DP}[y-1][x-M] \cdot \prod_{i=x-M}^{x-1} \theta_{(D_{y-1-c_1},\ldots,D_{y-1+c_2})}(S_i).
\end{aligned}$$

Skrátením ľubovoľného zarovnania z $C(y, x, M)$ o poslednú hodnotu dostaneme zarovnanie z $A(y, x-1, M)$, pričom opäť ide o bijekciu. Druhú sumu teda môžeme prepísať ako:

$$\sum_{a \in C(y,x,M)} \prod_{i=0}^{x-1} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i) =$$

$$= \sum_{a \in A(y,x-1,M)} \theta_{(D_{y-1-c_1},\ldots,D_{y-1+c_2})}(S_{x-1}) \prod_{i=0}^{x-2} \theta_{(D_{a(i)-c_1},\ldots,D_{a(i)+c_2})}(S_i) =$$

$$= \mathrm{DP}[y][x-1] \cdot \theta_{(D_{y-1-c_1},\ldots,D_{y-1+c_2})}(S_{x-1}).$$

Jednu hodnotu $\mathrm{DP}[y][x]$ teda vieme vypočítať z predošlých hodnôt v čase $O(M)$ ako:

$$\mathrm{DP}[y][x] = \mathrm{DP}[y-1][x-M] \cdot \prod_{i=x-M}^{x-1} \theta_{(D_{y-1-c_1},\ldots,D_{y-1+c_2})}(S_i) +$$

$$+ \mathrm{DP}[y][x-1] \cdot \theta_{D_{y-1-c_1},\ldots,D_{y-1+c_2}}(S_{x-1}).$$

Hodnota $\mathrm{DP}[n][m]$ je potom priamo podmienená pravdepodobnosť (preškálovaná koeficientom $|A(n,m,M)|/r^m$), ktorú sme chceli vypočítať. Aby sme sa pri výpočtoch vyhli problémom s presnosťou reálnej aritmetiky pri medzivýsledkoch veľmi blízkych nule, pracujeme v skutočnosti s ich logaritmami.

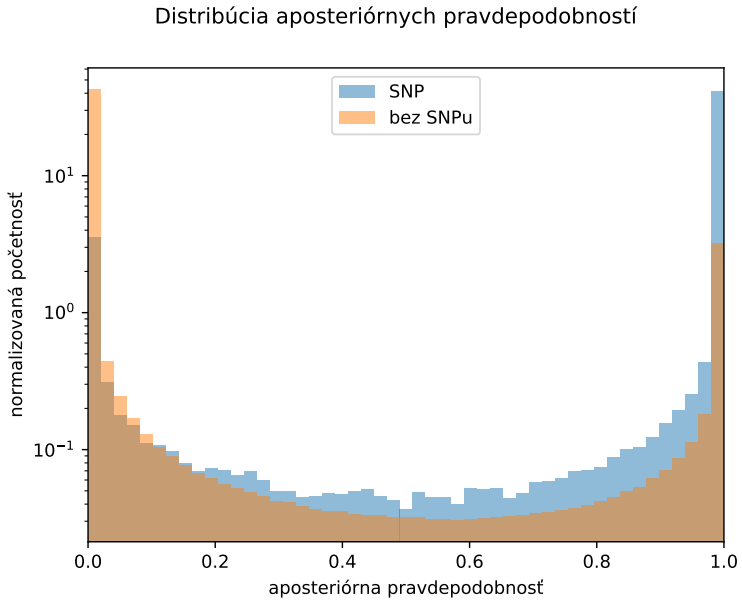
3.4 Apriórne a aposteriórne pravdepodobnosti

Ak máme nejaké očakávanie na množstvo SNPov v sekvenovanej postupnosti, môžeme každej našej hypotéze D_B priradiť apriórnu pravdepodobnosť $P(D_B)$. Ak očakávame, že podiel SNPov v sekvenovanej postupnosti (pomer počtu SNPov k dĺžke sekvencie) je p, apriórna pravdepodobnosť hypotézy hovoriacej, že na skúmanej pozícii nie je SNP, bude $1-p$. Pre jednoduchosť budeme predpokladať, že zvyšné tri hypotézy majú rovnakú apriórnu pravdepodobnosť $p/3$.

S použitím vypočítaných podmienených pravdepodobnosti následne môžeme určiť aposteriórne pravdepodobnosti pre jednotlivé hypotézy:

$$P(D_B \mid S) = \frac{P(S \mid D_B)P(D_B)}{\sum_{B' \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}} P(S \mid D_{B'})P(D_{B'})}.$$

Ako ilustruje graf na Obr. 3.1, v praxi dostaneme takýmto výpočtom väčšinu aposteriórnych pravdepodobností veľmi blízku 0 alebo 1. Relatívne často sa pritom stáva, že pravdepodobnosť blízku 1 získame pre nepravdivú hypotézu.

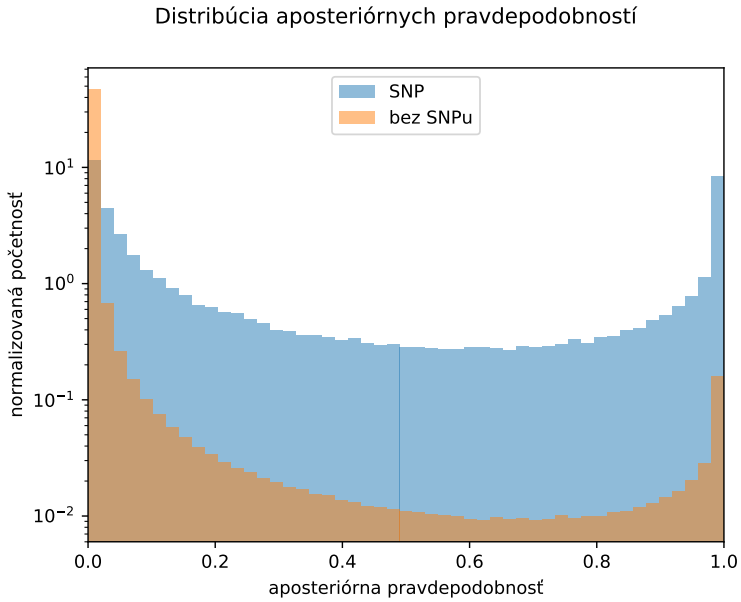


Obr. 3.1: Distribúcia aposteriórnych pravdepodobností pre pozície, na ktorých je SNP a pozície, na ktorých nie je SNP. Početnosť pozícií je normalizovaná a na logaritmickej škále.

Dá sa teda konštatovať, že náš model si je „priveľmi istý" svojimi predpoveďami. Tento efekt môže byť spôsobený nerealistickým predpokladom, že jednotlivé merania signálu sú nezávislé udalosti. V skutočnosti medzi meraniami pravdepodobne existujú závislosti: keďže meriame fyzikálnu veličinu v rôznych časoch tesne po sebe, dá sa napríklad očakávať, že v susedných meraniach väčšinou nameriame podobné hodnoty. Keď teda uvažujeme každé meranie ako samostatnú nezávislú udalosť, stáva sa, že podobnú (v podstate tú istú) informáciu model vo svojej predpovedi zohľadní viackrát.

Viacnásobné zohľadnenie jednej udalosti spôsobuje, že jej prikladá väčšiu výpovednú hodnotu, než táto udalosť v skutočnosti má. Pri finálnej predpovedi má potom väčšiu istotu, než by mal mať.

Aby sme získali menej extrémne čísla, pred výpočtom aposteriórnych pravdepodobností odmocníme vypočítané podmienené pravdepodobnosti desiatou odmocninou. Graf na Obr. 3.2 ukazuje, ako sa tým zmení distribúcia aposteriórnych pravdepodobností. Číslo 10 bolo zvolené na základe predstavy, že keďže jedna báza je v nanopóre v priemere zhruba počas 10 meraní, zhruba 10 po sebe idúcich meraní býva podobných a teda jednu udalosť zarátavame v našom modeli zhruba 10-krát. Tento dôvod je však čisto intuitívny; aposteriórne pravdepodobnosti, ktoré získame, je preto lepšie chápať iba ako skóre určené na porovnávanie dôveryhodnosti hypotéz, nie pravdepodobnosť, s ktorou by sa dali robiť ďalšie výpočty.



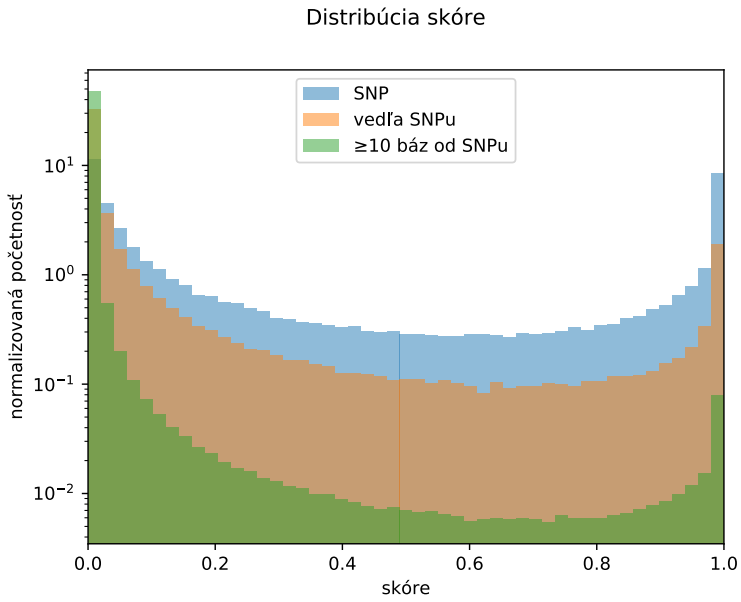
Obr. 3.2: Distribúcia aposteriórnych pravdepodobností pri použití desiatej odmocniny.

3.5 Pozície blízko SNPov

Postup, ktorý popisujeme v kapitole 3.3 máva často problémy pri pozíciách, na ktorých síce nie je SNP, ale nachádzajú sa blízko nejakého SNPu. V grafe na Obr. 3.3 vidíme, že pozície tesne vedľa SNPov majú vysoké skóre oveľa častejšie než pozície ďaleko od SNPov, kým pozície ďaleko od SNPov majú oveľa častejšie skóre blízke nule. Pri týchto pozíciách totiž žiadna zo štyroch uvažovaných hypotéz nezodpovedá realite, nakoľko všetky predpokladajú, že na pozíciách v okolí skúmanej pozície sa sekvenovaný genóm zhoduje s referenciou. Žiadna z hypotéz potom nedokáže dobre vysvetliť nameraný signál.

Tento problém môžeme vyriešiť tak, že pre každú pozíciu berieme do úvahy aj hypotézy, ktoré hovoria, že sa SNP nachádza na niektorej z okolitých pozícií. Za okolité pozície pritom považujeme všetky pozície vzdialené nanajvýš $k-1$ od tej skúmanej, teda všetky pozície, ktoré v našom modeli majú nejaký vplyv na úsek signálu ovplyvnený skúmanou pozíciou. Namiesto štyroch hypotéz teda budeme mať $3 \cdot (2k-1)+1$ hypotéz: jedna hovorí, že v sekvenovanej postupnosti nebol v okolí skúmanej pozície žiaden variant, tri hovoria, že na skúmanej pozícii je SNP a $3 \cdot (2k-2)$ hypotéz hovorí, že na niektorej z okolitých pozícií je SNP.

Pri určovaní apriórnych pravdepodobností hypotéz budeme pracovať s predpokladom, že to, či sa na jednotlivých pozíciách vyskytne SNP, sú nezávislé udalosti. Ak teda v sekvenovanej postupnosti očakávame podiel SNPov p, potom hypotéze, že žiadna z $2k-1$ okolitých pozícií neobsahuje variant, priradíme apriórnu pravdepodobnosť $(1-p)^{2k-1}$. Všetkým ostatným hypotézam priradíme apriórnu pravdepodobnosť $p/3 \cdot (1-p)^{2k-2}$, pretože predpokladajú na práve jednej z $2k-1$ pozícií inú bázu ako



Obr. 3.3: Distribúcia skóre pre pozície so SNPom, pozície tesne vedľa SNPu a pozície aspoň desať báz od najbližšieho SNPu.

referencia. Keďže zanedbávame možnosť, že medzi $2k-1$ bázami by mohol byť viac než jeden SNP, súčet takto vypočítaných apriórnych pravdepodobností pre naše hypotézy bude menší než 1. Preto ich nakoniec ešte znormalizujeme: všetky apriórne pravdepodobnosti prenásobíme rovnakým číslom tak, aby dávali súčet 1.

Zvýšením počtu uvažovaných hypotéz sme zvýšili aj čas potrebný na vypočítanie všetkých podmienených pravdepodobností. Preto navrhneme spôsob, ako náš výpočet zrýchliť.

Časť referencie, ktorá zodpovedá postupnosti sekvenovanej v práve spracovávanom čítaní, označme R. Jej bázy budeme označovať $R_1, R_2, \ldots$. Slovom *referencia* budeme v nasledujúcom texte označovať iba túto relevantnú časť referencie. Výsek z referencie s ktorým pracujeme, keď odhadujeme, či sa na i-tej pozícii v sekvenovanej DNA nachádza SNP, označme T^i. Úsek signálu, ktorý tomuto výseku zodpovedá podľa približného zarovnania označme S^i. Postupnosť, ktorú dostaneme, ak v T^i zmeníme j-tu bázu z referencie na B, označme ako $D^i_{j,B}$. Množinu všetkých pozícií, ktoré skúmame, označme I (pôjde o všetky pozície sekvenovanej postupnosti okrem zopár pozícií blízko začiatku a konca, pre ktoré nemáme dostatočne dlhý kontext).

V pôvodnom prístupe so štyrmi hypotézami sme pre každé $i \in I, B \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ potrebovali vypočítať $P(D^i_{i,B} \mid S_i)$. V novom prístupe potrebujeme pre každé $i \in I$ vypočítať $P(T^i \mid S_i)$ a pre každé $i \in I; j \in \{i-k+1, \ldots, i+k-1\}; B \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \setminus R_j$ potrebujeme vypočítať $P(D^i_{j,B} \mid S_i)$. To je $|I| \cdot (3(2k-1)+1)$ rôznych podmienených pravdepodobností.

Ak budeme pri každej pozícii ako T^i uvažovať celú referenciu R a ako S^i celý sig-

nál (označme ho S), niektoré podmienené pravdepodobnosti sa nám začnú opakovať. Pre každé $i \in I$ bude $P(T^i \mid S_i)$ rovné $P(R \mid S)$, túto podmienenú pravdepodobnosť nám teda stačí vypočítať len raz. Okrem toho potrebujeme pre každé $i \in I; j \in \{i-k+1, \ldots, i+k-1\}; B \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \setminus R_j$ vypočítať $P(D^i_{j,B} \mid S_i) = P(R_{j,B} \mid S)$, kde $R_{j,B}$ je postupnosť, ktorú dostaneme, ak v referencii R zmeníme j-tu bázu na B. Tieto podmienené pravdepodobnosti tiež nie sú všetky rôzne, stačí nám vlastne vypočítať $P(R_{j,B} \mid S)$ pre každé $j \in \{\min(I) - k + 1, \ldots, \max(I) + k - 1\}; B \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \setminus R_j$. Dokopy nám teda stačí vypočítať $3(|I| + 2k - 2) + 1$ rôznych podmienených pravdepodobností, čo je zhruba $2k$-krát menej.

Výpočet jednej podmienenej pravdepodobnosti sa nám však výrazne spomalil, keďže sme výšku tabuľky DP, ktorú potrebujeme vyplniť, zväčšili na dĺžku celej referencie R a šírku sme zväčšili na dĺžku celého signálu S. Keďže však pri všetkých podmienených pravdepodobnostiach, ktoré počítame, zarovnávame ten istý signál k takmer tej istej DNA postupnosti, výpočty si budú veľmi podobné. Tento fakt využijeme nasledujúcim spôsobom.

Dĺžku referencie označme n a dĺžku signálu označme m. Postupom z kapitoly 3.3.2 vypočítame pre referenciu R a signál S tabuľku $\mathrm{DP}_{R,S}$, ktorej prvok $\mathrm{DP}_{R,S}[y][x]$ je rovný podmienenej pravdepodobnosti, že by prvých y báz referencie vygenerovalo prvých x hodnôt signálu. Podobnú tabuľku vypočítame aj odzadu: pre každé $x \in \{0, 1, \ldots, m\}, y \in \{0, 1, \ldots, n\}$ vypočítame hodnotu $\overline{\mathrm{DP}_{R,S}}[y][x]$, ktorá bude rovná podmienenej pravdepodobnosti, že by posledných $n-y$ báz referencie vygenerovalo posledných $m-x$ hodnôt signálu. Pri výpočte $\overline{\mathrm{DP}_{R,S}}$ postupujeme analogicky s výpočtom $\mathrm{DP}_{R,S}$, akurát odzadu.

Keď počítame $P(R_{j,B} \mid S)$ pre nejaké j, B, vypočítame hodnoty $\mathrm{DP}_{R_{j,B},S}[y][x]$ iba pre $y \in \{0, 1, \ldots, j+c_1+1\}, x \in \{0, 1, \ldots, m\}$, teda iba prvých $j+c_1+2$ riadkov tabuľky. Okrem toho vypočítame ešte hodnoty $\overline{\mathrm{DP}_{R_{j,B},S}}[y][x]$ pre $y \in \{j + c_1 + 1, \ldots, n\}, x \in \{0, 1, \ldots, m\}$. Podmienenú pravdepodobnosť $P(R_{j,B} \mid S)$ potom vypočítame ako:

$$P(R_{j,B} \mid S) = \sum_{x=0}^{m} \mathrm{DP}_{R_{j,B},S}[j + c_1 + 1][x] \cdot \overline{\mathrm{DP}_{R_{j,B},S}}[j + c_1 + 1][x].$$

Pri výpočte $\mathrm{DP}_{R_{j,B},S}$ využijeme, že postupnosť $R_{j,B}$ sa v pozíciách 0 až $j-1$ zhoduje s referenciou R, pre pozície 0 až $j - c_2 - 1$ sa teda zhoduje aj k-mer, ktorý ovplyvňuje signál zarovnaný k danej pozícii. To znamená, že tabuľky $\mathrm{DP}_{R,S}$ a $\mathrm{DP}_{R_{j,B},S}$ budú mať prvých $j - c_2 + 1$ riadkov zhodných, teda pre každé $y \in \{0, \ldots, j - c_2\}, x \in \{0, \ldots, m\}$ bude platiť:

$$\mathrm{DP}_{R_{j,B},S}[y][x] = \mathrm{DP}_{R,S}[y][x].$$

Do tabuľky $\mathrm{DP}_{R_{j,B},S}$ nám teda v skutočnosti stačí dopočítať riadky s $y \in \{j - c_2 + 1, \ldots, j + c_1 + 1\}$, čo je k riadkov.

Pri výpočte $\overline{\mathrm{DP}_{R_{j,B},S}}$ využijeme, že pre všetky pozície od $j + c_1 + 1$ sa v R a $R_{j,B}$ zhoduje k-mer ovplyvňujúci túto pozíciu, teda tabuľky $\overline{\mathrm{DP}_{R,S}}$ a $\overline{\mathrm{DP}_{R_{j,B},S}}$ sa budú zhodovať v riadkoch s $y \in \{j + c_1 + 1, \ldots, n\}$. To znamená, že z tabuľky $\overline{\mathrm{DP}_{R_{j,B},S}}$ nemusíme počítať nič, všetko už máme vypočítané.

Pri výpočte jednej podmienenej pravdepodobnosti počítame iba k riadkov tabuľky, vyriešili sme teda spomalenie spôsobené zväčšením počtu riadkov tabuľky. Momentálna verzia nášho algoritmu uvažuje všetky možné zarovnania signálu S k sekvenovanej postupnosti. Väčšina zarovnaní je pritom veľmi nerealistických a do výslednej pravdepodobnosti prispeje len zanedbateľne. Aby sme náš algoritmus urýchlili, obmedzíme sa preto len na zarovnania blízke približnému zarovnaniu získanému nástrojom `resquiggle`.

Približné signálové zarovnanie, ktoré sme získali v kapitole 3.2 označme r. Hodnoty $\mathrm{DP}[y][x]$ a $\overline{\mathrm{DP}}[y][x]$ budeme počítať iba pre x, y, pre ktoré platí:

$$y \in \{r(x) - 5, \ldots, r(x) + 5\}.$$

Ostatné hodnoty v tabuľkách budeme považovať za nulové. Tým náš algoritmus obmedzíme na signálové zarovnania, v ktorých je každá hodnota signálu zarovnaná nanajvýš o 5 báz inam ako v zarovnaní r. Z každého riadku tabuľky teda budeme počítať iba malú časť, čím vyriešime spomalenie spôsobené zvýšením počtu stĺpcov tabuľky.

Takéto obmedzovanie počítanej časti tabuľky patrí medzi bežné techniky zrýchľovania dynamic time warpingu [10].

V kapitole 4.3.1 porovnávame tento prístup s pôvodným prístupom popísaným v 3.3.

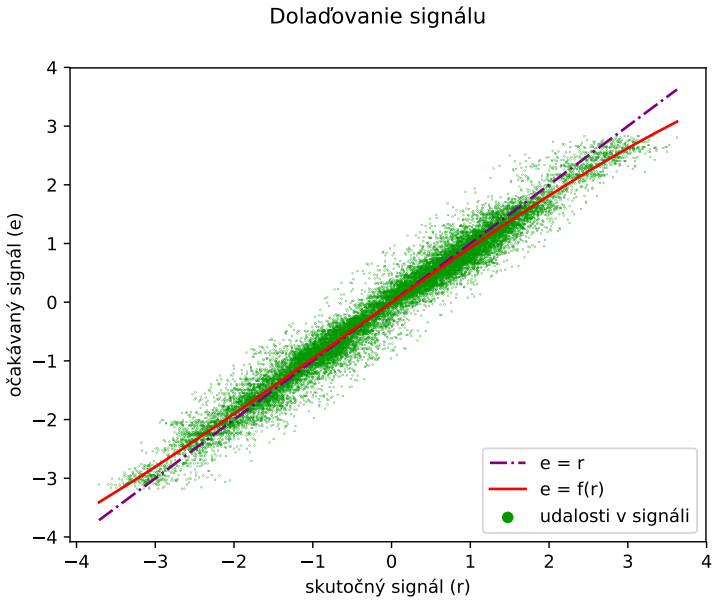
3.6 Vylepšenia zohľadňujúce špecifiká dát

Pri testovaní nášho modelu na reálnych dátach nastávali aj situácie, keď mal problém správne určiť prítomnosť SNPu na nejakej pozícii. Skúmaním takýchto situácií sme dospeli k dvom malým vylepšeniam, ktoré zvýšili presnosť modelu.

3.6.1 Doladenie normalizácie

Jedným z dôvodov, prečo náš model niekedy považoval správnu hypotézu za menej pravdepodobnú než niektorú z nesprávnych, bol systematický posun skutočného signálu od očakávaného, v rámci celého čítania. Tento problém sa snažíme riešiť doladením normalizácie signálu.

Na začiatku spracovania každého čítania, po znormalizovaní surového signálu tento znormalizovaný signál ešte mierne upravíme. Pre každú udalosť (úsek signálu) z približného zarovnania vypočítaného nástrojom `resquiggle` sa pozrieme na očakávanú



Obr. 3.4: Dvojice (r, e) z jedného čítania, funkcia interpolujúca tieto dvojice a identická funkcia.

priemernú hodnotu signálu v tejto udalosti e a na priemernú hodnotu skutočného signálu r. Dvojice (r, e), pre ktoré je rozdiel $|r - e|$ väčší ako prah $t = 1$ zahodíme. Ostatné dvojice interpolujeme, snažíme sa teda nájsť takú funkciu f, aby pre každú uvažovanú dvojicu (r, e) platilo $f(r) \approx e$, ale aby funkcia f zároveň bola dostatočne hladká. Následne každú hodnotu signálu s upravíme na $f(s)$.

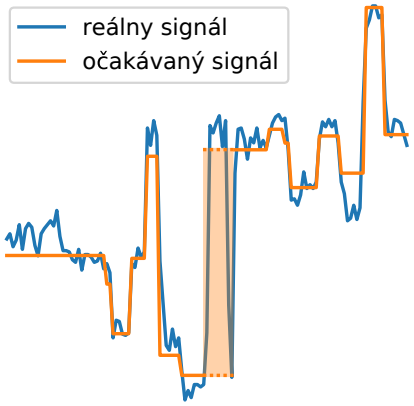
V praktickej implementácii sme na interpoláciu použili B-spline krivku vypočítanú pythonovskou funkciou `scipy.interpolate.splrep`. Graf na Obr. 3.4 zobrazuje dvojice (r, e) zo signálu jedného čítania, pre ktoré $|r - e| \leq 1$, funkciu f, ktorou sme tieto dvojice interpolovali a identickú funkciu.

Doladenie normalizácie mierne spresnilo náš model, toto spresnenie meriame v kapitole 4.3.2.

3.6.2 Modelovanie cúvania

V signáli sa občas vyskytuje nasledujúci vzor. Signál sa drží blízko nejakej hladiny h_1 a očakávame, že po posunutí ďalšej bázy do nanopóru prejde na nejakú inú hladinu h_2. Namiesto toho signál najprv na krátky čas prejde k hladine h_2, potom sa na chvíľu vráti na h_1 a až potom naozaj prejde na hladinu h_2, kde už zostane až do ďalšej zmeny spôsobenej posunutím sekvenovaného vlákna v nanopóre. Príklad takéhoto správania je na Obr. 3.5.

Toto správanie by mohlo byť spôsobené tým, že sekvenované vlákno sa krátku chvíľu hýbe opačným smerom, pričom sa na chvíľu do nanopóru vráti báza, ktorá z neho práve vyšla. Keďže náš pravdepodobnostný model s týmto správaním neráta, nevie



Obr. 3.5: Reálny a očakávaný signál v situácii, kde sa signál na chvíľu vrátil na hladinu, z ktorej už odišiel.

nájsť dobré signálové zarovnanie ani pre správnu hypotézu.

Preto mierne zmeníme predstavu, ako vlákno prechádza nanopórom. Stále musí platiť, že bázy prechádzajú cez nanopór postupne a každá je tam počas minimálne M po sebe idúcich meraní, medzi každými dvoma bázami však povolíme ľubovoľne dlhé prechodné obdobie. Od signálu počas prechodného obdobia očakávame, že bude v každom momente podobný buď signálu očakávanému tesne pred týmto prechodným obdobím, alebo signálu očakávanému tesne po ňom. Rozdelenie pravdepodobnosti pre signál počas prechodného obdobia definujeme zmiešaním (sčítaním a vydelením dvomi) rozdelenia pravdepodobnosti pre k-mer, ktorý bol v nanopóre pred týmto prechodným obdobím a rozdelenia pravdepodobnosti pre k-mer, ktorý bude v nanopóre po ňom.

Pozitívny efekt, ktorý mala táto zmena na presnosť nášho modelu meriame v kapitole 4.3.3.

Kapitola 4

Testovanie

Postupy popsísané v kapitole 3 sme implementovali v jazyku Python 3 s výpočtovo kritickými časťami implementovanými v C++ (Príloha A). V tejto kapitole náš postup testujeme na reálnych dátach.

4.1 Návrh experimentu

Na experiment budeme potrebovať dáta skladajúce sa z troch zložiek:

1. Postupnosť báz A.
2. Signál S zo zariadenia MinION získaný pri sekvenovaní nejakej sekvencie B, ktorá sa od A líši iba v niekoľkých SNPoch.
3. Postupnosť báz B, získaná nejakou presnou sekvenačnou metódou.

Počas testovania dáme nášmu modelu na vstup iba prvé dve zložky dát, teda signál S a postupnosť A ako referenciu. Model na základe signálu odhadne, kde sa v postupnosti B vyskytujú SNPy. Na základe skutočnej postupnosti B (zložka 3) sa potom vyhodnotí, aká presná je predpoveď modelu.

Dáta s vhodnou štruktúrou skonštruujeme tak, že začneme so zložkami 2 a 3 a vytvoríme k nim zložku 1. Z rôznych iných prác pracujúcich so sekvenátorom MinION existujú dáta, kde je rovnaká DNA sekvencia osekvenovaná dvakrát: raz sekvenátorom MinION a raz nejakou presnou metódou. Signál zo sekvenovania prístrojom MinION môžeme teda použiť ako S a výsledok presného sekvenovania ako B. Postupnosť A potom vytvoríme z postupnosti B tak, že v nej umelo vytvoríme niekoľko SNPov. Takýto postup nám dáva plnú kontrolu nad množstvom SNPov, v ktorých sa B líši od A, nášmu modelu teda dokážeme dať na vstup dobré apriórne pravdepodobnosti.

Pri každom experimente si určíme podiel SNPov p, ktorý chceme mať v našich dátach. Pri konštrukcii postupnosti A postupujeme tak, že pre každú bázu postupnosti

Tabuľka 4.1: Testovacie sady

skratka	zdroj	čítaní	po fáze 2	priemerná dĺžka	podiel SNPov
KLEBS_01	[16]	1 000	668	18 029	0,1%
KLEBS_03	[16]	1 000	666	17 895	0,3%
KLEBS_10	[16]	1 000	657	18 026	1%
KLEBS_30	[16]	1 000	626	18 086	3%
NANO_01	[8]	1 000	934	1 801	0,1%
NANO_03	[8]	1 000	934	1 803	0,3%
NANO_10	[8]	1 000	928	1 801	1%
NANO_30	[8]	1 000	888	1 798	3%

B sa nezávisle rozhodneme, či ju zmeníme. S pravdepodobnosťou $1-p$ bázu ponecháme, s pravdepodobnosťou p ju zmeníme na inú, pričom každá z troch možných iných báz má rovnakú pravdepodobnosť $p/3$.

V každom experimente testujeme náš model s rovnakými parametrami na viacerých čítaniach. Všetky čítania sú zo sekvenovania rovnakej postupnosti B, postupnosť A však ku každému čítaniu vytvoríme zvlášť. Podiel SNPov p je pre všetky čítania rovnaký a dávame ho modelu k dispozícii, aby si na základe neho vypočítal apriórne pravdepodobnosti.

Pri experimentoch používame podmnožinu dát použitých v [16] a podmnožinu dát z [8]. Z týchto dát sme vytvorili osem testovacích sád s rôznymi podielmi SNPov. V každej sade bolo aj niekoľko čítaní, pri ktorých zlyhala niektorá z prvých dvoch fáz spracovania – buď nástroj BWA-MEM nenašiel vhodné zarovnanie čítania k referencii, alebo nástroj `resquiggle` nenašiel vhodné približné zarovnanie signálu k jednotlivým bázam. V tabuľke 4.1 uvádzame základné charakteristiky testovacích sád.

4.2 Metriky

V našom modeli pri každej pozícii počítame aposteriórne pravdepodobnosti pre niekoľko hypotéz. Sčítaním pravdepodobností pre všetky tri hypotézy hovoriace, že na danej pozícii je iná báza ako v referencii, získame pre danú pozíciu nejaké skóre. Toto skóre vyjadruje, ako veľmi je náš model presvedčený, že na tejto pozícii je SNP. To, ako dobre tieto skóre popisujú realitu, budeme v jednotlivých experimentoch vyhodnocovať pomocou dvoch metrík: krivky ROC a úspešnosti identifikácie.

4.2.1 Krivka ROC

Náš model môžeme chápať ako klasifikátor, ktorý sa snaží pozície sekvenovanej postupnosti rozdeliť na dve skupiny: tie, na ktorých je SNP a tie, na ktorých nie je SNP. Presnosť klasifikátorov s dvoma triedami sa často vyjadruje pomocou krivky *Receiver operating characteristic* (ROC) [7].

Pozície, ktoré prehlásime za SNPy, môžeme vyberať pomocou prahu. Zvolíme si nejaký prah t a za SNPy budeme považovať práve tie pozície, ktoré majú skóre aspoň t. Voľba prahu bude ovplyvňovať, koľkých a akých chýb sa pri klasifikácii dopustíme. Pre každú možnú hodnotu prahu nás budú zaujímať dve čísla: *Frekvencia skutočných pozitívnych príkladov* a *Frekvencia falošných pozitívnych príkladov*.

Definícia 9.

Frekvencia skutočných pozitívnych príkladov (TP[1]) je pomer počtu SNPov, ktoré náš model správne označí ako SNPy, k počtu všetkých SNPov v sekvenovanej postupnosti.

Frekvencia falošných pozitívnych príkladov (FP[2]) je pomer počtu pozícií, ktoré náš model nesprávne označí ako SNPy, k počtu všetkých pozícií, ktoré nie sú SNPy.

Zvolením vhodného prahu vieme dosiahnuť ľubovoľné FP medzi 0 a 1, za znižovanie FP však musíme platiť znižovaním TP. Krivka ROC je graf závislosti TP od FP. Príklad krivky ROC je na Obr. 4.1.

Čím presnejší je model, tým bližšie je jeho krivka k ľavému hornému rohu grafu (teda k bodu $\mathrm{FP} = 0, \mathrm{TP} = 1$). Krivka ROC pre model, ktorý ignoruje dáta a všetkým pozíciám priradí náhodné skóre, vyzerá ako diagonála $\mathrm{TP} = \mathrm{FP}$.

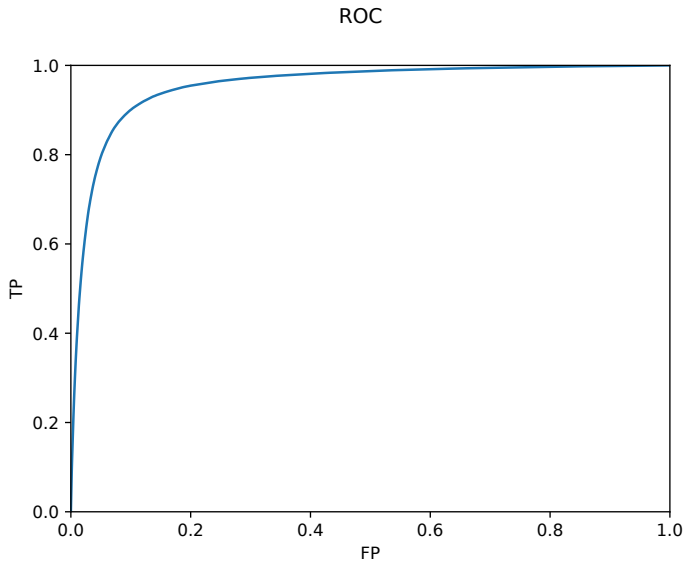
Keď počítame krivky ROC pre naše experimenty, vždy uvažujeme rovnaký prah pre všetky čítania v danom experimente. Neuvažujeme teda možnosť, že by sme v rôznych čítaniach v rámci jedného experimentu použili rôzny prah.

4.2.2 Úspešnosť identifikácie

Pri identifikácii je náš hlavný cieľ trochu iný, než správne pre čo najviac pozícií určiť, či ide o SNPy. Skôr je naším cieľom nájsť medzi všetkými pozíciami tých zopár, ktoré obsahujú SNPy. Tento cieľ sa snažíme reflektovať zavedením druhej metriky, ktorú nazývame *úspešnosť identifikácie*. Úspešnosť identifikácie meriame pre každé čítanie zvlášť.

[1] z anglického ***t**rue **p**ositive*

[2] z anglického ***f**alse **p**ositive*



Obr. 4.1: Príklad krivky ROC.

Definícia 10. Nech s je počet pozícií v čítaní, na ktorých je SNP. Nech i je počet SNPov, ktoré patria medzi s pozícií s najvyšším skóre. *Úspešnosť identifikácie* pre dané čítanie definujeme ako podiel i/s.

Úspešnosť identifikácie je teda jedno číslo z rozsahu 0 až 1. Úspešnosť identifikácie je menej robustná, než krivka ROC. V čítaniach, kde je väčší podiel SNPov, môžeme pri použití rovnakého modelu očakávať vyššiu úspešnosť identifikácie. V extrémnom prípade, keď sú na všetkých pozíciách SNPy, bude mať ľubovoľný model úspešnosť identifikácie 1.

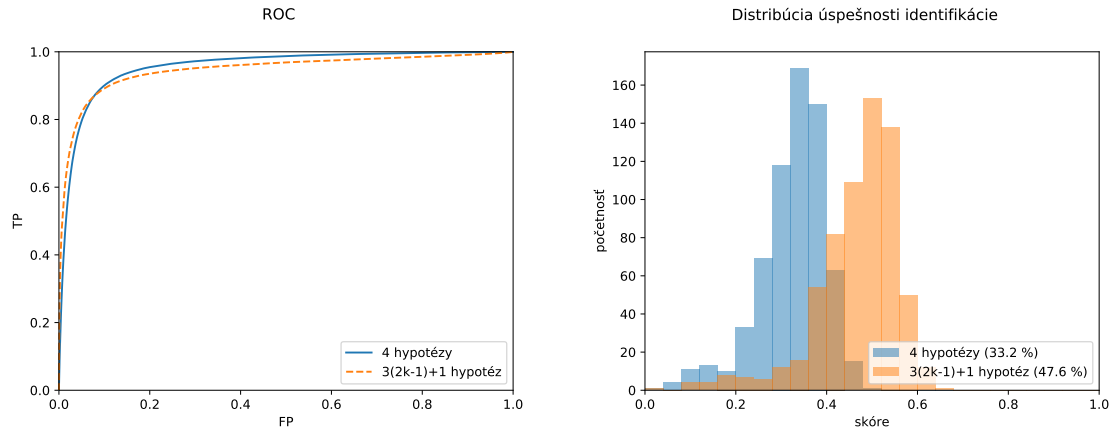
Ak podiel SNPov v čítaní označíme ako p, úspešnosť identifikácie je rovná hodnote TP v bode, kde $\mathrm{TP} \cdot p + \mathrm{FP} \cdot (1-p) = p$. Tento bod sa na krivke ROC typicky nachádza v časti s veľmi malým FP.

Pri vyhodnocovaní experimentu vypočítame úspešnosť identifikácie pre každé čítanie. Z týchto úspešností potom vypočítame vážený priemer, kde váha každého čítania je počet SNPov v ňom. Okrem tohto priemeru uvádzame aj distribúciu úspešnosti identifikácie pre jednotlivé čítania.

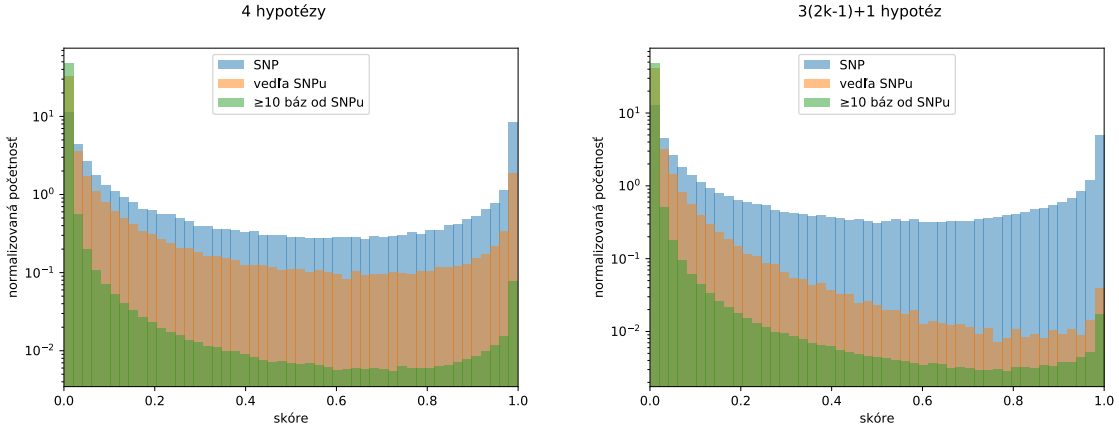
4.3 Výsledky

4.3.1 Porovnanie prístupu so štyroma hypotézami a prístupu s $3(2k-1)+1$ hypotézami

V tejto časti porovnávame verziu nášho modelu popísanú v kapitole 3.3 s verziou popísanou v kapitole 3.5. Obe verzie nášho modelu sme pustili na testovacej sade KLEBS_10



Obr. 4.2: Krivka ROC a distribúcia úspešnosti identifikácie pre prístup so štyroma hypotézami a pre prístup s $3(2k-1)+1$ hypotézami.

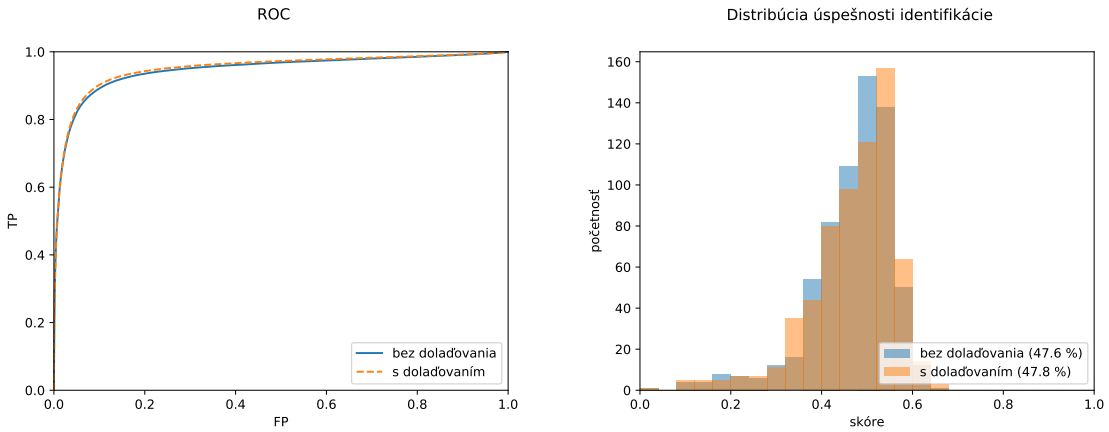


Obr. 4.3: Porovnanie distribúcie skóre pre pozície so SNPom, vedľa SNPu a ďaleko od SNPov pre model s štyroma hypotézami a model s $3(2k-1)+1$ hypotézami.

s parametrom $M = 2$ (minimálny počet hodnôt signálu zarovnaný k jednej báze). Vylepšenia popísané v kapitole 3.6 zatiaľ nepoužívame.

Výsledok testovania úvádzame v grafoch na Obr. 4.2. Prístup s $3(2k-1)+1$ hypotézami má priemernú úspešnosť identifikácie $47,6\%$, prístup so štyroma hypotézami má priemernú úspešnosť $33,2\%$. Na krivke ROC vidíme, že prístup s $3(2k-1)+1$ hypotézami je lepší pre nízke FP, kým prístup so štyroma hypotézami je lepší pre vysoké TP. V ďalších experimentoch používame prístup s $3(2k-1)+1$ hypotézami, keďže má výrazne lepšiu úspešnosť identifikácie.

Hlavným dôvodom zavedenia nových hypotéz bola snaha lepšie vyhodnocovať pozície blízko SNPov. Grafy na Obr. 4.3 ukazujú, že zavedením viacerých hypotéz sa dramaticky znížilo množstvo pozícií tesne vedľa SNPov s vysokým skóre.



Obr. 4.4: Krivka ROC a distribúcia úspešnosti identifikácie bez dolaďovania normalizácie a s dolaďovaním normalizácie.

4.3.2 Vplyv dolaďovania normalizácie

V tomto experimente meriamie vplyv dolaďovania normalizácie navrhovaného v kapitole 3.6.1. Náš model sme dvakrát pustili na testovacej sade KLEBS_10 : raz s dolaďovaním normalizácie a raz bez neho. V oboch prípadoch používame $M = 2$, modelovanie cúvania zatiaľ nepoužívame.

Výsledok testovania uvádzame v grafoch na Obr. 4.4. Bez dolaďovania normalizácie mal model priemernú úspešnosť identifikácie $47,6\%$, s dolaďovaním $47,8\%$. Na krivke ROC vidíme, že pre vysoké TP je model s dolaďovaním mierne presnejší. Keďže dolaďovanie normalizácie mierne zvýšilo presnosť modelu, v ďalších experimentoch používame model s dolaďovaním normalizácie.

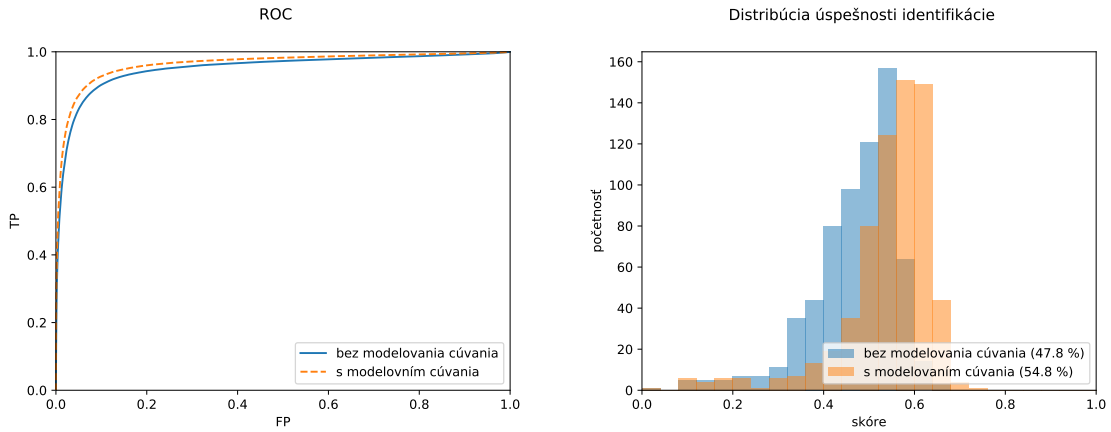
4.3.3 Vplyv modelovania cúvania

Tento experiment skúma vplyv modelovania cúvania navrhovaného v kapitole 3.6.2. Náš model sme opäť pustili testovacej sade KLEBS_10, raz bez modelovania cúvania a raz s ním. Opäť v oboch prípadoch používame $M = 2$.

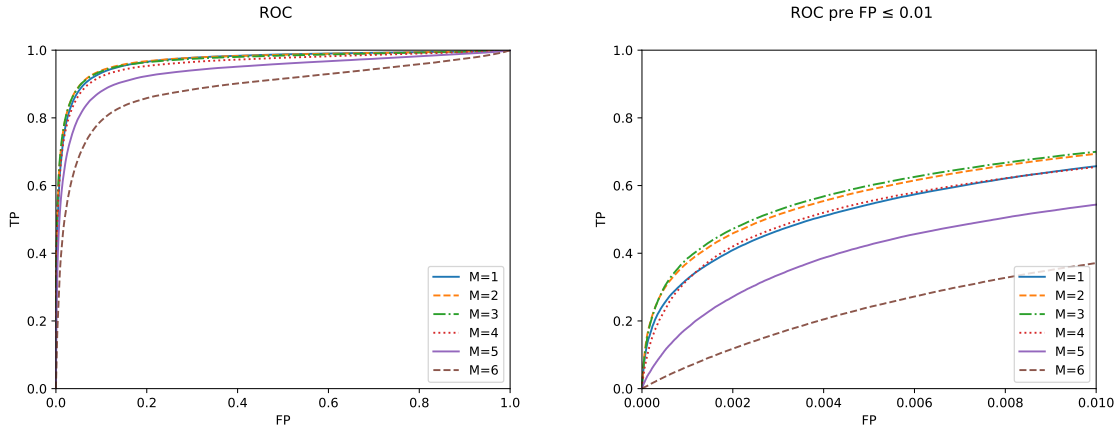
Grafy s výsledkami uvádzame na Obr. 4.5. Bez modelovania cúvania má náš model priemernú úspešnosť identifikácie $47,6\%$, s modelovaním cúvania má priemernú úspešnosť $54,8\%$. Keďže modelovanie cúvania spresňuje náš model, v ďalších experimentoch používame model s modelovanám cúvania.

4.3.4 Vplyv parametra M

V tomto experimente sme náš model pustili postupne pre $M = 1, 2, \ldots, 6$. Stále používame testovaciu sadu KLEBS_10. Krivky ROC uvádzame na Obr. 4.6. Priemernú úspešnosť identifikácie uvádzame v tabuľke 4.2. Vidíme, že náš model je najpresnejší



Obr. 4.5: Krivka ROC a distribúcia úspešnosti identifikácie bez modelovania cúvania a s modelovaním cúvania.



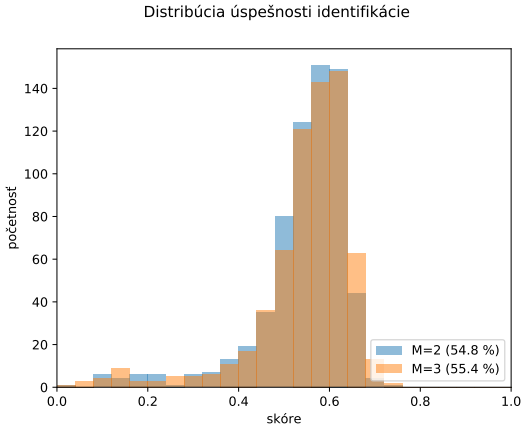
Obr. 4.6: Krivka ROC pre $M = 1, 2, \ldots, 6$. Graf vpravo zobrazuje časť kriviek pre FP $\leq 0,01$.

pre $M = 2$ a $M = 3$. Zvyšovanie M na 4 a viac je už kontraproduktívne. V ďalších experimentoch preto budeme používať iba $M = 2$ a $M = 3$. Distribúciu úspešnosti identifikácie pre $M = 2$ a $M = 3$ uvádzame v grafe na Obr. 4.7.

4.3.5 Porovnanie s priamočiarym prístupom

K identifikácii variantov sa dá pristupovať aj priamočiarejšie, než k nej pristupuje náš model: zo signálu by sme mohli pomocou prekladača báz určiť bázy a následne hľadať rozdiely medzi postupnosťou báz z prekladača a referenciou.

V našej práci sme sa snažili navrhnúť model, ktorý by bol presnejší, než takýto priamočiary prístup. V tejto časti preto navrhneme, ako identifikovať SNPy len na základe referencie a postupnosti, ktorú zo signálu vypočíta prekladač báz. Tento postup potom porovnáme s naším modelom.



Obr. 4.7: Distribúcia úspešnosti identifikácie pre $M = 2$ a $M = 3$.

Tabuľka 4.2: Priemerná úspešnosť identifikácie pre rôzne M

M	1	2	3	4	5	6
úspešnosť identifikácie	$54,3\%$	$57,4\%$	$58,1\%$	$54,5\%$	$45,7\%$	$33,6\%$

Priamočiary prístup

Postupnosť báz, ktorú nám dá prekladač, budeme volať *vzorka*. Keďže prekladač báz nevie, akú dlhú postupnosť sme sekvenovali, dĺžka vzorky sa nemusí zhodovať s dĺžkou zodpovedajúcej časti referencie. Na niektorých miestach nájde prekladač viac báz, než tam reálne je, na iných zasa niektoré bázy nenájde.

Na začiatku preto vzorku pomocou nástroja BWA-MEM zarovnáme k referencii. Toto zarovnanie nám určuje, ktoré časti referencie a vzorky si vzájomne zodpovedajú, pričom niektoré pozície z referencie nemusia zodpovedať žiadnym pozíciám zo vzorky, a obrátene. Zarovnanie v podstate hovorí, ktoré časti referencie a ktoré časti vzorky treba vynechať, aby k sebe to, čo zostane, čo najlepšie pasovalo.

Na základe tohto zarovnaia rozdelíme pozície v referencii na štyri skupiny:

Delécie. Pozície v referencii, ktorým nezodpovedá žiadna pozícia zo vzorky.

Substitúcie. Pozície v referencii, kde sa báza z referencie líši od zodpovedajúcej bázy vo vzorke.

Blízko inzercií. Pozície v referencii, ktorých zodpovedajúca pozícia vo vzorke susedí s časťou vzorky, ktorá nezodpovedá ničomu v referencii. Budeme pritom vyžadovať, aby nešlo o substitúcie.

Zhody: Pozície, kde sa báza v referencii zhoduje s bázou vo vzorke a nie sú blízko inzercií.

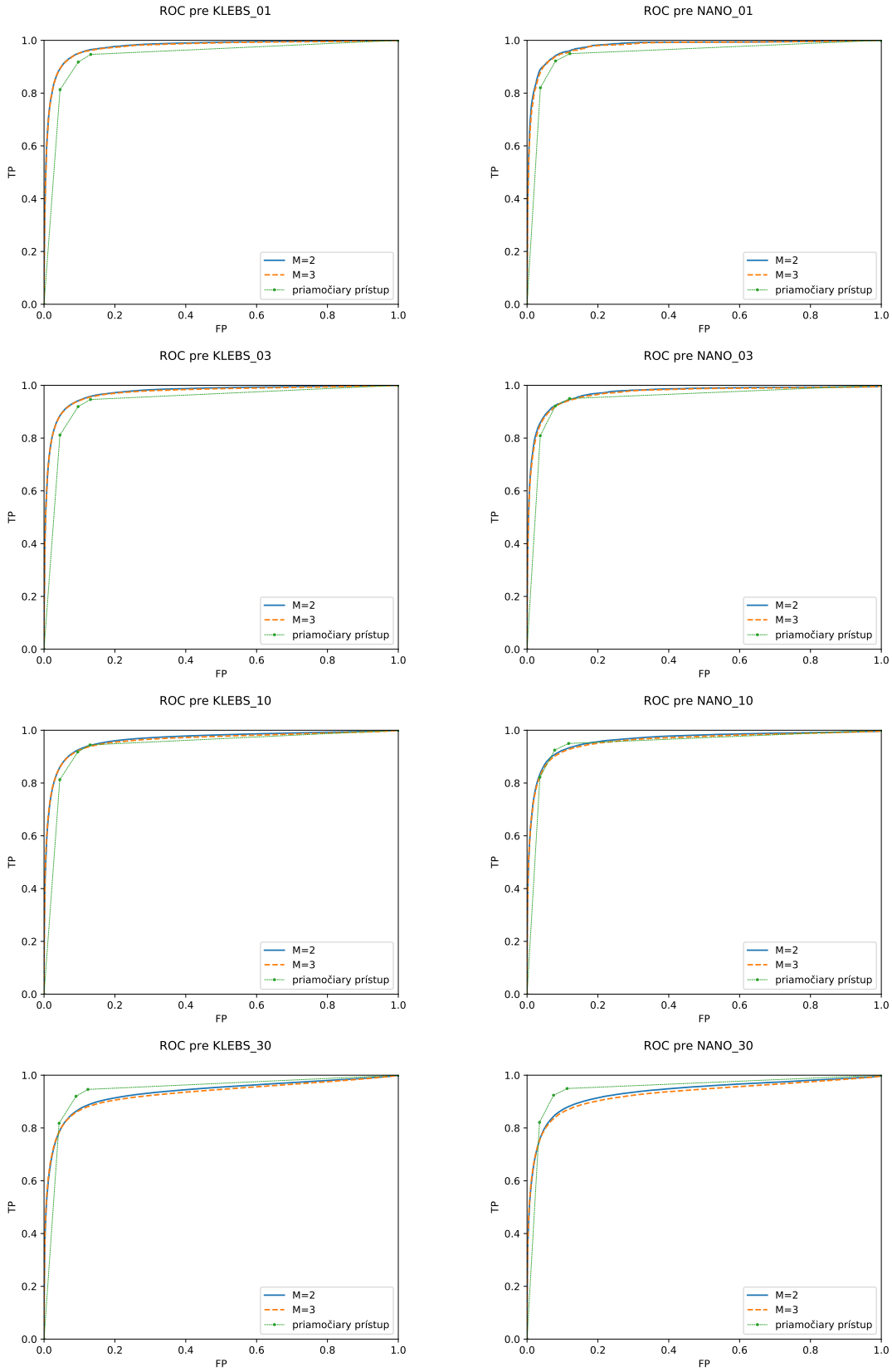
Podobne ako pri pravdepodobnostnom modeli, každej pozícii priradíme skóre vyjadrujúce, ako veľmi si myslíme, že ide o SNP. Toto skóre bude závisieť iba od toho, do ktorej zo štyroch skupín daná pozícia patrí. Budú teda existovať iba štyri možné hodnoty skóre, krivka ROC pre tento prístup teda bude mať iba 5 bodov (pričom jeden z nich bude $\mathrm{TP} = 0, \mathrm{FP} = 0$ a jeden bude $\mathrm{TP} = 1, \mathrm{FP} = 1$).

Tieto štyri možné hodnoty skóre určujeme nasledujúcim spôsobom. Na niekoľkých trénovacích čítaniach sme pre každú zo štyroch skupín G odmerali, aká časť všetkých pozícií, na ktorých je SNP, patrí do skupiny G (toto číslo označme $V(G)$) a aká časť všetkých pozícií bez SNPu patrí do skupiny G (toto číslo označme $N(G)$). Hodnotu $V(G)$ používame ako podmienenú pravdepodobnosť, že by pozícia so SNPom bola v skupine G a hodnotu $N(G)$ používame ako podmienenú pravdepodobnosť, že by pozícia bez variantu bola v skupine G. Pre každú pozíciu potom na základe týchto podmienených pravdepodobností a apriórnej pravdepodobnosti, že na danej pozícii je SNP, vypočítame aposteriórnu pravdepodobnosť, že je na tejto pozícii SNP. Táto aposteriórna pravdepodobnosti bude skóre pre danú pozíciu.

4.3.6 Experimenty

Priamočiary prístup porovnávame s naším modelom pre $M = 2$ a $M = 3$ na všetkých ôsmich testovacích sadách. Pri vyhodnocovaní úspešnosti identifikácie priamočiareho prístupu sa nám stáva, že nevieme jednoznačne vybrať s pozícií s najvyšším skóre, pretože veľa pozícií má rovnaké skóre. V takých prípadoch spomedzi kandidátov s rovnako vysokým skóre vyberáme náhodne.

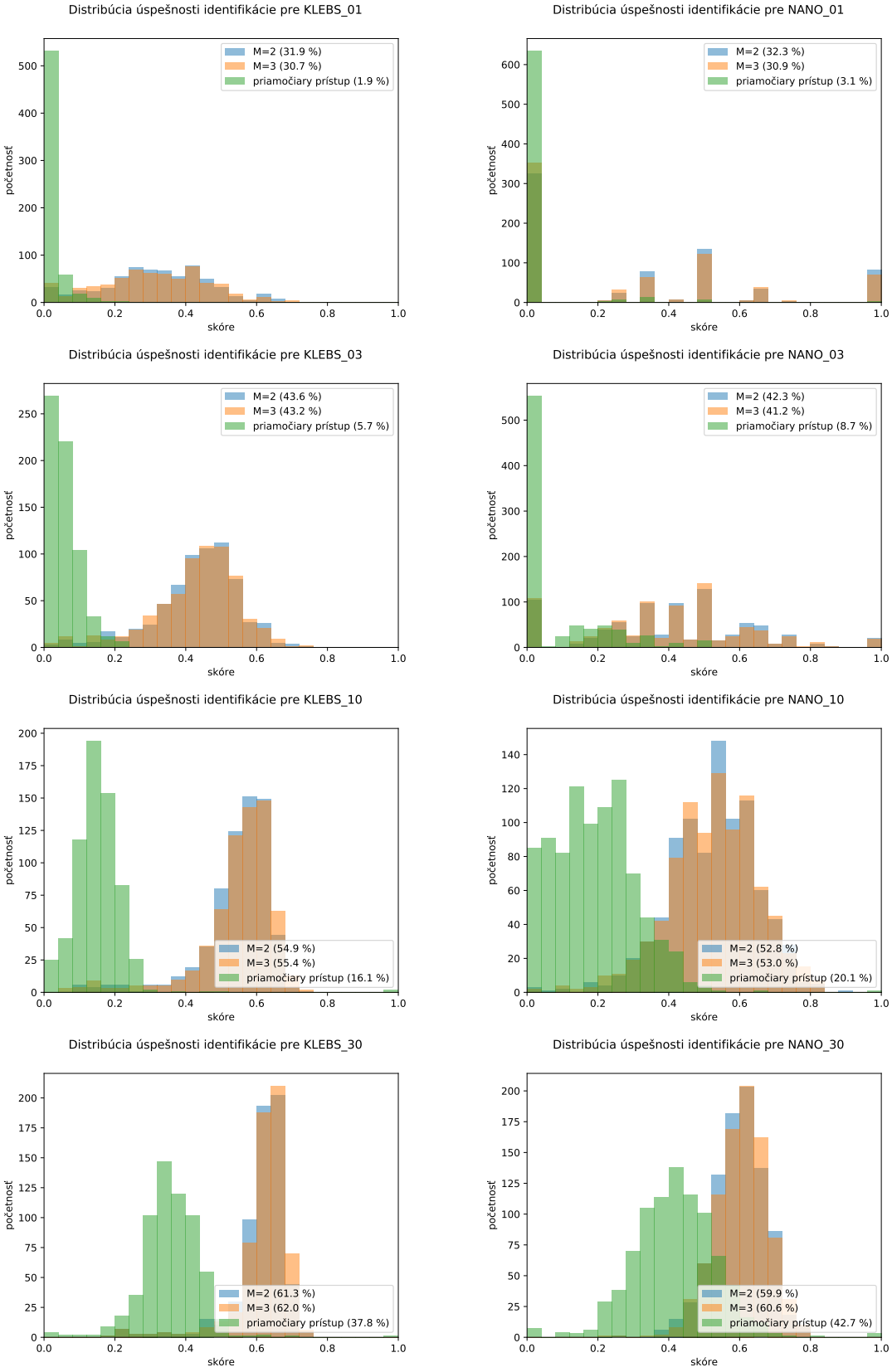
Na krivkách ROC (Obr. 4.8) vidíme, že v testovacích sadách s malým podielom SNPov je náš model lepší než priamočiary prístup. V testovacích sadách s väčším podielom SNPov je náš model lepší pre malé FP a horší pre veľké TP. Vidíme tiež, že kým priamočiary prístup je zhruba rovnako dobrý na všetkých sadách, nášmu modelu sa lepšie darí pri menšom podieli SNPov. Náš model má výrazne lepšiu úspešnosť identifikácie (Tabuľka 4.3 a Obr. 4.9) než priamočiary prístup. Pri vyšších podieloch SNPov dosahuje náš model aj priamočiary prístup vyššiu úspešnosť identifikácie, čo sa však pri tejto metrike dalo očakávať.



Obr. 4.8: Krivky ROC pre priamočiary prístup a pre náš model s $M = 2$ a $M = 3$.

Tabuľka 4.3: Priemerná úspešnosť identifikácie pre priamočiary prístup a pre náš model s $M = 2$ a $M = 3$.

testovacia sada	priamočiary prístup	$M = 2$	$M = 3$
KLEBS_01	$2,1\%$	$31,9\%$	$30,7\%$
KLEBS_03	$5,7\%$	$43,6\%$	$43,2\%$
KLEBS_10	$16,1\%$	$54,9\%$	$55,4\%$
KLEBS_30	$37,8\%$	$61,3\%$	$62,0\%$
NANO_01	$3,1\%$	$32,3\%$	$30,9\%$
NANO_03	$8,7\%$	$42,3\%$	$41,2\%$
NANO_10	$20,1\%$	$52,8\%$	$53,0\%$
NANO_30	$42,7\%$	$59,9\%$	$60,6\%$



Obr. 4.9: Distribúcia úspešnosti identifikácie pre priamočiary prístup a pre náš model s $M = 2$ a $M = 3$.

Záver

Cieľom tejto práce bolo navrhnúť nový algoritmus na identifikáciu variantov zo surových dát sekvenačného prístroja MinION, ktorý by bol presnejší, než porovnanie referencie s výstupom z prekladača báz. Toto spresnenie sme sa snažili dosiahnuť využitím informácie, ktorú prekladač báz nevyužíva: že sekvenovaná postupnosť sa musí podobať na referenciu.

Vytvorili a implementovali sme pravdepodobnostný model, ktorý identifikuje SNPy na základe signálu z jedného čítania. V ideálnych podmienkach, teda keď sekvenovaná postupnosť neobsahuje žiadne varianty okrem malého množstva SNPov a dopredu vieme, koľko SNPov máme očakávať, je náš model pri testovaní presnejší, než porovnanie výstupu z prekladača báz s referenciou. Hoci k presnej identifikácii variantov všetkých druhov máme ešte ďaleko, tento výsledok vzbudzuje nádej, že využitie znalosti referencie pri spracovaní signálu môže naozaj spresniť identifikáciu variantov.

Pravdepodobnostný model, ktorý sme navrhli, by sa dal v budúcnosti rozšíriť dvoma spôsobmi. Môžeme ho rozšíriť na model, ktorý identifikuje SNPy aj z viacerých čítaní jednej sekvencie. Pre niektoré hypotézy dostaneme podmienené pravdepodobnosti z viacerých čítaní. Aposteriórnu pravdepodobnosť danej hypotézy budeme potom počítať z apriórnej pravdepodobnosti a všetkých týchto podmienených pravdepodobností.

Ďalšou možnosťou je rozšíriť model, aby bolo pomocou neho možné identifikovať aj krátke inzercie a delécie. Stačí pre jednotlivé pozície začať uvažovať aj hypotézy hovoriace, že na danej pozícii je inzercia, prípadne delécia. Rozšírenie na dlhšie inzercie je problematické, lebo s dĺžkou inzercie exponenciálne rastie počet možností, aké bázy boli do postupnosti vložené a tým aj počet hypotéz.

Presnejšie, než rozšírenia nášho modelu, však možno budú fungovať úplne iné prístupy. Najlepšie súčasné prekladače báz pre prístroj MinION sú založené na neurónových sieťach, je teda dosť dobre možné, že aj identifikácia variantov sa bude dať riešiť presnejšie pomocou neurónovej siete.

Literatúra

[1] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE*, 12(6):1–13, 06 2017.

[2] Phelim Bradley et al. Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis. *Nat Commun*, 6:10063, Dec 2015.

[3] Daniel Branton et al. The potential and challenges of nanopore sequencing. *Nat Biotechnol*, 26(10):1146–1153, Oct 2008.

[4] Matei David et al. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, 33(1):49–55, 2017.

[5] Thomas Laver et al. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular detection and quantification*, 3:1–8, 2015.

[6] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*, March 2013.

[7] Charles E Metz. Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[8] Yannick Rondelez, 2018. Osobná komunikácia.

[9] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, Feb 1978.

[10] Stan Salvador and Philip Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. *KDD Workshop on Mining Temporal and Sequential Data*, pages 70–80, 2004.

[11] Marcus H Stoiber et al. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv* 094672, 2017.

[12] Oxford Nanopore Technologies. kmer_models. `https://github.com/nanoporetech/kmer_models`.

[13] Oxford Nanopore Technologies. Re-squiggle algorithm. `https://nanoporetech.github.io/tombo/resquiggle.html#algorithm-details`.

[14] Oxford Nanopore Technologies. Tombo is a suite of tools primarily for the identification of modified nucleotides from raw nanopore sequencing data. `https://github.com/nanoporetech/tombo`.

[15] Haotian Teng et al. Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. *bioRxiv* 179531, 2017.

[16] Ryan Wick. A comparison of different Oxford Nanopore basecallers. `https://github.com/rrwick/Basecalling-comparison`.

Príloha A

Implementácia nášho postupu je na priloženom CD.