VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

SIMULÁTOR ČTENÍ PRO BAKTERIÁLNÍ RNA-SEQ

READ SIMULATOR FOR BACTERIAL RNA-SEQ

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — Adéla Fialová

AUTHOR

VEDOUCÍ PRÁCE — doc. Mgr. Ing. Karel Sedlář, Ph.D.

SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Adéla Fialová ***ID:*** 240499

Ročník: 3 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Simulátor čtení pro bakteriální RNA-Seq

POKYNY PRO VYPRACOVÁNÍ:

1) Prostudujte princip exprese genetické informace u bakterií a způsob jejího měření se zaměřením na techniky využívající sekvenaci DNA/RNA. 2) Vypracujte literární rešerši o nástrojích pro simulaci sekvenačních dat. 3) Data generovaná vybraným nebo vybranými nástroji srovnejte s experimentálně naměřenými daty pro libovolnou bakterii s publikovaným transkriptomem. 4) Navrhněte a ve vhodném jazyce implementujte vlastní metodu pro simulaci RNA-Seq dat, přičemž bude postačovat, pokud bude generovat pouze single-end čtení. 5) Navržený výpočetní nástroj doplňte o funkce zohledňující specifika exprese genetické informace u bakterií, především existenci operonů. 6) Srovnejte navržený simulátor s nástroji již existujícími i s reálnými daty a výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] FRAZEE, Alyssa C., Andrew E. JAFFE, Ben LANGMEAD a Jeffrey T. LEEK. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015, 31(17), 2778-2784

[2] GERARD, David. Data-based RNA-seq simulations by binomial thinning. BMC Bioinformatics, 2020, 21(1), 206

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 29.5.2024

Vedoucí práce: doc. Mgr. Ing. Karel Sedlář, Ph.D.

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

UPOZORNĚNÍ:

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

ABSTRAKT

Tato práce se zaměřuje na analýzu bakteriálního genomu a jeho exprese prostřednictvím technologie RNA-Seq, s důrazem na simulaci RNA-Seq dat. První část práce poskytuje teoretický základ o struktuře bakteriálního genomu, jeho expresi a metodách jeho studia, včetně moderních sekvenačních technik. Pozornost je věnována i několika vybraným simulátorům RNA-Seq dat. V druhé části je představena vlastní implementace simulátoru, navržená pro zohlednění charakteristik bakteriálního genomu, především přítomnosti operonů.

KLÍČOVÁ SLOVA

RNA-Seq, RNA, simulace, sekvenování, bakterie, genová exprese

ABSTRACT

This thesis is dedicated to the analysis of the bacterial genome and its expression through RNA-Seq technology, with emphasis on the simulation of RNA-Seq data. The first part of the thesis provides a theoretical background on the structure of the bacterial genome, its expression and methods to study it, including modern sequencing techniques. Several selected simulators of RNA-Seq data are also mentioned. In the second part, the actual simulator implementation, designed to account for the characteristics of the bacterial genome, especially the presence of operons, is presented.

KEYWORDS

RNA-Seq, RNA, simulation, sequencing, bacteria, gene expression

Vysázeno pomocí balíčku `thesis` verze 4.09; `https://latex.fekt.vut.cz/`

FIALOVÁ, Adéla. *Simulátor čtení pro bakteriální RNA-Seq*. Bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024. Vedoucí práce: doc. Mgr. Ing. Karel Sedlář, PhD.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Adéla Fialová
VUT ID autora:	240499
Typ práce:	Bakalářská práce
Akademický rok:	2023/24
Téma závěrečné práce:	Simulátor čtení pro bakteriální RNA-Seq

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

Jako autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucímu práce panu doc. Mgr. Ing. Karlu Sedlářovi, Ph.D. za odborné vedení, konzultace, trpělivost a ochotnou pomoc při psaní práce. Za trpělivost a podporu bych ráda poděkovala i všem svým blízkým.

Obsah

Seznam obrázků

Úvod

RNA-Seq představuje revoluční metodu v měření genové exprese a molekulární biologii, která umožňuje podrobně analyzovat a kvantifikovat transkriptomy organismů na velmi přesné úrovni. Tato technologie poskytuje bohatý přehled o tom, jaké geny jsou v daném čase a v daném buněčném typu aktivní, a umožňuje tak hlubší porozumění biologickým funkcím a regulacím.

Simulace RNA-Seq dat se proto staly významným prvkem bioinformatického výzkumu a vývoje. Důvodů pro simulaci dat RNA-Seq je několik. Díky simulacím je možné provádět komplexní testy na různých algoritmech bez potřeby reálných biologických dat, což může významně snížit náklady a zjednodušit experimentální plánování. Kromě testování softwaru simulace také napodobují různé experimentální scénáře, což umožňuje předem ověřit, jaké faktory mohou ovlivnit výsledky studií a jak optimálně nastavit parametry experimentů.

Navzdory pokrokům v oblasti simulací RNA-Seq existuje však mezera ve specifikách simulace bakteriálních transkriptomů. Bakteriální genomy se vyznačují několika unikátními rysy, jako je přítomnost operonů – skupin genů transkribovaných jako jediná molekula mRNA, což může významně ovlivnit interpretaci dat.

V minulosti bylo vyvinuto několik simulátorů RNA-Seq, z nichž každý nabízí jedinečné funkce a možnosti. Všechny tyto simulátory mají však jeden společný nedostatek - žádný z nich nenabízí zohlednění specifik bakteriálního genomu, zejména výskytu operonů v bakteriálním genomu nebo různých délek nepřekládaných oblastí. Proto je cílem práce vyvinout simulátor RNA-Seq, který by reflektoval unikátní aspekty bakteriálních genomů. Implementace takového nástroje by poskytla nástroj pro analýzu a interpretaci bakteriální genové exprese a umožnila by lépe pochopit, jak bakteriální geny interagují a regulují se v rámci operonů.

První kapitola práce se věnuje popisu bakterií, jejich genomu a specifikům jako jsou operony a plazmidy. Detailně je zde popsána exprese genetické informace. Další kapitola poskytuje přehled moderních sekvenačních platforem, jako jsou Illumina, IonTorrent, Oxford Nanopore a Pacific Bioscience, a jejich využití ve studiu genetické exprese. Ve třetí kapitole jsou popsány různé metody měření genetické exprese, včetně microarrays, RT-qPCR a RNA-Seq. Přehled vybraných nástrojů pro simulaci RNA-Seq dat a popis rozdílů mezi nimi je uveden ve čtvrté kapitole. Pátá kapitola práce je věnována vývoji a implementaci vlastního simulátoru RNA-Seq dat, BacSeqeru, který zohledňuje specifika bakteriálního genomu, zejména existenci operonů. V poslední kapitole jsou diskutovány výsledky a možnosti dalšího vývoje simulátoru pro zvýšení jeho efektivity a užitečnosti.

Záměrem práce je tedy poskytnout teoretický přehled o bakteriálním genomu, způsobech měření exprese genetické informace, především metody RNA-Seq, o ná-

strojích sloužících k simulaci RNA-Seq dat, a v neposlední řadě provést implementaci vlastního nástroje zaměřujícího se na specifika bakteriálního genomu, konkrétně výskytu operonů.

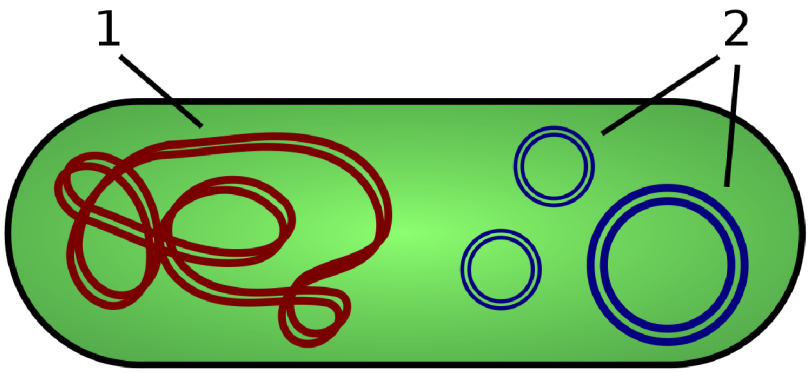
1 Bakterie

Bakterie představují rozmanitou a adaptabilní skupinu jednobuněčných organismů, které hrají důležitou roli v mnoha ekologických procesech a vyskytují se téměř v každém prostředí planety Země. [1]

Bakterie, jako prokaryotické organismy, mají několik charakteristických rysů, které je odlišují od eukaryot [2]. Jedním z hlavních rozdílů je absence obaleného jádra. Na rozdíl od eukaryot, kde je genetický materiál uchován v jádru obklopeném dvojitou membránou, je bakteriální DNA volně uložena v cytoplazmě v oblasti zvané nukleoid. Tento nukleoid není obklopen membránou. Dalším významným rozdílem je absence některých membránových organel v prokaryotických buňkách. Bakterie typicky nemají složité organely jako jsou mitochondrie nebo endoplazmatické retikulum. Pro většinu biochemických procesů bakterie používají svou buněčnou membránu a cytoplazmu. Většina bakterií má také charakteristickou buněčnou stěnu, která chrání buňku a udržuje její tvar.

1.1 Bakteriální genom

Genetická informace bakterií je uložena buď ve formě chromozomu, nebo na extrachromozomálních útvarech. Velikost bakteriálního genomu je variabilní, a to od tisíců až po miliony párů bází (bp - base pairs) [3]. Například, *Mycoplasma genitalium* má jeden z nejmenších známých bakteriálních genomů s přibližně 0.58 milionu bp, což odpovídá přibližně 500 genům [4]. Na druhé straně, větší bakterie jako *Sorangium cellulosum* mají genom o velikosti až 13 milionů bp [5].



Obr. 1.1: Uspořádání bakterálního genomu. 1 - Bakteriální DNA. 2 - Plazmid. [6]

1.1.1 Nukleoid

Nukleoid [7] představuje intracelulární oblast bakteriální buňky, která je zásadní pro udržení a funkci genetického materiálu. Zatímco eukaryotické buňky obsahují jádro obklopené membránou, prokaryoty takové jádro nemají. Místo toho je jejich DNA (deoxyribonukleová kyselina) lokalizována v oblasti zvané nukleoid. Tato oblast sice není ohraničena membránou, ale je vizuálně rozlišitelná mikroskopicky díky své husté struktuře.

Nukleoid obvykle obsahuje jediný kruhový chromozom, který je složen z dvouvláknové DNA. Tato DNA je mnohem kompaktnější než u eukaryot, což je důsledkem superkondenzace pomocí proteinů, které se vážou na DNA. U některých bakterií, jako je *Borrelia burgdorferi*, se může vyskytovat i lineární nukleoid [8].

Nukleoid obsahuje enzymy nezbytné pro replikaci DNA a transkripci RNA. DNA polymeráza, která katalyzuje syntézu nových DNA řetězců během replikace, a RNA polymeráza, zodpovědná za syntézu RNA z DNA, jsou lokalizovány právě v nukleoidu [9]. Kromě polymeráz nukleoid obsahuje řadu dalších molekul, jako jsou RNA (ribonukleové kyseliny) a různé regulační proteiny, které modulují transkripci a ostatní genetické procesy. Tyto komponenty hrají významnou roli v regulaci genové exprese, umožňují bakteriím adaptaci na různé podmínky a zajišťují přežití v proměnlivých podmínkách.

1.1.2 Plazmidy

Plazmidy jsou malé, autonomní molekuly DNA, které se běžně vyskytují ve vnějším chromozomálním prostoru bakterií. Mají zpravidla kruhový tvar, existují ale i lineární formy. Plazmidy jsou specifické svou schopností replikace nezávisle na bakteriálním chromozomu, což jim umožňuje reprodukovat se a přenášet mezi buňkami bez ovlivnění hlavního genomu hostitele. [7]

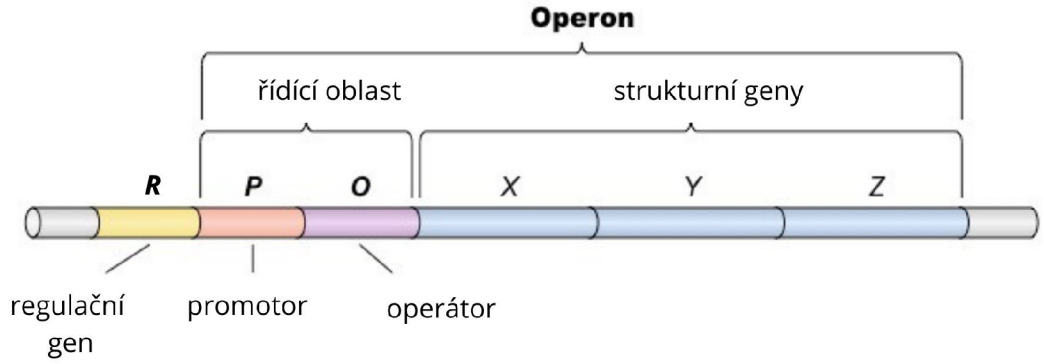
Plazmidy nejsou pro základní přežití bakterií nezbytné, ale mohou poskytovat významné selektivní výhody v určitých prostředích. Například, mnohé plazmidy obsahují geny, které kódují enzymy schopné rozkládat neobvyklé nebo toxické organické sloučeniny [10]. Další důležitou funkcí plazmidů je jejich role v antimikrobiální rezistenci. Geny odpovědné za odolnost proti antibiotikům [11], které jsou často lokalizovány na plazmidech, umožňují bakteriím přežít v prostředích s vysokými koncentracemi antimikrobiálních látek.

Jednou z nejzajímavějších vlastností plazmidů je jejich schopnost být přenášeny mezi buňkami prostřednictvím procesu zvaného konjugace [12]. Tento proces umožňuje plazmidům přesunout se z jedné bakterie do druhé a šířit tak genetické informace napříč bakteriálními populacemi, což zahrnuje i přenos genů pro antibiotickou rezistenci.

1.1.3 Operony

Bakteriální genom je charakteristický výskytem operonů [13]. Operony představují jedinečný a účinný způsob regulace genové exprese u bakterií, který umožňuje současnou aktivaci nebo deaktivaci skupiny genů reagujících na specifické environmentální signály. Tyto genetické struktury jsou zásadní pro adaptabilitu a přežití bakterií v různých podmínkách.

Operony mají velmi specifickou strukturu, která umožňuje koordinovanou regulaci genů. Společný promotor slouží jako místo vázání RNA polymerázy a iniciování transkripce, zatímco operátor je specifickým místem, kde se vážou regulační proteiny, které mohou stimulovat nebo potlačit transkripci celé skupiny genů v operonu. To umožňuje bakteriím rychle reagovat na změny v prostředí. Struktura operonu je znázorněna na obrázku 1.2.

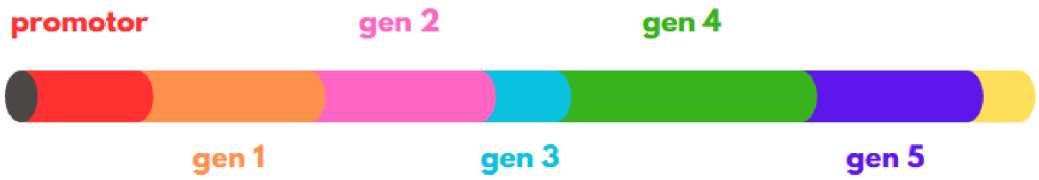


Obr. 1.2: Schéma uspořádání operonu [14]. Upraveno.

Operony mohou být rozděleny na katabolické, které se zapojují do rozkladu substrátů pro získání energie a základních stavebních bloků, a anabolické, které se podílejí na syntéze komplexních molekul, jako jsou aminokyseliny a nukleotidy. Příklad katabolického operonu je lac operon, který reguluje rozklad laktózy v *E. coli*. Tryptofanový operon [15] je příkladem anabolického operonu, neboť zajišťuje syntézu tryptofanu. Obsahuje geny nezbytné pro biosyntézu této esenciální aminokyseliny. Regulace tryptofanového operonu závisí na koncentraci tryptofanu v okolním prostředí, což bakteriím umožňuje účinně přizpůsobit svou schopnost syntetizovat tryptofan aktuálním potřebám.

Výsledným produktem transkripce operonu je polycistronní RNA [16], znázorněná na obrázku 1.3. Tato polycistronní RNA je použita jako šablona pro syntézu několika různých proteinů, což je zásadní rozdíl oproti eukaryotickému systému, kde jedna mRNA obvykle nese informaci pro syntézu jediného proteinu. Každý z genů v operonu má v rámci polycistronní mRNA vlastní ribozomové vazebné místo,

které umožňuje ribozomům navázat a zahájit translaci specifického segmentu RNA. Toto umožňuje nezávislou syntézu proteinů z různých částí téže mRNA. Expresní mechanismy polycistronních RNA umožňují bakteriím efektivně regulovat produkci proteinů v závislosti na metabolických potřebách a okolních podmínkách. Například, v odpovědi na nutriční signály může dojít ke změnám ve vazbě ribozomů a efektivitě translace jednotlivých genů v operonu.

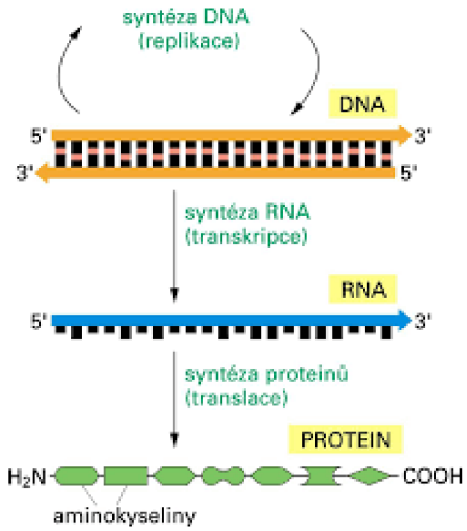


Obr. 1.3: Schéma polycistronní RNA.

1.2 Exprese genetické informace

Genetická exprese, neboli přenos genetické informace obsažené v DNA do funkčních produktů, umožňuje buňkám vykonávat své životní funkce. Zahrnuje tři hlavní fáze: replikaci, kdy z jediné molekuly DNA vzniknou dvě identické kopie, transkripci, kdy se z DNA vytváří messenger RNA (mRNA), a translaci, během které se mRNA používá jako šablona pro syntézu proteinů. Tyto proteiny následně plní různé strukturní a enzymatické funkce v buňce. Fáze exprese genetické informace jsou znázorněny na obrázku 1.4.

Bakterie se vyznačují jednoduchým, ale efektivním mechanismem genové exprese, který jim umožňuje rychle reagovat na změny v prostředí. Díky přítomnosti operonů mohou bakterie koordinovaně regulovat exprese více genů. Absence intronů [17] v bakteriální DNA dále zjednodušuje proces genové exprese a zvyšuje rychlost syntézy proteinů. To bakteriím poskytuje značnou výhodu v prostředích, kde jsou zdroje omezené a rychlá adaptace je klíčem k přežití.



Obr. 1.4: Schéma fází exprese genetické informace. [18].

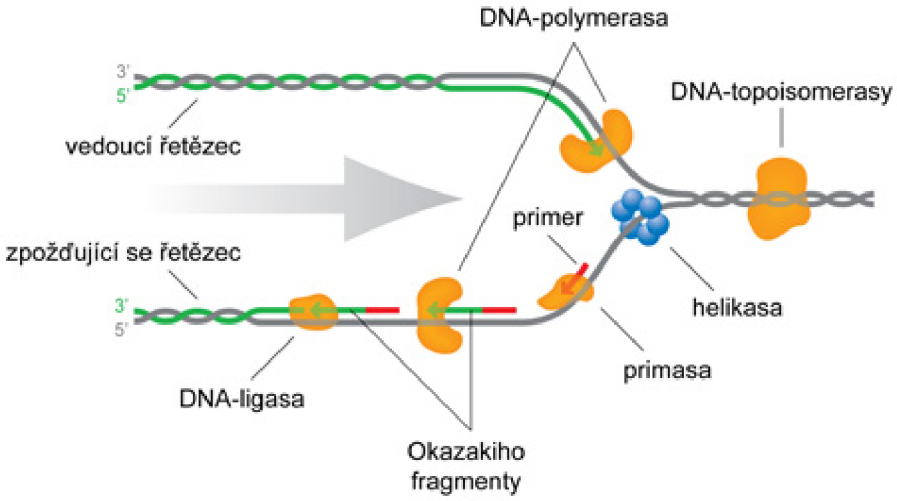
1.2.1 Replikace

Replikace [19, 20] slouží k přenosu genetické informace z mateřské molekuly DNA do dceřiné. Probíhá ve třech fázích – iniciace, elongace, terminace. Fáze iniciace začíná od místa ori („origin of replication“), kde se vytvoří replikační vidlice – místo v DNA, kde dochází k rozestupu komplementárních řetězců v důsledku uvolnění vodíkových vazeb pomocí enzymu DNA helikázy.

Následně se vstupuje do fáze elongace. Enzym primáza syntetizuje RNA primery – krátké úseky RNA sloužící jako startovací místa pro enzym DNA polymerázu. DNA polymeráza se naváže k primeru a začne syntetizovat nový řetězec DNA připojováním nukleotidů ke stávajícím řetězcům na základě komplementarity bází. Elongace probíhá jak ve směru od 5‘ konce k 3‘ konci, což dává za vznik kontinuálnímu řetězci, tak v opačném směru, což vede k tvorbě Okazakiho fragmentů, diskontinuálního řetězce. Okazakiho fragmenty jsou později spojeny DNA ligázou.

Poslední fází je terminace. Při dosažení specifického bodu na chromozomu zvaného „terminus“ dojde k ukončení replikace. Existují dva hlavní mechanismy terminace. Prvním je „theta replikace“ vyskytující se u kruhové bakteriální DNA. Dochází k oddělení dvou replikovaných kruhů posunem replikační vidlice. Druhým mechanismem je „rolling-circle replikace“ lineární bakteriální DNA [21]. V tomto případě dochází k odtržení jednoho konce replikované DNA a jeho obtáčení kolem mateřské kruhové DNA. Vytvoří se tak dlouhé lineární vlákno, které se po dokončení syntézy uzavře do kruhu.

Schéma průběhu replikace je znázorněno na obrázku 1.5.



Obr. 1.5: Schéma replikace [22].

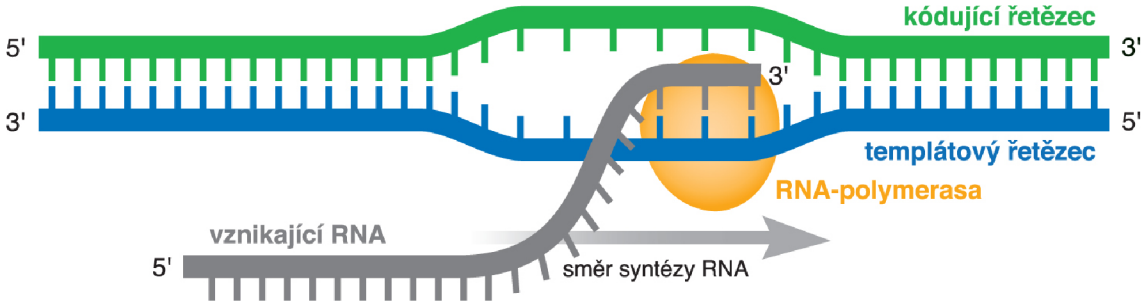
1.2.2 Transkripce

Prvním krokem k realizaci genetické informace je transkripce [19]. Transkripce u bakterií převádí DNA do RNA a stejně jako replikace probíhá ve třech stejnojmenných fázích – iniciace, elongace a terminace.

Hlavním aktérem iniciační fáze je enzym RNA polymeráza. RNA polymeráza rozezná pomocí enzymové podjednotky zvané sigma faktor startovní sekvenci nukleotidů, tzv. promotor. RNA polymeráza spolu se sigma faktorem a promotorem tvoří iniciační komplex. Vlákna DNA jsou rozpojena metodou „negative supercoiling“, kdy jsou vlákna stáčena v opačném směru, než bylo jejich původní uspořádání [23]. Po rozpoznání promotoru a rozvázání vodíkových můstků následuje elongace.

Fáze elongace označuje prodlužování řetězce RNA. RNA polymeráza se posouvá po řetězci a vytváří oblast zvanou transkripční bublina, ve které je prováděna syntéza RNA, znázorněno na obrázku 1.6. RNA polymeráza přidává nukleotidy do rostoucího řetězce na základě komplementarity k jednomu z vláken, kdy je místo nukleotidu T (thymin) připojován nukleotid U (uracil). Vždy dochází k přepisu pouze jednoho vlákna – pracovního vlákna. Druhé vlákno pro transkripci význam nemá – vlákno paměťové.

Terminace transkripce probíhá dvěma možnými způsoby. Prvním je terminace závislá na specifickém proteinu Rho faktoru [24]. Ve chvíli, kdy RNA polymeráza dosáhne sekvence zvané Rho-terminátor, naváže se Rho faktor na nově syntetizovanou RNA. Poté Rho faktor migruje po RNA řetězci směrem k RNA polymeráze a tím pomáhá urychlit oddělení RNA od DNA. Při terminaci nezávislé na Rho hrají důležitou roli terminátorové sekvence na konci genu. Tyto sekvence obsahují palindromické úseky (sekvence identické při čtení zleva i zprava) vytvářející specifickou



Obr. 1.6: Schéma transkripce [22].

sekundární strukturu, při které se RNA řetězec skládá do smyčky. Tato struktura tvoří slabé vazby mezi RNA a DNA destabilizující vazbu RNA polymerázy k DNA, což vede k uvolnění nově vytvořené mRNA (mediátorová RNA).

Transkripce bakteriálního genomu se vyznačuje jednodušším a efektivnějším průběhem ve srovnání s eukaryotickými organismy, což lze vysvětlit několika klíčovými faktory. Prvním z těchto faktorů je lokalizace procesů transkripce a translace v bakteriální cytoplazmě, díky čemuž mohou tyto dva klíčové kroky probíhat souběžně. Dále je podstatným aspektem skutečnost, že bakterie využívají pouze jeden typ RNA polymerázy, na rozdíl od eukaryot, které disponují třemi různými typy tohoto enzymu. Existence operonů, skupin genů, které sdílejí společný promotor a kódují jeden genový produkt, představuje další klíčový prvek. Je tak umožněno přepisovat více genů do jedné polycistronní molekuly RNA. V neposlední řadě je nutno zmínit vzácnost přítomnosti intronů, která eliminuje potřebu posttranskripčních úprav a umožňuje rychlejší a plynulejší průběh genové exprese.

1.2.3 Translace

Translace genetické informace [19], znázorněná na obrázku 1.7, umožňuje přeměnu informace zapsané v RNA na funkční proteiny, nezbytné pro různé buněčné funkce a procesy. Tento proces se odehrává v ribozomech.

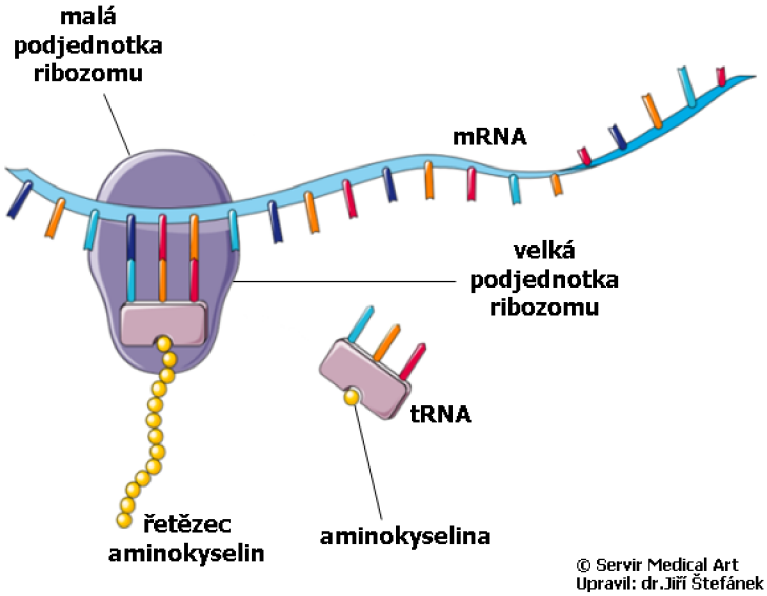
V první fázi translace, iniciaci, dochází k sestavení ribozomálního komplexu. Tento krok začíná vázáním menší ribozomální podjednotky na mRNA v blízkosti speciálního startovacího kodonu, typicky AUG, který kóduje aminokyselinu methionin. K tomuto komplexu se připojí tRNA (transferová RNA) s methioninem, která rozpozná startovací kodon. Následně se k sestavenému komplexu připojí větší ribozomální podjednotka, čímž se vytvoří kompletní funkční ribozom připravený k syntéze proteinu.

Během elongace, druhé fáze translace, se ribozom posouvá po mRNA, čte její kodony a připojuje odpovídající aminokyseliny, které jsou donášeny specifickými tRNA molekulami. Každá tRNA nese specifickou aminokyselinu, která odpovídá

kodonu na mRNA. Aminokyseliny jsou spojovány peptidovými vazbami, čímž vzniká rostoucí polypeptidový řetězec. Proces syntézy proteinu pokračuje, dokud ribozom nenarazí na stop kodon, který signalizuje konec translace. Tyto stop kodony (UAA, UAG, UGA) nemají odpovídající tRNA a způsobují, že ribozomální komplex se rozpadne a uvolní syntetizovaný protein.

Třetí fáze, terminace, zahrnuje uvolnění kompletního polypeptidu z ribozomu a rozpad ribozomálního komplexu. Po syntéze protein podstupuje různé post-translační modifikace, které mohou zahrnovat přidání sacharidů, fosfátů nebo jiných molekul, které modifikují a aktivují protein, nebo jej adresují do specifických částí buňky.

Translace je úzce kontrolována a regulována různými buněčnými signály a mechanismy [25], které zajišťují, že proteiny jsou produkovány v pravý čas a v potřebných množstvích. Poruchy v translaci mohou vést ke zdravotním problémům, včetně rakoviny, dědičných onemocnění a infekcí.



Obr. 1.7: Schéma translace [26].

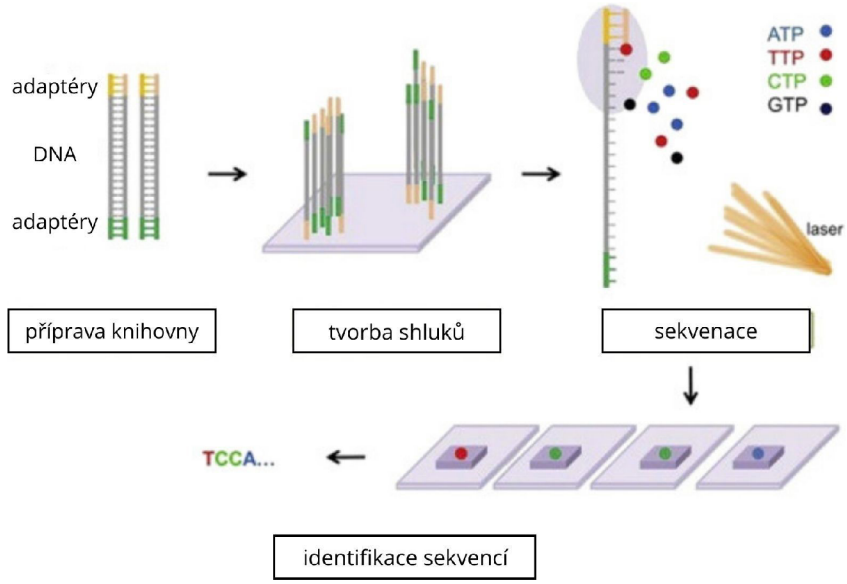
2 Sekvenační platformy

Sekvenování DNA je základní metodou molekulární biologie a genetiky, která umožňuje získat přesné informace o sekvenci nukleotidů v DNA nebo RNA. Znalost pořadí nukleotidů poskytuje informace o dědičných a biochemických faktorech organismu [27]. Metody sekvenování se v průběhu let výrazně vyvinuly, od původně manuálních technik, jako je metoda Sanger, až po moderní technologie sekvenování nové generace (NGS), které umožňují analyzovat celé genomy rychle a s vysokou přesností.

Tato kapitola poslouží jako nezbytný rámec pro porozumění metodě RNA-Seq, která je v následující části podrobně rozebrána.

2.1 Illumina

Technologie sekvenování Illumina [28] je předním přístupem v sekvenování nové generace a umožňuje provádět paralelní analýzu DNA a RNA. Tato metoda se vyznačuje schopností generovat obrovské množství krátkých čtení, typicky o délce 100-150 párů bazí (bp).



Obr. 2.1: Schéma sekvenace pomocí platformy Illumina [45]. Upraveno.

Proces začíná rozštěpením vstupního vzorku DNA na krátké fragmenty. Tyto fragmenty jsou následně ligovány k adaptérům, které umožňují fragmentům připevnit se na speciálně připravenou destičku - flow cell. Tato destička umožňuje, aby se fragmenty DNA selektivně vázaly a rozmístily pro optimální sekvenování. Na destičce je provedena polymerázová řetězová reakce (PCR) pro amplifikaci jednotlivých fragmentů, což vede ke vzniku clusterů; každý cluster obsahuje mnoho kopií

stejného fragmentu DNA. Tento krok zajišťuje dostatečné množství DNA pro detekci během sekvenování.

Během sekvenování jsou k destičce postupně přidávány fluorescenčně značené nukleotidy. Každá ze čtyř bází (A, T, C, G) má svou vlastní specifickou barvu. Technologie Illumina využívá terminátory, které zajišťují, že je během každého cyklu přidán pouze jeden nukleotid. Po přidání nukleotidu a jeho inkorporaci do rostoucího řetězce DNA je z destičky pořízen snímek. Fluorescenční signály z jednotlivých clusterů jsou detekovány a zaznamenány, což umožňuje identifikaci konkrétních nukleotidů přidaných v každém cyklu.

Po každém snímkovém cyklu jsou terminátory odstraněny, aby umožnily přidání dalšího nukleotidu, a fluorescenční značky jsou také odstraněny, aby se předešlo kontaminaci signálu v následujících cyklech. Tento proces se opakuje pro požadovaný počet cyklů, čímž se postupně buduje kompletní sekvence každého fragmentu DNA na destičce.

Data získaná z každého cyklu jsou analyzována pomocí softwaru, který detekuje a sestavuje jednotlivé báze do kompletních sekvencí. Výsledné sekvence jsou poté připraveny pro další bioinformatickou analýzu a interpretaci v různých aplikacích.

Schéma sekvenace pomocí platformy Illumina je uvedeno na obrázku 2.1.

2.2 IonTorrent

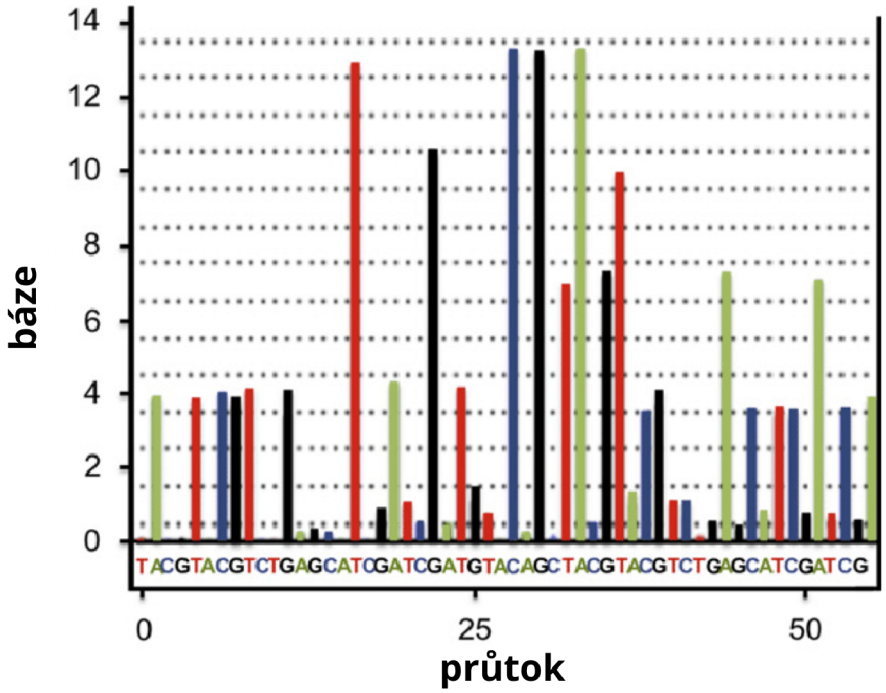
Ion Torrent [29] je inovativní technologie sekvenace DNA, která se odlišuje od tradičních postupů, mezi které se řadí i sekvenace pomocí Illumina.

Celý proces začíná přípravou knihovny DNA, ve které jsou fragmenty označeny adaptéry pro následné sekvenování. Klíčovým krokem je emulzní polymerázová řetězová reakce (emPCR), při níž je sekvenční knihovna amplifikována v mikrokapičkách s polymerázou a perličkovými částicemi. Amplifikované fragmenty DNA jsou pečlivě umístěny na povrch tzv. Ion Torrent čipu. Každý čip obsahuje miliony mikroskopických iontových senzorů, z nichž každý může obsahovat jediný cluster amplifikované DNA.

Během sekvenačního procesu jsou na čip postupně dodávány jednotlivé typy nukleotidů. Při začlenění nukleotidu do rostoucího řetězce DNA uvolňuje enzym DNA polymeráza proton jako vedlejší produkt syntézy. Množství uvolněných protonů je následně měřeno iontovými senzory, což umožňuje detekci pořadí nukleotidů.

Výsledné sekvence jsou zaznamenávány v reálném čase a výstupní datový soubor obsahuje informace o sekvencích a jejich kvalitě. Příklad výstupu je uveden na obrázku 2.2. IonTorrent je známý svou schopností generovat krátké sekvence velmi rychle, navíc má v porovnání s jinými sekvenačními technologiemi relativně nízké náklady na opakování. Je však důležité zdůraznit, že IonTorrent má omezenou délku

čtení, což může omezit jeho schopnost sestavit velké genomy nebo analyzovat velké repetitivní oblasti.



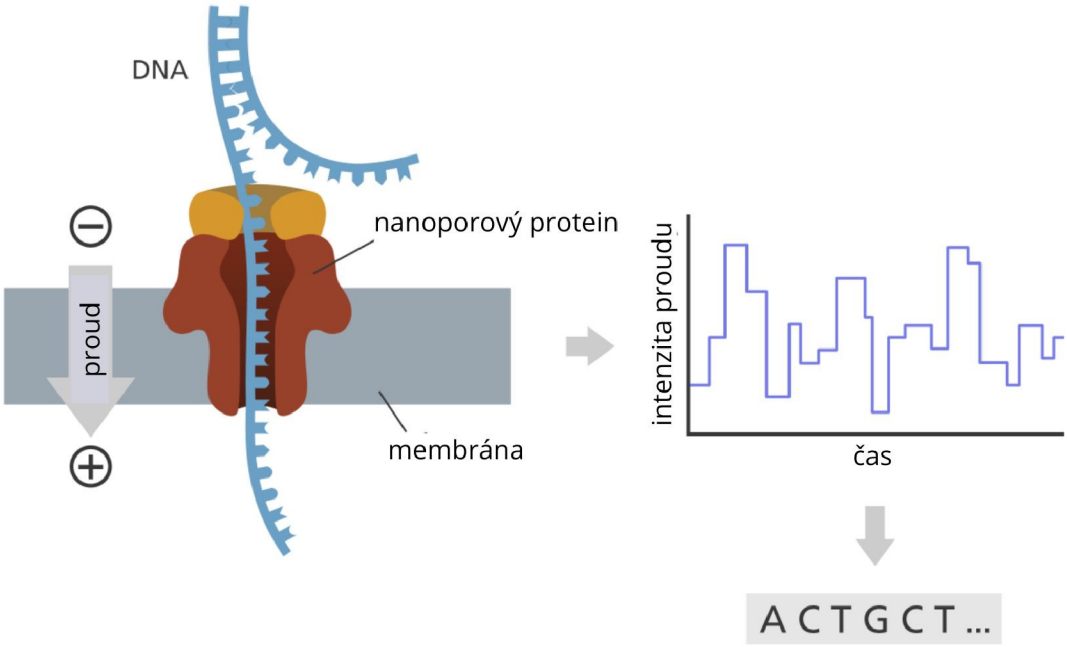
Obr. 2.2: Příklad výstupu sekvenování pomocí platformy IonTorrent [30]. Upraveno.

2.3 Oxford Nanopore

Oxford Nanopore Technologies (ONT) [31] je metoda sekvenace známá jako nanoporová sekvenace. Tato technologie je specifická svým přístupem k sekvenaci jednotlivých molekul s využitím nanoporů.

Princip fungování ONT spočívá v průchodu jednotlivých molekul DNA nebo RNA nanoporem, což je malý kanál umístěný v membráně. Nanopor je vyrobený z biokompatibilního materiálu a je velmi tenký, umožňující průchod jediné molekuly nukleové kyseliny. Na obou stranách membrány jsou elektrody, které vytvářejí elektrický potenciál napříč nanoporem.

Během procesu sekvenace jsou molekuly DNA nebo RNA zavedeny na jeden ze vstupů nanoporu. Když se molekula pohybuje skrze nanopor, nukleotidy vytvářející molekulu narušují elektrický proud. Tato změna v elektrickém proudu je detekována senzory a přiřazena k odpovídajícím nukleotidům, což umožňuje rekonstrukci sekvence.



Obr. 2.3: Schéma sekvenace pomocí platformy Oxford Nanopore [32]. Upraveno.

Jedním z hlavních benefitů této technologie je schopnost generovat dlouhá čtení, což je významné zejména pro analýzu komplexních oblastí genomu, dlouhých repetitivních úseků, nebo pro překonání obtíží spojených s opakováním. Při využití pro RNA-Seq přináší Oxford Nanopore také informace o exprese genů a alternativním splicingu.

Schéma sekvenace s využitím platformy ONT je znázorněno na obrázku 2.3.

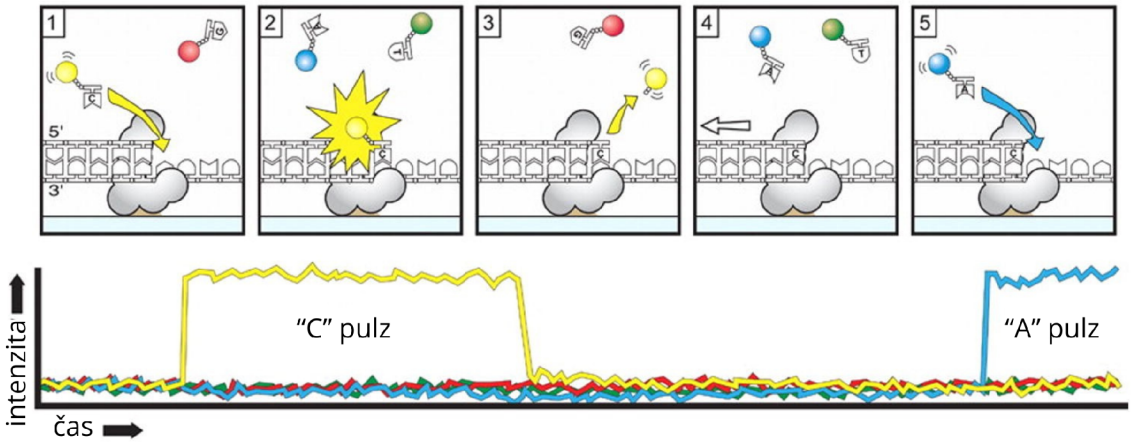
2.4 Pacific Bioscience

Pacific Biosciences (PacBio) [33] nabízí sekvenační technologii známou jako Single Molecule, Real-Time sequencing (SMRT). Tato metoda se vyznačuje schopností přímé sekvenace jednotlivých molekul DNA v reálném čase.

V prvním kroku je molekula DNA fixována na dno Zero Mode Waveguide (ZMW). Jedná se o mikroskopické nanostruktury, které umožňují monitorování enzymatické aktivity v extrémně malém objemu. DNA je připravena tak, aby byla fixována pomocí adaptéru vázajícího se na DNA polymerázu.

Každý nukleotid přidaný do sekvenační reakce je označen specifickou fluorescenční značkou, která odpovídá jednomu ze čtyř nukleotidů. Při inkorporaci zbarveného nukleotidu do rostoucího řetězce DNA polymeráza tuto fluorescenční značku uvolňuje, což vyvolává emisi světla. Intenzita a barva emitovaného světla jsou detekovány

vysoce citlivými kamerami a analyzovány. Tento proces je znázorněn na obrázku 2.4.



Obr. 2.4: Schéma sekvenace pomocí platformy Pacific Bioscience [33]. Upraveno.

Díky PacBio je možné generovat dlouhá čtení, SMRT umožňuje dosahovat průměrné délky čtení v řádu tisíců nebo dokonce desítek tisíc nukleotidů. PacBio se také vyznačuje minimální chybou během sekvenace.

3 Měření exprese bakteriálního genomu

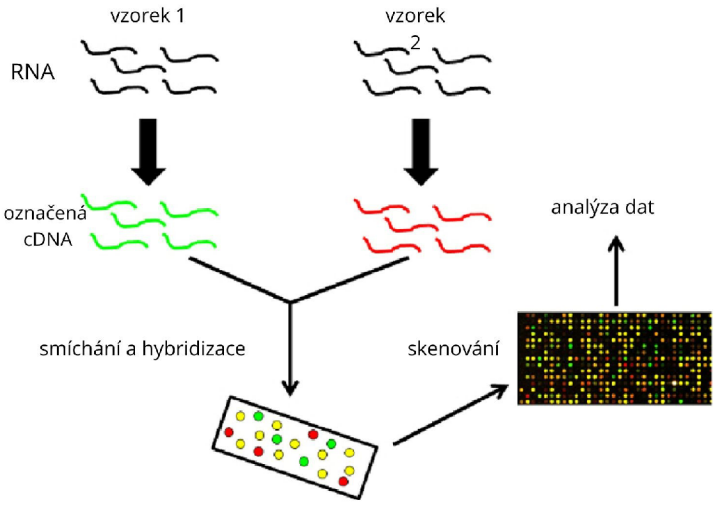
Analýza genové exprese bakteriálního genomu se typicky zaměřuje na studium transkripce RNA, protože samotná DNA není přímo zodpovědná za genovou expresi. Po provedení transkripce vzniká v bakterii velké množství molekul mRNA, které slouží jako přenašeče genetické informace pro syntézu bílkovin. Hodnocení množství a charakteristik mRNA umožňuje posoudit úroveň exprese konkrétních genů.

V bakteriální buňce má mRNA krátký poločas rozpadu, je náchylná k rychlé degradaci [34]. Pro zachování informací o transkripci a genové expresi je proto nezbytné převést mRNA na stabilnější formu, například pomocí reverzní transkripce, která vytváří komplementární DNA (cDNA). Tato transformace umožňuje následné studium genové exprese s větší stabilitou a přesností.

Následující kapitoly podrobněji rozeberou technologie využívané k měření exprese bakteriálního genomu, přičemž se nejvíce zaměří na metodu RNA-Seq.

3.1 Microarrays

Technologie mikročipů [35] umožňuje rychlé a efektivní analyzování tisíců genů najednou na velmi malém prostoru. Tato technologie využívá mikroskopické čipy, často vyrobené ze skla nebo silikonu, na kterých jsou systematicky uspořádány miliony různých oligonukleotidových sond. Tyto sondy slouží jako specifické záchytné body pro komplementární sekvence DNA nebo RNA z biologických vzorků [36].



Obr. 3.1: Schéma procesu technologie mikročipů [37]. Upraveno.

Čipy jsou navrženy tak, aby identifikovaly a kvantifikovaly množství specifických mRNA molekul v biologických vzorcích, poskytují tak informace o úrovni genové

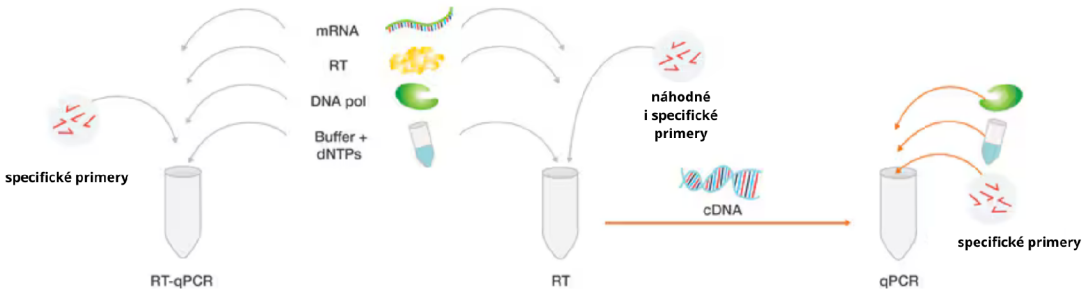
exprese v daných buňkách nebo tkáních při specifických podmínkách. Získaná data se využívají pro výzkum onemocnění, reakce na léčiva nebo různé environmentální stresory.

Analýza, znázorněná na obrázku 3.1, začíná přípravou vzorků, kde jsou cílové DNA nebo RNA molekuly označeny fluorescenčními značkami. Toto označení umožňuje následnou vizuální detekci těchto molekul na čipu. Vzorky jsou poté přivedeny na čip, kde dochází k hybridizaci, tj. specifickému párování komplementárních nukleotidových sekvencí mezi sondou na čipu a cílovou molekulou v vzorku. Efektivita hybridizace se měří na základě intenzity fluorescenčního signálu, který je přímo úměrný množství hybridizované nukleové kyseliny.

Po hybridizaci je čip skenován speciálním skenerem, který detekuje a kvantifikuje fluorescenční signály z jednotlivých políček čipu. Intenzita těchto signálů je následně analyzována pomocí bioinformatických nástrojů, které umožňují srovnání genové exprese mezi různými vzorky, identifikaci genetických variant nebo změn v genomu. Tato intenzita přímo koreluje s množstvím provedené hybridizace.

3.2 RT-qPCR

Real-Time Quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR) [38] jednou z nejcitlivějších a nejpřesnějších metod pro kvantifikaci genové exprese. Metoda kombinuje principy reverzní transkripce (RT), která převádí RNA na cDNA, s kvantitativní PCR (qPCR), která amplifikuje a detekuje cílovou DNA v reálném čase.



Obr. 3.2: Schéma procesu technologie RT-qPCR [39]. Upraveno.

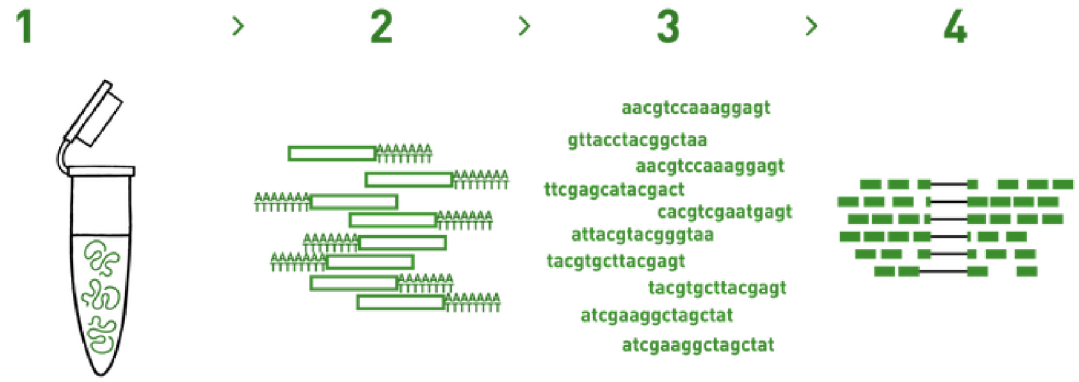
Prvním krokem procesu znázorněného na obrázku 3.2 v RT-qPCR je reverzní transkripce, kde se mRNA z biologických vzorků převádí na cDNA pomocí enzymu reverzní transkriptázy, neboť DNA polymerázy použité v následující fázi PCR mohou amplifikovat pouze DNA. Po syntéze cDNA se přidávají specifické primery, které umožňují zaměření na konkrétní geny nebo sekvence genetického materiálu.

Amplifikace cDNA se provádí pomocí PCR, kde se cyklicky opakuje denaturace, navázání specifických primerů a elongace. Během těchto cyklů se používají fluorescenční sondy nebo interkalující barviva, která se vážou na dvojvláknovou DNA nebo jsou inkorporovány do DNA během elongace. Fluorescenční signál se měří po každém cyklu, což umožňuje sledovat exponenciální nárůst množství DNA v reálném čase.

Fluorescenční signály jsou zaznamenávány a vyhodnocovány pomocí softwaru generujícího amplifikační křivky. Tyto křivky zobrazují intenzitu fluorescenčního signálu v závislosti na počtu cyklů a poskytují důležité informace o množství cílové DNA v vzorku. Z křivek lze získat hodnoty CT (Cycle Threshold), což je počet cyklů potřebných k dosažení určitého prahu fluorescenčního signálu. Nižší CT hodnota obvykle indikuje vyšší počáteční množství cílového RNA v vzorku, což naznačuje silnější expresi daného genu.

3.3 RNA-Seq

RNA-Seq [40] , neboli RNA sekvenování, je metoda sekvenace nové generace (NGS), která umožňuje komplexní analýzu transkriptu buněk a významně přispívá k našemu porozumění biologických a patologických procesů. Metoda RNA-Seq nabízí detailní pohled na celý transkriptom – soubor všech RNA molekul – což zahrnuje mRNA, sRNA (small RNA), ncRNA (non-coding RNA) a další. Díky vysoké citlivosti této technologie lze nejen kvantifikovat množství a typy RNA v biologickém vzorku, ale i studovat post-transkripční modifikace, alternativní sestřih RNA a interakce RNA s proteiny. Průběh procesu RNA-Seq jsou znázorněny na obrázku 3.3.



Obr. 3.3: Kroky metody RNA-Seq. 1 - Izolace a příprava vzorku. 2 - Příprava knihovny (fragmentace a ligace adaptérů). 3 - Vzniklá čtení. 4 - Mapování čtení.

Prvním krokem v procesu RNA-Seq je izolace RNA z biologického materiálu, což zahrnuje použití chemických látek, membránových filtrů nebo magnetických částic pro oddělení a čištění RNA. Jak již bylo zmíněno v úvodu, metody měření exprese genů spočívají v syntéze cDNA - tato izolovaná RNA je proto následně převedena na stabilnější formu cDNA prostřednictvím procesu zvaného reverzní transkripce. Tento krok je velmi podstatný, protože většina sekvenačních platforem je primárně navržena pro práci s DNA.

Součástí transkripce je mapování míst iniciování transkripce s rozlišením na nukleotidové úrovni. Mapování je nezbytné pro úplné definování produktů RNA a identifikaci přilehlých promotorových oblastí, které regulují expresi každého transkriptu. Jedním z prvních způsobů mapování byla metoda cAGe („cap analysis of gene expression“), původně navržená pro Sangerovo sekvenování, zahrnující sekvenování klonovaných cDNA odvozených od RNA s neporušenými 5‘ konci [41]. Tato metoda však vyžadovala vysoké množství vstupní RNA a generovala pouze krátká čtení o délce okolo 20 nukleotidů. Metoda cAGe byla proto později adaptována pro platformy sekvenátorů druhé generace. Nově je tak možné mapovat i malé množství vstupní RNA, přiřadit iniciační místa ke konkrétním transkriptům, navíc byly odstraněny obtíže se zarovnáváním jednotlivých krátkých opakujících se čtení. Obtíže provázející výše zmíněné metody mapování zahrnují jejich závislost na syntéze cDNA, na krocích hybridizace, jejichž efektivita závisí na sekvenci a struktuře RNA, a náročnost analýzy krátkodobých transkriptů, jako je například mikroRNA – hojně transkribovaná molekula RNA vzácná pro rychlou degradaci.

Na jeden nebo oba konce fragmentů transkribované cDNA jsou následně přidány adaptéry [42]. Adaptéry označují původ sekvencí a slouží jako prevence vzniku PCR duplikátů. Na rozdíl od sekvenování celého genomu, kde se předpokládá, že duplikovaná čtení jsou způsobena technickými zkresleními a jsou proto odstraněna, jsou taková čtení u RNA-Seq považována za indikaci skutečného biologického signálu a zůstávají zachována. Přidání adaptérů k molekulám cDNA před amplifikací tedy umožňuje identifikaci a počítačové odstranění PCR duplikátů, přičemž se zachovávají skutečné biologické duplicity. Aby byla dvojice sekvencí identifikována jako technický duplikát, musí obsahovat identický adaptér a být mapována na stejné místo v transkriptomu.

Pokud koncentrace cDNA vzorku po inicializaci knihovny a přidání adaptérů není dostatečná pro efektivní sekvenaci, je nezbytné provést amplifikaci cDNA [43]. Tento proces vede k zajištění dostatečného množství materiálu potřebného k detailní a přesné analýze sekvenace. Amplifikace se obvykle provádí pomocí polymerázové řetězové reakce (PCR). V rámci amplifikace PCR se využívají specifické primery, které se vážou na adaptérové sekvence připojené na koncích cDNA fragmentů. Tyto primery slouží jako startovací body pro syntézu nových DNA řetězců, čímž se zvýší

celkový počet kopií každého fragmentu cDNA v knihovně. Existuje několik variant PCR metod, které mohou být využity v závislosti na požadavcích experimentu, včetně kvantitativní PCR (qPCR) pro kvantifikaci a verifikaci úrovně amplifikace, nebo speciálních technik, jako je multiplex PCR, která umožňuje amplifikaci více cílů současně. Po dokončení amplifikace je nutné knihovnu cDNA vyčistit od nežádoucích produktů, jako jsou neamplifikované primery, nepoužité nukleotidy a krátké fragmenty DNA, které mohou znehodnotit sekvenační data. Čištění se obvykle provádí pomocí kolon s gelovou filtrací nebo magnetických kuliček vázajících DNA, což umožňuje oddělit požadované fragmenty cDNA od ostatních složek směsi.

Nakonec je provedena kontrola kvality amplifikované cDNA knihovny, obvykle s využitím metod jako jsou agarózová gelová elektroforéza, která ověří velikostní distribuci fragmentů, nebo pokročilejší techniky jako kapilární elektroforéza. Kontrola kvality slouží k potvrzení, že knihovna obsahuje fragmenty vhodné velikosti a koncentrace, a že je připravena pro vysokoproduktivní sekvenační analýzu.

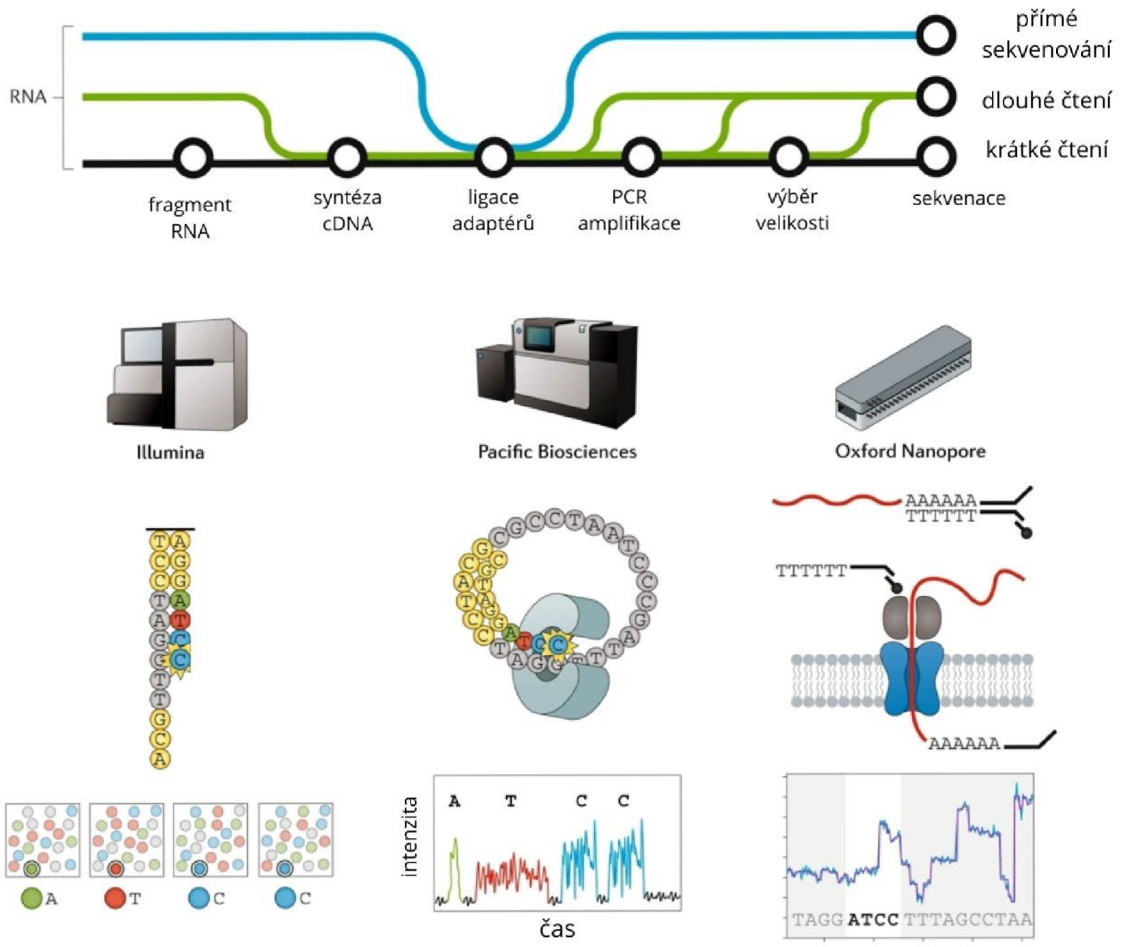
Pro pozdější návaznost je třeba zmínit i strand-specific RNA-Seq [44]. Jedná se o pokročilou metodu sekvenování, která odhaluje, ze kterého vlákna DNA jsou RNA molekuly transkribovány, což poskytuje přesné informace o směru transkripce. Tato technika zahrnuje speciální protokoly přípravy knihoven, jako je použití RNA adaptérů nebo specifických enzymů, umožňuje výzkumníkům rozlišit mezi dopřednými a reverzními transkripty a lépe mapovat transkripční aktivity genomu.

Příprava knihovny pro sekvenaci RNA-Seq se podstatně liší v závislosti na platformě sekvenace a požadavcích na délku čtení, což ovlivňuje typ a rozsah získaných dat [45]. Platformy jako Illumina a IonTorrent jsou preferovány pro sekvenaci krátkých čtení, zatímco Pacific Biosciences (PacBio) a Oxford Nanopore Technologies se specializují na dlouhá čtení.

Technologie jako Illumina a IonTorrent [46]vyžadují přípravu knihovny s fragmenty DNA typicky menšími než 200 bp. Tyto platformy využívají adaptérově zprostředkovanou amplifikaci DNA, kde jsou krátké adaptéry připojeny na oba konce cDNA fragmentů před amplifikací. Příprava knihovny pro krátká čtení umožňuje detailní analýzu exprese genů, kvantifikaci transkriptů a studium krátkých regulačních RNA (sRNA), využívá se také pro strukturální analýzu RNA a zkoumání interakcí mezi RNA a proteiny.

Technologie jako PacBio a Oxford Nanopore [46] umožňují sekvenaci dlouhých čtení. Platformy tak poskytují možnost analýzy celých transkriptů, ideální pro studium struktury a funkce nekódujících RNA, jako jsou lncRNA ("long non-coding RNA") [47], které mohou hrát roli v regulačních sítích a jsou často spojeny s komplexními onemocněními. Dlouhá čtení jsou také užitečná pro analýzu de novo transkriptomu, identifikaci fúzních transkriptů a mapování alternativních míst sestřihu, které mohou být přehlédnuty u kratších čtení.

Metoda RNA-Seq s využitím různých platforem je znázorněna na následující straně na obrázku 3.4.



Obr. 3.4: Schéma metody RNA-Seq pomocí platformy Illumina, PacBio a Oxford Nanopore [43]. Upraveno.

4 Vybrané nástroje pro simulaci sekvenačních dat

Nástroje pro simulaci sekvenačních dat slouží k vytváření umělých datasetů, které simulují procesy objevující se při skutečném sekvenování RNA. Pomocí těchto simulací lze hodnotit výkonnost různých metod analýzy dat RNA-Seq, například porovnáním různých algoritmů na detekci genů nebo jejich exprese. Dále lze data simulací použít k vývoji a testování nových algoritmů před jejich použitím na skutečná experimentální data a k seznámení se s procesem analýzy dat RNA-Seq bez nutnosti skutečných experimentů. Níže jsou přiblíženy vybrané nástroje pro simulaci RNA-Seq dat.

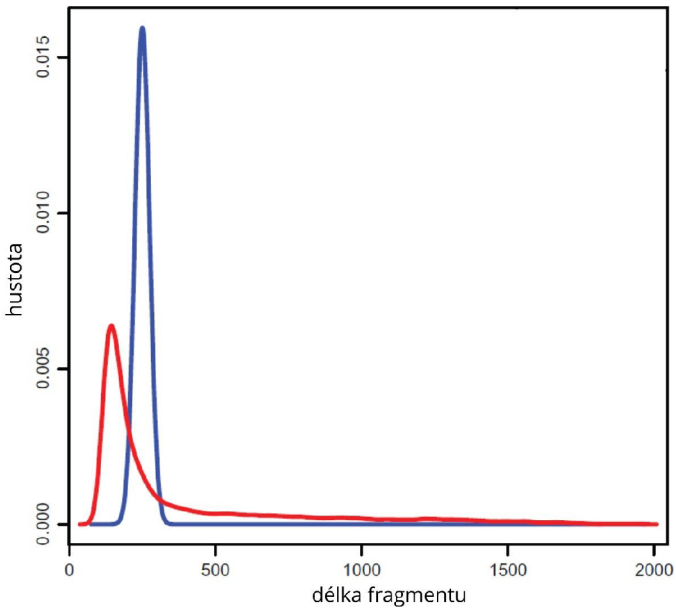
4.1 Polyester

Polyester [48] je balíček pro R (programovací jazyk pro statistické výpočty a grafiku) určený k simulaci sekvencí čtení RNA-Seq. Na základě vstupních souborů simuluje kroky RNA-Seq, tedy reverzní transkripci, fragmentaci a následné sekvenování. Jeho hlavní výhodou je, že uživatelé mohou simulovat sekvence čtení diferenciální exprese genů – analýzu rozdílů v četnosti výskytu genových transkriptů v transkriptomu. Cílem diferenciální exprese je určit, které geny jsou odlišně exprimovány. Klíčovým prvkem Polyesteru je kromě simulace diferenciální exprese také možnost nastavení síly jejího signálu. To je umožněno plnou kontrolou uživatele nad počtem čtení, která jsou generována z každého vstupního transkriptu. Polyester obsahuje vestavěný model pro diferenciální expresi založený na negativní binomické distribuci. Polyester dále také umožňuje generování surových čtení RNA-Seq, je transparentní a volně dostupný.

Vstupní data pro Polyester představují anotované transkripty. Tyto anotace jsou buď ve formátu FASTA, nebo doplněné formátem GTF. FASTA je formát souborů používaný pro reprezentaci sekvencí nukleových kyselin nebo proteinů. Začíná hlavičkou uvozenou znakem „>“, za kterým následuje identifikátor SeqID dané sekvence. Identifikátor musí být jedinečný pro každou sekvenci a nesmí obsahovat mezery. Dále obsahuje samotnou nukleotidovou nebo aminokyselinovou sekvenci. GTF, neboli „gene transfer format“, je 9 sloupcový textový soubor obsahující základní charakteristiky sekvence nukleotidů. Mezi charakteristiky patří například název sekvence, zdroj, ze kterého pocházejí anotační data, kategorie „feature“ popisující typ prvku, počáteční pozice, konečná pozice nebo čtecí rámec.

Simulace začíná fragmentací simulovanou tak, aby byla kompatibilní s postupem metody Illumina, délky fragmentů jsou odvozeny z normální distribuce. Poté jsou

fragmenty sekvenovány, přičemž si uživatel může vybrat z různých modelů chyb sekvenování. Výstupem Polyesteru jsou soubory FASTA obsahující simulovaná čtení a další informace využitelné pro pozdější analýzu.



Obr. 4.1: Rozložení délky fragmentů programu Polyester. Červená křivka zobrazuje distribuci délky fragmentů pro vybraná sekvenační čtení ze souboru dat RNA-seq od GEUVADIS; modrá křivka ukazuje normální rozdělení. [48] Upraveno.

4.2 Flux Simulator

Flux Simulator [49] je dalším z nástrojů zaměřujících se na modelování RNA-Seq dat. Sekvenční čtení jsou vytvářena z referenčního genomu podle anotovaných transkriptů, vstupní data Flux Simulator přijímá v již výše zmíněných formátech FASTA a GTF. Simulaci lze zvolit dle různých protokolů, jako je Illumina, Roche 454, Ion Torrent nebo PacBio. Jazykem implementace je JavaScript.

Simulace probíhá v modulech s nastavitelnými parametry založenými na reálných experimentech, což umožňuje věrnější reprodukci experimentálních dat. Flexibilita spojování jednotlivých modulů navíc umožňuje simulace libovolných protokolů. Prvním krokem je příprava knihovny, která začíná reverzní transkripcí RNA do cDNA. Následně dochází k fragmentaci transkriptů, jelikož současné platformy produkují spíše krátká čtení. Dále jsou připojeny adaptéry a je provedena amplifikace pomocí PCR. Nakonec je provedena samotná sekvenace. Sekvenční čtení prochází modifikacemi podle technických omezení odpovídající platformy.

Flux Simulator umožňuje simulace různých procesů fragmentace. Uživatel tak může využít enzymatickou fragmentaci, nebulizaci (mechanické štěpení) nebo hydrolýzu. Reverzní transkripci Flux modeluje odděleně pro syntézu prvního a druhého řetězce. Počáteční bod závisí na umístění primeru a volitelně na matici vah pozic. Co se týče velikosti fragmentů pozorovaných po elektroforéze, pracuje Flux Simulator s normálním nebo empirickým rozdělením. Simulována je i kinetika reakce procesu ligace adaptéru, a to zohledněním sekvencí preferovaných enzymů. Pro sekvenaci lze specifikovat počet a délku čtení, nicméně čtení nemůže být více než fragmentů v knihovně a žádné čtení nemůže být delší než fragment, ze kterého pochází. Orientace čtení je striktně stanovena na směr od 3‘ konce k 5‘ konci.

Výstupem Flux Simulatoru jsou soubory formátu FASTA. Rozdělení celého procesu na jednotlivé kroky s nastavitelnými parametry umožňuje uživatelům zkoumat a porozumět různým experimentálním faktorům a zkreslením, které ovlivňují výsledky RNA-Seq experimentů.

4.3 RSS - RNASeqReadSimulator

Pro simulaci RNA-Seq dat lze také využít nástroj RSS [50]. RSS byl vytvořen v jazyce Python – všechny skripty lze ovládat prostřednictvím příkazové řádky, což umožňuje uživatelům volat a navrhovat další funkce. Stejně jako předchozí nástroje umožňuje generovat sekvence čtení RNA-Seq, náhodně přiřadit úrovně exprese transkriptu a simulovat náhodné chyby čtení. RSS přijímá vstupní data ve formátu SAM (Sequence Alignment/Map; formát pro ukládání zarovnání čtení s referenčními sekvencemi, podporuje krátká i dlouhá čtení vytvořená různými sekvenačními platformami), nebo BED (Browser Extensible Data; textový soubor, kde každý řádek reprezentuje jednu anotační oblast).

4.4 SimSeq

SimSeq [51] je nástroj založený na postupech platformy Illumina, vstupem je matice vzorkovaných genů. SimSeq je specifický především pro možnost chimérických čtení - čtení v sekvencích DNA nebo RNA, která vznikla z fúze dvou nebo více původních fragmentů. Tato fúze může nastat během procesů přípravy vzorku, amplifikace nebo sekvenování, a výsledkem jsou čtení, která obsahují informace z více než jednoho místa v genomu nebo transkriptomu. Chimérická čtení mohou být problematická při analýze genomických nebo transkripčních dat, zejména pokud jsou interpretována nesprávně, což může vést k chybným závěrům o genomové struktuře

nebo exprese genů. Dále SimSeq umožňuje uživateli simulovat data nejen s předpokladem negativního binomického rozdělení, ale pracuje také s vlastním neparametrickým algoritmem. Použití neparametrického algoritmu SimSeq by mělo poskytovat přesnější obraz o účinnosti dané metody pro detekci diferenciální exprese.

4.5 BEERS2

Primárním vstupem pro BEERS2 [52] jsou sady molekul, které jsou výstupem programu CAMPAREE; přijme však i jakoukoli správně naformátovanou sadu molekul RNA. Vstupní molekuly transformuje na RNA-Seq čtení simulací jednoho z několika různých protokolů přípravy knihovny, které jsou následně podrobeny sekvenování. Tento proces zahrnuje depleci ribozomů, fragmentaci, různé kroky PCR, značení a následně syntézu čtení, což zahrnuje hybridizaci průtokových buněk, amplifikaci můstků a vyhodnocení bází.

Chyby a další charakteristiky sekvenování vznikají v různých fázích přípravy knihovny. Díky modularitě systému lze vytvářet nové protokoly přípravy knihovny nebo implementovat změny v existujících protokolech. Tato modulární povaha umožňuje oddělit jednotlivé aspekty přípravy knihovny a sekvenování a posoudit vliv těchto kroků na kvantifikaci. Jedním z omezení modelu BEERS2 je, že hodnoty jsou nastaveny podle toho, jak se chovají jednotlivé kroky. Skutečná data jsou však vždy výsledkem mnoha kroků provedených společně. Proto hodnoty konfigurace neodpovídají přímo naměřeným hodnotám.

4.6 ART: Adaptive Read-pair Technology for Sequencing Simulation

ART [53] je nástroj pro simulaci krátkých čtení, který napodobuje chybovost a variační rysy různých sekvenačních technologií, jako jsou Illumina, Roche 454, Helicos a další. Simuluje jak jednotlivá, tak párová čtení. ART efektivně modeluje specifické chyby každé platformy, včetně chyb ve čtení, vkládání a vypouštění nukleotidů. ART poskytuje flexibilitu v nastavení různých parametrů simulace, včetně hloubky pokrytí, délky čtení a frekvence chyb, což umožňuje simulaci široké škály sekvenačních scénářů a podmínek.

4.7 pIRS: Profile-based Illumina pair-end Reads Simulator

Simulátor pIRS [54] je zaměřený na napodobování specifických charakteristik sekvenování pomocí technologie Illumina. Tento nástroj je speciálně navržen pro generování párových čtení, které jsou kritické pro aplikace, jako je sestavování celých genomů, kde přesné spojení konců fragmentů může zásadně ovlivnit kvalitu a přesnost sestaveného genomu. pIRS simuluje kompletní proces sekvenace včetně tvorby PCR chimer, což jsou chyby, které vznikají během amplifikace vzorku PCR a mohou způsobovat problémy při bioinformatické analýze. Simulátor umožňuje detailní nastavení parametrů, jako jsou specifické chybové profily pro různé čtecí délky a sekvenační chemie, což uživatelům umožňuje přizpůsobit simulace svým konkrétním potřebám.

5 BacSeqer

Teoretická část práce poskytuje kromě jiného popis bakteriálního genomu a přehled několika používaných nástrojů pro simulaci RNA-Seq dat. Tyto nástroje jsou však zaměřeny především na eukaryotický transkriptom. Hlavním cílem této práce je implementace vlastního simulátoru RNA-Seq čtení se zaměřením na prokaryotický transkriptom. Proto byl vyvinut BacSeqer, nástroj pro RNA-Seq navržený speciálně pro bakteriální transkriptomy s cílem produkovat data, která se podobají datům získaným ve skutečných experimentech. Umožňuje zohlednit specifika bakteriálního genomu, zejména existenci exonů a jejich uspořádání v operonech, které se u eukaryot běžně nevyskytují. BacSeqer je volně k dispozici na adrese `https://github.com/adelafi/BacSeqer.git` a také v příloze práce.

5.1 Implementace

Hlavní myšlenkou BacSeqeru je vytvoření programu pro simulaci RNA-Seq experimentů; simulátory RNA-Seq jsou užitečné pro testování a ověřování bioinformatických nástrojů, nabízejí způsob, jak porovnat různé výpočetní algoritmy, a umožňují zkoumat různé sekvenační protokoly, nastavení parametrů a velikosti vzorků, aby se zajistilo co nejefektivnější využití zdrojů.

BacSeqer obsahuje různé funkce pro práci s anotovanými sekvencemi, na základě jejichž výstupů lze následně provádět simulaci. Samotná simulace probíhá s možností nastavení délky a počtu čtení, procentuálního obsahu GC, zohlednění operonových oblastí, procentuální kontaminace rRNA a výběru orientace vláken. Výsledkem simulace je soubor s požadovaným počtem vygenerovaných RNA-Seq čtení. Nástroj je implementován v jazyce Python.

5.1.1 Vstupní data

Simulátor BacSeqer přijímá jako vstup libovolné anotované sekvence (nejen) bakteriálního genomu. Ty mohou být poskytnuty jako sekvence ve formátu FASTA představující genom, doplněné anotací poskytnutou v souboru GTF (Gene Transfer Format) nebo GFF (General Feature Format).

Formáty souborů GTF a GFF jsou používány pro ukládání informací o genech a jiných genetických sekvencích, jako jsou lokace genů, exony, introny a další regulační oblasti na genomu. GTF soubory obsahují devět polí oddělených tabulátory, které zahrnují informace jako chromozom, zdroj (software nebo metoda generující anotaci), typ prvku (například gen nebo exon), začátek a konec lokace, skóre, strand, fázi a atributy. Atributy v GTF musí vždy obsahovat položky jako gene_id

a transcript_id, které jednoznačně identifikují každý transkript a gen. GFF formát je obecnější a má několik verzí, nejnovější a nejpoužívanější je verze GFF3. Stejně jako GTF, GFF3 obsahuje devět polí oddělených tabulátory, ale má flexibilnější formát pro atributy, které jsou uvedeny jako páry klíč=hodnota oddělené středníky a mohou zahrnovat různé dodatečné informace.

5.1.2 Funkce

Kromě funkcí pro načítání vstupních souborů (*extract_sequences*) a vytváření výstupních souborů (*write_output*) obsahuje simulátor několik funkcí pro zpracování podkladů pro sekvenační simulaci RNA-Seq dat. Je tak umožněno procházet procesem krok za krokem s možností nastavení některých parametrů. Jednotlivé funkce jsou níže blíže popsány, přehled funkcí včetně návodu k volání je dostupný také v dokumentaci v příloze práce.

Vyhledání operonových oblastí

Funkce *extract_operons* je určena k analýze souboru GTF nebo GFF a následné extrakci informací o operonech. Při zpracování souboru funkce prochází každý řádek, ignoruje hlavičky a prázdné řádky. Řádky jsou rozděleny na části a pokud řádek obsahuje méně než devět segmentů, je přeskočen. Podstatné informace jsou v osmém sloupci, funkce zde hledá zmínky o operonech pomocí regulárních výrazů v závislosti na formátu souboru.

Identifikované operony jsou zaznamenány do slovníku spolu s jejich lokacemi. Pokud operon již ve slovníku existuje, jeho nová lokace je přidána do seznamu lokací. Výsledný slovník poskytuje přehled o všech operonech a jejich pozicích.

Vyhledání kódujících oblastí

Funkce *extract_cds* slouží k vyhledání kódujících oblastí v souboru GTF nebo GFF. Identifikátory a odpovídající pozice kódujících oblastí jsou uloženy do slovníku, kde každý identifikátor slouží jako klíč a přidružené pozice jako hodnoty. Její výstup figuruje jako vstup hlavní simulační funkce v případě, že je požadováno simulovat čtení bez kontaminací, čili sestavit vstupní knihovnu pouze z mRNA.

Funkce umožňuje také rozšíření extrahovaných kódujících oblastí o dodatečné nukleotidy na obou koncích, čímž je umožněno zahrnutí 5' a 3' nepřekládaných oblastí (UTR - untranslated region) [55]. 5' UTR ovlivňuje translaci RNA překladem tím, že se podílí na určení efektivity navázání ribozomu a následné iniciace překladu. Tato oblast může obsahovat regulační sekvence, které modulují stabilitu a lokalizaci mRNA v buňce, což má přímý dopad na množství produkovaného proteinu. 3' UTR

obsahuje signály pro ukončení překladu a sekvence zapojené do polyadenylace, které jsou nezbytné pro správnou maturaci a export mRNA z jádra. Navíc, 3' UTR jsou často místa vázání pro regulační RNA molekuly, jako jsou mikroRNA, které mohou potlačit exprese cílového genu tím, že blokují translaci nebo urychlují degradaci mRNA.

Výpočet a úprava procentuálního zastoupení GC

Funkce *calculate_gc_content* vypočítá procentuální obsah GC v aktuálně zpracovávané sekvenci. Funkce používá metodu *count* na řetězci pro zjištění počtu výskytů 'G' a 'C'. Tyto hodnoty jsou sečteny, aby byl získán celkový počet guaninů a cytosinů v sekvenci. Pro určení obsahu GC je třeba vypočítat podíl sumy G a C vůči celkové délce sekvence. Tento podíl je následně vynásoben 100, tedy převeden na procentuální hodnotu. Výsledek je vyjádřen jako desetinné číslo (float).

Pokud vypočítaný obsah neodpovídá požadované hodnotě, je pomocí funkce *adjust_gc_content* upraven tak, aby se zachovala původní délka sekvence. Začátek procesu spočívá ve spočítání rozdílu mezi aktuálním obsahem GC a cílovým obsahem GC. Na základě tohoto rozdílu určuje, kolik bází A nebo T (adenin nebo thymin) je potřeba nahradit guaninem nebo cytosinem.

Pokud je třeba zvýšit obsah GC (rozdíl je kladný), funkce najde všechny pozice bází A a T v sekvenci a náhodně vybere dostatečný počet z nich pro nahrazení. Naopak, pokud je potřeba snížit obsah GC (rozdíl je záporný), funkce postupuje obdobně s pozicemi obsazenými bázemi G a C, které náhodně nahradí bázemi A nebo T. V případě, že aktuální obsah GC již odpovídá požadovanému, nejsou prováděny žádné změny.

Výsledkem funkce je upravená sekvence, ve které byl obsah GC přizpůsoben požadavkům uživatele. Příklad výstupu je uveden na obrázku 5.1.

```
ACGTAGCTGATGCTACGATGC
Obsah GC v původní sekvenci: 0.5238095238095238
Obsah GC nově nastaven na 15%. Nová sekvence:
AAGTAAATGATGTTAAGATAT
```

Obr. 5.1: Příklad úpravy procentuálního obsahu GC v sekvenci.

Vyhledání oblastí ribozomální RNA

Funkce *extract_rrna* slouží k analýze souboru GTF nebo GFF za účelem vyhledání oblastí ribozomální RNA. Každý zpracovaný řádek se rozdělí na devět částí, které

odpovídají standardní struktuře polí v GFF/GTF souborech. Funkce se zaměřuje na řádky, kde *feature_type* je rRNA, což značí, že řádek obsahuje informace o ribosomální RNA. Z atributů řádku funkce dále extrahuje identifikátor genu (gene_id). Pokud je *gene_id* úspěšně extrahován, funkce použije tento identifikátor společně s indexem, který označuje, o kolikátý výskyt tohoto genu se jedná. Tento index je ukládán a aktualizován v pomocném slovníku. Každý unikátní identifikátor genu se společně s jeho indexem stává klíčem v hlavním slovníku, a jeho hodnotou je další slovník obsahující počáteční a koncové pozice rRNA. Výstup funkce opět slouží jako vstupní parametr pro hlavní simulační funkci v případě, že je požadována kontaminace rRNA ve vstupní knihovně.

Výpočet procentuálního zastoupení rRNA ve vstupních datech

Cílem funkce *calculate_rrna_percentage* je spočítat, jaký procentuální podíl z celkové délky sekvencí ve vstupním FASTA souboru představují sekvence rRNA, identifikované pomocí anotací v GTF nebo GFF souboru - využívá se výstup funkce *extract_rrna*. Výpočet procenta rRNA se provádí jako podíl celkové délky rRNA sekvencí a celkové délky všech sekvencí získaných z FASTA souboru, převedený na procenta.

Zjištění orientace vláken

Funkce *parse_for_strand* zpracovává vstupní soubory a vytvoří slovník obsahující názvy genů nebo ID a jim odpovídající informace o směru, ve kterém jsou geny transkribovány, tedy zda na dopředném ('+') nebo reverzním ('-') vláknu DNA.

K zajištění potřebných informací je procházen GTF nebo GFF soubor. Pro každý platný řádek, který obsahuje dostatečné informace (minimálně 9 polí rozdělených tabulátorem), funkce získá informace o vláknu z šestého pole a identifikaci genu z osmého pole. Extrakce názvu genu z pole atributů je realizována pomocí pomocné funkce extract_gene_name, která respektuje rozdíly mezi vstupními formáty. Výsledný název genu je použit jako klíč ve výstupním slovníku, kam je přiřazen také směr vlákna. Ukázka slovníku je uvedena na obrázku 5.2.

```
{'cds-ABR32225.1': '+', 'gene-Cbei_0036': '+', 'cds-ABR32226.1': '+',
'gene-Cbei_0037': '-', 'cds-ABR32227.1': '-', 'gene-Cbei_0038': '-',
'cds-ABR32228.1': '-', 'id-ABR32228.1:1..28': '-', 'gene-Cbei_0039': '+',
'cds-ABR32229.1': '+', 'gene-Cbei_0040': '+', 'cds-ABR32230.1': '+',
'gene-Cbei_0041': '-', 'cds-ABR32231.1': '-', 'gene-Cbei_0042': '+',
'cds-ABR32232.1': '+', 'gene-Cbei_0043': '+', ...}
```

Obr. 5.2: Příklad výstupního slovníku funkce *parse_for_strand*.

Tato funkce je využitelná při hlavní simulaci, zejména při poskytování informací o orientaci vláken pro každou sekvenci. Při simulaci čtení se bere informace o vláknu v úvahu, zejména při simulaci strand-specific čtení.

Simulace RNA-Seq čtení

Nejvýznamnější funkcí programu je funkce *simulate_reads*, která je určena k simulaci čtení ze vstupních sekvencí. Funkce iteruje nad náhodně zvolenou sekvencí ze vstupní knihovny a generuje zadaný počet čtení. Přijímá několik vstupních parametrů; základními parametry jsou již zpracovaná vstupní data, neboli knihovna, požadovaný počet a délka simulovaných čtení. Další parametry jsou volitelné.

Po zahájení simulace funkce nejprve načte sekvence z poskytnutého slovníku sekvencí, kde klíče jsou identifikátory sekvencí a hodnoty jsou samotné sekvence. Na základě dodaných informací o lokacích kódujících oblastí, operonů a rRNA funkce upravuje a doplňuje vstupní knihovnu. Například, pokud jsou specifikovány kódující oblasti, knihovna se upraví tak, aby obsahovala pouze tyto kódující oblasti. Stejně tak, pokud je požadováno určité procento rRNA, funkce zabezpečí, že jej knihovna bude obsahovat. K simulaci čtení RNA sekvencí založených na operonech je využita pomocná funkce *create_long_mrna*, která umožňuje rekonstrukci molekul polycistronní RNA. Funkce prochází každý operon a vyhledá pro něj odpovídající sekvenci DNA pomocí identifikátoru genu. Nalezené DNA sekvence se postupně přidávají do jedné dlouhé sekvence.

Proces generování čtení začíná náhodným výběrem sekvence a určením startovního bodu pro čtení. Toto čtení je pak vystřiženo z celkové sekvence s ohledem na požadovanou délku. Pokud obsah GC není explicitně uveden, nemění se. Při uvedení požadovaného obsahu GC funkce *adjust_gc_content* iterované čtení upravuje výše zmíněným postupem.

Pokud je zadána specifická orientace vlákna, funkce kontroluje, zda aktuální čtení této orientaci odpovídá. Pokud ne, přepisuje čtení na jeho reverzní komplement.

Výstupem funkce je seznam čtení, kde každé čtení je reprezentováno jako dvojice obsahující identifikátor sekvence a samotné čtení.

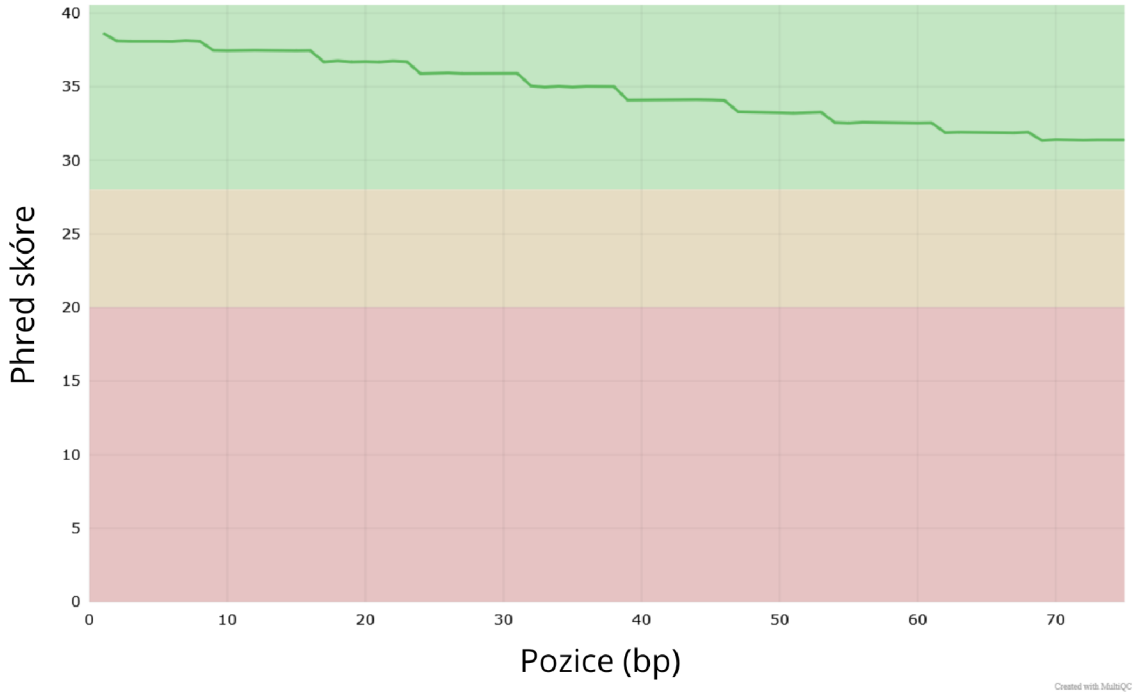
Hlavička funkce se všemi vstupními parametry je uvedena na obrázku 5.3.

```
simulate_reads(seqs, read_length, num_reads, gc_content, cds_locations,
operon_locations, rrna_locations, rrna_percentage, strand_ori, strand_info)
```

Obr. 5.3: Ukázka vstupních parametrů funkce *simulate_reads*.

Vytvoření výstupního souboru

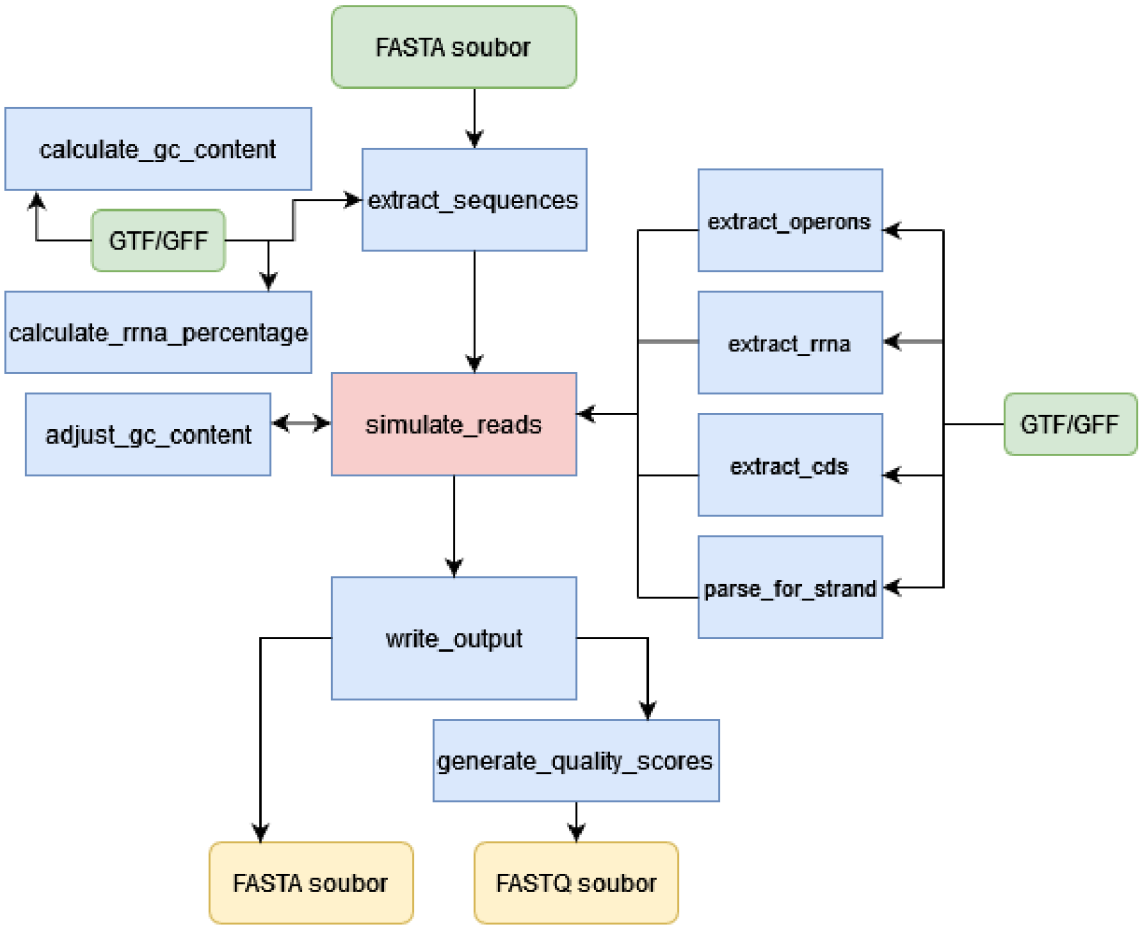
Funkce *write_output* pro zápis výstupu zaznamenává simulovaná čtení do souboru ve formátu FASTA nebo FASTQ. Pro formát FASTQ umožňuje nastavit maximální a minimální skóre kvality a v případě potřeby je vygenerovat pomocí funkce *generate_quality_scores*; tato funkce respektuje sestupnou posloupnost skóre kvality Phred reprezentovaného znaky ASCII. Pokud nejsou maximální a minimální hodnoty zadány, použije se výchozí hodnota ASCII založená na skórech kvality typických pro platformu Illumina. Pomocnou funkcí pro generování skóre kvality je funkce *phred_to_ascii*, která převádí skóre kvality Phred na odpovídající znak ASCII. Graf zobrazující klesající kvalitu čtení v rozsahu od 32 do 28 je uveden na obrázku 5.4.



Obr. 5.4: Graf zobrazující klesající kvalitu čtení v simulovaném datasetu.

5.1.3 Blokové schéma BacSeqeru

Přehled funkcí zmíněných v předchozí kapitole je znázorněn pomocí blokového schéma na obrázku 5.5



Obr. 5.5: Blokové schéma BacSeqeru.

5.2 Výsledky

Tato kapitola se zaměřuje na prezentaci výsledků BacSeqeru, včetně srovnání s reálnými daty. Jsou zmíněny různé možnosti nastavení parametrů a jejich vliv na výstupy.

K vyhodnocení simulovaných dat a jejich porovnání s výsledky skutečných experimentů byly použity nástroje FastQC a MultiQC. FastQC [56] je bioinformatický nástroj pro kontrolu kvality surových sekvenačních dat, generuje zprávy o metrikách, jako je kvalita sekvencí a duplikace, a pomáhá tak rychle identifikovat nejasnosti v projektech vysokokapacitního sekvenování. MultiQC [57] sdružuje data z různých bioinformatických nástrojů do jediné zprávy, čímž poskytuje komplexní přehled o více analýzách v přístupném formátu.

5.2.1 Využitá data

Soubor dat z experimentální analýzy RNA-Seq, dále zmiňováno jako "reálná data"nebo "experimentální data", který zkoumal transkripční odpovědi bakterie *Clostridium beijerinckii* kmene NRRL B-598 na butanolový šok, byl získán z databáze NCBI Sequence Read Archive pod přístupovým číslem SRP033480 [58]. Jako vstupní data pro simulaci pomocí vlastního nástroje byla použita nejnovější verze genomu *Clostridium beijerinckii* NRRL B-598 dostupná z databáze NCBI GenBank pod přístupovým číslem CP011966.3.

Bakterie *Clostridium beijerinckii*

Clostridium beijerinckii je anaerobní bakterie náležející do rodu *Clostridium*, který se vyznačuje schopností produkovat různé metabolity. Je známá svými výraznými fermentačními schopnostmi, kdy přeměňuje organické látky na produkty, jako jsou kyselina máselná, aceton, butanol a ethanol. *Clostridium beijerincki* i proto hraje významnou roli v průmyslových aplikacích, zejména v oblasti biotechnologie [59].

5.2.2 Popis výstupů

V průběhu kapitoly bude pozornost věnována analýze simulovaných dat, která jsou generována na základě stálého, výše zmíněného referenčního genomu bakterie *Clostridium beijerincki*. Při simulacích jsou měněny pouze parametry pro demonstraci změn ve výsledcích.

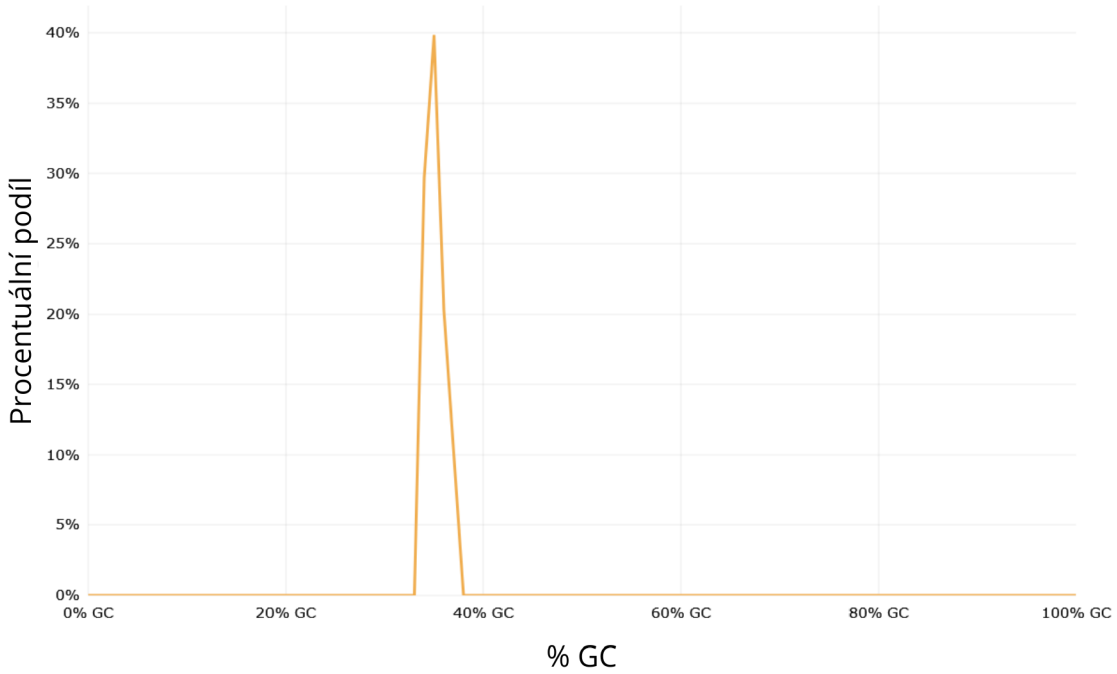
Obsah GC

Jedním z nastavitelných parametrů simulace je procentuální obsah GC. Procentuální obsah GC v DNA je poměr počtu bází guaninu (G) a cytosinu (C) vzhledem k celkovému počtu bází v molekule. Vyjadřuje se jako procento a může být ovlivněn různými faktory. Klíčovou roli hrají například genetické vlastnosti; různé druhy a dokonce i různé oblasti genomu mají různý obsah GC. Bakteriální genom tedy může mít vyšší obsah GC než genom eukaryotních organismů. Navíc v bakteriálním genomu může být vyšší obsah GC v genech, které jsou aktivní za určitých podmínek, jako je stres nebo vyšší teplota. Organismy vystavené extrémním podmínkám, jako je vysoká teplota nebo extrémní pH, mohou vyvinout vyšší obsah GC, což jim může poskytnout výhodu pro přežití. Obsah GC může být také důležitý pro určité funkce DNA, jeho vyšší hodnoty můžou zvýšit stabilitu sekundární struktury DNA. [60]

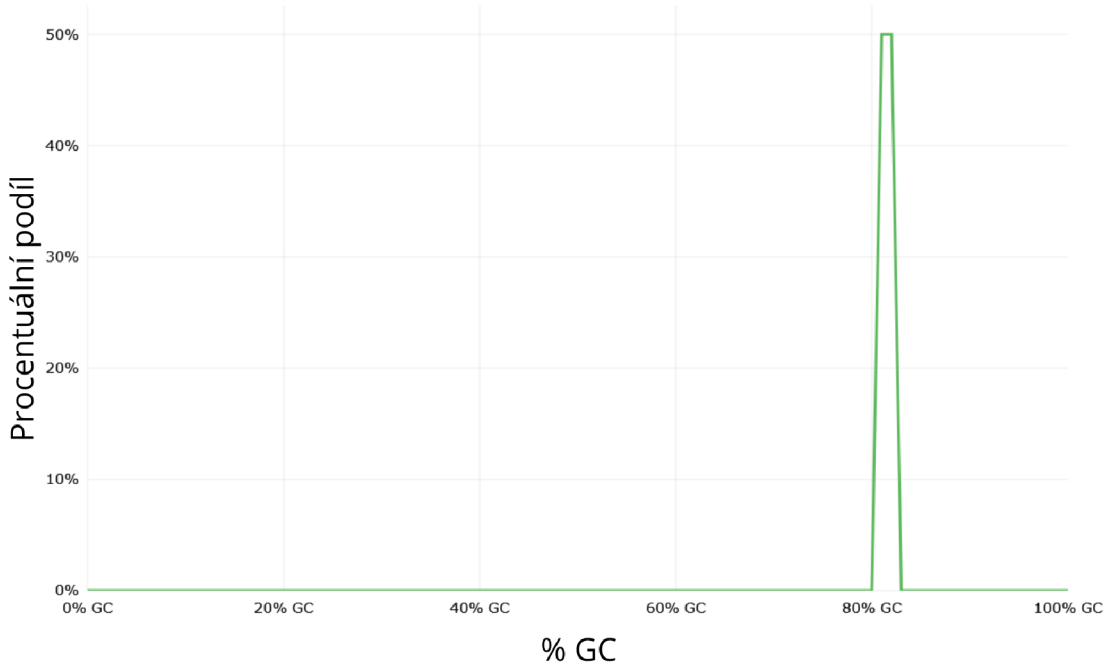
Nastavení %GC uživatelem v simulátoru přináší kontrolu nad procesem simulace. Tato funkce umožňuje uživateli lépe modelovat různorodost RNA sekvencí podle specifických podmínek organismů nebo experimentů. Na následujících stránkách lze vidět na obrázcích 5.6 a 5.8 výstup simulátoru s nastaveným procentem GC na hodnotu 35, na obrázcích 5.7 a 5.9 na hodnotu 82. Základní parametry byly nastaveny na délku čtení 75, počet čtení 10 000, zbývající volitelné parametry nebyly využity.

Histogramy obsahu GC (na obrázcích 5.6 a 5.7) zobrazují rozdělení procenta guaninu a cytosinu ve všech sekvencích ve vzorku. Neobvykle vysoký nebo nízký GC obsah může naznačovat kontaminaci vzorku DNA z jiných organismů s odlišným GC profilem.

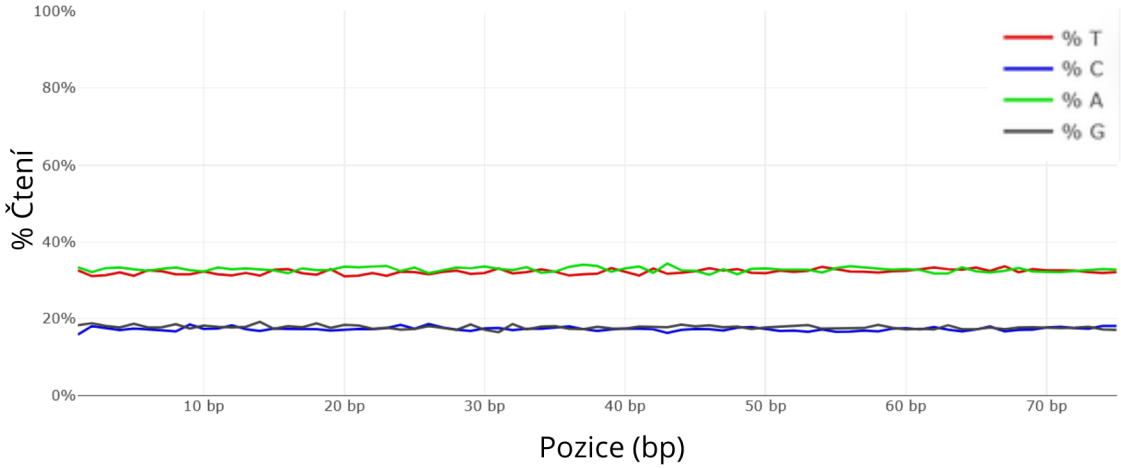
Graf obsahu sekvence bází (na obrázcích 5.8 a 5.9) ukazují variabilitu zastoupení jednotlivých bází na každé pozici ve všech sekvencích. Ideálně by procenta A, T, G a C měla být na každé pozici konstantní, což by naznačovalo absenci biasu při sekvenování (systematické zkreslení v sekvenačních datech, které může vést k nerovnoměrnému nebo nepřesnému zastoupení některých sekvencí nebo regionů genomu). Výrazné odchylky v těchto grafech mohou indikovat technické problémy, jako je degradace vzorku, problémy s sekvenováním nebo přítomnost kontaminace. Také se často využívají k detekci adaptérových sekvencí, které mohou být přítomny na koncích čtecích cyklů a způsobovat zvýšení specifických bází.



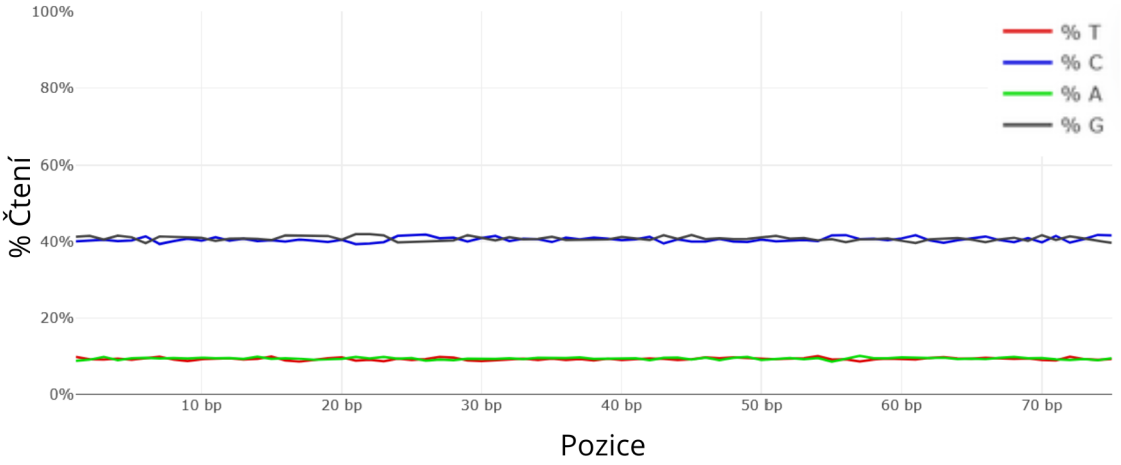
Obr. 5.6: Histogram distribuce GC s obsahem 35%.



Obr. 5.7: Histogram distribuce GC s obsahem 82%.



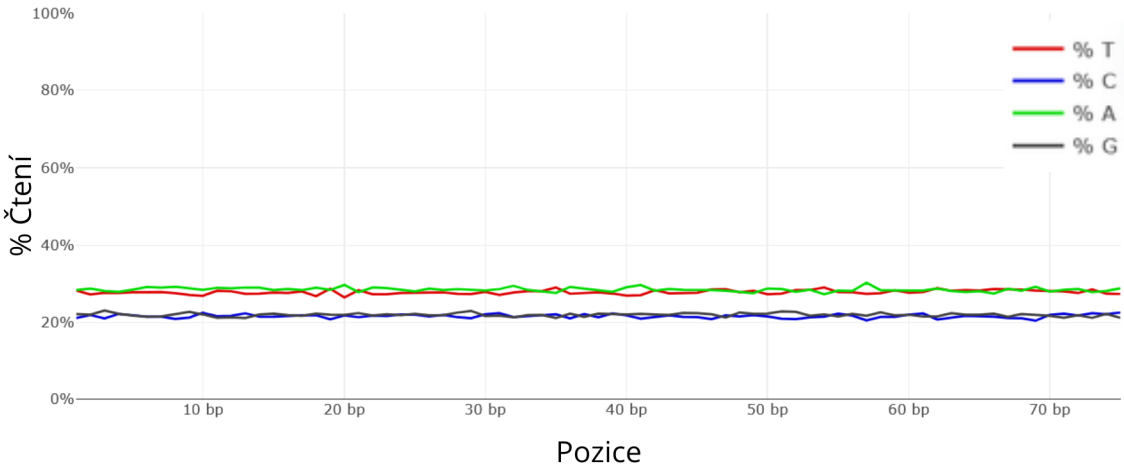
Obr. 5.8: Graf obsahu sekvence bází pro hodnotu GC 35%.



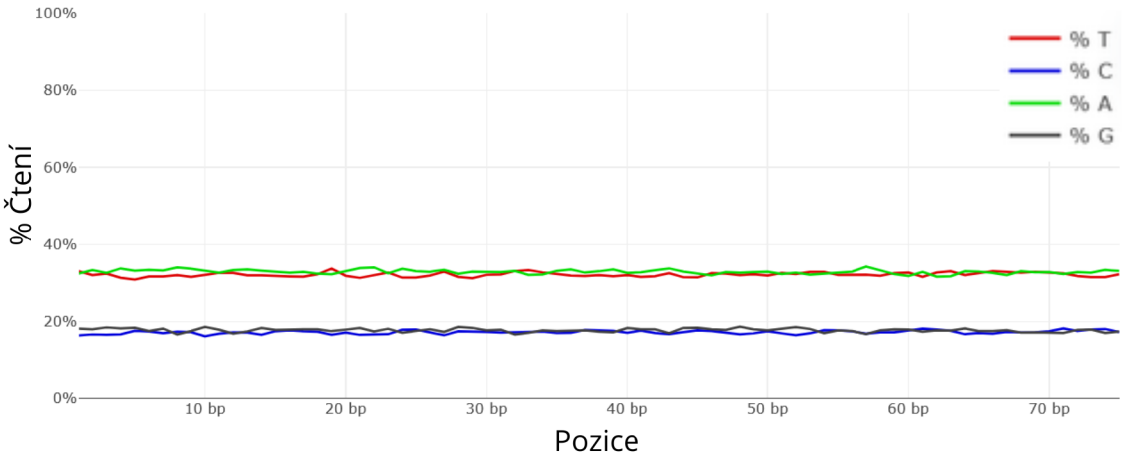
Obr. 5.9: Graf obsahu sekvence bází pro hodnotu GC 82%.

Operony

Zohlednění výskytu operonů při simulaci RNA-Seq dat pro bakteriální genom je důležité pro modelování skutečného chování bakteriálních genů a pro pochopení regulačních mechanismů v bakteriálních genomech. Jak již bylo zmíněno v teoretické části práce, operony jsou skupiny genů, které jsou transkribovány jako jediná transkripční jednotka. Zohlednění operonů může také pomoci při analýze dat z RNA-Seq experimentu. Získané výsledky mohou být zkresleny, pokud není brán v úvahu vztah mezi geny v rámci operonů. Cílí se tak na přesnější predikci genové exprese. Na obrázcích 5.10 a 5.11 lze vidět čtení generovaná s hodnotami: délka čtení 75, počet čtení 10 000, obsah GC 44%. Jediným rozdílem je ten, že při jedné ze simulací byla zohledněna přítomnost operonů ve vstupním genomu využitím volitelného parametru *operon_locations*. Lze si tak povšimnout drobného rozdílu v grafech, především v rozestupu jednotlivých dvojic křivek – výskyt operonů tedy není neúčastný a způsobuje odchylky ve výsledcích.



Obr. 5.10: Graf obsahu sekvence bází pro simulaci se zohledněním operonů.



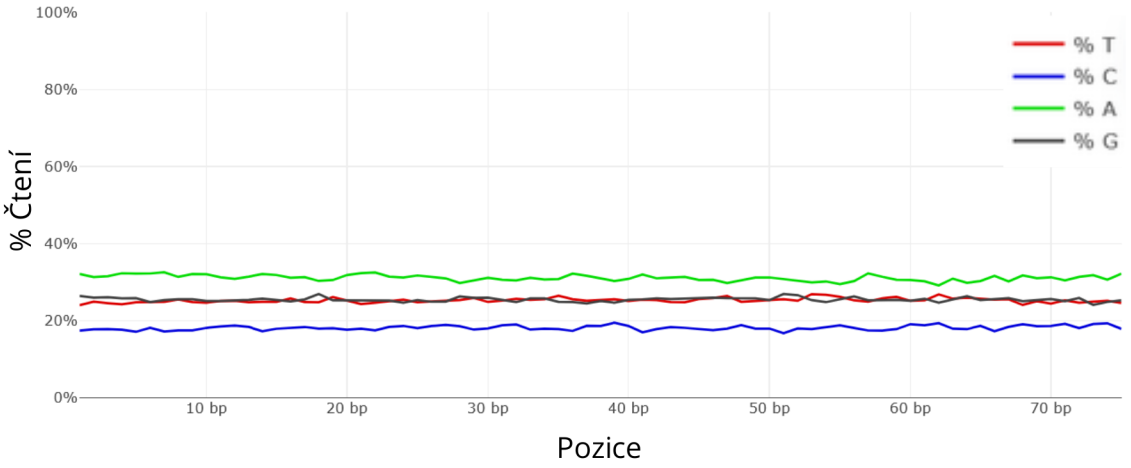
Obr. 5.11: Graf obsahu sekvence bází pro simulaci bez zohlednění operonů.

Duplikace

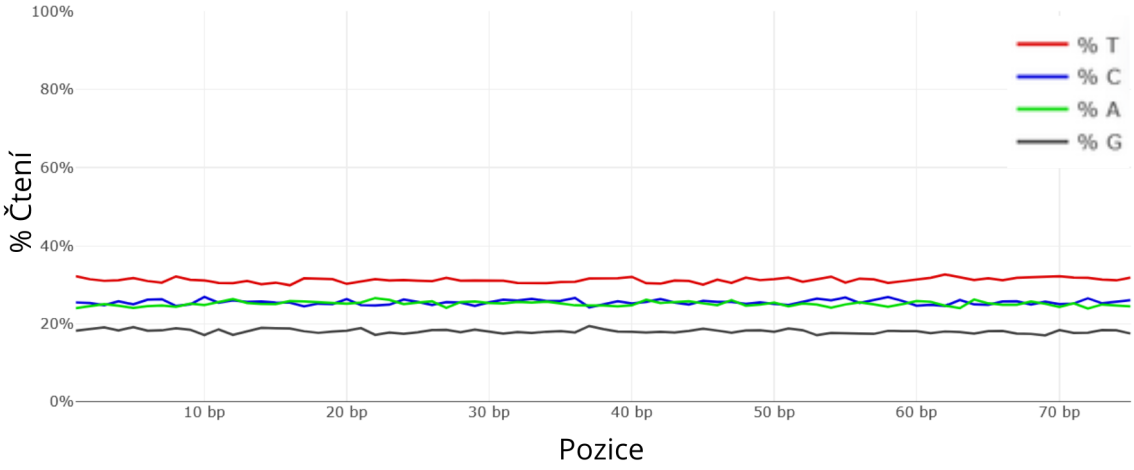
V datech RNA-Seq experimentů se často vyskytují duplikace. Bývají záměrně zahrnuty z důvodu zajištění spolehlivosti a přesnosti výsledků, statistické analýzy a odhadu variability mezi replikacemi, navíc umožňují minimalizovat náhodné chyby. Při využití knihovny se zahrnutými operony ve formě uměle vytvořených molekul polycistronní RNA došlo ke vzniku duplikací, a to v počtu 6 duplikovaných čtení vůči 9994 unikátním. Při větších datových objemech tedy lze očekávat vyšší zastoupení duplikací respektující reálné podoby výsledků.

Vláknová orientace

Zvláštním typem RNA-Seq experimentů jsou strand-specific experimenty, které umožňují určit, ze kterého řetězce DNA pochází každá transkripční sekvence. To znamená, že experimenty dokáží rozlišit mezi transkripty produkovanými z dopředného a reverzního vlákna. Tento typ experimentu se provádí přidáním adaptérů specifických pro každý směr transkriptu, což umožňuje zachovat informaci o směru transkripce. Strand-specific experimenty jsou užitečné pro přesnější mapování transkriptů a analýzu. Pomáhají odhalit diferenciální exprese genů a přesněji určit počátek transkriptů a exony, mohou tak vést k pochopení regulace genové exprese, identifikaci nových genů a ncRNA (non-coding RNA), jsou užitečné i pro studium přesné struktury transkripčních jednotek v genomu. Při simulaci je proto umožněno vytvořit si vstupní knihovnu tvořenou pouze dopřednými, nebo reverzními vlákny. Mezi grafy na obrázcích 5.12 a 5.13 je jasně viditelný rozdíl v "pořadí"křivek jednotlivých nukleotidů, reprezentující orientaci vláken tvořících vstupní knihovnu.



Obr. 5.12: Graf obsahu sekvence bází pro simulaci s dopřednými vlákny.



Obr. 5.13: Graf obsahu sekvence bází pro simulaci s reverzními vlákny.

Napodobení reálných dat

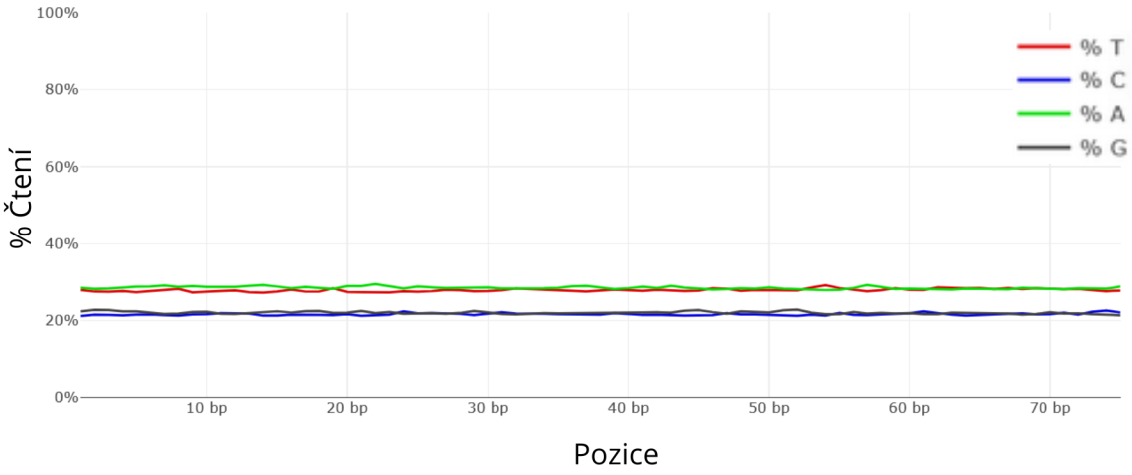
Pro otestování schopnosti simulátoru přiblížit se svými výstupy přinejmenším vizuálně reálným datům byla simulace provedena nejprve pouze s využitím základních parametrů simulace. Data byla simulována ve větším objemu s využitím Metacentra [61], které poskytlo potřebné výpočetní zdroje a infrastrukturu umožňující efektivně zpracovat velké množství sekvenačních dat. Použitými parametry byla délka čtení 75 bp, počet čtení 4 000 000, obsah GC 44%.

Měření	**Hodnota pro exp. data**	**Hodnota pro sim. data**
Název souboru	SRR1055.6741.fastq.gz	outputMC.fastq
Kódování	Sanger / Illumina 1.9	Sanger / Illumina 1.9
Celkový počet sekvencí	43414028	4000000
Celkový počet bazí	3.2 Gbp	300 Mbp
Délka čtení	75	75
%GC	44	44

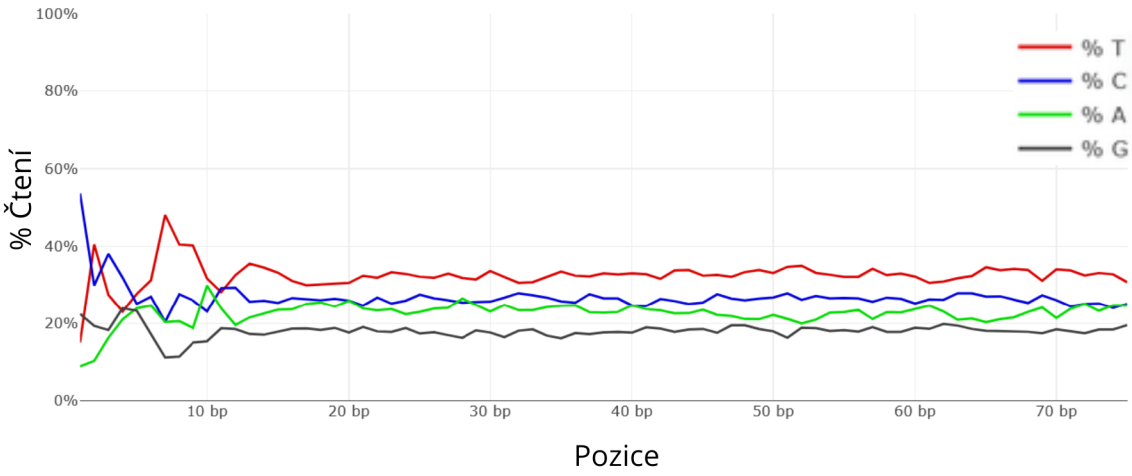
Tab. 5.1: Tabulka se základním přehledem pro experimentální a simulovaná data.

V tabulce 5.1 je uvedeno, že experimentální data mají celkový počet bází 3.2 Gbp (gigabáze), což je typické pro hluboké sekvenování, kde je potřeba získat velké množství dat pro detailní genetické analýzy. Simulovaná data mají celkový počet bází 300 Mbp (megabáze). Toto množství je dostatečné pro testování nebo validaci experimentálních metodik, nicméně by nebylo dostatečně rozsáhlé pro komplexní genetické studie. Oba datasety si zachovávají stejné hodnoty pro délku čtení (75 bp) a procento GC obsahu (44 %).

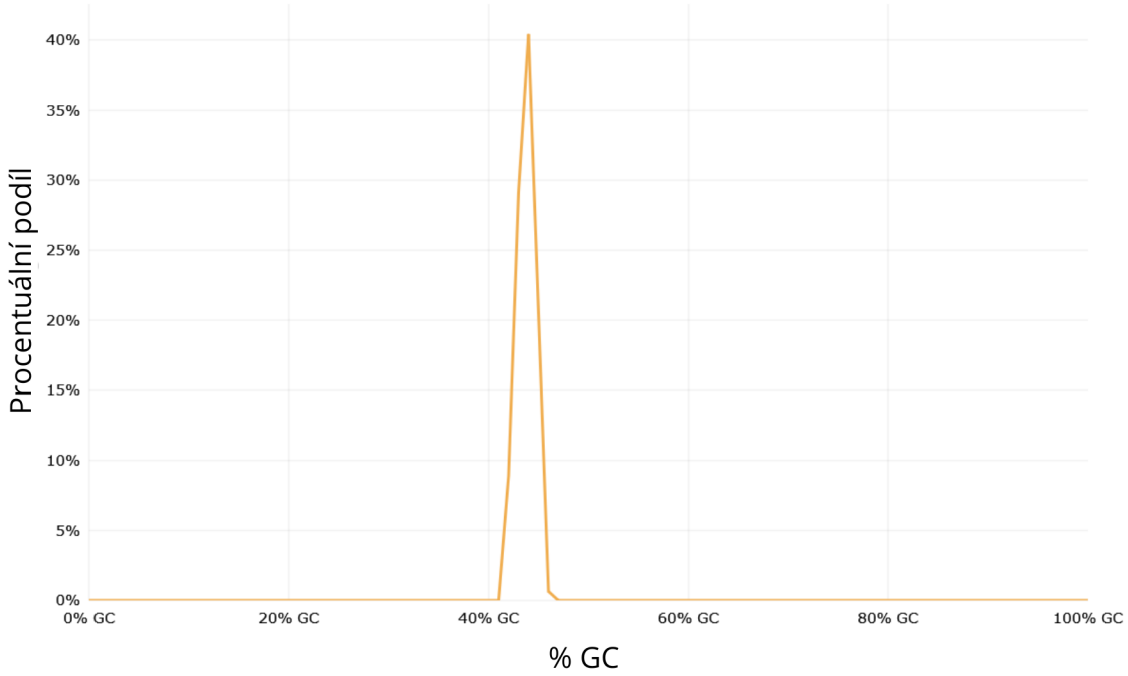
Jak bylo zmíněno výše - při větších datových objemech lze od simulovaných dat očekávat vyšší zastoupení duplikací. Vzniklo tak 53 367 duplikovaných čtení na 3 946 633 unikátních. Došlo však k víceméně teoretické distribuci obsahu GC, chybějící kontaminace navíc vedly k jakémusi „vyhlazení“ křivek pro jednotlivé nukleotidy oproti experimentálním datům - vyobrazeno v grafech na obrázcích 5.14, 5.15 a 5.16 na následujících stranách. Z tohoto důvodu byly pro lepší viditelnost rozdílů všechny demonstrační grafy v předchozích kapitolách generovány pro menší počet čtení, a to 10 000.



Obr. 5.14: Graf obsahu sekvence bází pro simulovaná data generovaná v počtu 4 milionů čtení.

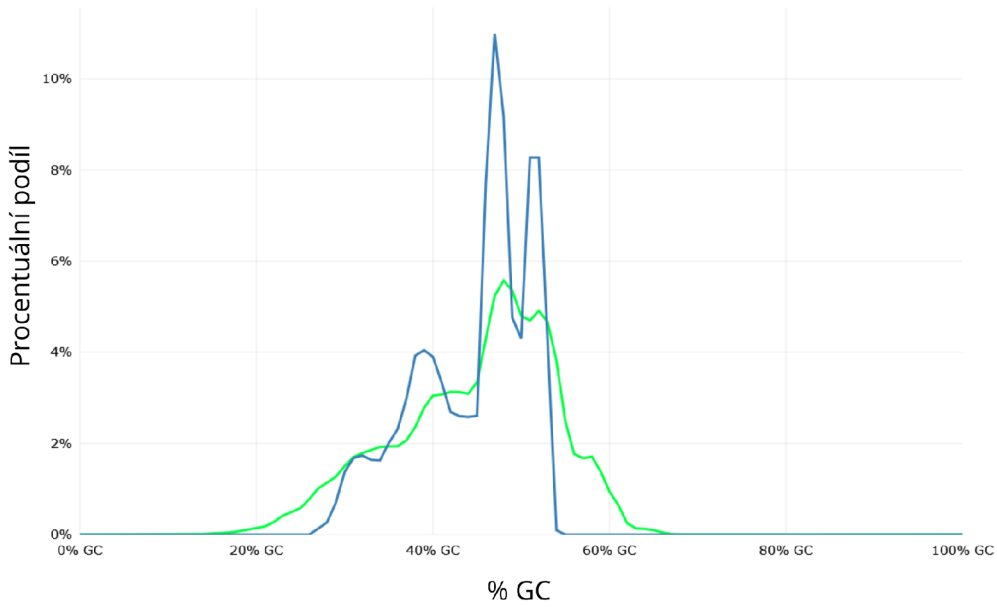


Obr. 5.15: Graf obsahu sekvence bází pro experimentální data.



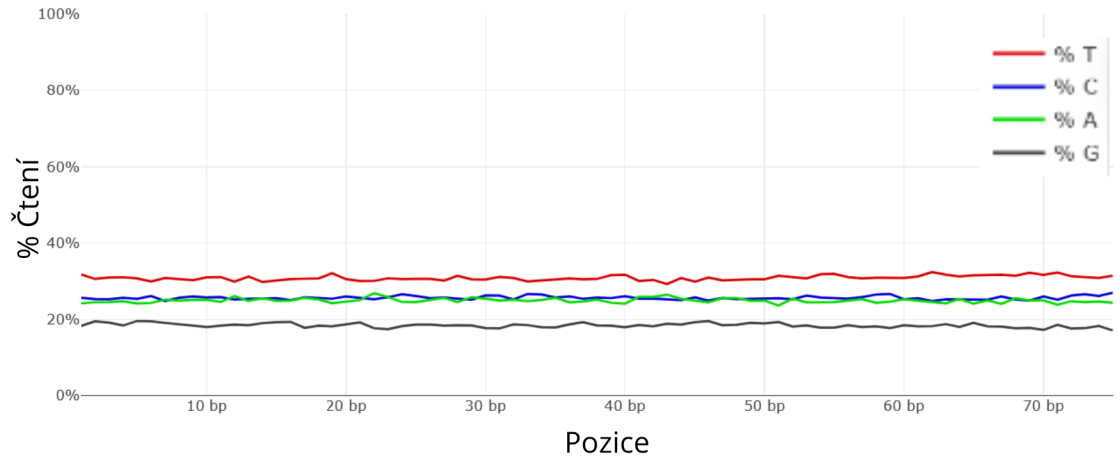
Obr. 5.16: Histogram distribuce GC pro simulovaná data generovaná v počtu 4 milionů čtení.

Nastavením volitelných parametrů simulace a průběžnou kontrolou výstupů je však možné se podobě reálných dat ještě více přiblížit. Na obrázku 5.17 je znázorněn graf, který ukazuje distribuce obsahu GC napříč experimentálními i simulovanými čteními při simulaci s odlišnými hodnotami parametrů, než tomu bylo ve výše uvedeném příkladu - přibylo zohlednění kódujících a operonových oblastí. Modrá křivka reprezentující simulovaná data má vrchol distribuce přibližně ve středu, průměrný GC obsah sekvencí je tedy kolem 50%. Zelená křivka (experimentální data) má podobný tvar distribuce, nicméně zahrnuje mnohem větší počet sekvencí. Distribuce GC obsahu je podobná, s vrcholem také kolem 50%. GC obsah je v obou datasetech rozložen kolem stejného průměru s mírně odlišnou variabilitou. Vytvořením většího množství různých souborů s rozmanitými parametry by bylo možné se ještě více přiblížit podobě experimentálních dat.

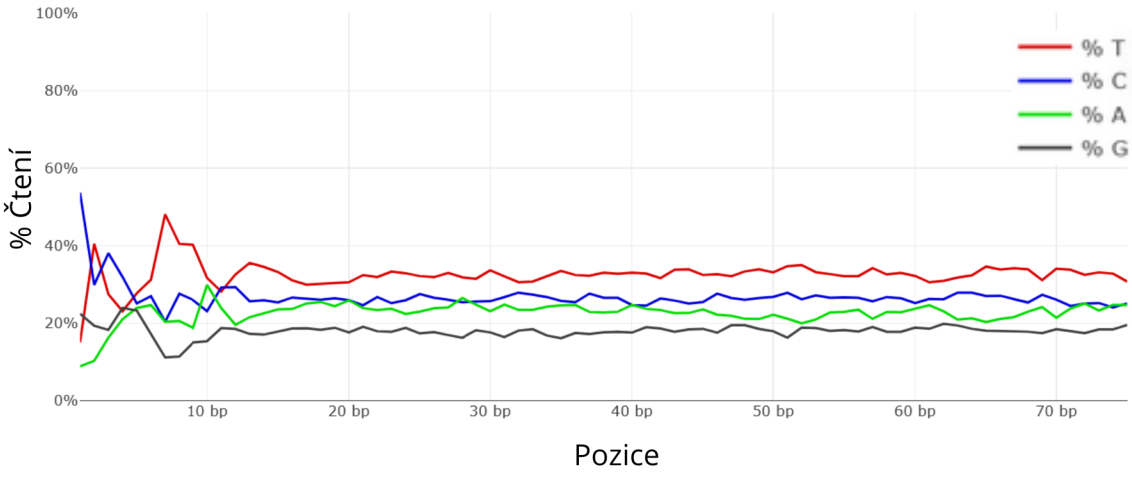


Obr. 5.17: Histogram distribuce GC pro simulovaná (modrá křivka) a experimentální (zelená křivka) data.

Po vizuálním zhodnocení všech předchozích výsledků byl program doplněn o novou funkci k úpravě obsahu GC v sekvenci s možnou nastavitelnou procentuální odchylkou. Následně byla data simulována tak, aby výsledky věrněji odpovídaly těm experimentálním (s využitím menšího množství čtení). Bylo vygenerováno několik výstupních souborů s různým počtem čtení (v rozmezí 1000 až 5000) a různými rozsahy %GC (v rozmezí 20 až 60 %), nově byl použit volitelný parametr *strand_ori* s výběrem dopředných vláken. Výstupní soubory byly následně spojeny do jednoho výstupního souboru. Srovnání dosažených výsledků je uvedeno na následujících grafech.



Obr. 5.18: Graf obsahu sekvence bází pro simulovaná data.



Obr. 5.19: Graf obsahu sekvence bází pro experimentální data.

Graf na obrázku 5.18 ukazuje velmi stabilní složení nukleotidů po celé délce čtení simulovaných dat. Druhý graf reprezentující experimentální data na obrázku 5.19 na-

opak vykazuje výrazné výkyvy na začátku čtení (pozice 1-15). Výkyvy mohou být důsledkem sekvenačního biasu či specifického složení některých regionů v sekvencích, také adaptérové ligace nebo PCR amplifikace mohou vést k nerovnoměrnému zastoupení nukleotidů na koncích sekvencí. Tento jev může ovlivnit výsledky analýz, pokud nejsou tyto počáteční variace náležitě zohledněny. Oba grafy však sdílí stejné pořadí křivek jednotlivých nukleotidů s podobnými rozestupy.

Výstupní soubory

Obrázek 5.20 zobrazuje útržek výstupního FASTQ souboru vygenerovaného BacSeqerem. Každý záznam v FASTQ souboru se skládá ze čtyř řádků: první řádek začíná znakem '@' a obsahuje identifikaci sekvence, druhý řádek zobrazuje samotnou sekvenci nukleotidů, třetí řádek s oddělovačem '+' může být následován identifikátorem sekvence, a čtvrtý řádek obsahuje ASCII znaky odpovídající kvalitním skóre pro každý nukleotid z druhého řádku.

```
@gene-Cbei_0549_read0
AGAACTATACAAAGAGTTTGCGGAGATAGGTGGAATTGTTCAGGCACATTCAACTTGGGCTACTATATGGGCGCA
+
FIGDDIGDBIGIGEBFFBDIIIAICHE@@GIDFI@AF?BF??C?B??GF@?C@?C@CBECCEC???D??D????B
@cds-ABR33454.1_read1
TTACTACTGCGCACATGGGTGGAATCCGAATTTTAGAAGTACCGTAAGCAATACTGGCAACACAAGAGTCTATGC
+
IEGIIICHHEIIGDFIHIAIADIEEBB@I@DAGHDA?A?D?A?FHC??ECGC@FBAC?CB?B@CB?ADB?????@
@gene-Cbei_3169_read2
GCATATATAACTTCTCCCTGTGTGTGCACACCAATTGAGCTTAATGAAGTTTCAAGTAGCCCACCATCCCACAGT
+
FIEIIIFCGIICBIEIGBBGICBEDIHIDIDBA@B?@@G?EDED???AD?@B@DA?D?FCA?B??EA?DB?B@BC
@gene-Cbei_5074_read3
ACCTGTAGGCCCGATACCGTTTTTGCCTCTAGCCTTTTCTTTTAATTTGTCTAGTACCTTCTGGGATGGCAGTAT
+
IIDIIIGHFIFIIFBDHHFAAGHHGCICGGIBFGHIEDCDCBCDH?CBFGGAB@BFCD?AECE?A?C???B?BB?
@gene-Cbei_0777_read4
AACTTATGGGGTTACTTTTACCATGTTGGAGAAAATTGGTGTGACTGGAACGAGCGCACATCCGCTACACGAATT
+
FICDIGIIHIIHIFDDFEFIIBIIGB@EHHBGDDABI@B@@EB?@EBBC?A?F@?E@?AAF??EDCCD???DCD?
```

Obr. 5.20: Ukázka výstupního souboru.

Simulace pomocí nástroje Polyester

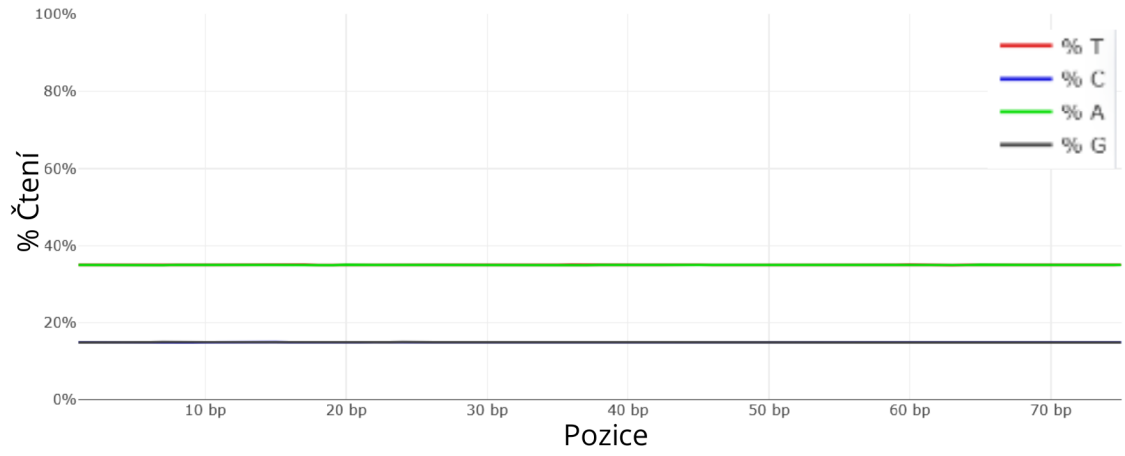
Pro komplexní srovnání výsledků BacSeqeru by bylo vhodné zahrnout data také z jiných simulačních nástrojů. Při zajišťování těchto výsledků se však vyskytlo hned několik překážek. První významnou překážkou byla nedostupnost některých populárních nástrojů. Dále, většina těchto nástrojů je implementována v programovacích jazycích, které vyžadují technické znalosti pro jejich spuštění a účinné nastavení. Nástroj, který je impementován taktéž v Pythonu, bohužel požaduje specifické a obtížně získatelné formáty vstupních dat. Z těchto důvodů byla simulace provedena pomocí nástroje Polyester [48], který je implementovaný v jazyce R. Výsledky této simulace lze vidět na přiložené tabulce a obrázcích – tabulka 5.2 poskytuje základní přehled dat, na obrázku 5.21 lze vidět zastoupení jednotlivých nukleotidů, obrázek 5.22 ukazuje distribuci obsahu GC. GC obsah zůstává na 29%, jelikož se nepodařilo dohledat a využít vhodný model simulátoru pro jeho přenastavení.

Měření	**Hodnota pro exp. data**	**Hodnota pro sim. data**
Název souboru	SRR1055.6741.fastq.gz	sample_01_1_output.fastq
Kódování	Sanger / Illumina 1.9	Illumina 1.5
Celkový počet sekvencí	43414028	2000000
Celkový počet bazí	3.2 Gbp	150 Mbp
Délka čtení	75	75
%GC	44	29

Tab. 5.2: Tabulka se základním přehledem pro experimentální data a data simulovaná pomocí nástroje Polyester.

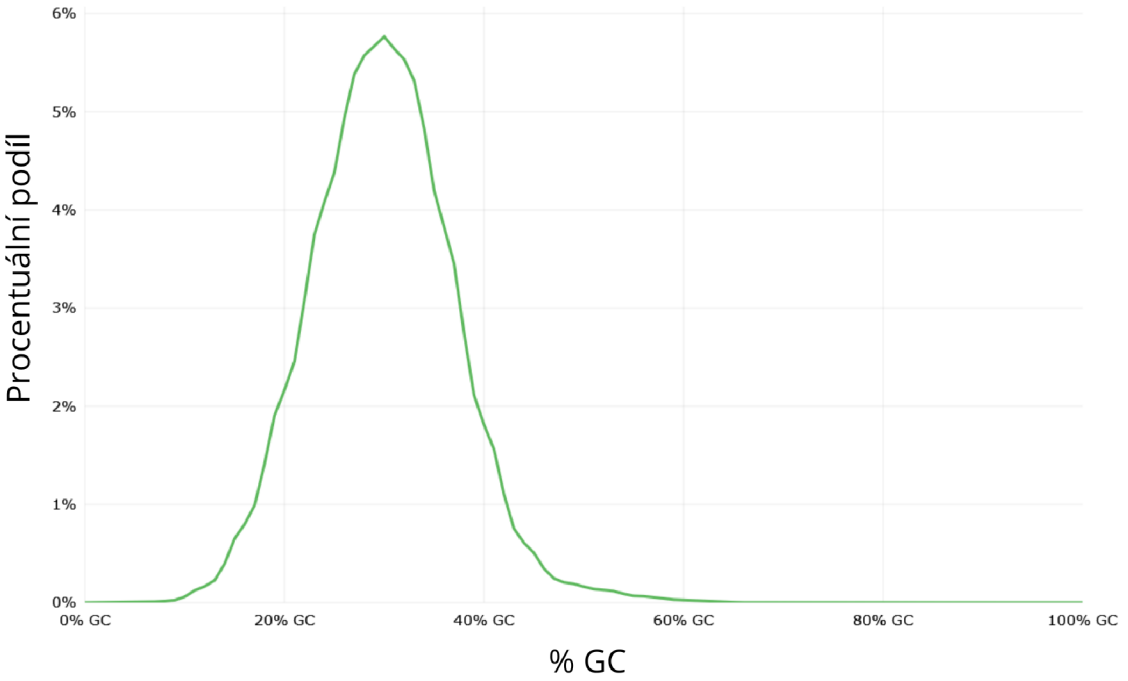
Z tabulky 5.2 lze vyčíst, že experimentální data využila Phred+33 kódování pro kvalitu sekvencí (Sanger / Illumina 1.9). Simulovaná data mají kódování specifikováno jako Illumina 1.5, což značí starší Phred+64 kódování. Tento rozdíl by mohl ovlivnit interpretaci kvality sekvencí při analýze.

Z grafu na obrázku 5.21 je patrné, že zastoupení všech čtyř nukleotidů je konstantní po celou délku čtení, což by byl ideální stav, a naznačuje, že nejsou přítomna žádná významná zkreslení způsobená sekvenační technologií nebo přípravou vzorku. Taková rovnoměrnost se však v experimentálních datech běžně nevyskytuje.



Obr. 5.21: Graf obsahu sekvence bází pro simulovaná data simulovaná pomocí nástroje Polyester.

V grafu na obrázku 5.22 křivka pro simulovaná data víceméně kopíruje teoretickou distribuci.



Obr. 5.22: Histogram distribuce GC pro data simulovaná pomocí nástroje Polyester.

5.3 Možnosti optimalizace

Další vývoj BacSeqeru přesahující hranice zadání práce by mohl vést k tvorbě velmi užitečného nástroje pro simulaci RNA-Seq dat.

Tento vývoj by měl zahrnovat rozšíření funkčností simulátoru pro lepší simulaci variabilit vlivem stresorů prostředí, mutačních rychlostí a adaptivních změn. Zahrnutí posttranskripčních úprav, jako jsou rozdíly v polyadenylaci a stabilizaci mRNA, by zlepšilo modelování dynamiky mRNA a přesnost předpovědí chování RNA. Modelování interakcí mezi různými typy RNA, včetně sRNA a ncRNA, které ovlivňují stabilitu a translaci mRNA, by poskytlo hlubší vhledy do regulace genové exprese.

Technické aspekty sekvenace, jako jsou sekvenační chyby, duplicity čtení a bias způsobený při přípravě knihovny, by pomohly uživateli lépe rozumět a korigovat tyto artefakty ve skutečných datech.

V neposlední řadě by bylo vhodné zvážit také vytvoření grafického uživatelského rozhraní pro snadnější přístup a manipulaci s nástrojem, stejně jako podporu pro různé formáty vstupních a výstupních dat.

Závěr

Záměrem práce bylo poskytnout přehled o bakteriálním genomu, způsobech měření exprese genetické informace, především metody RNA-Seq, o nástrojích sloužících k simulaci RNA-Seq dat, a v neposlední řadě provést implementaci vlastního nástroje zaměřujícího se na specifika bakteriálního genomu, konkrétně výskytu operonů.

První kapitola práce se věnuje popisu bakterií a jejich genomu, je zde vysvětlena problematika operonů. Obsahuje také základní popis exprese genetické informace.

Druhá kapitola je zaměřena na přehled moderních sekvenačních platforem úzce související s následující, třetí kapitolou o způsobech měření genetické exprese. Jedná se o metody microarrays a RTq-PCR, největší pozornost je však věnována metodě RNA-Seq.

Čtvrtá sekce poskytuje přehled některých dostupných nástrojů pro simulaci RNA-Seq dat. V kapitole jsou zmíněny zásadní rozdíly mezi jednotlivými nástroji, například jejich specifické funkce a vlastnosti, požadované vstupní formáty či jazyk implementace.

Hlavním cílem této práce bylo vyvinout nástroj, který umožňuje simulovat data RNA-Seq s ohledem na specifika bakteriálního genomu, zejména na přítomnost operonů. Tomuto nástroji je věnována pátá kapitola práce. BacSeqer obsahuje několik funkcí umožňujících simulaci právě takových dat, přičemž vstupní formáty jsou široce dostupné a běžné. Výstup simulátoru byl porovnán s výsledky skutečného experimentu pomocí nástroje FastQC, přičemž oba soubory dat obsahovaly sekvence genomu bakterie *Clostridium beijerinckii*. V kapitole jsou uvedeny příklady výsledků simulace s nastavením různých parametrů.

Poslední kapitola zmiňuje konkrétní možnosti dalšího vývoje simulátoru, které by vedly ke zvýšení efektivity simulace a užitečnosti výsledků.

Přínosem této práce je především možnost použití BacSeqeru v oblastech, kde je experimentální měření nákladné nebo technicky náročné. Simulátor může sloužit k testování hypotéz o genové expresi, k optimalizaci experimentálních podmínek nebo k přípravě analýz, které by mohly identifikovat nové regulátorové sekvence nebo nově exprimované geny.

Literatura

[1] KADNER, Robert J. a ROGERS, Kara. Bacteria. Online. *Encyclopedia Britannica* 2023. Dostupné z: https://www.britannica.com/science/bacteria. [cit. 2023-11-11].

[2] VELLAI, T. a VIDA, G. The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. Online. *Proceedings of the Royal Society of London. Series B: Biological Sciences*. 1999, roč. 266, č. 1428, s. 1571-1577. ISSN 0962-8452. Dostupné z: https://doi.org/10.1098/rspb.1999.0817. [cit. 2023-11-11].

[3] BOBAY, Louis-Marie a OCHMAN, Howard. The Evolution of Bacterial Genome Architecture. Online. *Frontiers in Genetics* 2017, roč. 8. ISSN 1664-8021. Dostupné z: https://doi.org/10.3389/fgene.2017.00072. [cit. 2023-11-11].

[4] SU, C J a BASEMAN, J B. Genome size of Mycoplasma genitalium. Online. *Journal of Bacteriology* 1990, roč. 172, č. 8, s. 4705-4707. ISSN 0021-9193. Dostupné z: https://doi.org/10.1128/jb.172.8.4705-4707.1990. [cit. 2023-11-11].

[5] SCHNEIKER, Susanne; PERLOVA, Olena; KAISER, Olaf; GERTH, Klaus; ALICI, Aysel et al. Complete genome sequence of the myxobacterium Sorangium cellulosum. Online. *Nature Biotechnology* 2007, roč. 25, č. 11, s. 1281-1289. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/nbt1354. [cit. 2023-11-11].

[6] Přispěvatelé Wikipedie, Plazmid. online. *Wikipedie: Otevřená encyklopedie*. 2023. DOstupné z: https://cs.wikipedia.org/w/index.php?title=Plazmid&oldid=22774336. [cit. 2023-12-26]

[7] DORMAN, Charles J. Structure and Function of the Bacterial Genome. Online. *Wiley*. 2020. ISBN 9781119308799. Dostupné z: https://doi.org/10.1002/9781119309697. [cit. 2023-11-30].

[8] BARIL, C.; RICHAUD, C.; BARANTON, G. a GIRONS, I.Saint. Linear chromosome of Borrelia burgdorferi. Online. *Research in Microbiology*. 1989, roč. 140, č. 7, s. 507-516. ISSN 09232508. Dostupné z: https://doi.org/10.1016/0923-2508(89)90083-1. [cit. 2023-12-06].

[9] DILLON, Shane C. a DORMAN, Charles J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Online. *Nature Reviews Microbiology* 2010, roč. 8, č. 3, s. 185-195. ISSN 1740-1526. Dostupné z: https://doi.org/10.1038/nrmicro2261. [cit. 2023-11-11].

[10] LI, Jihong; ADAMS, Vicki; BANNAM, Trudi L.; MIYAMOTO, Kazuaki; GARCIA, Jorge P. et al. Toxin Plasmids of Clostridium perfringens. Online. *Microbiology and Molecular Biology Reviews*. 2013, roč. 77, č. 2, s. 208-233. ISSN 1092-2172. Dostupné z: https://doi.org/10.1128/MMBR.00062-12.. [cit. 2023-12-06].

[11] BENNETT, P M. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. Online. *British Journal of Pharmacology*. 2008, roč. 153, č. S1. ISSN 0007-1188. Dostupné z: https://doi.org/10.1038/sj.bjp.0707607. [cit. 2023-12-06].

[12] THOMAS, Christopher M. a NIELSEN, Kaare M. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Online. *Nature Reviews Microbiology*. 2005, roč. 3, č. 9, s. 711-721. ISSN 1740-1526. Dostupné z: https://doi.org/10.1038/nrmicro1234. [cit. 2023-02-27].

[13] OSBOURN, Anne E. a FIELD, Ben. Operons. Online. *Cellular and Molecular Life Sciences* 2009, roč. 66, č. 23, s. 3755-3775. ISSN 1420-682X. Dostupné z: https://doi.org/10.1007/s00018-009-0114-3. [cit. 2023-11-11].

[14] Operons. Online. *BioNinja*. Dostupné z: https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/72-transcription-and-gene/operons.html. [cit. 2023-12-27]

[15] YANOFSKY, C.; PLATT, T.; CRAWFORD, I.P.; NICHOLS, B.P.; CHRISTIE, G.E. et al. The complete nucleotide sequence of the tryptophan operon of Escherichia coli. Online. *Nucleic Acids Research* 1981, roč. 9, č. 24, s. 6647-6668. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/9.24.6647. [cit. 2023-11-11].

[16] GRISWOLD, A. Genome packaging in prokaryotes: the circular chromosome of E. coli. Online. *Nature*. 2008. Dostupné z: https://www.nature.com/scitable/topicpage/genome-packaging-in-prokaryotes-the-circular-chromosome-9113/#. [cit. 2023-11-11].

[17] TORO, Nicolás; JIMÉNEZ-ZURDO, José Ignacio a GARCÍA-RODRÍGUEZ, Fernando Manuel. Bacterial group II introns: not just splicing. Online. *FEMS Microbiology Reviews*. 2007, roč. 31, č. 3, s. 342-358. ISSN 1574-6976. Dostupné z: https://doi.org/10.1111/j.1574-6976.2007.00068.x. [cit. 2023-12-27]

[18] Přednáška Veterinární Univerzity Brno. EXPRESE GENŮ A JEJÍ REGULACE. Online. *VFU Brno*. Dostupné z: https://www.vfu.cz/files/upload/1458/Biologie-prednaska4.2022-2023.pdf. [cit. 2023-12-27]

[19] TURNER, P. C. *Molecular biology. 2nd ed. Instant notes.* Liverpool: BIOS Scientific Publishers. 2000. ISBN 1-85996-152-5. [cit. 2023-11-14].

[20] ROSYPAL, Stanislav. *Úvod do molekulární biologie*. 4., (inovované) vyd. Brno: Stanislav Rosypal, 2005. ISBN 80-902562-5-2. [cit. 2023-11-14].

[21] KHAN, S A. Rolling-circle replication of bacterial plasmids. Online. *Microbiology and Molecular Biology Reviews*. 1997, roč. 61, č. 4, s. 442-455. ISSN 1092-2172. Dostupné z: https://doi.org/10.1128/mmbr.61.4.442-455.1997. [cit. 2023-12-06].

[22] E-Learning VŠCHT. Biochemie II. Online. Dostupné z: https://e-learning.vscht.cz/mod/glossary/view.php?id=60908. [cit. 2023-12-06].

[23] DREW, H.R.; WEEKS, J.R. a TRAVERS, A.A. Negative supercoiling induces spontaneous unwinding of a bacterial promoter. Online. *The EMBO Journal*. 1985, roč. 4, č. 4, s. 1025-1032. ISSN 02614189. Dostupné z: https://doi.org/10.1002/j.1460-2075.1985.tb03734.x. [cit. 2023-12-06].

[24] CIAMPI, M. Sofia. Rho-dependent terminators and transcription termination. Online. *Microbiology*. 2006, roč. 152, č. 9, s. 2515-2528. ISSN 1350-0872. Dostupné z: https://doi.org/10.1099/mic.0.28982-0. [cit. 2023-12-06].

[25] GEBAUER, Fátima a HENTZE, Matthias W. Molecular mechanisms of translational control. Online. *Nature Reviews Molecular Cell Biology*. 2004, roč. 5, č. 10, s. 827-835. ISSN 1471-0072. Dostupné z: https://doi.org/10.1038/nrm1488. [cit. 2023-12-06].

[26] ŠTEFÁNEK, Jiří. Ribozom. Online. *Medicína, nemoci, studium na 1. LF UK*. Dostupné z: https://www.stefajir.cz/ribozom. [cit. 2023-12-06].

[27] HEATHER, James M. a CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. Online. *Genomics*. 2016, roč. 107, č. 1, s. 1-8. ISSN 08887543. Dostupné z: https://doi.org/10.1016/j.ygeno.2015.11.003. [cit. 2023-11-11].

[28] HUERTA, Laura; BURKE, Melissa. Functional genomics II. Online. 2020. Dostupné z: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/. [cit. 2023-11-12].

[29] LAHENS, Nicholas F.; RICCIOTTI, Emanuela; SMIRNOVA, Olga; TOORENS, Erik; KIM, Eun Ji et al. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. Online. *BMC Genomics*. 2017, roč. 18, č. 1. ISSN 1471-2164. Dostupné z: https://doi.org/10.1186/s12864-017-4011-0. [cit. 2023-12-25].

[30] ANASAGASTI, Ander; IRIGOYEN, Cristina; BARANDIKA, Olatz; LÓPEZ DE MUNAIN, Adolfo a RUIZ-EDERRA, Javier. Current mutation discovery approaches in Retinitis Pigmentosa. Online. *Vision Research*. 2012, roč. 75, s. 117-129. ISSN 00426989. Dostupné z: https://doi.org/10.1016/j.visres.2012.09.012. [cit. 2024-05-17].

[31] WANG, Yunhao; ZHAO, Yue; BOLLAS, Audrey; WANG, Yuru a AU, Kin Fai. Nanopore sequencing technology, bioinformatics and applications. Online. *Nature Biotechnology*. 2021, roč. 39, č. 11, s. 1348-1365. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/s41587-021-01108-x. [cit. 2023-12-25].

[32] What is Oxford Nanopore Technology (ONT) sequencing? Online. *YourGenome*. https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/. [cit. 2023-12-27]

[33] RHOADS, Anthony a AU, Kin Fai. PacBio Sequencing and Its Applications. Online. *Genomics, Proteomics & Bioinformatics*. 2015, roč. 13, č. 5, s. 278-289. ISSN 16720229. Dostupné z: https://doi.org/10.1016/j.gpb.2015.08.002. [cit. 2023-12-25].

[34] RAUHUT, Reinhard a KLUG, Gabriele. MRNA degradation in bacteria. Online. FEMS *Microbiology Reviews*. 1999, roč. 23, č. 3, s. 353-370. ISSN 1574-6976. Dostupné z: https://doi.org/10.1111/j.1574-6976.1999.tb00404.x. [cit. 2023-12-06].

[35] GOVINDARAJAN, Rajeshwar; DURAIYAN, Jeyapradha; KALIYAPPAN, Karunakaran a PALANISAMY, Murugesan. Microarray and its applications. Online. *Journal of Pharmacy And Bioallied Sciences*. 2012, roč. 4, č. 6. ISSN 0975-7406. Dostupné z: https://doi.org/10.4103/0975-7406.100283. [cit. 2023-11-30].

[36] LIU, Hongfang. Microarray probes and probe sets. Online. *Frontiers in Bioscience*. 2010, roč. E2, č. 1, s. 325-338. ISSN 1945-0494. Dostupné z: https://doi.org/10.2741/e93. [cit. 2023-12-06].

[37] AFZAL, Muhammad; MANZOOR, Irfan a KUIPERS, Oscar P. A Fast and Reliable Pipeline for Bacterial Transcriptome Analysis Case study: Serine-dependent Gene Regulation in Streptococcus pneumoniae. Online. *Journal of Visualized Experiments*. 2015, č. 98. ISSN 1940-087X. Dostupné z: https://doi.org/10.3791/52649. [cit. 2023-12-27].

[38] HEID, C A; STEVENS, J; LIVAK, K J a WILLIAMS, P M. Real time quantitative PCR. Online. *Genome Research*. 1996, roč. 6, č. 10, s. 986-994. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.6.10.986. [cit. 2023-12-01].

[39] ThermoFisher Scientific. Basic Principles of RT-qPCR. Online. Dostupné z: https://www.thermofisher.com/cz/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/basic-principles-rt-qpcr.html. [cit. 2023-12-01].

[40] OZSOLAK, Fatih a MILOS, Patrice M. RNA sequencing: advances, challenges and opportunities. Online. *Nature Reviews Genetics*. 2011, roč. 12, č. 2, s. 87-98. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/nrg2934. [cit. 2023-11-14].

[41] KODZIUS, Rimantas; KOJIMA, Miki; NISHIYORI, Hiromi; NAKAMURA, Mari; FUKUDA, Shiro et al. CAGE: cap analysis of gene expression. Online. *Nature Methods*. 2006, roč. 3, č. 3, s. 211-222. ISSN 1548-7091. Dostupné z: https://doi.org/10.1038/nmeth0306-211. [cit. 2023-12-06].

[42] HAQUE, Ashraful; ENGEL, Jessica; TEICHMANN, Sarah A. a LÖNNBERG, Tapio. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Online. *Genome Medicine*. 2017, roč. 9, č. 1. ISSN 1756-994X. Dostupné z: https://doi.org/10.1186/s13073-017-0467-4. [cit. 2023-11-14].

[43] STARK, Rory; GRZELAK, Marta a HADFIELD, James. RNA sequencing: the teenage years. Online. *Nature Reviews Genetics*. 2019, roč. 20, č. 11, s. 631-656. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/s41576-019-0150-2. [cit. 2023-11-14].

[44] LEVIN, Joshua Z; YASSOUR, Moran; ADICONIS, Xian; NUSBAUM, Chad; THOMPSON, Dawn Anne et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Online. *Nature Methods*. 2010, roč. 7, č. 9, s.

709-715. ISSN 1548-7091. Dostupné z: https://doi.org/10.1038/nmeth.1491. [cit. 2023-11-14].

[45] RAGHAVENDRA, Pongali; PULLAIAH, Thammineni. Illumina Dye Sequencing. Online. *Advances in Cell and Molecular Diagnostics*. 2018. Dostupné z: https://www.sciencedirect.com/topics/medicine-and-dentistry/illumina-dye-sequencing. [cit. 2023-11-14].

[46] METZKER, Michael L. Sequencing technologies — the next generation. Online. *Nature Reviews Genetics*. 2010, roč. 11, č. 1, s. 31-46. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/nrg2626. [cit. 2023-11-14].

[47] WANG, Kevin C.; CHANG, Howard Y. Molecular mechanisms of long noncoding RNAs. Online. *Molecular cell*. 2011, 43.6: 904-914. Dostupné z: https://www.cell.com/molecular-cell/pdf/S1097-2765(11)00636-8.pdf. [cit. 2023-11-14].

[48] FRAZEE, Alyssa C.; JAFFE, Andrew E.; LANGMEAD, Ben a LEEK, Jeffrey T. Polyester: simulating RNA-seq datasets with differential transcript expression. Online. *Bioinformatics*. 2015, roč. 31, č. 17, s. 2778-2784. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btv272. [cit. 2023-11-18].

[49] GRIEBEL, Thasso; ZACHER, Benedikt; RIBECA, Paolo; RAINERI, Emanuele; LACROIX, Vincent et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Online. *Nucleic Acids Research*. 2012, roč. 40, č. 20, s. 10073-10083. ISSN 1362-4962. Dostupné z: https://doi.org/10.1093/nar/gks666. [cit. 2023-11-25].

[50] RNASeqReadSimulator [software]. Online. Dostupné z: http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html#history. [cit. 2023-11-28].

[51] BENIDT, Sam a NETTLETON, Dan. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. Online. *Bioinformatics*. 2015, roč. 31, č. 13, s. 2131-2140. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btv124. [cit. 2023-12-06]

[52] BROOKS, Thomas G.; LAHENS, Nicholas F.; MRČELA, Antonijo; SARANTOPOULOU, Dimitra; NAYAK, Soumyashant; NAIK, Amruta; SENGUPTA, Shaon; CHOI, Peter S.; GRANT, Gregory R. BEERS2: RNA-Seq simulation through high fidelity in silico modeling. Online. *PubMed Central*. 2023. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10168222/. [cit. 2023-12-25].

[53] HUANG, Weichun; LI, Leping; MYERS, Jason R. a MARTH, Gabor T. ART: a next-generation sequencing read simulator. Online. *Bioinformatics*. 2012, roč. 28, č. 4, s. 593-594. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btr708. [cit. 2024-05-16].

[54] HU, Xuesong; YUAN, Jianying; SHI, Yujian; LU, Jianliang; LIU, Binghang et al. PIRS: Profile-based Illumina pair-end reads simulator. Online. *Bioinformatics*. 2012, roč. 28, č. 11, s. 1533-1535. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/bts187. [cit. 2024-05-16].

[55] WILKIE, Gavin S.; DICKSON, Kirsten S. a GRAY, Nicola K. Regulation of mRNA translation by 5- and 3-UTR-binding factors. Online. *Trends in Biochemical Sciences*. 2003, roč. 28, č. 4, s. 182-188. ISSN 09680004. Dostupné z: https://doi.org/10.1016/S0968-0004(03)00051-3. [cit. 2024-05-17].

[56] FastQC: A quality control tool for high throughput sequence data. Dostupné z: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [cit. 2024-03-31].

[57] EWELS, Philip; MAGNUSSON, Mans; LUNDIN, Sverker; KÄLLER, Max. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. Dostupné z: https://doi.org/10.1093/bioinformatics/btw354. [cit. 2024-03-31].

[58] SEDLAR, Karel; KOLEK, Jan; GRUBER, Markus; JURECKOVA, Katerina; BRANSKA, Barbora et al. A transcriptional response of Clostridium beijerinckii NRRL B-598 to a butanol shock. Online. *Biotechnology for Biofuels*. 2019, roč. 12, č. 1. ISSN 1754-6834. Dostupné z: https://doi.org/10.1186/s13068-019-1584-7. [cit. 2023-12-23].

[59] EZEJI, Thaddeus; QURESHI, Nasib; BLASCHEK, Hans P. Butanol production from agricultural residues: Impact of degradation products on Clostridium beijerinckii growth and butanol fermentation. Online. *Biotechnology and Bioengineering*. 2007. Dostupné z: https://doi.org/10.1002/bit.21373. [cit. 2023-12-25].

[60] HILDEBRAND, Falk; MEYER, Axel; EYRE-WALKER, Adam a NACHMAN, Michael W. Evidence of Selection upon Genomic GC-Content in Bacteria. Online. *PLoS Genetics*. 2010, roč. 6, č. 9. ISSN 1553-7404. Dostupné z: https://doi.org/10.1371/journal.pgen.1001107. [cit. 2024-05-16].

[61] Metacentrum, Výpočetní a datová infrastruktura. Dostupné z: https://www.metacentrum.cz. [cit. 2024-05-16].

Seznam symbolů a zkratek

ART	Adaptive Read-pair Technology for Sequencing Simulation
BED	Browser Extensible Data
bp	páry bází - base pairs
cAGe	Cap Analysis of Gene Expression
cDNA	komplementární deoxyribonukleová kyselina - complementary DNA
CT	Cycle Treshold - prahová hodnota cyklu
DNA	deoxyribonukleová kyselina
emPCR	emulzní polymerázová řetězová reakce
FASTQC	Fast Quality Control
Gbp	gigabáze
GTF	Gene Transfer Format
GFF	General Feature Format
lncRNA	long non-coding RNA
Mbp	megabáze
mRNA	mediátorová ribonukleová kyselina
NCBI	National Center for Biotechnology Information
ncRNA	nekódující ribonukleová kyselina - non coding RNA
NGS	sekvenování nové generace - New Generation Sequencing
PCR	polymerázová řetězová reakce - Polymerase Chain Reaction
pIRS	Profile-based Illumina pair-end Reads Simulator
qPCR	kvantitativní polymerázová řetězová reakce - Quantitative Polymerase Chain Reaction

RNA	ribonukleová kyselina
RSS	RNASeqReadSimulator
RT	reverzní transkripce
RT-qPCR	Real-Time Quantitative Reverse Transcription Polymerase Chain Reaction
SAM	Sequence Alignment/Map
SMRT	Single-Molecule, Real-Time Sequencing
sRNA	malá nekódující ribonukleová kyselina - small RNA
tRNA	transferová RNA
UTR	Untranslated Region