VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2021 Barbora Pomykalová

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

IDENTIFIKACE NEKÓDUJÍCÍ RNA U CLOSTRIDIUM BEIJERINCKII NRRL B-598 POMOCÍ RNA-SEQ DAT

IDENTIFICATION OF NON-CODING RNAS OF CLOSTRIDIUM BEIJERINCKII NRRL B-598 USING RNA-SEQ DATA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR
Barbora Pomykalová

VEDOUCÍ PRÁCE
SUPERVISOR
Ing. Kateřina Jurečková

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Barbora Pomykalová ***ID:*** 211209

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

NÁZEV TÉMATU:

Identifikace nekódující RNA u Clostridium beijerinckii NRRL B-598 pomocí RNA-Seq dat

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma význam nekódující RNA u bakterií. 2) Prostudujte a popište laboratorní techniky pro stanovení genové exprese a zejména technologii RNA-Seq. 3) Seznamte se s dostupnými RNA-Seq daty Clostridium beijerinckii NRRL B-598, navrhněte postup pro identifikaci nekódující RNA a postup otestujte. 4) Implementujte navrženou metodu v libovolném programovacím jazyce a podrobně ji popište. 5) Algoritmus otestujte na dostupných RNA-Seq datech Clostridium beijerinckii NRRL B-598. 6) Proveďte vyhodnocení a diskutujte výsledky.

DOPORUČENÁ LITERATURA:

[1] STORZ, Gisela, Jörg VOGEL a Karen M. WASSARMAN. Regulation by Small RNAs in Bacteria: Expanding Frontiers. Molecular Cell. 2011, 43(6), 880–891. ISSN 10972765. DOI:10.1016/j.molcel.2011.08.022

[2] CHO, Seung Hee, Katie HANING, Wei SHEN, Cameron BLOME, Runxia LI, Shihui YANG a Lydia M. CONTRERAS. Identification and characterization of 5' untranslated regions (5'utrs) in zymomonas mobilis as regulatory biological parts. Frontiers in Microbiology. 2017, 8(DEC). ISSN 1664302X. DOI:10.3389/fmicb.2017.02432

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

Vedoucí práce: Ing. Kateřina Jurečková

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

Abstrakt

Tato bakalářská práce obsahuje stručný úvod do problematiky nekódujících malých RNA u bakterií (sRNA). Zaměřuje se na jejich vlastnosti a funkce v organismu a to konktrétně pro bakterii *Clostridium beijerinckii* NRRL B-598. Dále obsahuje popis laboratorních metod pro stanovení genové exprese a navrhuje postup pro identifikaci malých nekódujících RNA z dat získaných metodou RNA-Seq, pro zkoumanou bakterii *Clostridium beijerinckii* NRRL B-598. V neposlední řadě dochází k implementaci navrženého postupu v prostředí MATLAB a zhodnocení výsledků získaných touto metodou.

Klíčová Slova

RNA, sRNA, RNA-Seq, sekvenování, genová exprese, bakterie, detekce, MATLAB

Abstract

This bachelor thesis contains short introduction into bacterial small non-coding RNA problematic. It is oriented on their features and functions in organisms, especially in bacteria *Clostridium beijerinckii* NRRL B-598. Bachelor thesis also contains description of various laboratory methods for gene expression determination and suggests a detection method for small non-coding RNA in bacteria *Clostridium beijerinckii* NRRL B-598. Suggested method works with data, which were obtained by RNA-Seq technology. Within the framework of the bachelor thesis was suggested method implemented in programming and numeric computing platform MATLAB and its results were discussed.

Keywords

RNA, sRNA, RNA-Seq, sequencing, gene expression, bacteria, detection, MATLAB

Bibliografická citace

POMYKALOVÁ, Barbora. *Identifikace nekódující RNA u Clostridium beijerinckii NRRL B-598 pomocí RNA-Seq dat.* Brno, 2021. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/134378. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce Kateřina Jurečková.

Prohlášení autora o původnosti díla

Jméno a příjmení studenta:	Barbora Pomykalová
VUT ID studenta:	211209
Typ práce:	Bakalářské práce
Akademický rok:	2020/21
Téma závěrečné práce:	Identifikace nekódující RNA u *Clostridium beijerinckii* NRRL B-598 pomocí RNA-Seq dat

Prohlašuji, že svou bakalářskou práci na téma Identifikace nekódující RNA u *Clostridium beijerinckii* NRRL B-598 pomocí RNA-Seq dat jsem vypracovala samostatně pod vedením vedoucí bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu citace použitých zdrojů na konci práce.

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

V Brně dne: 28. května 2021

.....

Podpis autora

Poděkování

Děkuji vedoucí bakalářské práce Ing. Kateřině Jurečkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce. Také chci poděkovat organizaci MetaCentrum, které mi poskytlo bezplatné využití výpočetní kapacity pro mou bakalářskou práci.

V Brně dne: 28. května 2021 ……

Podpis autora

Obsah

Seznam obrázků

Seznam tabulek

Úvod

V posledních letech se, díky velikému posunu v oblasti detekce, analýzy a zpracovávání dat na genetické úrovni, otevírá mnoho dveří pro zkoumání metabolických pochodů a interakcí ve všech živých organizmech.

Sekvenování za využití platforem Illumina, PacBio, Oxford Nanopore a dalších, umožňuje zisk nových, dosud nedetekovaných úseků genetické informace, a tím docílit většího přehledu o celkové struktuře genetické informace daných organizmů. Díky těmto sekvenačním postupům byla objevena řada nových informací o nekódujících úsecích RNA, které hrají komplexní a spletitou roli v regulaci genové exprese daného organizmu.

Pro každý organizmus je genetická informace unikátní a je tedy složité při vysokém počtu nově získaných dat přesně detekovat neznámé úseky genetické informace a určovat jejich funkce v organizmu. Pro tuto detekci vzniká v posledních letech mnoho nových postupů a výpočetních metod, které mají za úkol detekovat právě dosud nedetekované úseky a nalézt v těchto nově obdržených datech ty úseky, které známé již jsou.

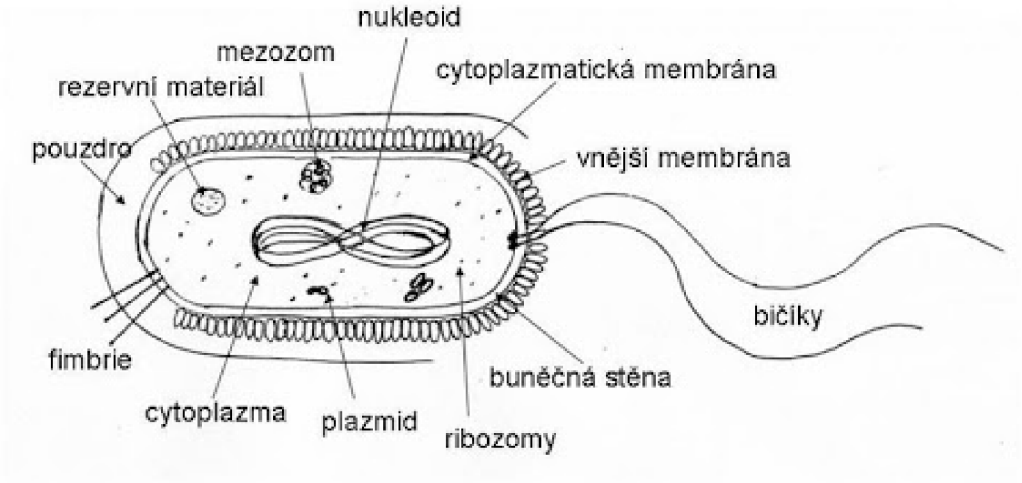
Tato bakalářská práce nastiňuje základní znalosti o bakteriích nekódujících RNA, převážně o bakteriálních malých nekódujících RNA (sRNA) a o základních dostupných metodách sekvenování. Také se tato práce zabývá návrhem algoritmu, který z dat získaných metodou RNA-Seq detekuje malé nekódující bakteriální RNA (sRNA) u *Clostridium beijerinckii* NRRL B-598 a jeho následnou implementací. Navržený implementovaný algoritmus byl vyzkoušen na dostupných datech o *Clostridium beijerinckii* NRRL B-598 a na závěr byly vyhodnoceny výsledky detekce.

1 Bakterie

Počet formálně detekovaných druhů bakterií byl v roce 2011 stanoven na třicet tisíc [1]. Odhaduje se, že na Zemi existuje přes 1 milion druhů bakterií. Bakterie se řadí mezi jednobuněčné prokaryotické organismy. Jejich buňky neobsahují pravé jádro s jadernou membránou. Bakterie patří mezi mikroorganismy, jejich velikost se pohybuje nejčastěji v maximálně desítkách mikrometrů. [2]

1.1 Stavba bakteriální buňky

Tvar buňky bakterie je buď kulovitý (kloky), nebo protáhlý (tyčinky). Základní bakteriální buňka je složena z buněčné stěny, cytoplazmatické membrány, cytoplazmy, nukleoidu, plazmidů, ribozomů, vakuol, či granul a inkluzních tělísek [2]. Stavba bakteriální buňky je vyobrazena na Obr. 1-1.



Obr. 1-1: Stavba bakteriální buňky. [3]

Buněčná stěna je pevný, tuhý útvar, který má za úkol chránit obsah buňky bakterie a držet její tvar. Hlavní složkou buněčné stěny je peptidoglykan. Na základě složení a stavby buněčné stěny jsme schopni určit, jestli se jedná o grampozitivní (tlusté buněčné stěny, barví se ireversibilně), či gramnegativní bakterii (tenčí, komplexnější buněčná stěna, odbarvitelné). U některých buněk se na povrchu buněčné stěny nacházejí polymery, které tvoří kolem buňky pouzdra či slizovou vrstvu (tvorba bakteriálních kolonií). Z povrchu buňky bakterie mohou vyúsťovat další struktury – bičíky a fimbrie. Bičíky jsou dlouhé, velmi tenké útvary, díky kterým je bakterie schopna pohybu. Fimbrie, nebo taktéž pili, jsou tenké výběžky, které bakterie využívá pro jednodušší přilnutí k podkladu. [2]

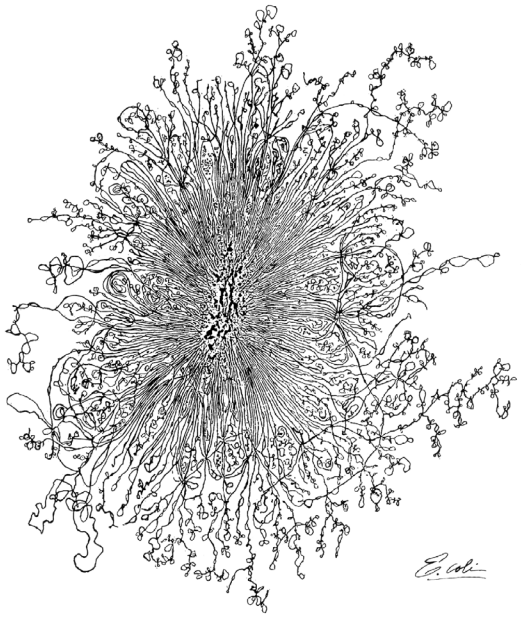
Cytoplazma vyplňuje prostor bakteriální buňky. Je ohraničena cytoplazmatickou membránou, která má velmi podobné složení jako u eukaryotických buněk. Cytoplazma obsahuje několik nerozpustných složek. Mezi tyto složky patří nukleoid (hlavní chromozom) neboli nepravé jádro, které je složeno z dvoušroubovicové kruhové molekuly DNA. Tato DNA nese genetickou informaci dané bakterie. Dalším útvarem je mezozom (u grampozitivních bakterií), který je využit při replikaci deoxyribonukleové kyselině (DNA) v buňce. [2]

Bakteriální ribozomy jsou oproti eukaryotickým ribozomům menší, jsou složeny z ribonukleové kyseliny (RNA) a tvoří zhruba 40 % cytoplazmy. Plazmidy jsou malé kruhovité části DNA umístěné volně v cytoplazmě, a jejich replikace je nezávislá na nukleoidu. Plazmidy nemají existenciální vliv na přežití bakterie, ale mohou obsahovat geny pro rezistenci vůči některým antibiotikům. [2]

Pro uchovávání energie a živin slouží u bakterií různé inkluze (např. inkluze glykogenu či lipidů). Fototropní bakterie zase obsahují chromatofory, což jsou útvary obsahující pigmenty, které jsou schopné absorbovat sluneční záření. Některé bakterie také obsahují vakuoly (plynové), což jsou drobné váčky propustné pro plyny a vodu. [2]

1.2 DNA u bakterií

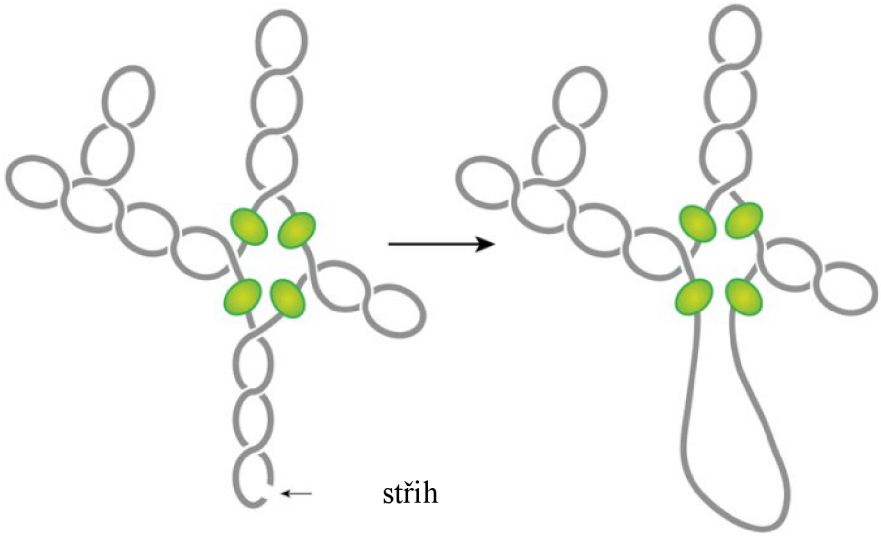
Deoxyribonukleová kyselina (DNA) je makromolekula skládající se z podjednotek – nukleotidů. Nukleotid je složen z jedné purinové (adenin, guanin), nebo pyrimidinové (cytosin, tymin) báze, fosfátové skupiny (zbytku kyseliny fosforečné) a pentózy (deoxyribóza) [4]. Jedná se obvykle o dvouvláknovou molekulu. Názorná ukázka chromozomu u bakterie *E. coli* je vyobrazena na Obr. 1-2.



Obr. 1-2: Snímek chromozomu *E. coli* z elektronové mikroskopie. [5]

Každá molekula DNA tvořící hlavní chromozom u bakterie se musí nacházet ve velmi kondenzovaném stavu tzv. svinutém genomu. V takto kondenzovaném stavu musí být kvůli velikosti bakterie, která je průměrně v jednotkách mikrometrů. Naopak délka chromozomální kruhové molekuly DNA dosahuje velikosti přes tisíc mikrometrů.

Molekula DNA je tedy uspořádána do přibližně 50 jednotlivých smyček (domén), které jsou spiralizovány, viz Obr. 1-3. Na udržení tohoto svinutého genomu se podílejí některé proteiny a RNA. Ty mohou být uvolňovány působením deoxyribonukleázy (DNázy), která uvolňuje spiralizaci jednotlivých domén. Také mohou být rozvolněny působením ribonukleázy (RNázy), která uvolňuje jednotlivé RNA z DNA, jež upevňují jednotlivé domény [4].



Obr. 1-3: Znázornění čtyř domén bakteriální DNA. Upraveno. [6]

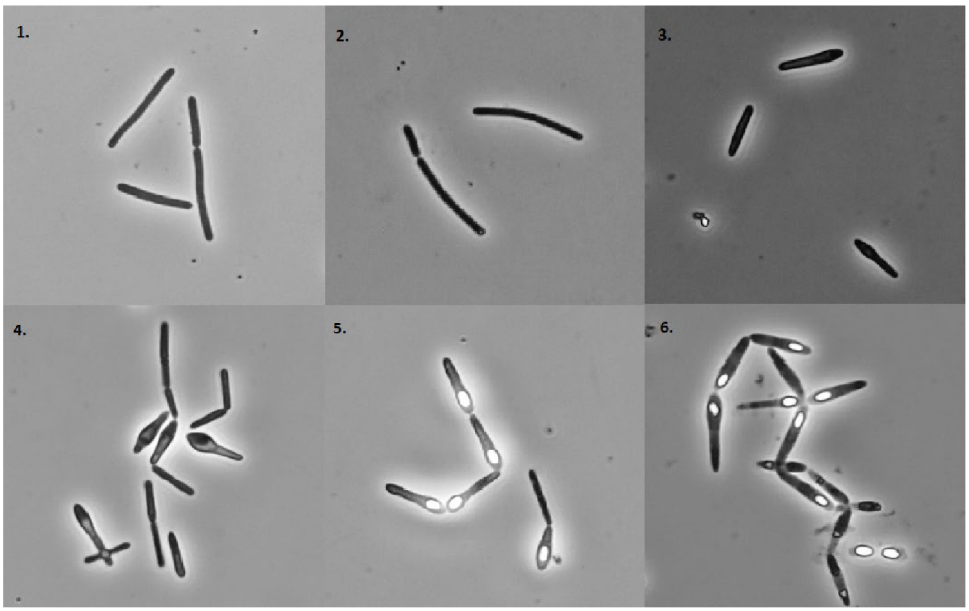
1.3 RNA u bakterií

Ribonukleová kyselina (RNA) je makromolekula skládající se, stejně jako DNA, z podjednotek – nukleotidů. Nukleotid je složen z jedné purinové (adenin, guanin), nebo pyrimidinové (cytosin, uracil) báze, fosfátové skupiny (zbytku kyseliny fosforečné) a pentózy (ribóza). V pyrimidinových bázích se RNA liší od DNA, kdy namísto thyminu obsahuje bázi uracil. Také se liší v použité pentóze. Pro genovou expresi rozdělujeme RNA do pěti základních typů: transferová RNA (tRNA), mediátorová RNA (mRNA), ribozomální RNA (rRNA), malá bakteriální RNA (sRNA). Všechny tyto RNA vznikají transkripcí z DNA. Kromě mRNA jsou jejími konečnými produkty. Pouze mRNA dále podléhá translaci za vzniku proteinů. [4]

2 *Clostridium beijerinckii* NRRL B-598

Rod *Clostridium* je jeden z velkých bakteriálních rodů čítající zhruba 150 druhů, jež některé mají obrovský potenciál pro biotechnologické zkoumání [7]. Vyskytují se v půdách, vodních sedimentech a v trávicím traktu řady organizmů. Jsou účastníci v hnilobných procesech a díky jejich schopnosti tvořit silně rezistentní spory, jsou vysoce odolné vůči nepříznivým podmínkám. Řada bakterií z rodu *Clostridium* má významné patologické účinky. Jedná se o relativně velké sporulující bakterie, často širší než 0,5 μm, které zaujímají tvar tyčinek s bičíky [8]. Jsou grampozitivní, ale svou grampozitivitu mohou postupně ztrácet. Většina druhů roste za silně anaerobních podmínek. Výjimku tvoří řada bakterií (*C. perfringens, C. histolyticum*), které jsou schopny přežít malé množství kyslíku v prostředí [9]. Díky jejich silně anaerobním vlastnostem bylo až donedávna těžké s nimi manipulovat a využívat je tak pro zkoumání na genetické úrovni.

Clostridium beijerinckii NRRL B-598, původně pojmenovaná jako *Clostridium pasteurianum* NRRL B-598 [10], je sporulující, kyslík tolerující bakterie schopná produkovat aceton-butanol a vodík. Schopnost růst v levném jednoduchém prostředí, stabilita degenerace kmene, dobrá adaptace na kontinuální procesy a hlavně schopnost produkce aceton-butanolu určují bakterii *C. beijerinckii,* jako skvělého adepta pro širší biotechnologické zkoumání [7]. Ukázka *C. beijerinckii* je zobrazena na Obr. 2-1.



Obr. 2-1: Sporulující *Clostridium beijerinckii* NRRL B-598:
1. T1 (3,5 h) – acidogenní fáze; 2. T2 (6 h) – přechod z acidogenní fáze do solventogenní; 3. T3 (8,5 h) – solventogenní fáze; 4. T4 (13 h) – solventogeneze a sporulace; 5. T5 (18 h) – solventogeneze a sporulace; 6. T6 (23 h) – solventogeneze a sporulace.

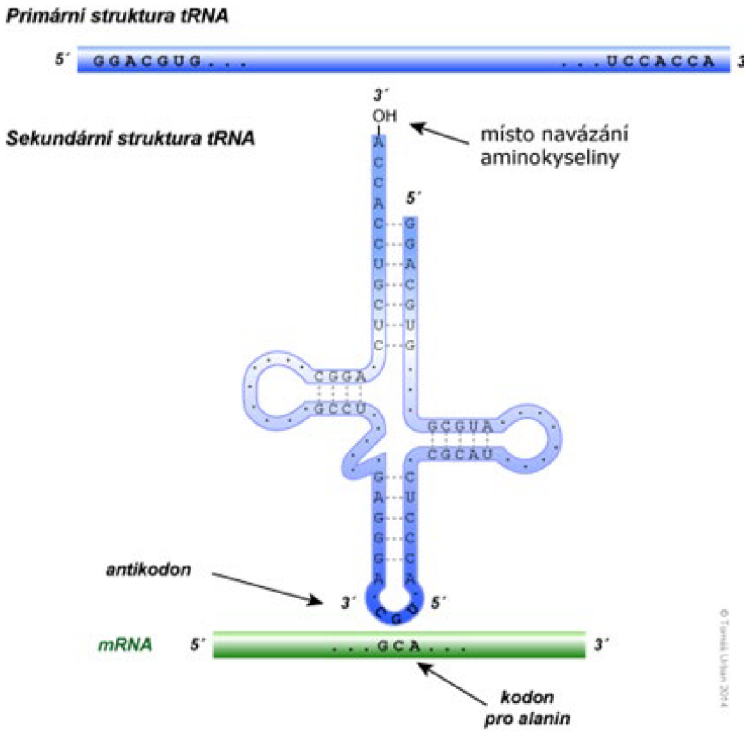
3 Nekódující RNA

Nekódující RNA (ncRNA) jsou funkční RNA molekuly vznikající transkripcí z templátové DNA, které nepodléhají translaci na protein. Mezi hlavní nekódující RNA řadíme transferovou RNA, ribozomální RNA, malou jadernou RNA, mikro RNA a další. U bakterií jsou tato krátká nekódující RNA označovány jako malé RNA [4].

3.1 Transferová RNA (tRNA)

Transferová RNA neboli tRNA slouží jako zprostředkovatel, díky němuž se při translaci na ribozomech jednotlivé aminokyseliny správně vážou do polipeptidového řetězce podle kodonů mRNA. tRNA jsou složena z 70-95 nukleotidů a vytvářejí tzv. strukturu jetelového listu, viz Obr. 3-1. Právě díky této struktuře má tRNA jedinečnou vlastnost vázat na sebe patřičnou aminokyselinu za přítomnosti enzymu aminoacyl-tRNA syntetáza. Také díky tomuto tvaru má tRNA oblast s antikodonem pro rozeznání kodonu mRNA na základě komplementarity bází a mít místo k uchycení na ribozom. [4]

Každý ribozom má tři místa pro uchycení jednotlivé tRNA. Jedná se o místa A, P a E. Místo A slouží k navázání tRNA nesoucí aminokyselinu, místo P slouží k navázání této aminokyseliny do polypeptidového řetězce a místo E, kde se tRNA, která již nenese aminokyselinu, odpojuje od ribozomu. [4]

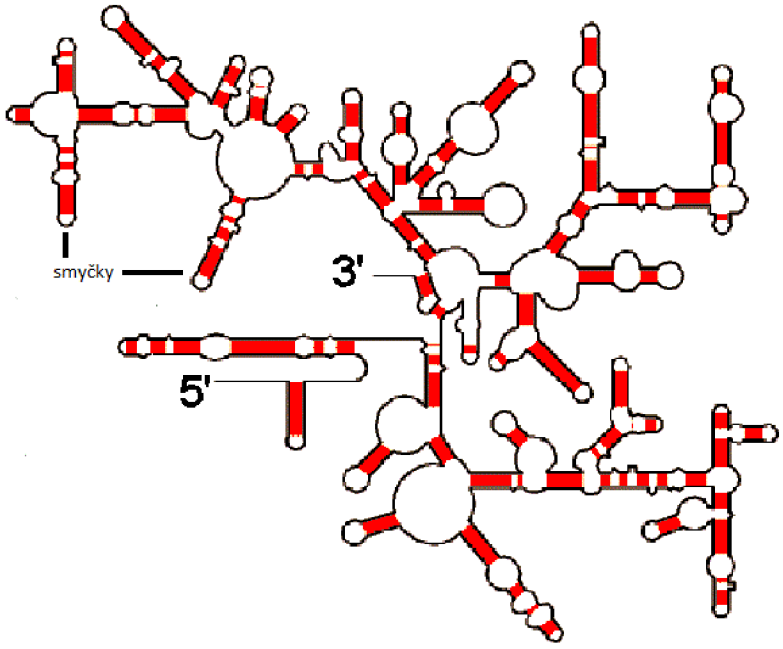


Obr. 3-1: Schéma primární a sekundární struktury tRNA. [11]

3.2 Ribozomální RNA (rRNA)

Ribozomální RNA neboli rRNA slouží jako funkční stavební části ribozomů, na kterých dochází k translaci a předpokládá se, že hraje důležitou roli ve tvorbě peptidových vazeb jako peptidyltransferáza [12]. rRNA je jednovláknová a jen v některých částech tvoří strukturu dvojité šroubovice [4]. Sekundární struktura jedné z rRNA je zobrazena na Obr. 3-2.

Ribozomy jsou složeny zhruba z jedné poloviny právě z rRNA a druhou polovinu tvoří proteiny. Ribozomy se skládají ze dvou podjednotek – z malé a velké podjednotky. Tyto podjednotky se sestavují při iniciaci translace, po ukončení translace se znovu rozpadají a čekají na další iniciaci [4].



Obr. 3-2: Sekundární struktura 5.8S rRNA. Upraveno. [13]

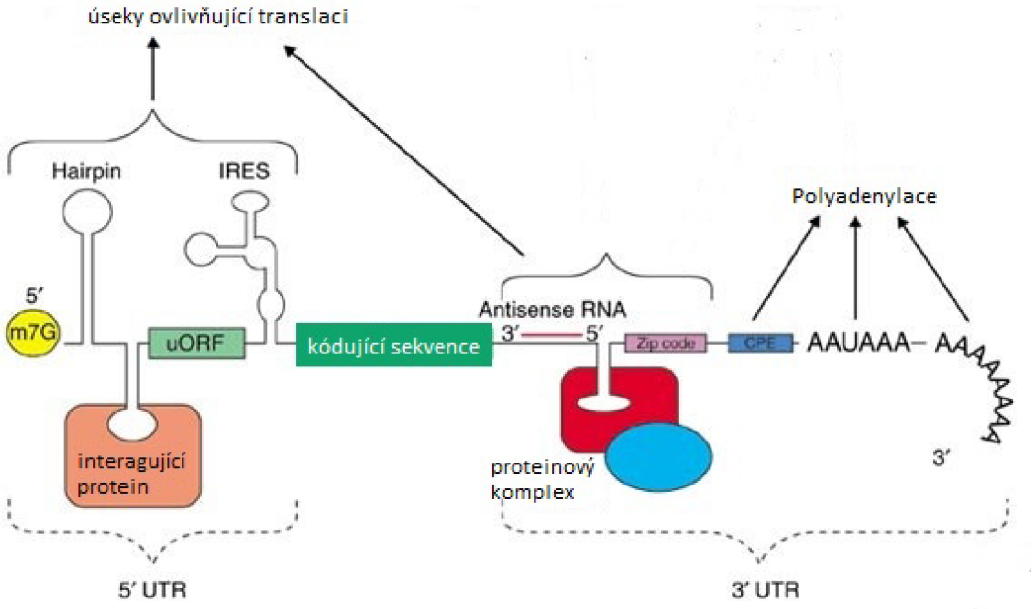
3.3 5' a 3' nepřekládaná oblast, riboswitch

5' nepřekládaná oblast neboli 5' untranslated region (UTR) se nachází v oblasti od prvního nukleotidu mRNA až po iniciační kodon (AUG) této mRNA podléhající translaci. 5' UTR tedy nepodléhají translaci, ale mají naopak regulační funkci, kterou translaci ovlivňují. U prokaryotních organizmů se v oblasti 5' UTR nachází vazebné místo pro ribozom tzv. Shine-Dalgarnova sekvence (AGGAGGU) a její délka bývá zhruba od 3-10 nukleotidů. U eukaryotických organizmů délka sahá až k tisícům bází [14]. 5' UTR obsahují riboswitche, které regulují jak transkripci, tak translaci a celkovou stabilitu mRNA [15].

Riboswitch (RNA přepínač) je struktura obsahující aptamer tj. RNA, která je schopná na sebe navázat specifickou malou molekulu, a tzv. expression platform, jež

mění svou konformaci na základě vazby dané molekuly na aptamer. Na základě změny této struktury závisí jednotlivé regulační vlastnosti riboswitche na úrovni transkripce i translace a také má vliv na celkovou stabilitu mRNA. [15]

3' nepřekládaná oblast neboli 3' UTR se nacházejí na mRNA ihned po terminačním kodonu (UGA, UAA, UAG). 3' UTR tedy nepodléhají translaci, ale obsahují oblasti regulující translaci. Mají celkový vliv na stabilitu dané mRNA a také mají vliv na polyadenylaci, tj. procesu, kdy se na mRNA navazuje poly-A konec. Poly-A konec se na konec mRNA navazuje na základě sekvence AAUAAA, kterou 3' UTR obsahuje. 3' UTR oblast také obsahuje oblasti, které jsou schopné navázat některé bakteriální sRNA, které následně ovlivňují genovou expresi [16]. Schéma eukaryotické mRNA s výše zmíněnými strukturami je znázorněna na Obr. 3-3.



Obr. 3-3: Schéma eukaryotické mRNA s 3' UTR, 5' UTR a poly-A koncem. Upraveno. [17]

3.4 Malá jaderná RNA

Malá jaderná RNA neboli small nuclear RNA (snRNA) jsou krátké nekódující RNA, jejich funkcí je převážně ovlivňovat posttranskripční děje u eukaryotických organizmů. S proteiny tvoří snRNA komplexní strukturu zvanou spliceozom, který provádí splicing – umožňuje vystřižení nekódujících částí – intronů, z primárního transkriptu (pre-mRNA) [4].

Mikro RNA neboli miRNA je malá nekódující jednovláknová RNA (délka 21-23 nukleotidů), vyskytující se převážně u eukaryotických organizmů. Svou funkcí je schopna ovlivňovat genovou expresi tak, že se naváže na cílovou mRNA a zastaví translaci, nebo zapříčiní rozpad této mRNA [18].

3.5 Malé bakteriální RNA

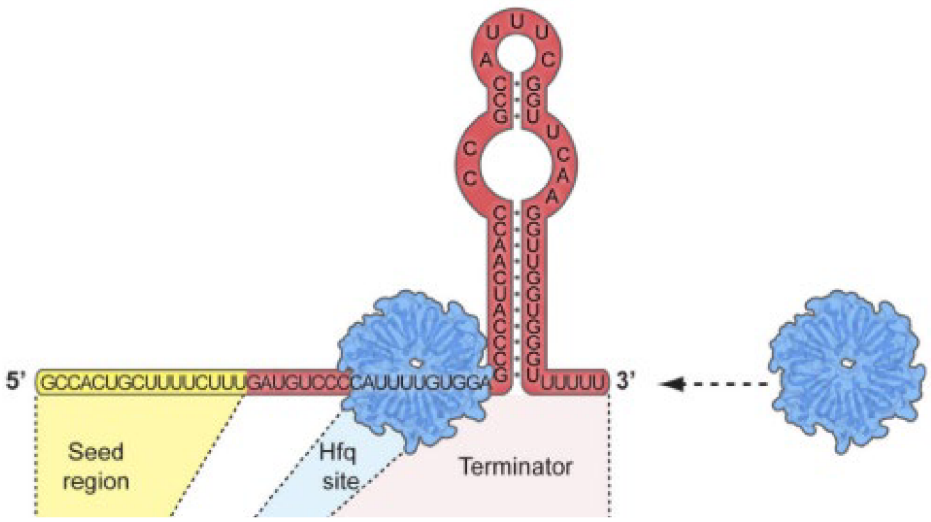
Malé bakteriální RNA neboli small RNA (sRNA) jsou RNA složená z přibližně 50 až 300 nukleotidů. Jejich struktura je komplexní a obsahuje několik kmenových smyček [19, 20]. sRNA mají různé posttraskripční regulační účinky působící na nezávislých cílech genové exprese. Nejenom díky těmto regulačním účinkům sRNA jsou bakterie schopné rychle reagovat na změnu podmínek okolního prostředí, ovlivnit růstové procesy nebo řídit a ovlivňovat další posttraskripční faktory [20].

První skupinka sRNA, které regulují mRNA tím, že se na ni navážou krátkým nedokonalým párování bází, jsou nazývány také jako transkódované sRNA. Mnoho z těchto sRNA se párují buď v těsné blízkosti, nebo přímo na ribozomální vazebné místo (RBS) – oblast, kterou se mRNA váže na ribozom, a tím zabraňují dané mRNA translaci na ribozomech [20].

Další z těchto sRNA naopak nasedají na příslušnou mRNA dál od RBS a znemožňují tak vytvoření sekundární struktury, která by měla za následek zabránění translace, nebo svou vazbou zvyšují či snižují stabilitu dané mRNA [20].

Jak tyto sRNA rozpoznají svá cílová místa na mRNA, i přes možnou komplementaritu se stovkami dalších míst podél celého genomu, není doposud přesně známo. Některé výzkumy naznačují, že určité sRNA využívají své vyšší struktury (sekundární, terciální) pro detekci cílového místa. Většina sRNA se však váže na jednořetězcové úrovni a díky tomu se na tyto struktury nemůžeme vždy spoléhat. [20]

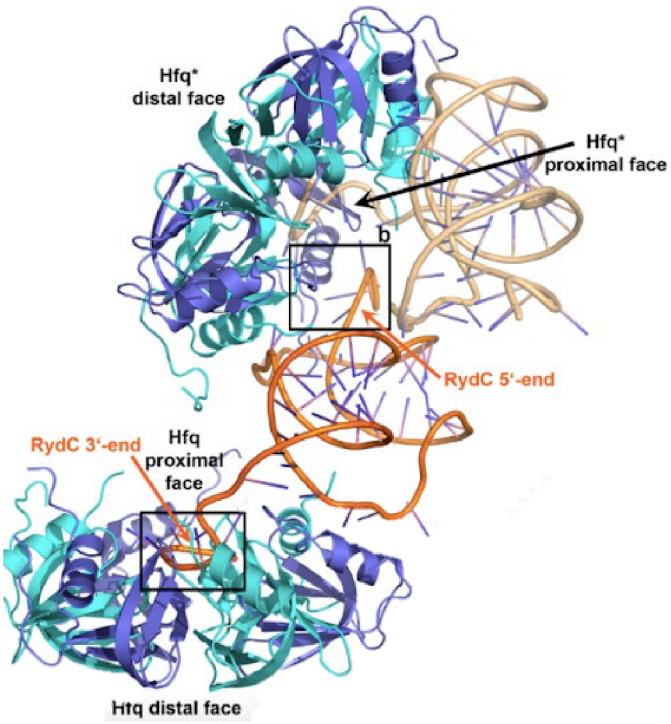
sRNA se vyznačují svými poly-U 3' koncem, který slouží jako ochrana vůči exonukleázám, enzymům schopných štěpit jednotlivé nukleové kyseliny od konců molekul a tím je znehodnocovat. K vazbě na odpovídající mRNA dochází nejčastěji na 5' konci sRNA, v oblasti zvané „seed region“, znázorněné na Obr. 3-4. [20]



Obr. 3-4: Znázornění struktury sRNA s proteinem Hfq. [20]

U gramnegativních druhů bakterií sRNA vyžadují pro svou správnou funkci a stabilitu přítomnost proteinu Hfq. Protein Hfq má tendence k vazbě v oblastech blízkých kmenové smyčce sRNA bohatých na A a U a jsou schopny detekovat 3' konec polyU dané sRNA. Jednotlivá RNA se na protein Hfq vážou jak na proximální, tak i na distální část. Proto se na jediný Hfq protein může vázat zároveň více jednotlivých RNA. Díky této vlastnosti Hfq může urychlovat vazbu sRNA na mRNA, stabilizovat vazbu sRNA-mRNA či měnit strukturu některé z těchto RNA. [20]

U grampozitivních bakterií se zdálo, že je protein Hfq postradatelný, ale například studie o bakterii *Listeria monocytogenes* dokazuje, že i u této skupiny bakterií může Hfq zaujmout svou roli v regulaci [21]. Také u naší sledované bakterie *Clostridium beijerinckii* NRRL B-598 ukazují sekvenační studie, že obsahuje homologickou sekvenci pro protein Hfq, a tedy lze předpokládat účinky tohoto proteinu na vazbu mezi sRNA a mRNA [22]. Existují i další proteiny (YbeY protein), které mají vliv na funkci a strukturu sRNA a u bakterií nedisponujících genem pro Hfq protein [20]. Vazba Hfq proteinu s sRNA je znázorněna na Obr. 3-5.



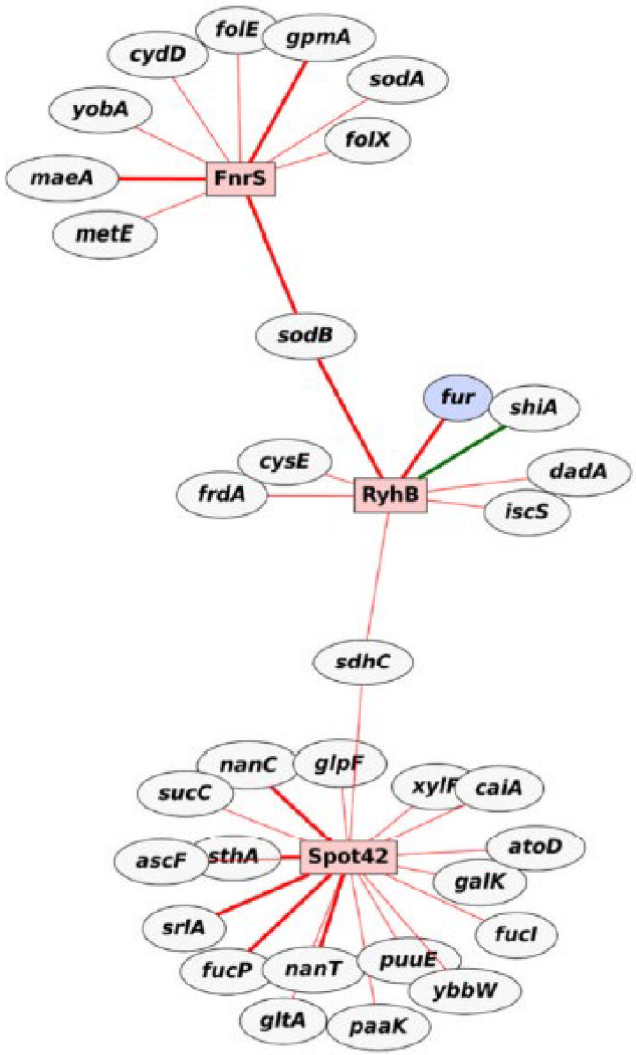
Obr. 3-5: Vazba Hfq proteinu s sRNA RydC vyskytující se u *E. coli.* [23]

Bakteriální malé RNA hrají také důležitou roli ve stabilitě mRNA, kdy zabraňují jednotlivým RNázám navazovat se na 5' konce tím, že se na tyto oblasti navážou sami. Tím nedochází ke štěpení mRNA a zvyšuje se tak její stabilita. Tato vlastnost byla pozorována při studii bakterií *Streptococcus*, kde tuto funkci zastupuje sRNA FasX. Některé sRNA také ovlivňují svou vazbou na specifické místo mRNA účinky RNáz, které na základě této vazby sRNA-mRNA štěpí mRNA na určitém místě. Naopak

některé RNázy mají schopnost štěpit sRNA navázané na mRNA a tím ovlivňovat jejich regulační vlastnosti. [20]

Další vlastností některých sRNA je interakce se specifickými proteiny. Jak už bylo uvedeno výše, protein Hfq je protein napomáhající ke správnému navázání jednotlivých sRNA na cílová místa mRNA. Existují však i jiné interakce mezi sRNA a proteiny, které svou vazbou ovlivňují aktivitu těchto proteinů a u některých má dopad na jejich enzymatickou aktivitu. Jiné sRNA mohou také spojovat jednotlivé proteiny k sobě a tím vytvářet složitější struktury. [20]

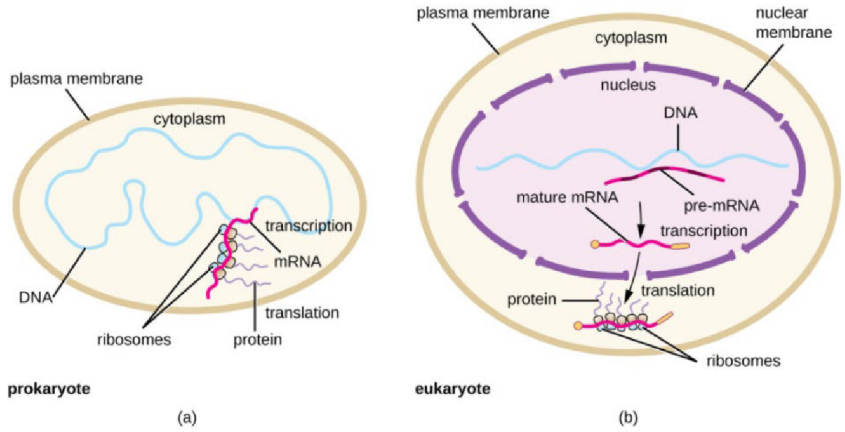
Část sRNA regulují více konečných cílů a některé mRNA či proteiny jsou ovlivňovány více než jednou sRNA. Tím dochází k tvorbě velmi spletitých a komplexních regulačních sítí, viz Obr. 3-6, které ovlivňují celkovou genovou expresi v organizmu. [20]



Obr. 3-6: Část regulační sítě pro *E. coli*. Obdélníky představují jednotlivé sRNA a ovály odpovídají regulovaným cílovým mRNA. Upraveno. [20]

4 Laboratorní techniky pro stanovení genové exprese

Existují dva hlavní důvody proč se zabývat studiem genové exprese. Prvním z nich je určení, které tkáně exprimují jaké geny. Tyto informace mohou pomoci při indikaci jednotlivých fyziologických funkcí daného kódovaného proteinu. Druhým, ne méně důležitým důvodem, je detekce a určení funkcí jednotlivých regulátorů genové exprese. V dnešní době se ke stanovení genové exprese využívá několik metod, které si představíme v následujících kapitolách. Schéma genové exprese je znázorněno na Obr. 4-1.



Obr. 4-1: Schéma genové exprese. [24]
(a) Znázorňuje GE u prokaryot, kdy transkripce i translace probíhá zároveň v cytoplazmě buňky, (b) znázorňuje GE u eukaryot, kde transkripce probíhá v jádře buňky a translace na ribozomech v cytoplazmě.

Genová exprese (GE) je proces, při kterém je genetická informace z DNA přenesena do polypeptidového řetězce proteinu. Genová exprese je prováděna ve dvou základních krocích – transkripcí a translací [4].

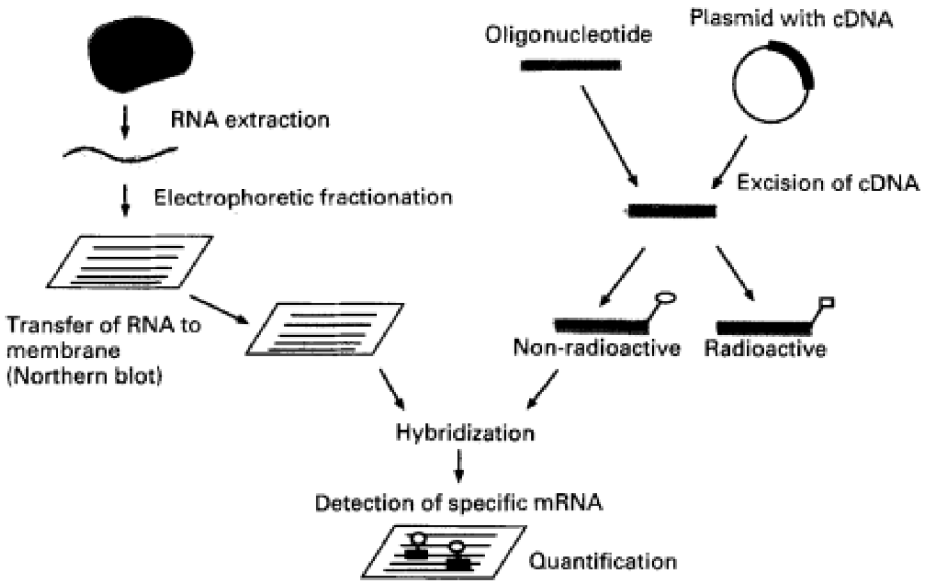
Při transkripci dochází k přepisu genetické informace obsažené v sekvenci nukleotidů DNA do sekvence nukleotidů RNA na základě komplementarity bází. Této RNA se říká mediátorová RNA (mRNA). U eukaryotických organismů probíhá transkripce v buněčném jádře za vzniku pre-mRNA. Posléze prochází tato mRNA ještě posttranskripčními úpravami, kdy jsou vystřiženy za účasti snRNA z pre-mRNA introny (části, které nenesou informace o proteinech) a jsou upraveny jejich 3' a 5' konce. Tím získáváme finální molekulu mRNA, která vystupuje z buněčného jádra do cytoplazmy a podléhá translaci. U prokaryotických organismů probíhá transkripce volně v cytoplazmě a je získán primární transkript, který odpovídá výsledné molekule mRNA, která následně podléhá translaci. [4]

Translace je děj, který probíhá na ribozomech (makromolekulární struktury složené z 3-5 rRNA a 50-90 proteinů) umístěných v cytoplazmě. Při translaci se informace ze sekvence nukleotidů mRNA překládá do polypetidového řetězce výsledného proteinu. Translace se účastní i transferové RNA (tRNA), které mají za úkol přivádět jednotlivé aminokyseliny na ribozom. Připojení správných aminokyselin na jednotlivé tRNA probíhá za pomoci enzymů nazývajících se aminoacyl-tRNA-syntetázy. Každé aminokyselině připadá jednou až čtyř tRNA na základě tzv. genetického kódu. [4]

4.1 Northern blot

Northern blot je metoda určená k měření velikosti a množství transkriptu RNA. Základním principem je dělení RNA dle velikosti a detekce na membráně za pomocí hybridizační sondy se sekvencí báze komplementární k celé, nebo k části cílové RNA. [25]

Prvním krokem je extrakce celkové RNA z buňky za užití chaotropního činidla (látka schopná narušovat vodíkové vazby). Tyto činidla denaturují proteiny, včetně RNáz. Dále můžeme zahrnout do postupu izolaci jednotlivé mRNA. K izolaci se využívá poly-A+ konce mRNA za užití celulózové chromatografie, kde jsou použity oligo-T kolony. Tyto vzorky RNA oddělujeme gelovou elektroforézou na agarovém gelu podle velikosti. Poté následuje proces zvaný blotování, kdy je daná RNA přenesena z agarového gelu na nylonovou membránu. Využívá se především kladně nabitých nylonových membrán, díky jejich vysoké afinitě k záporným nukleovým kyselinám mRNA a vyšší odolnosti. [26] Schéma metody Northern blot je vyobrazeno na Obr. 4-2.



Obr. 4-2: Schéma zobrazující postup metody Northern blot. [26]

Pro blotování se využívají dvě hlavní metody – kapilární a vakuová. Tradičnější a na výbavu méně náročná metoda je kapilární blotování. Její nevýhodou je však délka trvání (4-18 h). Na rozdíl u technicky náročnějšího vakuového blotování je čas výrazně kratší (1-2 h), a proto se v poslední době využívá ve větší míře. Výsledný blot na nylonové membráně obsahuje přesný obraz rozdělených mRNA z agarového gelu. [26]

Po blotování následuje ukotvení mRNA na nylonovou membránu. K tomuto se využívá buď UV světlo, nebo vyšší teploty, které zapříčiní vznik kovalentní vazby mezi mRNA a membránou. [26]

Pro přípravu hybridizační sondy je základem, aby byla sonda částečně, nebo kompletně komplementární ke sledované mRNA. Pro sondy se využívají komplementární DNA (cDNA), RNA nebo oligonukleotidy, které mají se sledovanou mRNA minimálně 25 komplementárních bází. Sonda je nadále označena buď radioaktivními izotopy (^{32}P), nebo chemiluminiscencí. Po označení je sonda hybridizována s mRNA na membráně. Poté dochází k postupnému promývání, kdy jsou odplaveny přebytečné zbytky sondy tak, aby na membráně zbyly jen označené části mRNA. Výsledný obraz je získán za použití rentgenového filmu a výsledky kvantifikovány denzitometrií. [26]

4.2 SAGE

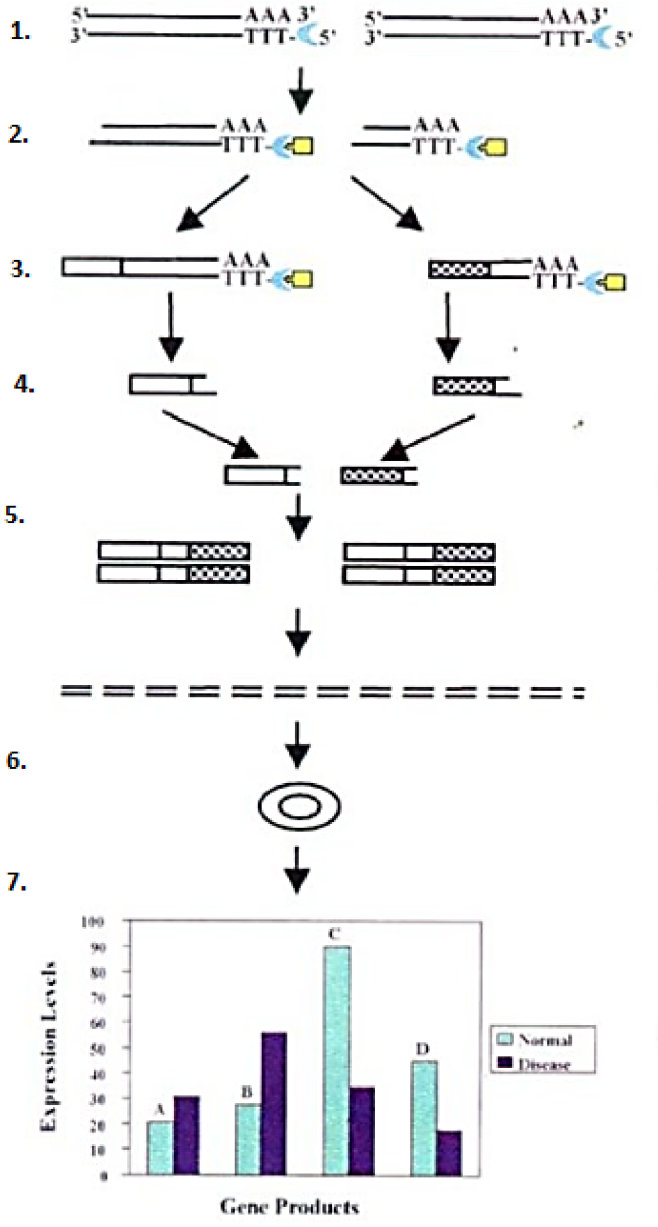
SAGE (serial analysis of gene expression) je metoda, jenž umožňuje určit absolutní četnosti jednotlivých transkriptů exprimovaných v populaci buněk. [27]

Metoda SAGE je založena na dvou předpokladech. První z předpokladů využívá toho, že označené nukleotidové sekvence o délce 9-10 nukleotidů jsou dostatečným identifikátorem jednotlivých transkriptů, neboť sekvence o délce 10 nukleotidů je schopna rozpoznat až 1048576 jednotlivých transkriptů (4^{10}). Druhým předpokladem je, že krátké označené sekvence mohou být zřetězeny a mohou tak být sériově sekvenovány v jednom klonu. U této sekvenace je však důležité rozeznat začátky a konce jednotlivých označení daných sekvencí. [27]

Prvním krokem u metody SAGE pro analýzu exprese mRNA je syntéza cDNA z mRNA za použití biotin-oligo(dT) primeru. Tato nově vzniklá dvouřetězcová cDNA je nadále štěpena za použití restrikční endonukleázy. Restrikční endonukleáza je kotvící enzym, který je schopný štěpit DNA na specifických místech. [27]

Dále jsou části, které byly štěpeny restrikční endonukleázou nejblíže k poly-A konci, navázány na streptadivin. Tyto části cDNA navázaných na streptadivin jsou rozděleny na dvě poloviny a každá z těchto polovin je označena enzymem (linkerem) navázaným na místo cDNA, kde proběhla restrikce. Následně dochází k dalšímu štěpení, kdy je odstřižen od označených cDNA streptadivin. Tyto úseky uvolněné od streptadivinu s jednotlivými linkery se spojí v jeden dlouhý řetězec. [27]

Tento řetězec nadále slouží jako templát pro PCR (polymerázová řetězová reakce) se specifickým označením (primery) na obou stranách. Výsledkem amplifikace (namnožení) jsou dvě sekvence spojené svými konci, ohraničené místy pro kotvící enzym a neobsahuje linkery. Tímto je vytvořen tzv. konkatemer – dlouhá molekula DNA obsahující mnohonásobné kopie stejné části DNA, která je následně klonována a sekvenována. [27] Celý tento postup je znázorněn na Obr. 4-3.



Obr. 4-3: Schéma znázorňující metodu SAGE. Upraveno. [28]
1. Syntéza cDNA pomocí biotin-oligo(dT) primeru; 2. Štěpení kotvícím enzymem a navázání streptadivinu; 3. Rozdělení na dvě části a označení jednotlivých částí linkery; 4. Odstřižení streptadivinu; 5. Spojení řetězců částí s odlišnými linkery v jeden řetězec; 6. PCR, sekvenování; 7. Analýza dat.

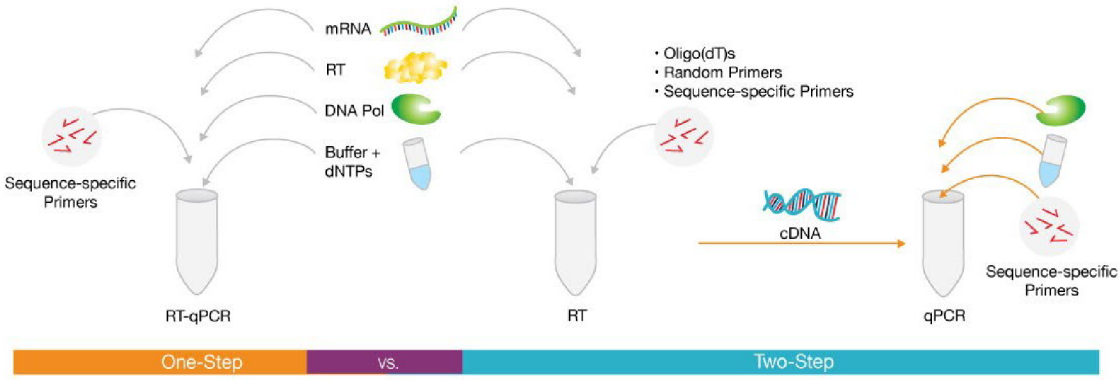
4.3 Reversně transkripční kvantitativní PCR

Reverzně transkripční kvantitativní PCR (RT-qPCR) je v dnešní době jedna z nejrozšířenějších a nejdostupnějších metod pro stanovení kvantity mRNA. Jedná se o fluorescenční metodu, která pracuje v reálném čase. [29]

Prvním krokem je u RT-qPCR právě reverzní transkripce, kdy je buď celková RNA, nebo cílová mRNA transkriptována do cDNA. Pro tento přepis se nejčastěji využívá směs oligo(dT), náhodných či specifických primerů. Tyto primery nasedají na vlákno RNA. Enzymy reverzní transkriptázy (RT) jsou potřebné k zahájení syntézy. Často používanými enzymy jsou enzymy s relativně vysokou tepelnou stabilitou. Využívá se reverzní transkriptáza získaná z viru ptačí myeloblastózy nebo Molonayova myšího leukemického viru [30]. Následně je využito funkce RNázy H, která odstřihne RNA z duplexu RNA-DNA a dovolí tak dokončit syntézu dvouvláknové cDNA. [29]

Poté probíhá zahřátí směsi na vysokou teplotu (okolo 70 °C) tak, aby došlo k inaktivaci RT. Takto získaná cDNA je následně použita jako templát pro kvantitativní PCR, kde je požadovaná RNA amplifikována a v jednotlivých cyklech detekována. V dnešní době se pro PCR používají fluorescenčně značené nukleové kyseliny, které umožňují real-time detekci. [29]

Existuje jednostupňová a dvoustupňová RT-qPCR. Jednostupňová RT-qPCR kombinuje oba hlavní kroky (RT i PCR) v jediné zkumavce, ve které jsou přítomny všechny potřebné látky. Naopak dvoustupňová RT-qPCR probíhá jednotlivě ve dvou zkumavkách se specifickými primery a postupy, jak je znázorněno na Obr. 4-4. [29]



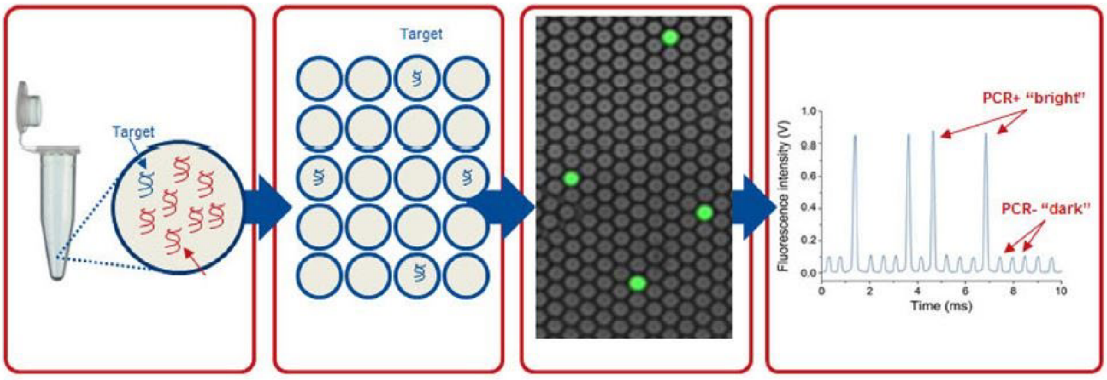
Obr. 4-4: Schéma jednostupňové a dvoustupňové RT-qPCR. [29]

4.4 Digitální PCR (dPCR)

Metoda digitální PCR (dPCR) je jedna z novějších metod, která umožňuje analýzu z malého množství materiálu s vysokou kvantifikací získaného výsledku bez nutnosti využití kalibračních křivek. Při dPCR je reakční směs obsahující analyzovaný vzorek rozdělena do velkého počtu reakčních podílů. V každém z těchto podílů následně probíhá vlastní amplifikace pomocí PCR. dPCR využívá stejné složky, jako jsou

využity u RT-qPCR, včetně fluorescenčního značení. Při vyhodnocování je získána jako odpověď sérii jedniček a nul. Jednička nám značí pozitivní nález v daném reakčním podílu, nula značí, že v daném reakčním podílu nebyl detekován dostatek analyzované DNA (detekce nepřekročila stanovenou detekční hranici). Díky tomuto binárnímu zápisu výsledku získala tato metoda svůj název „digitální PCR“. [31]

Pro rozdělení reakční směsi do jednotlivých reakčních podílů existují dvě metody. První metoda využívá mikročipy – čipová dPCR (cdPCR). Čipy rozdělují směs do přesně daného množství mikrojamek (komůrek). V dnešní době se jedná o poloautomatické či automatické přístroje. Druhá metoda – kapičková dPCR (ddPCR), využívá emulgace vzorku. Reakční směs je smíchána s olejem a stabilizátory. Následně je reakční směs rozdělena na mikrokapičky za pomocí tzv. generátoru kapek, tyto mikrokapičky jsou přeneseny na mikrotitrační destičky a probíhá na nich emulzní PCR. Výsledky jsou snímány čtečkou kapek, která je schopná analyzovat zhruba 1000 kapek za jednu sekundu. [32] Schéma metody dPCR je zobrazena na Obr. 4-5.



Obr. 4-5: Schéma metody dPCR. [32]

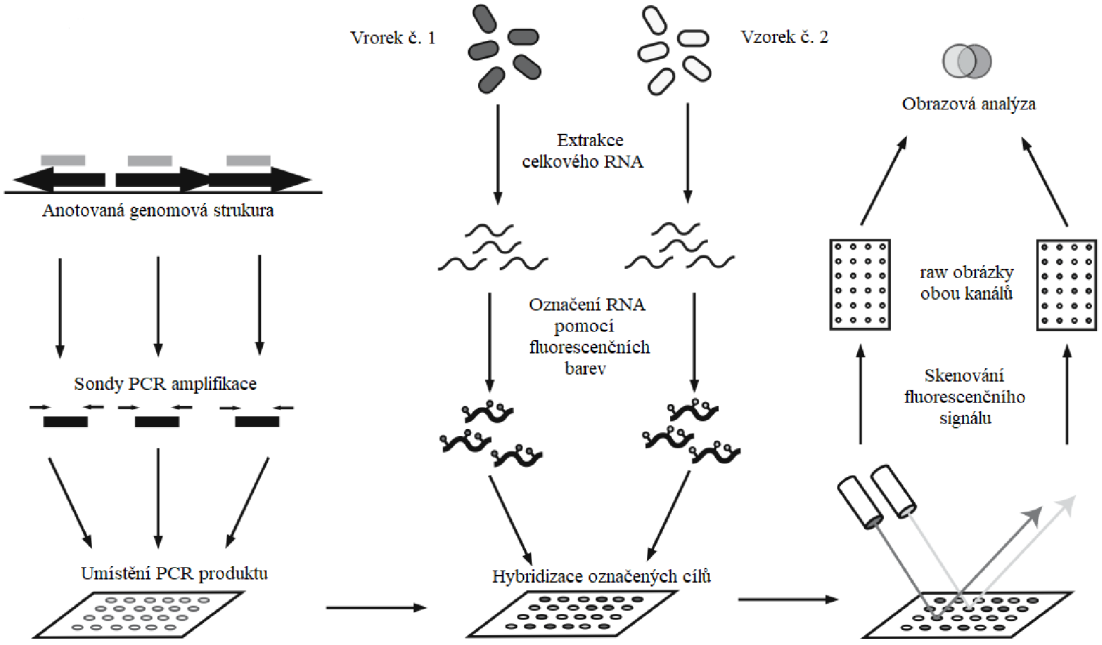
Počet kapiček, respektive počet komůrek a jejich objemy jsou zásadní pro určení detekčních limitů dPCR. Důležitá je také náhodné rozdělení reakční směsi do jednotlivých reakčních podílů. Pro výsledný výpočet koncentrace analyzované DNA je nutné znát objemy reakčních podílů a jejich počet a z nich poté znát počet pozitivních, nebo negativních podílů. [32]

4.5 DNA microarray

Metoda DNA microarray se stala jednou z nejpoužívanějších metod pro určení relativní koncentrace nukleových kyselin ve směsi a to díky skvělému poměru: počet sekvencí/cena. V dnešní době existuje mnoho různých metod microarray (DNA, proteinová, …). Zde se však zaměříme na DNA microarray. [33]

Principem dvouvláknové metody DNA čipu je tzv. paralelní hybridizace cílové směsi (značená analyzovaná cDNA) s tisíce sondami (reprezentovány částí nebo celým genomem). Každá z těchto sond má určené přesné místo na čipu, jehož povrch je rozdělen mřížkou na jednotlivé pole. Sondy jsou nejprve amplifikovány metodou PCR a následně jsou kovalentně přichyceny na povrch čipu (nejčastěji ze skla, plastu či silikonu). Analyzovanou cDNA získáváme reverzní transkripcí z analyzované mRNA, kterou při syntéze značíme fluorescenční barvou. Jinou fluorescenční barvou označíme i tzv. kontrolní cDNA sloužící pro komparativní hybridizaci. [34]

Tyto cDNA jsou následně zahřáty v horké lázni, rychle schlazeny a naneseny na předem připravený čip. Čip s nanesenou cDNA je poté uložen v hybridizační komůrce ve vodní lázni. Po hybridizaci dochází k promytí čipu a následnému skenování. Přístroj určený pro skenování je složen ze dvou laserů o různých vlnových délkách. Lasery je ozářen čip a fluorescenční barva, kterou byla označena cDNA, emituje charakteristické záření, které je detekováno a následně analyzováno. Celý postup metody DNA microarray je znázorněn na Obr. 4-6. [34]



Obr. 4-6: Schéma metody DNA microarray. Upraveno. [34]

4.6 RNA-Seq

Metoda sekvenování RNA (RNA-Seq) v posledních letech postupně nahrazuje využití ostatních metod při analýze genové exprese. Jedná se o užitečný nástroj pro analýzu celého transkriptomu – nejen tedy mRNA, ale i ostatní nekódující RNA. RNA-Seq je

založena na platformách pro sekvenování nové generace (NGS). Zatím bylo odvozeno přibližně 100 různých protokolů pro RNA-Seq. Základním předpokladem pro úspěšnou analýzu metodou RNA-Seq je správný výběr knihovny, hloubky sekvenování a počet replikátů. [35]

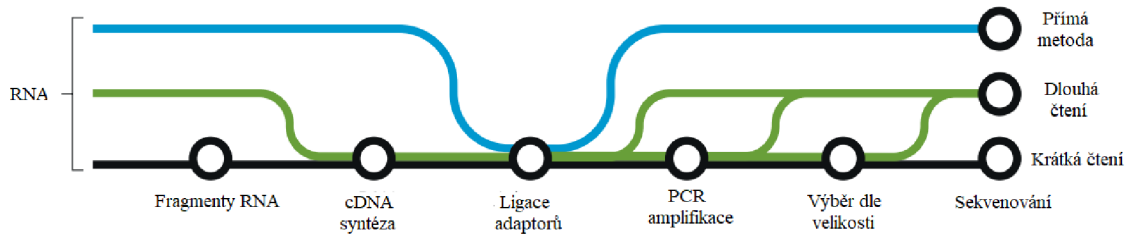
Principem této metody je připravení cDNA knihovny, pro krátká (Illumina), nebo dlouhá (PacBio) čtení, určená pro následovné sekvenování. První krok pro přípravu knihovny pro krátká čtení spočívá v extrakci požadované RNA z buňky. V některých případech je ze vzorku RNA odstraněna rRNA (ta se ve vzorku vyskytuje ve více než 90% zastoupení). Jestliže bude analyzována pouze mRNA, lze pro její izolaci využít selekci na základě poly-A konců za použití oligo(dT). Pokud bude zkoumán celý transkriptom, mimo rRNA, je nutné rRNA ze vzorku odstranit. [35]

Následně dochází k fragmentaci – ta může být enzymatická (využití endonukleáz) nebo fyzikální (teplo). Poté dochází k syntéze reverzní transkriptázou, čímž vzniká cDNA knihovna. Jednotlivé cDNA jsou označeny z jedné nebo obou stran adaptéry (jejich charakter záleží na zvolené metodě následného sekvenování, mohou mít i funkci tzv. barcodes – označují původ sekvencí). [35]

Jestliže vzorek nemá dostatečnou koncentraci, je nutné tuto cDNA nejdříve amplifikovat (i zde záleží na zvolené metodě sekvenování). K tomu se využívají různé PCR metody (emulzní, můstková, …). Tímhle způsobem lze získat fragmenty o velikosti kratší než 200 bp. Pro odstranění fragmentů menší jak 150 bp a větších jak 200 bp jsou používány speciální sondy. [35]

Příprava knihovny pro dlouhá čtení se příliš neliší od přípravy knihovny pro krátká čtení, nedochází zde však k fragmentaci na tak krátké úseky. Levnější krátká čtení se obvykle využívají pro studium genové exprese u dobře anotovaných organizmů, naopak delší čtení je vhodnější pro studium špatně anotovaných transkriptů. Připravená amplifikovaná knihovna cDNA je následně sekvenována. Počet čtení a jejich hloubka záleží na vybrané metodě sekvenování (Illumina, PacBio, …). [35]

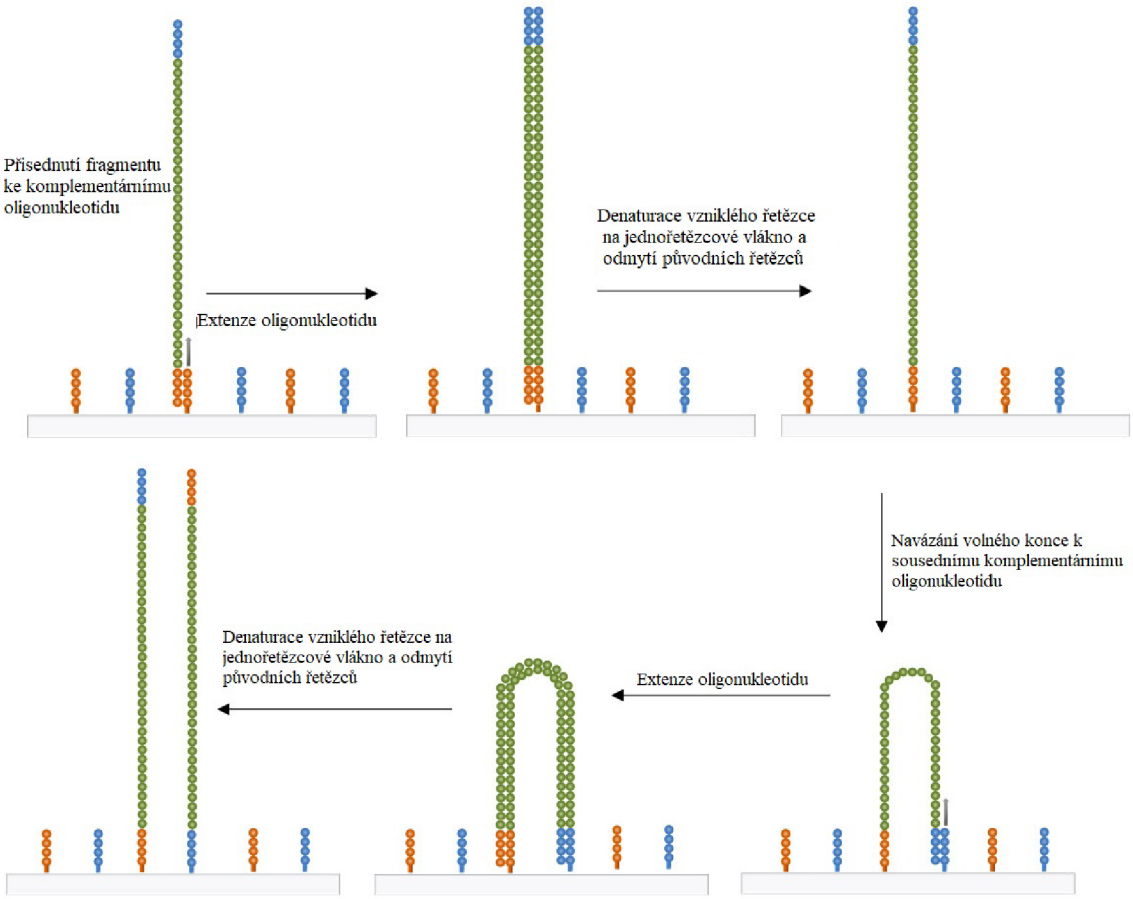
Následně jsou jednotlivá čtení zpracovávána bioinformatickými postupy – zarovnávání, sestavování jednotlivých čtení (využití referenčního genomu, nebo sestavování nových neznámých transkriptů) a statistické analýzy změn v genové expresi. [36] Základní schéma metody RNA-Seq je nastíněno na Obr. 4-7.



Obr. 4-7: Schéma znázorňující základní kroky metody RNA-Seq. Upraveno. [35]

4.6.1 Illumina

Illumina (dříve SOLEXA) je metoda sekvenace, která pracuje na základě sekvenace krátkých čtení (úseky do velikosti 250 bp). Po fragmentaci a následném navázání odlišných adaptorů na jednotlivé konce, jsou fragmenty denaturovány a jsou komplementárně navázány pomocí adaptorů na oligonukleotidy, které jsou přichyceny na povrchu tzv. flow cell (reakční komůrky). Následně dochází k amplifikaci můstkovou PCR. [37] Průběh můstkové PCR je vyobrazen na Obr. 4-8.



Obr. 4-8: Schéma můstkové PCR. [37]

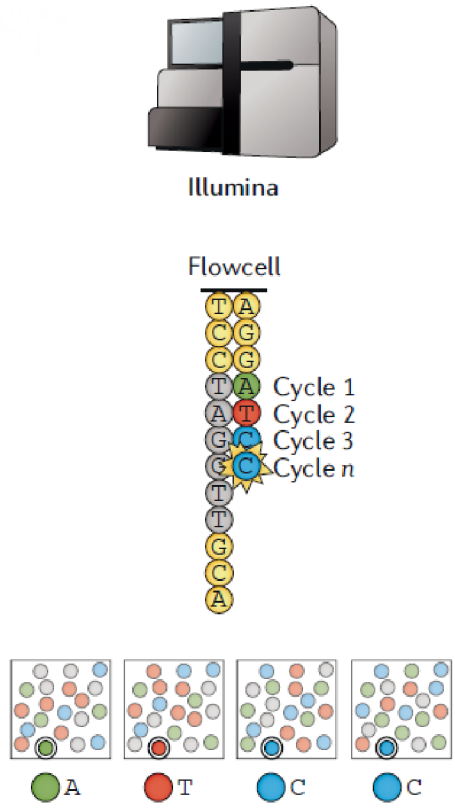
Prvním krokem pro můstkovou PCR je syntéza druhého vlákna dle původní templátové cDNA. Po dokončení této syntézy je dvouřetězcová molekula denaturována a původní templátová cDNA je omyta pryč. Na flow cell zůstává pouze nově dosyntetizovaná cDNA, která je na povrch přichycena kovalentně. Následně dochází k uchycení volného konce cDNA ke komplementárnímu oligonukleotidu. Takto vzniká jednořetězcový můstek. [37]

Poté dochází k dosyntetizování oligonukleotidu a tím ke vzniku dvouřetězcového můstku, jenž je následně denaturován za vzniku dvou jednořetězcových molekul. Každá

z těchto jednořetězcových molekul má jeden konec volný. S těmito volnými konci jsou znovu uchyceny dle komplementarity bází na blízké oligonukleotidy a celý proces se opakuje. Tímto způsobem vznikají na jednotlivých místech reakční komůrky tzv. clustery (místa, kde jsou amplifikovány stejné molekuly cDNA). [37]

Po úspěšné amplifikaci můstkovou PCR dochází k sekvenování. Sekvenování u metody Illumina je započato odstraněním reverzních vláken a přidání primerů. Následně jsou přidány všechny typy nukleotidů, které jsou označeny různými fluofory. Také je do směsi přidána DNA polymeráza a reverzibilní terminátory – ty zapříčiňují, že je v každém cyklu sekvenace navázán na vlákno pouze jeden označený nukleotid. Po uchycení označeného nukleotidu dochází k excitaci fluoroforu laserem a detekci emitovaného světla. Nakonec jsou odstraněny reverzibilní terminátory a celý cyklus se opakuje, dokud není sekvenovaný celý fragment cDNA. [37]

V dnešní době poskytuje firma Illumina několik typů přístrojů pro sekvenaci – Illumina MiniSeq, Illumina MiSeq, Illumina NextSeq 550 Series, Illumina NextSeq 1000 & 2000, Illumina NovaSeq 6000. Každý z těchto přístrojů se liší v počtu čtení a délce fragmentů analyzované DNA. [38] Zjednodušené schéma sekvenace metodou Illumina je na Obr. 4-9.



Obr. 4-9: Schéma znázorňující princip sekvenování metodou Illumina. [35]

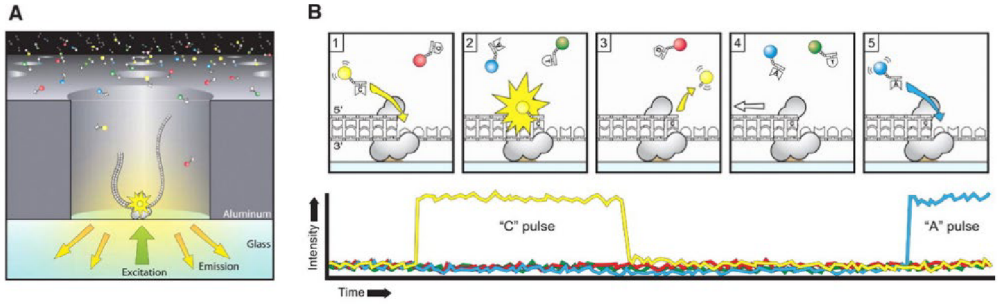
4.6.2 PacBio

Metoda PacBio, vyvinutá společností Pacific BioSciences, nabízí jedinečnou sekvenační analýzu založenou na sekvenování jediné molekuly v reálném čase (single molecule real time, SMRT), což nevyžaduje pauzy mezi jednotlivými čtení a není zde zapotřebí PCR amplifikace. Tato metoda se řadí do skupiny označované sekvenování třetí generace (TGS). Hlavní výhodou jsou dlouhé délky čtení a celkově rychlejší proces, naopak je zde vyšší chybovost a především cena. [39]

Metoda PacBio zachycuje informace o dané sekvenci při replikaci cílové molekuly DNA. Nejprve je vytvořen templát – analyzovaná DNA, která je dvouřetězcová (dsDNA) a na jejichž koncích jsou ligované vlásenkové adaptéry vytvářející smyčku. Tomuto templátu se také říká SMRTbell a je znázorněný na Obr. 4-10. [39]

Obr. 4-10: SMRTbell [39]

Následně je tato SMRTbell nanesena na SMRT cell, což je čip, který obsahuje zhruba 150 000 vlnovodů v nulovém režimu (zero-mode waveguides, ZMW). ZMW jsou jamky s průměrem 70 nm a hloubce 100 nm a poskytují tak nejmenší dostupný objem pro světelnou detekci. Každá ZMW obsahuje ve spodní části imobilizovanou polymerázu, na kterou se může daná SMRTbell navázat pomocí vlásenkového adaptoru. Následně dochází k zahájení replikace. Do SMRT cell jsou přidány fluorescenčně značené nukleotidy, které při navázání v průběhu replikace generují odlišná emisní spektra. Tyto záblesky jsou zaznamenávány pro každou ZMW filmem, jenž tak nese informaci o posloupnosti daných bází v průběhu replikace. Replikace probíhá jak na templátovém, tak i následně na komplementárním řetězci. [39] Schéma metody PacBio je uvedeno na Obr. 4-11.

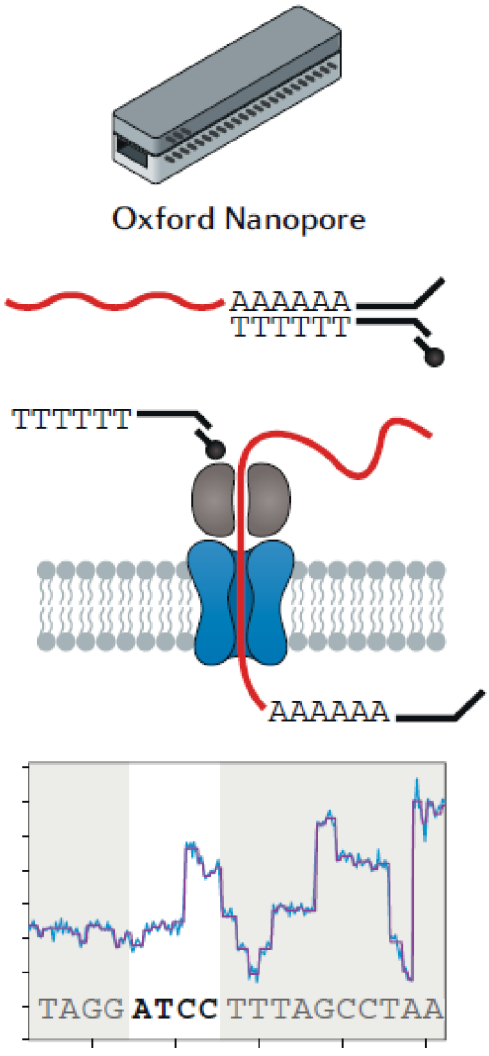


Obr. 4-11: Schéma PacBio metody sekvenování. [39]
(A) Znázorňuje schéma SMRTbell navázané v ZMW, (B) schéma průběhu replikace a detekce emitovaného světla při navázání jednotlivých fluorescenčně značených nukleotidů.

4.6.3 Oxford Nanopore

Metoda vyvinutá společností Oxford Nanopore Technologies, stejně jako předchozí PacBio, nabízí možnost sekvenování jediné molekuly DNA v reálném čase za použití zařízení MinION. Zde však metoda nevyužívá fluorescence, ale změny elektrického proudu při průchodu jednotlivých nukleotidů. Metoda se řadí do TGS, umožňuje vysoké délky čtení (až stovky tisíc párů bází), ale také má vyšší chybovost (až 12 %) spojenou s horší propustností metody. [40]

Přístroj MinION je kompaktní (rozměry 10 x 3 x 2 cm) a cenově dostupnější než zbylé metody pro sekvenování. Velkou výhodou této metody je možnost jej zapojit do stolního počítače nebo notebooku přes USB port a analyzovat tak daný vzorek kdekoliv v terénu. [40]



Obr. 4-12: Schéma metody Oxford Nanopore [35]

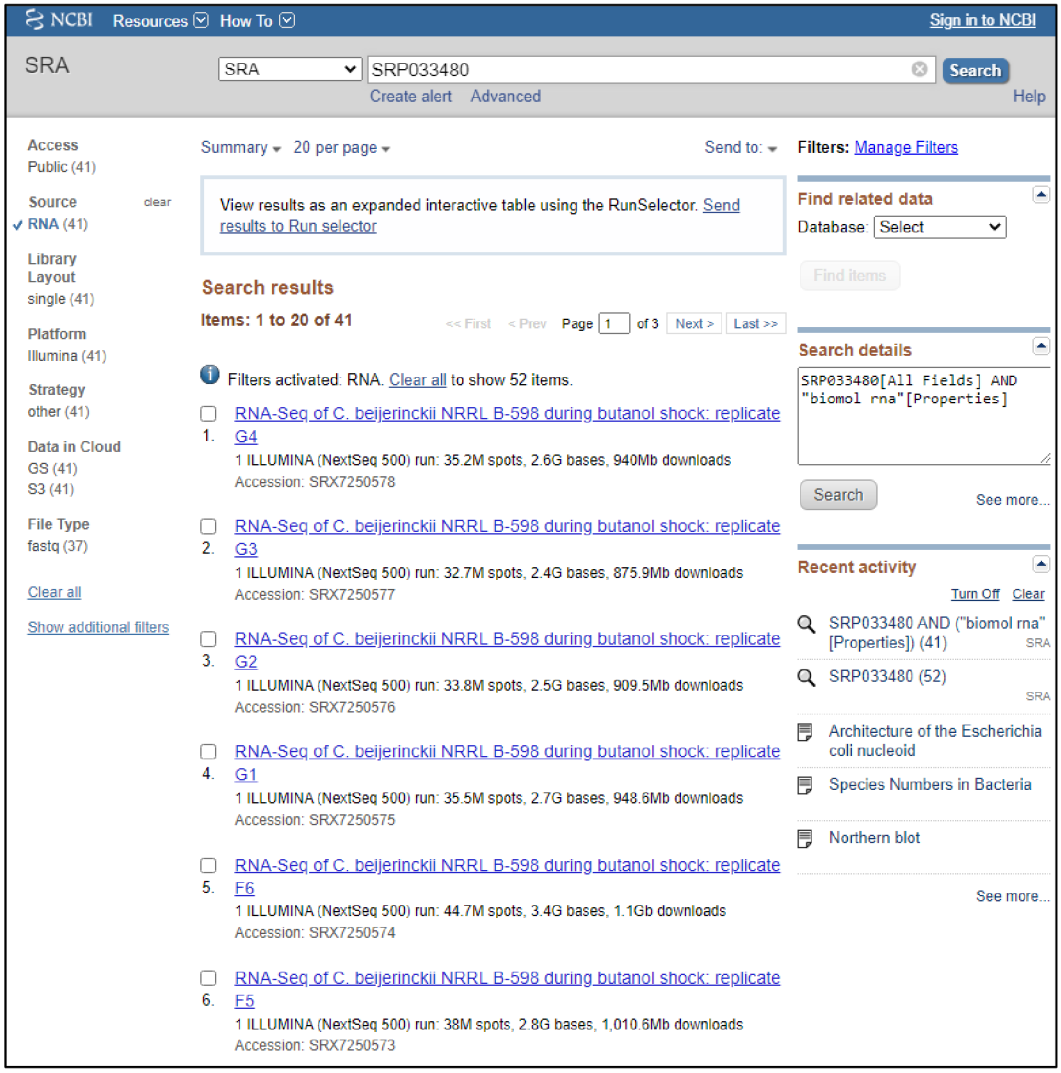
Vzorek DNA, který je zkoumán, je získán pomocí některé ze standartních metod pro extrakci a následné čištění DNA. Poté probíhá příprava knihovny, která se skládá z dlouhých dvouvláknových DNA (dsDNA) tak, aby mohla proběhnout sekvenace obou těchto řetězců. Na každém konci dsDNA je ligován adaptér. Na prvním konci je ligován vedoucí adaptér tzv. Y adaptér, pojmenovaný díky své struktuře ve tvaru „Y“. Na druhém konci je ligován vlásenkový (hairpin, HP) adaptér. Samotná sekvenace začíná na Y adaptéru, za kterým následuje templátová DNA, HP adaptér a nakonec komplementární řetězec. [40]

Fragment sledované DNA je veden přes kanál (pór), který je tvořený proteiny. Tyto proteiny mají za následek rozbalení dsDNA a postupný průchod jednovláknových sekvencí na základě přivedeného napětí na pór. Proteinový kanál charakteristicky ovlivňuje změnou své konformace elektrický proud podle druhu procházejícího nukleotidu. Tento elektrický proud je detekován a převáděn na signál, který je následně analyzován. Zpracovávání dat provádí software zvaný MinKNOW. [40] Zjednodušené schéma metody Oxford Nanopore je vyobrazeno na Obr. 4-12.

5 Postup pro identifikaci nekódující RNA

5.1 Data potřebná pro identifikaci

Pro návrh postupu detekující nekódujících bakteriálních sRNA se v této práci využívají data dostupná z veřejné databáze Národního centra pro biotechnologické informace (NCBI): https://www.ncbi.nlm.nih.gov/. Nejprve došlo k vyhledání dostupných dat pro *Clostridium beijerinckii* NRRL B-598 získaných metodou RNA-Seq a vytvoření databáze těchto dat. Celkově se v databázi NCBI Sequence Read Archive (SRA) nachází 41 záznamů o této bakterii s využitím metody RNA-Seq souhrnně pod přístupovým kódem SRP033480. Jednotlivé záznamy se mezi sebou liší délkou čtení, přípravou vzorku, prostředím kultivace atd. Ukázka úvodní stránky z databáze NCBI je vyobrazena na Obr. 5-1.



Obr. 5-1: Úvodní stránka z databáze NCBI při hledání RNA-Seq dat pro *Clostridium beijerinckii* NRRL B-598

Dále byl vyhledán anotovaný genom a referenční sekvence pro *Clostridium beijerinckii* NRRL B-598 z GenBank databáze NCBI (CP011966.3). Část anotovaného úseku je vyobrazen na Obr. 5-2.

```
gene            500..1849
                /gene="dnaA"
                /locus_tag="X276_26820"
CDS             500..1849
                /gene="dnaA"
                /locus_tag="X276_26820"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:WP_017210024.1"
                /codon_start=1
                /transl_table=11
                /product="chromosomal replication initiator protein DnaA"
                /protein_id="ALB48608.1"
                /translation="MDADLKNLWDKTLDIIKSELSEVSFNTWIKSCEPLSISSNTLKI
                SVPNSFTQDILDKRYKDLVANSIKAVCSKLYTIEFIIMSEIYEKEEIKSSSNQKPKAI
                VVNDEMSSTLNPKYTFNSFVIGNSNRFAHAASLAVAESPAKAYNPLFIYGGVGLGKTH
                LMHAIGHYILDGNPNAKVVYVSSEKFTNELINAIKDDKNEEFRNKYRNVDILLIDDIQ
                FIAGKERTQEEFFHTFNALHDANKQIILSSDRPPKEIPTLEDRLRSRFEWGLIADIQV
                PDFETRMAILKKKADVENLNVANEVMGYIATKIKSNIRELEGALIRIIAYSSLTNREV
                TVDLATEALKDIISKKQGKHVTIDLIQDVVSSYFNLRVEDLKSQRRTRNVAYPRQIAM
                YLSRKLTDMSLPKIGEEFGGRDHTTVIHAYEKISENLKTDDSLQNTVNDITKKLTQN"
gene            2111..3211
                /locus_tag="X276_26815"
CDS             2111..3211
                /locus_tag="X276_26815"
                /EC_number="2.7.7.7"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:WP_011967367.1"
                /codon_start=1
                /transl_table=11
                /product="DNA polymerase III subunit beta"
                /protein_id="ALB48607.1"
                /translation="MIFTCEKQKILEGISIVQKAITGRSTMPILEGIYINASNSTITL
                IGSDMDVSIQTLVDATIMEEGSIVIDAKIFGEIIRKLPNSTIKIETMENQLIKITCEK
                SIFDVVYMNTNEFPELPEINENLKISVNQNILKNMIKGTSFAIAQDETRPILQGILFE
                VRNKNLNLVALDGYRLAIKSEFLDTDIDIEVVIPGKTLNEVSKILEDIDEIVDITFTN
                NHILFNLKRTKIISRLLEGKFINYKSLLPQEHKLFVNVNRQELQNAIERASLMAKDGN
                TNLIKLDLHQDNLVITSNSQLGKVRDEISIKLQGDEIEIAFNSKYLLDVLKNMEDNEV
                VMKMTSGISPCVIEENSNENAKYLVLPVRLMR"
```

Obr. 5-2: Část anotované sekvence pro *Clostridium beijerinckii* NRRL B-598 z webových stránek GenBank databáze (CP011966.3)

Na závěr je pro postup detekce sRNA důležitá znalost o dosud detekovaných sRNA u bakterií. Tyto informace jsou volně dostupné v databázi Rfam (RNA families): http://rfam.xfam.org/. Pomocí této domény také bylo provedeno závěrečné porovnání detekovaných potencionálních sRNA s již známými sRNA.

Pro testování navrženého postupu detekce sRNA byla využita transkriptomická data získaná během 23 hodin při aceton-butanol-ethanol (ABE) fermentaci, v šesti časových krocích tak, aby pokryly významné části fermentačního cyklu *C. beijerinckii* NRRL B-598:

1. T1 (3,5 h) – acidogenní fáze: bakterie produkuje kyseliny (kyselina máselná, kyselina mléčná) a spotřebovává glukózu;
2. T2 (6 hod) – přechod z acidogenní fáze do solventogenní;
3. T3 (8,5 h) – solventogenní fáze: bakterie spotřebovávají kyseliny a produkují rozpouštědla, akumuluje se granulóza a probíhá počáteční fáze sporulace;
4. T4 (13 hod) – solventogeneze a sporulace;
5. T5 (18 hod) – solventogeneze a sporulace;
6. T6 (23 hod) – solventogeneze a sporulace.

Z kultivace bylo v jednotlivých časech T1-T6 odebráno šest vorků. Z každého vzorku byly získány dva replikáty označované označované B1 až B6 a D1 až D6 (tedy celkem 12 souborů dat). RNA získaná ze vzorků byla osekvenována na platformě Illumina NextSeq a výsledná čtení mají délku 75 bp. Během předzpracování těchto dat byly odstraněny adaptéry a nekvalitní čtení. Průměrná kvalita čtení dosahovala PHRED skóre přibližně hodnoty 35. Z dat byly také odstraněny čtení odpovídající zbytkové rRNA. Počet RNA čtení po filtraci se pohyboval od 7,3 do 20,5 milionů čtení na vzorek. Předzpracovaná čtení byla namapována k referenčnímu genomu bakterie *Clostridium beijerinckii* NRRL B-598 (CP011966.3). Získaná zarovnaná čtení v Sequence Alingment/Map (SAM) formátu byla následně indexována a komprimována do Binary Alignment/Map (BAM) formátu. [43]

5.2 Formát souboru BAM

Formát BAM (*.bam) je komprimovaná binární verze souboru SAM (*.sam). Oba tyto formáty slouží k zaznamenání zarovnaných sekvencí k referenci (mapování), kdy BAM soubor má velikost do 128 MB a obsahuje počáteční indexy namapovaných čtení, které slouží pro další rychlé zpracovávání dat. Formát SAM je textový soubor se sloupci, které jsou oddělené tabulátory a obsahují potřebné informace o daných čteních. [41]

Název souboru formátu BAM je formátován jako *NázevVzorku_S*.bam*, kde * je číslo vzorku. Tento formát obsahuje dvě části – hlavičku a část se zarovnáním. Hlavička jako taková obsahuje informaci o celém souboru – název vzorku, délka vzorku a způsob zarovnání. [41]

Část se zarovnáním obsahuje povinná pole, která musejí být vždy vyplněná. Mohou však nabývat hodnot „0“ pro číselné hodnoty a „*“ pro řetězce, jestliže jejich hodnotu neznáme. Celkem BAM soubor může obsahovat 11 a více polí. [41]

Pole `QueryName` obsahuje název templátu, ke kterému se čtení zarovnalo. Tedy čtení obsahující stejné `QueryName` jsou zarovnaná vůči stejné části templátu. Pole `Flag` obsahuje 16-ti bitové příznaky, které nesou informaci o daném čtení. Jejich význam je vypsán v Tabulce 5-1. [41]

Tabulka 5-1: Popis bitových informací pro BAM soubor [41]

Bit		Popis
hex	dec	
0x1	1	templát s více segmenty v sekvenování
0x2	2	každý segment byl správně zarovnán k referenční sekvenci
0x4	4	nezarovnaný segment k referenční sekvenci
0x8	8	následující segment není zarovnaný k sekvenci
0x10	16	SEQ se doplňuje reverzně
0x20	32	SEQ dalšího segmentu se doplňuje reverzně
0x40	64	první segment v templátu
0x80	128	poslední segment v templátu
0x100	256	záznam je sekundárním zarovnáním segmentu
0x200	512	segment neprošel kontrolou kvality
0x400	1024	segment je duplikát (PCR nebo optický)
0x800	2048	dodatečné zarovnání

Dále obsahuje BAM soubor pole `Position`, které nese informaci o počátečních indexech vůči referenci. Pole `MappingQuality` obsahuje informaci o kvalitě namapování daného čtení – tj. zaznamenává pravděpodobnost, jak je pozice namapování čtení chybná. Hodnota je udávána v PHRED skórování a pohybuje se v rozmezí od 0 do 255. Hodnota 255 znamená, že hodnota `MappingQuality` není známá. S tímto dosti úzce souvisí i pole `Quality`, které obsahuje řetězec PHRED skóre pro všechny báze v daném čtení. Z toho plyne, že má tento řetězec stejnou délku, jako čtení. [41]

Ve výčtu polí nechybí ani pole `Sequence`, které obsahuje řetězec s celou sekvencí pro dané čtení. `CigarString` je pole, ve kterém je zapsán řetězec obsahující informace o shodnosti zarovnání s referenční sekvencí. Je zapsána do posloupnosti dvojic – vždy písmeno (operace) a číslo (délka operace). Součet délek všech operací se musí shodovat s délkou dané sekvence. Základní operace a jejich popis jsou vypsané v Tabulce 5-2. [41]

Dále část se zarovnáním obsahuje pole `Tags`, které obsahuje strukturální data. V tomto poli jsou vypsané všechny tzv. standartní značky, které nesou další dodatečné informace o daném čtení (počet čteních namapovaných v daném úseku, skóre zarovnání, …) [42]. BAM soubor poté obsahuje další pole, jako je například `MatePosition`, `InsertSize`, `ReferenceIndex` a `MateReferenceIndex`. [41]

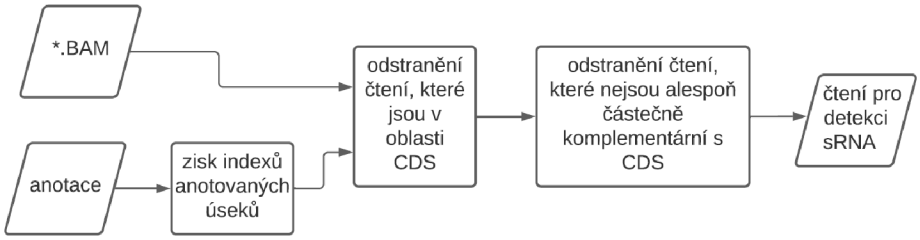
Tabulka 5-2: Popis operací, které se mohou vyskytovat v CigarString [41]

Operace	Popis
M	shoda v zarovnání
I	vložení do referenční sekvence
D	vymazání z referenční sekvence
N	přeskočení části referenční sekvence
S	měkké ořezání
H	tvrdé ořezání
P	vyplnění (mazání z reference)
=	shoda sekvence s referenční sekvencí
X	neshoda sekvence s referenční sekvencí

5.3 Návrh postupu pro detekci sRNA

Navržený postup slouží k detekci jednoho typu sRNA – sRNA, která je komplementární, nebo alespoň částečně komplementární k mRNA. S tímto přístupem může být předpokládáno a poté určeny pouze ta místa, kde by se dané sRNA mohly nacházet.

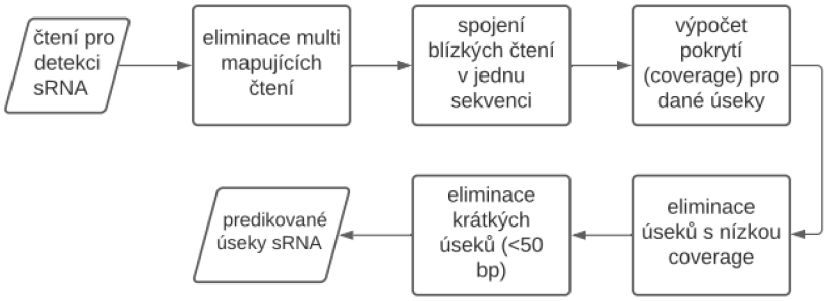
Postup detekce sRNA počítá s využitím dat ve formátu BAM, ve kterém jsou obsažena jednotlivá čtení. Tato čtení jsou již namapované k referenci *Clostridium beijerinckii* NRRL B-598. Dále je využito anotovaných úseků pro tuto bakterii. V postupu dochází nejprve k eliminaci čtení, která se nachází ve stejném směru a ve stejných pozicích jako již anotované úseky. Poté je využito faktu, že je požadováno získat sRNA alespoň částečně komplementární s nějakou mRNA. V postupu budou tedy eliminována ta čtení, která nemají na opačném vlákně v pozici, kde se namapovala, nějaký anotovaný gen tedy kódující sekvenci (CDS). Touto eliminací zůstanou pouze ta čtení, jež jsou v daném směru komplementární s CDS v opačném směru. Tento postup je nastíněn schématem na Obr. 5-3.



Obr. 5-3: Schéma první části navrženého postupu

Poté je u těchto čtení uvažována unikátnost mapování na daný úsek a čtení, která nebyla unikátně namapována, jsou eliminována. Také je uvažována míra pokrytí (tzv.

coverage) v daném úseku a jsou odstraněny ty úseky, které mají hloubku pokrytí příliš nízkou. Nakonec jsou vyřazeny ze seznamu potencionálních sRNA příliš krátké úseky. Tyto kroky jsou znázorněny na Obr. 5-4.



Obr. 5-4: Schéma druhé části navrženého postupu

Následně jsou detekované potencionální úseky sRNA pro bakterii *Clostridium beijerinckii* NRRL B-598 porovnávány s již známými sRNA napříč bakteriální sférou z Rfam databáze. Tímto se zvyšuje pravděpodobnost, že daný detekovaný úsek je skutečná sRNA a bylo by vhodné jej i nadále zkoumat. Tato část postupu je znázorněna na Obr. 5-5.

Obr. 5-5: Schéma poslední části navrženého postupu

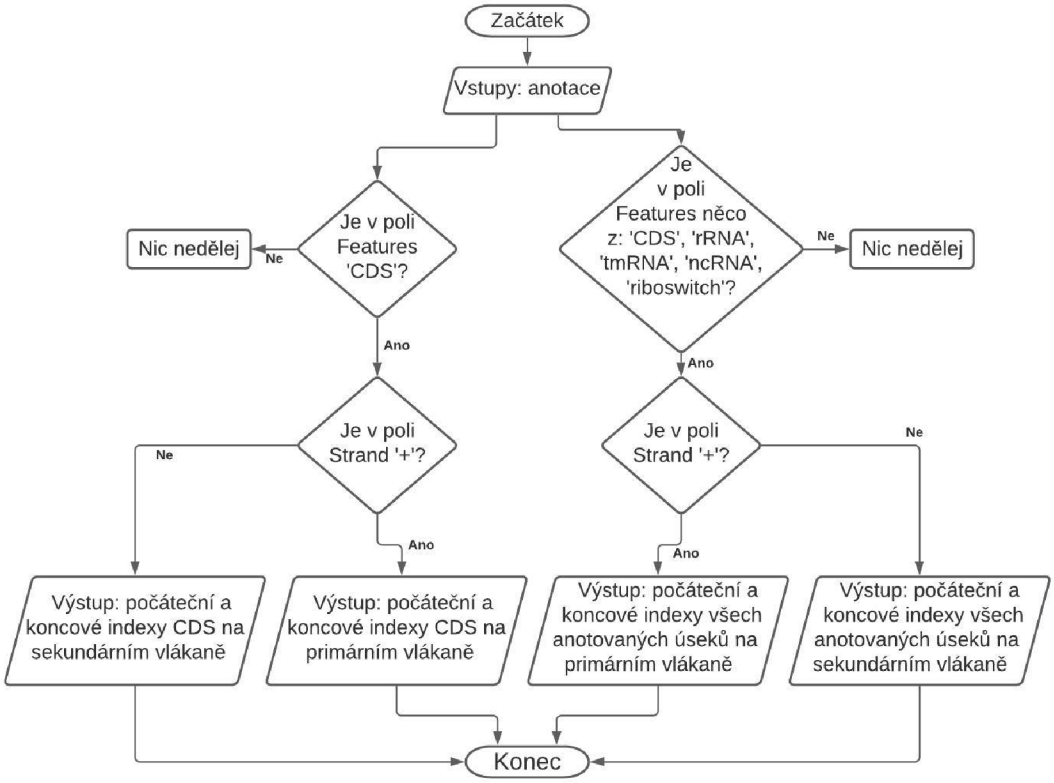
5.4 Implementace návrhu postupu pro detekci sRNA v prostředí MATLAB

Navržený postup z kapitoly 5.3 byl implementován v programovacím prostředí MATLAB. Celý postup se nachází ve skriptu s názvem `bp_detekce_sRNA.m` a byl rozdělen do 11 kratších bloků, které budou podrobněji popsány v následujících odstavcích. Některé z bloků jsou tvořeny novými funkcemi, které zlepšují efektivitu a rychlost výpočtu.

V první části kódu dochází k načtení všech potřebných dat. Jedná se o referenční sekvenci bakterie *Clostridium beijerinckii* NRRL B-598 (CP011966.3) ve formátu FASTA, anotaci k této sekvenci ve formátu gff3 (Generic Feature Format v.3) a BAM soubor obsahující namapovaná čtení k referenci.

V druhé části kódu jsou z anotace získány informace o počátečních a koncových indexech všech anotovaných úseků a jednotlivých CDS. Obě kategorie jsou rozděleny

také podle toho, zda je daný úsek anotován na primárním či sekundárním vlákně. Tímto je získáno osm nových proměnných – počáteční a koncové indexy všech známých anotovaných úseků a počáteční a koncové indexy všech CDS a to pro obě vlákna. První dvě proměnné pro dané vlákno obsahují informace o CDS, které jsou i v druhých proměnných, ale ty navíc obsahují informace o ostatních anotovaných úsecích (rRNA, tRNA, riboswitch, atd.). Tato část kódu je obsažena ve funkci `getindex` a jak je patrné z předchozího textu, vstupem této funkce je anotace a výstupem 8 nových proměnných obsahující informace o start a stop pozicích anotovaných úseků. Schéma funkce `getindex` je znázorněno na Obr. 5-6.



Obr. 5-6: Schéma funkce `getindex`

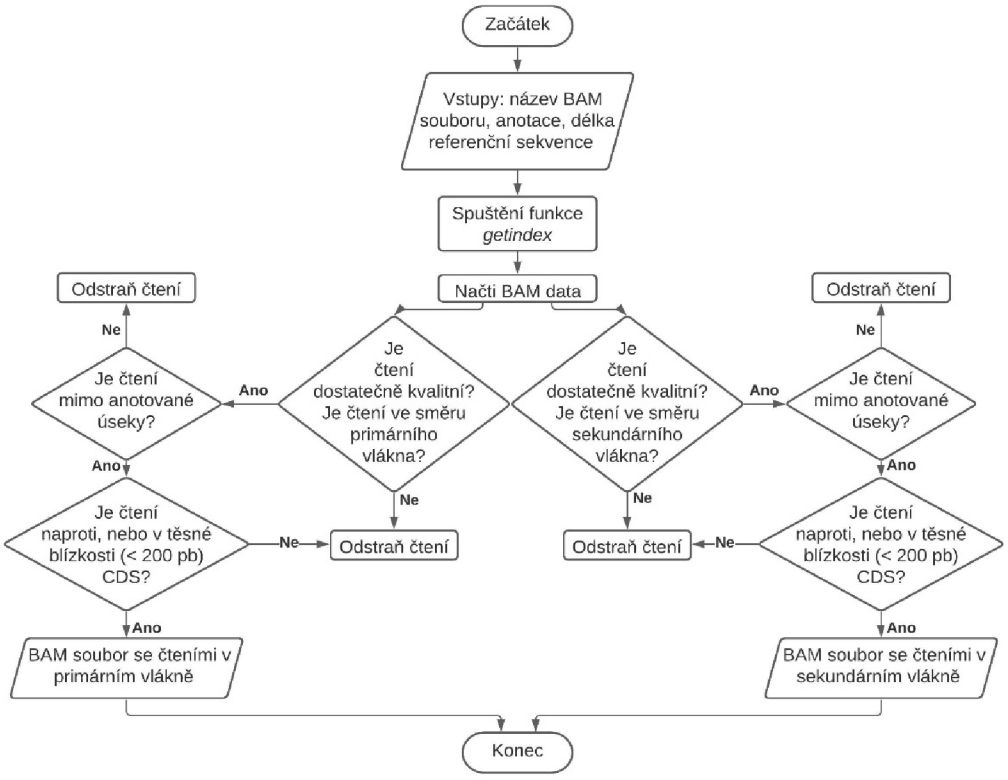
Další část je zaměřena na práci s BAM souborem a dochází zde k eliminaci určitých čtení. I pro tuto část byla vytvořena nová funkce s názvem `getreads`. Vstupem této funkce je název pro BAM soubor, který má být zpracován, anotace a délka referenční sekvence. Tato funkce nemá výstup, jelikož jsou v rámci její struktury vytvářeny a ukládány nové filtrované BAM soubory. Vstupní proměnná obsahující informaci o anotaci je nezbytná, neboť v rámci této funkce je znovu volána funkce `getindex`, která byla popsána výše. Ve funkci `getreads`, po proběhnutí funkce `geitindex`, dochází k načtení BAM dat pomocí inicializace instance třídy `BioMap`, která je implementovaná v prostředí MATLAB. Vstupním argumentem této třídy je název pro BAM soubor a výstupem je objekt třídy `BioMap` obsahující kompletní data o jednotlivých čteních namapovaných k referenci uložených v BAM souboru.

Následně jsou rozdělena čtení, která jsou obsažena v objektu z BAM dat, podle toho, jestli se namapovala k primárnímu či sekundárnímu vláknu. Také jsou odstraněna všechna čtení, která nebyla unikátně namapována. K rozdělení čtení pro jednotlivá vlákna je využito informací, které BAM formát obsahuje, a to v tzv. flags. Význam jednotlivých flags jsou uvedeny v kapitole 5.2. Po této úpravě je ve zbylé části kódu pracováno odděleně se čteními pro primární a sekundární vlákno.

V dalším kroku jsou pro jednotlivá vlákna eliminována ta čtení, která se nachází na stejném vlákně v již známé a anotované oblasti. Jejich funkce je již objasněná a nemusí být tedy uvažována pro další analýzu a detekci sRNA.

Jelikož se celý postup detekce sRNA zaměřuje na typ sRNA, která jsou alespoň částečně komplementární k nějakému CDS, jsou v další části funkce eliminována i ta čtení, jež nemají na protějším vlákně v jejich rozsahu, nebo alespoň v těsné blízkosti (do 200 bp, rozptyl byl zvolen díky předpokládaným velikostem sRNA, které jsou uvedeny v kapitole 3.5), některou anotovanou CDS.

Na závěr jsou vytvářeny a ukládány nové BAM soubory, které obsahují pouze ta čtení, která by mohla být potencionálními čteními pro sRNA. K uložení těchto BAM souborů dochází pomocí dostupné funkce MATLAB `write`. Funkce `write` má dva vstupní argumenty. Prvním argumentem jsou data, která budou uložena. Druhý vstupní argument určuje název souboru a také jeho formát. Tímto krokem jsou získány dva BAM soubory, jeden pro primární a druhý pro sekundární vlákno. Schéma funkce `getreads` je znázorněno na Obr. 5-7.



Obr. 5-7: Schéma funkce `getreads`

Ve čtvrté části dochází pouze k načítání nově vytvořených BAM souborů z předchozího kroku pomocí již implementované funkce v MATLAB – funkce `bamread`. Struktura takto načtených dat ve formátu BAM je znázorněn na Obr. 5-8.

204307x1 struct with 12 fields

Fields	QueryName	Flag	Position	MappingQ	CigarS	Sequence	Quality	Tags
1	'NB501229:74:...	16	26502	255	'75M'	'ATAATAAG...	'EEEEEEEEEE...	1x1 struct
2	'NB501229:74:...	16	26508	255	'75M'	'AGAAAAGC...	'EAEEEE<EE...	1x1 struct
3	'NB501229:74:...	16	26596	255	'75M'	'AAAAATCC...	'AEEEEAEEE...	1x1 struct
4	'NB501229:74:...	16	26596	255	'75M'	'AAAAATCC...	'EEEEEEAEE...	1x1 struct
5	'NB501229:74:...	16	26598	255	'75M'	'AAATCCAT...	'EEEAA<AE...	1x1 struct
6	'NB501229:74:...	16	26630	255	'75M'	'ATACAGCC...	'EEEEEEEEE...	1x1 struct

Obr. 5-8: Část BAM souboru po načtení pomocí funkce `bamread`

Následně se pokračuje do páté části kódu. V tomto bloku jsou tvořeny nové struktury z jednotlivých čtení. Nově vytvořená struktura obsahuje podstruktury obsahující pouze ta čtení, která jsou od sebe v maximální vzdálenosti 75 bp (tento rozsah byl zvolen díky známé délce čtení sekvenace). Tímto jsou vytvářeny potencionální „shluky“ všech čtení nacházejících se ve stejné oblasti. Následně dochází k eliminaci struktur mající čtyři a méně čtení, neboť už z tohoto může být usuzováno, že celkové pokrytí, které je v dalších krocích požadováno, bude nízké. Celý tento krok je proveden pro obě vlákna. Názorná ukázka nově vytvořených struktur je vyobrazena na Obr. 5-9.

1x2755 struct with 1 field

Fields	poradi
1	1x16 struct
2	1x34 struct
3	1x447 str...
4	1x8 struct
5	1x47 struct
6	1x15 struct
7	1x6 struct
8	1x8 struct
9	1x15 struct

VARIABLE SELECTION EDIT

dopredne(4).poradi

Fields	QueryName	Flag	Position	MappingQuality	CigarString	Seq
1	'NB501229:74:...	16	32660	255	'75M'	'AATAT
2	'NB501229:74:...	16	32661	255	'75M'	'ATATT
3	'NB501229:74:...	16	32661	255	'75M'	'ATATT
4	'NB501229:74:...	16	32661	255	'75M'	'ATATT
5	'NB501229:74:...	16	32661	255	'75M'	'ATATT
6	'NB501229:74:...	16	32663	255	'75M'	'ATTAA
7	'NB501229:74:...	16	32665	255	'74M'	'TAATT
8	'NB501229:74:...	16	32720	255	'75M'	'AAATA

Obr. 5-9: Nově vytvořená struktura obsahující další struktury se čteními

V šestém bloku se pro každý „shluk“ vytvořený v páté části určuje odpovídající referenční sekvence. Tato referenční sekvence je získána na základě indexů, kde se jednotlivá čtení ze „shluku“ namapovala k referenci. Referenční úsek, který je získán, je následně využit k výpočtu míry pokrytí podél celé jeho délky. Dále je určen počet čtení namapovaných v tomto úseku, a nakonec je vypočítán počet čtení na milion kilobází (RPKM). Hodnota RPKM je definována vztahem

$$RPMK = \frac{10^9 * PČKT}{PVČ * DT}, \tag{6.1}$$

kde $PČKT$ je počet čtení namapovaných k danému úseku, $PVČ$ počet čtení namapovaných k celé referenci a DT je délka daného úseku. Jedná se o hodnotu, která normalizuje míru exprese daného úseku [44].

Takto získané informace jsou ukládány do nové struktury, která obsahuje referenční úsek, vypočítanou míru pokrytí, počáteční a koncový index, kde byl úsek na referenci namapován a hodnotu RPMK. Ukázku struktury nově vytvořené proměnné v této části kódu je vyobrazen na Obr. 5-10. Na závěr této části jsou ještě eliminovány ty úseky, které mají RPMK rovno nebo nižší 10 a odpovídají tedy pravděpodobně biologickému nebo technickému šumu. Se zbylými daty poté pracuje sedmý blok kódu.

1x537 struct with 6 fields

Fields	sekvence	pokryti	start	stop	pocet_cteni	RPKM
1	'CTAAAAAT...	1x1248 dou...	29583	30830	447	104
2	'TGATTCTAT...	1x263 double	32890	33152	47	11
3	'GAACCAAT...	1x828 double	45618	46445	172	40
4	'AAGCAGTT...	1x2309 dou...	53000	55308	1065	248
5	'ATGAATAA...	1x899 double	57861	58759	126	29
6	'AAATTTGA...	1x1059 dou...	83844	84902	753	175
7	'TCAAGTCA...	1x1105 dou...	151672	152776	87	20

Obr. 5-10: Struktura nově vytvořené proměnné v šestém úseku kódu

V sedmém bloku jsou odstraněny z potencionálních úseků získaných v předchozím kroku nukleotidy s pokrytím nižším než 3. Také dochází k rozdělení úseku na dvě a více částí, jestliže se uprostřed původního úseku nachází oblast obsahující alespoň 30 nukleotidů bezprostředně za sebou, které mají míru pokrytí nižší než 3. Nově získaná data jsou znovu uloženy do nové proměnné ve stejném tvaru, jako v předchozí části kódu. Jediným rozdílem je to, že obsahuje navíc pole, které nese informaci o průměrném pokrytí napříč celým úsekem (hloubka pokrytí) a postrádá pole s počtem čtení a hodnotou RPMK. Nově vytvořená proměnná je ukázána na Obr. 5-11.

1x786 struct with 5 fields

Fields	sekvence	pokryti	start	stop	depth
1	'TAATTTTTA...	1x1231 dou...	29597	30827	26.8829
2	'TATAGTTTG...	1x236 double	32903	33138	14.0638
3	'ATTAGGAT...	1x799 double	45629	46427	15.7569
4	'CAGTTCCC...	1x2302 dou...	53003	55304	34.0409
5	'TAGTTGAT...	1x759 double	57929	58687	11.8786
6	'ATGCATGA...	1x800 double	84101	84900	12.4788
7	'AAATTTGA...	1x195 double	83844	84038	233.8918

Obr. 5-11: Struktura nové proměnné vytvořené v sedmém úseku kódu

Jelikož v předchozím bloku došlo k rozdělení některých úseků na kratší, mohlo nastat, že nově vytvořené úseky budou příliš krátké pro případná sRNA. Proto dochází v osmé části k eliminaci těch úseků, které jsou kratší jak 50 bp a delší než 500 bp. Tato eliminace je založena na znalosti délek sRNA, které jsou uvedeny v kapitole 3.5. Také jsou v této části eliminovány ty úseky, které mají hloubku pokrytí nižší nebo rovnu 5. Tato hodnota byla nastavena deterministicky.

Také mohlo dojít, v rámci rozdělení úseku v sedmé části, k vytvoření úseků, které již nejsou ani částečně komplementární vůči CDS na protějším vlákně. Proto dochází v deváté části kódu k eliminaci těchto úseků. Jsou zde znovu využity indexy získané v druhém kroku pomocí funkce `getindex`.

V desáté části se ke zbylým úsekům dohledávají potřebné informace z anotace a jsou ukládány do struktury, popsané v předchozí části kódu, do nových polí. Jsou zde doplněny informace o tom, jaká CDS se nachází naproti danému úseku a také jakou funkci tato CDS má. Část této struktury je vyobrazen na Obr. 5-12.

1x335 struct with 7 fields

Fields	sekvence	pokryti	start	stop	dept	CDS	produkt
1	'TATAGTTTG...	*1x236 double*	32903	33138	14.0638	"X276_26670"	"ATP-binding cassette domain-co...
2	'AATGCAAG...	*1x271 double*	151689	151959	6.0593	"X276_26110"	"M20 family peptidase"
3	'TTAAAGAT...	*1x491 double*	163334	163824	68.5816	"X276_26040"	"glutaredoxin family protein"
4	'AAAAATAA...	*1x96 double*	294336	294431	12.8421	"X276_25370"	"HutD-family protein"
5	'TTTTCACCT...	*1x141 double*	310509	310649	22.4894	"X276_25305"	"[FeFe] hydrogenase H-cluster radi...
6	'TCATTTATA...	*1x99 double*	310830	310928	59.8283	"X276_25305"	"[FeFe] hydrogenase H-cluster radi...
7	'TTATTTATC...	*1x347 double*	342714	343060	7.5562	"X276_25185"	"XRE family transcriptional regulat...
8	'TTTGAGCT...	*1x219 double*	384230	384448	15.0913	"X276_25000"	"hypothetical protein"

Obr. 5-12: Doplněné informace o komplementární CDS a její funkci

Na závěr celého kódu jsou spojeny potencionální úseky jak primárního, tak i sekundárního vlákna, které byly získány předchozími kroky, do jediné struktury. U této nové struktury se mění i formát. Nachází se zde pouze dvě pole. První pole je nazváno jako `Header` a obsahuje pro jednotlivé úseky počáteční indexy vůči referenci. Pole `Sequence` obsahuje poté danou sekvenci. Ukázku této struktury je na Obr. 5-13.

1x786 struct with 2 fields

Fields	Header	Sequence
1	'32903'	'TATAGTTTGAACGAACTGATGGTCATGAGAAGCAAATAATATATTACTGTTAT...
2	'151689'	'AATGCAAGAGCTGTAGCTGCTATATTGCATTGTTTTAAAGCAAATTCAGATGT...
3	'163334'	'TTAAAGATGAGAATATAAATCAGATATTAATTGATATAAGCCAAAATAAAAG...
4	'294336'	'AAAAATAAAAAGGAGTTATCTCAAATTAGAAGTAATTTGAGATGAAGCCTTTT...
5	'310509'	'TTTTCACCTAAACTTAAGGTTAAAGCGACATTTAATTTCTTTATCTCTTTTACTA...
6	'310830'	'TCATTTATATCATCGTTTTGTAGTAATTCAAGTATTTCATCTTTATTTAAGTCATG...
7	'342714'	'TTATTTATCAAATCTATATATTTCTCATTCAACTTTGGTAATTTTGTTACTATTTC...

Obr. 5-13: Struktura nové proměnné vytvořené v desátém úseku kódu

Důvodem pro tvorbu této struktury byl finální krok implementovaného postupu, a to uložení těchto potencionálních úseků do souboru formátu FASTA. K tomuto byla využita funkce dostupná v MATLAB a to funkce `fastawrite`. Vzniklý FASTA soubor byl následně porovnán mimo prostředí MATLAB s již známými sRNA na doméně Rfam (RNA families): http://rfam.xfam.org/.

6 Výsledky

Po implementaci výše uvedeného návrhu v prostředí MATLAB byly detekovány potencionální sRNA pro bakterii *Clostridium beijerinckii* NRRL B-598 pro jednotlivé replikáty datasetu. Bylo analyzováno dvanáct replikátů, které byly získány osekvenováním šesti vzorků pocházející ze stejné kultivace. Tyto vzorky byly odebrány v šesti různých časových okamžicích s dvakrát osekvenovány, jak už bylo zmíněno v kapitole 5.1. Počty výsledně detekovaných potencionálních úseků sRNA pro replikáty B a D jsou znázorněny v Tabulce 6-1 a Tabulce 6-2.

Tabulka 6-1: Počet detekovaných sRNA pro replikáty B v jednotlivých časech

Replikát	Počet detekovaných úseků na primárním vlákně	Počet detekovaných úseků na sekundárním vlákně
B1	195	192
B2	159	212
B3	126	161
B4	335	451
B5	210	249
B6	265	397

Tabulka 6-2: Počet detekovaných sRNA pro replikáty D v jednotlivých časech

Replikát	Počet detekovaných úseků na primárním vlákně	Počet detekovaných úseků na sekundárním vlákně
D1	78	69
D2	75	86
D3	101	99
D4	160	193
D5	225	211
D6	194	206

Analýza pomocí dostupné databáze Rfam neodhalila žádnou již známou sRNA, která by odpovídala zde detekovaným úsekům. Nicméně získané výsledky obsahují několik zajímavých genů, ke kterým byly nalezeny komplementární úseky ve všech časech pro oba vzorky. Pro vyhodnocení, zda by se mohlo jednat o případné sRNA, bylo pozorováno, jak se v průběhu fermentace *Clostridium beijerinckii* NRRL B-598

měnily oblasti detekovaných potencionálních sRNA. Převážně je sledováno, jak se mění hloubka pokrytí a jejich délka.

Prvním genem, který byl vybrán, je gen X276_23205 nacházející se na primárním vlákně na pozicích od 776221 do 777135. V bakterii má funkci jakožto transpozázový protein, který je nezbytný pro účinnou transpozici DNA [45]. Podrobnější informace o detekovaných úsecích ve všech časech vůči danému genu jsou uvedeny v Tabulce 6-3 a Tabulce 6-4. Z tabulek lze vyčíst, jak se v průběhu sporulace mění i genová exprese detekovaného úseku.

Je patrné, že je hloubka pokrytí u replikátu B je vyšší než u replikátu D. Přesto lze vidět, že změny v časech u obou replikátů mezi sebou korelují. V obou případech je míra pokrytí nejvyšší v čase T4. Naopak nejnižší je v časech T2 a T6. U replikátu D dokonce nedošlo k detekci v čase T6. Tato skutečnost může být přisouzena možnému přísnému požadavku na hloubku pokrytí v detekci. Je předpokládáno, že při snížení požadavků na menší hloubku pokrytí by i zde byl detekován podobný úsek. Průměrná délka úseku je 136 bp a modus je roven 145 bp, což je optimální délka pro sRNA.

Tabulka 6-3: Informace o detekovaných úsecích replikátu B genu X276_23205

X276_23205, replikát B				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
B1	776090	776234	145	48.49
B2	776092	776233	142	23.59
B3	776090	776234	145	33.90
B4	776090	776234	145	53.89
B5	776093	776233	141	40.66
B6	776113	776233	121	23.53

Tabulka 6-4: Informace o detekovaných úsecích replikátu D genu X276_23205

X276_23205, replikát D				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
D1	776091	776227	137	27.03
D2	776103	776222	120	11.11
D3	776090	776229	140	29.84
D4	776092	776231	140	57.04
D5	776107	776223	117	22.14
D6	-	-	-	-

Druhým zvoleným genem byl gen X276_15585 nacházející se na primárním vlákně na pozicích od 2522254 do 2522439. Jedná se o gen nesoucí informaci o proteinu patřící do rodiny YvrJ proteinů, jejichž funkce dosud nebyla objasněna [46]. V každém časovém kroku pro oba replikáty byl detekován právě jeden úsek odpovídající potencionální sRNA. Podrobnější informace o detekovaných úsecích v prvním i v dalších časech vůči danému genu jsou uvedeny v Tabulce 6-5 a Tabulce 6-6.

Lze pozorovat, že i zde hloubka pokrytí mezi replikáty do určité míry souvisí. Obdobně, jako u prvního zvoleného genu, i zde je v replikátu B hloubka pokrytí ve všech časech vyšší než v replikátu D. Také lze pozorovat, že nejvyšší hloubka pokrytí je pro oba replikáty v čase T1 a nejnižší naopak v čase T6. Také si lze povšimnout, že v čase T6 je obrovský pokles délky detekovaného úseku. Po bližším prozkoumání bylo zjištěno, že časy T1-T5 po celé své délce obsahují dva úseky, které mají velmi vysoké pokrytí a jsou rozděleny úsekem s mnohem nižším pokrytí. Proto se lze domnívat, že je pro zde navrženou detekci, zajímavý pouze úsek v replikátech B i D v čase T6. Tento úsek by i svou délkou více odpovídal hledané sRNA než celý úsek v ostatních časech.

Tabulka 6-5: Informace o detekovaných úsecích replikátu B genu X276_15585

X276_15585, replikát B				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
B1	2522301	2522731	431	37.82
B2	2522272	2522721	450	22.93
B3	2522269	2522719	451	32.96
B4	2522282	2522717	436	28.12
B5	2522307	2522715	409	18.07
B6	2522327	2522464	138	14.85

Tabulka 6-6: Informace o detekovaných úsecích replikátu D genu X276_15585

X276_15585, replikát D				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
D1	2522299	2522705	407	25.06
D2	2522299	2522701	403	19.16
D3	2522300	2522705	406	20.00
D4	2522279	2522701	423	24.00
D5	2522297	2522685	389	11.42
D6	2522305	2522454	150	10.30

Dalším, třetím, genem byl pro zhodnocení úspěšnosti detekce vybrán gen X276_04065 nacházející se na primárním vlákně na pozicích od 5170085 do 5171293. Jedná se o gen pro protein, který je součástí tzv. bičíkového spínače a ovlivňuje aktivitu fosfoproteinové fosfatázy [47]. V každém časovém kroku pro oba replikáty B a D byl detekován právě jeden úsek odpovídající potencionální sRNA. Podrobnější informace o detekovaných úsecích v prvním i v dalších časech vůči danému genu jsou uvedeny v Tabulce 6-7 a Tabulce 6-8.

Je patrné, že naproti oběma úsekům, které byly vybrány v předchozím popisu, je hloubka pokrytí ve všech časech vyšší u replikátu D a u replikátu B je nižší. Stejně jako u předchozích potencionálních sRNA, i zde hloubky pokrytí v čase mezi vzorky souvisí. Nejvyšší hloubka pokrytí se nachází opět v čase T1. Nejnižší opět v čase T6. Také je možné pozorovat, že v replikátu B má detekovaný úsek ve všech časech stejný jak počáteční, tak i koncový index, a tedy i stejnou délku. V replikátu D je délka o 2 bp delší. Po bližším zkoumání bylo však zjištěno, že tyto dva nukleotidy, které jsou oproti replikátu B „navíc“ mají mnohem nižší pokrytí a proto je do finálního predikovaného úseku není třeba uvažovat.

Tabulka 6-7: Informace o detekovaných úsecích replikátu B genu X276_04065

X276_04065, replikát B				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
B1	5171131	5171205	75	26.47
B2	5171131	5171205	75	16.30
B3	5171131	5171205	75	16.86
B4	5171131	5171205	75	18.39
B5	5171131	5171205	75	14.86
B6	5171131	5171205	75	5.78

Tabulka 6-8: Informace o detekovaných úsecích replikátu D genu X276_04065

X276_04065, replikát D				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
D1	5171131	5171207	77	77.29
D2	5171131	5171207	77	45.72
D3	5171131	5171207	77	57.33
D4	5171131	5171207	77	50.25
D5	5171131	5171207	77	36.30
D6	5171131	5171207	77	16.71

Další úseky na primárním vlákně zaznamenané, jako potencionální sRNA, jsou komplementární vůči následujícím CDS: X276_10525, X276_13890, X276_02670. Úseky naproti těmto CDS byly detekovány ve většině časových kroků u replikátů B i D.

Na sekundárním vlákně bylo taktéž vybráno pár úseků, které byly získány pomocí zde implementované detekce. První gen, který byl ze sekundárního vlákna vybrán, byl gen X276_26950 nacházející se na referenci v pozicích od 2432535 do 2432687. V bakterii nese tento gen informace o aspartyl-fosfát fosfatázovém proteinu z rodiny Spo0E. Pro oba replikáty B i D, byl detekován jeden úsek pro potencionální sRNA. Podrobnější informace o úsecích jsou uvedeny v Tabulce 6-9 a Tabulce 6-10.

Z tabulek lze vyčíst, že délka i hloubka pokrytí se v časech u obou replikátů liší. Přesto je možné určit nějaké spojitosti i zde. Nejnižší hodnota hloubky pokrytí je v čase T6 u obou replikátů. Nejvyšší hodnoty jsou v časech T3 a T4. U replikátu B v čase T2 má detekovaný úsek délku 188 bp a u replikátu D v čase T6 má délku 191 bp, což je oproti ostatním časovým krokům rapidní rozdíl. Po bližším zkoumání bylo zjištěno, že míra pokrytí v oblastech odpovídající B2 a D6 je mnohonásobně vyšší než míra pokrytí v okolí. Dá se tedy uvažovat, že hledaná sRNA je pouze v úseku odpovídající B2 a D6.

Tabulka 6-9: Informace o detekovaných úsecích replikátu B genu X276_26950

X276_26950, replikát B				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
B1	2432320	2432613	294	27.96
B2	2432423	2432610	188	26.56
B3	2432327	2432611	285	33.38
B4	2432313	2432612	300	31.47
B5	2432323	2432606	284	15.19
B6	2432321	2432610	290	14.87

Tabulka 6-10: Informace o detekovaných úsecích replikátu D genu X276_26950

X276_26950, replikát D				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
D1	2432361	2432620	260	24.70
D2	2432350	2432620	271	20.55
D3	2432341	2432628	288	37.74
D4	2432319	2432630	312	44.97
D5	2432332	2432620	289	22.38
D6	2432428	2432618	191	18.91

Dalším genem, který obsahuje zajímavé komplementární úseky získané pomocí detekce, je gen X276_27735. Tento gen se nachází na referenci v rozsahu od 4691843 do 4691962. Jedná se o gen nesoucí informace o cyklickém laktonovém autoindukčním proteinu. Podrobnější informace o detekovaných úsecích ve všech časových krocích vůči genu X276_27735 jsou uvedeny v Tabulce 6-11 a Tabulce 6-12.

Délky se v jednotlivých časech dost výrazně mění. Tím dochází i k ovlivnění hloubky pokrytí, neboť se pro výpočet hloubky pokrytí u delších úseků započítávají i úseky s nižším pokrytím. Po detailnějším prozkoumání bylo zjištěno, že nejoptimálnější úsek odpovídá replikátu B2 a D2, kde je míra pokrytí relativně konstantní a délka také odpovídá více sRNA. I v ostatních časových krocích se nachází v tomto úseku vysoké pokrytí. I přestože jsou délky úseků různé, můžeme pozorovat u replikátu B i D relativně stejné změny hloubky pokrytí v čase.

Tabulka 6-11: Informace o detekovaných úsecích replikátu B genu X276_27735

X276_27735, replikát B				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
B1	4691627	4691924	298	22.42
B2	4691719	4691907	189	16.62
B3	4691627	4691927	301	24.81
B4	4691433	4691910	478	22.29
B5	4691718	4691909	192	18.29
B6	4691719	4691907	189	9.29

Tabulka 6-12: Informace o detekovaných úsecích replikátu D genu X276_27735

X276_27735, replikát D				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
D1	4691665	4691919	255	15.34
D2	4691726	4691913	188	14.21
D3	4691720	4691936	217	15.55
D4	4691663	4691911	249	22.34
D5	4691664	4691911	248	9.69
D6	4691826	4691920	95	10.00

Poslední gen, který byl zvolen pro zhodnocení detekce hledaných sRNA u bakterie *Clostridium beijerinckii* NRRL B-598, je gen nesoucí název X276_27750. Tento gen slouží v bakterii pro tvorbu dvousložkového senzoru histidinkinázy a nachází se na sekundárním vlákně v pozicích od 5477808 do 5477956. Veškeré zjištěné informace

o úsecích detekovaných v časech T1-T6 pro oba replikáty B i D jsou vypsány v Tabulce 6-13 a Tabulce 6-14.

Celková délka všech úseků je zhruba kolem 330 bp, což je delší, než jaký je u hledaných sRNA požadováno. Po bližším prozkoumání lze zjistit, že úsek o délce zhruba 90 bp má mnohonásobně vyšší pokrytí ve všech časových krocích. Jedná se o úsek, který je i nadále komplementární vůči danému genu a jeho délka je optimálnější pro případnou sRNA. Díky tomu, že jsou porovnávány zhruba stejné délky, může být posouzena i hloubku pokrytí v průběhu časových kroků. Opět může být pozorováno i to, že u obou replikátů je nejvyšší hloubka pokrytí a tedy i míra exprese v dané oblasti v čase T4 a nejnižší v čase T6. To napovídá, že se jedná o úsek, jehož exprese závisí na stavu, ve kterém se bakterie zrovna nachází a tedy, že by se mohlo jednat o případnou sRNA.

Tabulka 6-13: Informace o detekovaných úsecích replikátu B genu X276_27750

X276_27750, replikát B				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
B1	5477838	5478171	334	83.06
B2	5477837	5478171	335	89.71
B3	5477829	5478171	343	126.63
B4	5477828	5478172	345	185.47
B5	5477832	5478172	344	61.54
B6	5477845	5478169	325	29.90

Tabulka 6-14: Informace o detekovaných úsecích replikátu D genu X276_27750

X276_27750, replikát D				
Replikát	**Start**	**Stop**	**Délka [bp]**	**Hloubka pokrytí**
D1	5477848	5478172	325	35.41
D2	5477845	5478173	329	55.84
D3	5477845	5478172	328	43.20
D4	5477839	5478173	335	142.14
D5	5477845	5478173	329	26.12
D6	5477845	5478171	327	15.42

V datech získaných naší detekcí se nacházeli i další zajímavé úseky, které by stáli za bližší prozkoumání, jedná se o úseky komplementární ke genům: X276_27300, X276_08755, X276_27255, X276_13430, X276_14645, X276_14670 a další. Všechny

detekované úseky jsou uloženy v souboru `vysledky_B_D.mat`. V tomto souboru jsou detekované úseky uloženy do jednotlivých struktur podle toho, jestli se jednalo o úseky replikátu B či D, primárního či sekundárního vlákna a v jakém časovém kroku se replikát nacházel. Tyto výsledky jsou také uloženy do dvou tabulek vytvořených v tabulkovém procesoru Excel. Všechny tyto soubory jsou součástí odevzdaných elektronických příloh. Soupis těchto příloh se nachází v Příloze 1.

Závěr

Cílem této bakalářské práce bylo uvedení do problematiky týkající se malých nekódujících RNA u bakterií (sRNA), konkrétně u bakterie *Clostridium beijerinckii* NRRL B-598. Byly vytyčeny základní znalosti o bakteriích a o RNA všeobecně. Došlo k seznámení s vlastnostmi, strukturou a esenciální rolí v celé genové expresi organizmu pro malá nekódující sRNA.

Také byla představena genová exprese a došlo k popisu některých z mnoha laboratorních metod pro zkoumání genové exprese, jako jsou Northern blot, SAGE, RT-qPCR, dPCR a zejména metodu RNA-Seq, neboť právě s daty získanými touto metodou se v této bakalářské práci pracuje.

Následně byl v práci navržen postup pro identifikaci sRNA u *Clostridium beijerinckii* NRRL B-598. Postup byl zaměřen pouze na detekci sRNA, které jsou alespoň částečně komplementární k určitým CDS. Navržený postup byl následně implementován v programovacím prostředí MATLAB.

Data získaná implementovaným postupem pro detekci byla subjektivně zhodnocena a bylo vybráno několik potencionálních úseků ke zhodnocení funkčnosti navržené detekce. Veškeré informace o těchto zvolených úsecích jsou obsaženy v kapitole 6. U některých takto vybraných úseků bylo možné pozorovat podstatné změny genové exprese v čase, a to pro oba zkoumané replikáty. Také tyto úseky po bližším prozkoumání odpovídali svou délkou délkám typickým pro sRNA. Proto je předpokládáno, že postup, kterým se tato práce zabývá, je do jisté míry úspěšný pro detekci sRNA. Bohužel díky neznalosti žádných sRNA u bakterie *Clostridium beijerinckii* NRRL B-598 nelze stoprocentně určit, zda se opravdu jedná o sRNA. Bylo by tedy vhodné tyto úseky do budoucna více prozkoumat.

Přestože metoda detekce našla některé potencionální úseky, již z výsledků může být nadále uvažováno o dalších změnách a úpravách postupu detekce. Ze zjištěných dat je jasné, že by bylo vhodné zakomponovat do rozhodování míru pokrytí pouze na daném úseku a nastavit hranice minimálního pokrytí po úsecích. Touto úpravou bychom získaly pouze ty úseky, které jsou v dané oblasti nejvíce zastoupené, nehledě na to, jakou míru pokrytí mají vůči celé sekvenci. Také by bylo vhodné zakomponovat i některé strukturální informace o sRNA a detekci upravit pomocí nich, nebo predikovat sekundární strukturu daného úseku a rozhodovat podle ní.

Samozřejmě by bylo také vhodné celý algoritmus detekce rozšířit na hledání i ostatních sRNA, která nejsou komplementární k CDS. K tomuto účelu bychom mohli využít právě znalosti o primární a sekundární struktuře, jak bylo uvedeno výše. Na závěr by bylo také zajímavé pozorovat nejenom míru exprese pro dané úseky, ale také míru exprese daného genu v průběhu časových kroků.

Celkově lze považovat postup detekce sRNA za úspěšný, neboť byly nalezeny úseky odpovídající typickým vlastnostem sRNA u bakterie. Také je metoda zhodnocena

jako potencionálně výhodná metoda pro zkoumání a detekci sRNA i u dalších bakterií, neboť se neodkazuje na přesné informace o bakterii *Clostridium beijerinckii* NRRL B-598, ale všechny použité informace mohou být nahrazeny informacemi o jiné bakterii.

Citace použitých zdrojů

[1] DYKHUIZEN, Daniel. Species Numbers in Bacteria. *Proc California Academy of Sciences*. 2011, vol. 56,6 Suppl 1: 62-71. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3160642/

[2] VOTAVA, Miroslav. *Lékařská mikrobiologie obecná*. 2. přeprac. vyd. Brno: Neptun, 2005. ISBN 80-86850-00-5.

[3] Fytopatologie cvičení: Fytopatogenní prokaryota. *Web2.mendelu.cz* [online]. [cit. 2020-11-19]. Dostupné z: https://web2.mendelu.cz/af_291_projekty2/vseo/print.php?page=130&typ=html

[4] SNUSTAD, D. Peter, SIMMONS, Michael J., RELICHOVÁ, Jiřina. *Genetika*. Druhé, aktualizované vydání. Brno: Masarykova univerzita, 2017. ISBN 978-80-210-8613-5.

[5] E. Coli. *Pitt.edu* [online]. [cit. 2020-11-19] Dostupné z: http://www.pitt.edu/~mcs2/ecoli.html

[6] VERMA, Subhash C., QIAN, Zhong, ADHYA, Sankar L. Architecture of the Escherichia coli nucleoid. *PLoS Genetics*. 2019, 15(12): e1008456. Dostupné z: https://doi.org/10.1371/journal.pgen.1008456

[7] BRANSKA, Barbora, PECHACOVA, Zora, KOLEK, Jan, VASYLKIVSKA, Maryna, PATAKOVA, Petra. Flow cytometry analysis of Clostridium beijerinckii NRRL B-598 populations exhibiting different phenotypes induced by changes in cultivation conditions. *Biotechnology for Biofuels*. 2018, 11(1), 1–16. Dostupné z: https://doi.org/10.1186/s13068-018-1096-x

[8] VOTAVA, Miroslav. *Lékařská mikrobiologie speciální*. Brno: Neptun, 2003. ISBN 80-902896-6-5.

[9] GOLDMAN, Emanuel, GREEN, Lorrence H. *Practical Handbook of Microbiology*. Boca Raton: CRC Press, 2015. ISBN 9780429168932.

[10] SEDLAR, Karel, KOLEK, Jan, PROVAZNIK, Ivo, PATAKOVA, Petra. Reclassification of non-type strain Clostridium pasteurianum NRRL B-598 as Clostridium beijerinckii NRRL B-598. *Journal of Biotechnology*. 2017, 244, 1–3. Dostupné z: https://doi.org/10.1016/j.jbiotec.2017.01.003

[11] Genetika zvířat: Exprese genu. *Web2.mendelu.cz* [online]. [cit. 2020-11-20]. Dostupné z: https://web2.mendelu.cz/af_291_projekty2/vseo/print.php?page=1483&typ=html

[12] MORAN, Larry. Ribosomal RNA Genes in Bacteria. *Sandwalk* [online]. 2008 [cit. 2020-11-21]. Dostupné z: https://sandwalk.blogspot.com/2008/01/ribosomal-rna-genes-in-bacteria.html

[13] Stems & Loops in rRNA. *Memorial University | Newfoundland and Labrador's University | Memorial University of Newfoundland* [online]. 2011 [cit. 2020-11-21]. Dostupné z: https://www.mun.ca/biology/scarr/rRNA_folding.html

[14] LODISH, Harvey F., BERK, Arnold, MATSUDAIRA, Paul, DARNELL, James E., KAISER, Chris A., KRIEGER, Monty, SCOTT, Mathew P., ZIPURSKY, S. Lawrence. *Molecular cell biology*. 5th ed. New York: W. H. Freeman and Company, 2003. ISBN 978-0-7167-4366-8.

[15] VERHOUNIG, Andreas, KARCHER, Daniel, BOCK, Ralph. Inducible gene expression from the plastid genome by a synthetic riboswitch. *Proceedings of the National Academy of Sciences of the United States of America*. 2010, 107(14), 6204–6209. Dostupné z: https://doi.org/10.1073/pnas.0914423107

[16] REN, Gai-Xian, GUO, Xiao-Peng, SUN, Yi-Cheng. Regulatory 3' untranslated regions of bacterial mRNAs. *Frontiers in Microbiology*. 2017, 8(JUL), 1–6. Dostupné z: https://doi.org/10.3389/fmicb.2017.01276

[17] MIGNOME, Flavio, GISSI, Carmela, LIUNI, Sabino, PESOLE, Graziano. Untranslated regions of mRNAs. *Genome Biology*. 2002, 3, reviews0004.1. Dostupné z: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-3-reviews0004

[18] TRAN, T. H., MONTANO, M. A. MicroRNAs: Mirrors of Health and Disease. In: *Translating MicroRNAs to the Clinic*. Academic Press, 2017. s. 1-15. ISBN 9780128005538.

[19] VIEGAS, Sandra C., ARRAIANO, Cecília M. Regulating the regulators: How ribonucleases dictate the rules in the control of small non-coding RNAs. *RNA Biology*. 2008, 5(4), 230–243. Dostupné z: https://doi.org/10.4161/rna.6915

[20] STORZ, Gisela, VOGEL, Jörg, WASSARMAN, Karen M. Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Molecular Cell*. 2011, 43(6), 880–891. Dostupné z: https://doi.org/10.1016/j.molcel.2011.08.022

[21] NIELSEN, Jesper S., LEI, Lisbeth K., EBERSBACH, Tine, OLSEN, Anders Steno, KLITGAARD, Janne K., VALENTIN-HANSEN, Poul, KALLIPOLITIS, Birgitte H. Defining a role for Hfq in Gram-positive bacteria: Evidence for Hfq-dependent antisense regulation in *Listeria monocytogenes*. *Nucleic Acids Research*. 2009, 38(3), 907–919. Dostupné z: https://doi.org/10.1093/nar/gkp1081

[22] A0A0K2MED1 | SWISS-MODEL Repository. *SWISS-MODEL* [online]. [cit. 2020-12-22]. Dostupné z: https://swissmodel.expasy.org/repository/uniprot/A0A0K2MED1

[23] Escherichia Coli Hfq. *Kenyon* [online]. [cit. 2020-12-22]. Dostupné z: http://biology.kenyon.edu/BMB/jsmol2016/Hfq/index.html

[24] Prokaryotic Transcription and Translation | Biology for Majors I. *Lumen Learning – Simple Book Production* [online]. [cit. 2020-12-22]. Dostupné z: https://courses.lumenlearning.com/wmopen-biology1/chapter/prokaryotic-transcription-and-translation/

[25] HE, Shan L., GREEN, Rachel. Northern blot. *Methods in enzymology*. 2013, 530, 75–87. Dostupné z: https://doi.org/10.1016/B978-0-12-420037-1.00003-8

[26] TRAYHURN, Paul. Northern blotting. *Proceedings of the Nutrition Society*. 1996, 55, 583-5891996, 583–589. Dostupné z: https://www.cambridge.org/core/journals/proceedings-of-the-nutrition-society/article/northern-blotting/5A8D7C070BF170BB7288E1BD08F798ED

[27] VELCULESCU, Victor E., ZHANG, Lin, VOGELSTEIN, Bert, KINZLER, W. Kenneth. Serial analysis of gene expression. *Nature Protocols*. 1995, 1(4). Dostupné z: https://doi.org/10.1126/science.270.5235.484

[28] YE, Shui Q., LAVOIE, Tera, USHER, David C., ZHANG, Li Q. Microarray, SAGE and their applications to cardiovascular diseases. *Cell Research*. 2002, 12(2), 105–115. Dostupné z: https://doi.org/10.1038/sj.cr.7290116

[29] Basic Principles of RT-qPCR. *Thermo Fisher Scientific*. [online]. [cit. 2020-12-22]. 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

[30] TICHOPAD, Ales, POLSTER, Jürgen, PECEN, Ladislav, PFAFFL, Michael W. Model of inhibition of *Thermus aquaticus* polymerase and *Moloney murine leukemia* virus reverse transcriptase by tea polyphenols (+)-catechin and (-)-epigallocatechin-3-gallate. *Journal of Ethnopharmacology*. 2005, 99(2):221-7. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/15894131/

[31] FIALOVÁ, Eliška, ZDEŇKOVÁ, Kamila, JABLONSKÁ, Eva, DEMNEROVÁ, Kateřina, OVESNÁ, Jaroslava. Digitální PCR: Princip a Aplikace. *Chemické Listy*. 2019, 113, 545–552. Dostupné z: http://www.chemicke-listy.cz/ojs3/index.php/chemicke-listy/article/view/3452

[32] PREDIGER, Ellen. Digital PCR (dPCR) - What is it and why use it?. *Integrated DNA technologies* [online]. 2013 [cit. 2020-12-22]. Dostupné z: https://eu.idtdna.com/pages/education/decoded/article/digital-pcr-(dpcr)-what-is-it-and-why-use-it-

[33] HELLER, Michael J. DNA microarray technology: Devices, systems, and applications. *Annual Review of Biomedical Engineering*. 2002, 4, 129–153. Dostupné z: https://doi.org/10.1146/annurev.bioeng.4.020702.153438

[34] EHRENREICH, Armin. DNA microarray technology for the microbiologist: An overview. *Applied Microbiology and Biotechnology*. 2006, 73(2), 255–273. Dostupné z: https://doi.org/10.1007/s00253-006-0584-2

[35] STARK, Rory, GRZELAK, Marta, HANDFIELD, James. RNA sequencing: the teenage years. *Nature Reviews Genetics*. 2019, 20(11), 631–656. Dostupné z: https://doi.org/10.1038/s41576-019-0150-2

[36] CONESA, Ana, MADRIGAL, Pedro, TARAZONA, Sonia, GOMEZ-CABRERO, David, CERVERA, Alejandra, MCPHERSON, Andrew, SZCZEŚNIAK, M. Wojciech, GAFFNEY, Daniel J., ELO, Laura L., ZHANG, Xuegong, MORTAZAVI, Ali. A survey of best practices for RNA-seq data analysis. *Genome Biology*. 2016, 17(1), 1–19. Dostupné z: https://doi.org/10.1186/s13059-016-0881-8

[37] SOLEXA (ILLUMINA). *LabGuide* [online]. [cit. 2020-12-22]. Dostupné z: https://labguide.cz/solexa-illumina/

[38] Illumina Inc. An introduction to Next-Generation Sequencing Technology. *Illumina sequencing introduction*. 2017. Dostupné z: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf

[39] RHOADS, Anthony, AU, Kin F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. 2015, 13(5), 278–289. Dostupné z: https://doi.org/10.1016/j.gpb.2015.08.002

[40] LU, Hengyun, GIORDANO, Francesca, NING, Zemin. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics and Bioinformatics*. 2016, Vol. 14, Issue 5. Dostupné z: https://doi.org/10.1016/j.gpb.2016.05.004

[41] The SAM/BAM Format Specification Working Group. Sequence Alignment / Map Format Specification. 2021, 1–16. Dostupné z: https://samtools.github.io/hts-specs/SAMv1.pdf

[42] The SAM/BAM Format Specification Working Group. Sequence Alignment / Map Optional Fields Specification. 2020, 1–5. Dostupné z: https://samtools.github.io/hts-specs/SAMtags.pdf

[43] SEDLAR, Karel, KOSCOVA, Pavlina, VASYLKIVSKA, Maryna, BRANSKA, Barbora, KOLEK, Jan, KUPKOVA, Kristyna, PATAKOVA, Petra, PROVAZNIK, Ivo. Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. BMC Genomics. 2018, 19(1), 1–13. Dostupné z: https://doi.org/10.1186/s12864-018-4805-8

[44] ZHAO, Shanrong, YE, Zhan, STANTON, Robert. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*. 2020, 26(8), 903–909. Dostupné z: https://doi.org/10.1261/RNA.074922.120

[45] Transposase IS3/IS911 family. *InterPro – The European Bioinformatics Institute* [online]. [cit. 2021-05-05]. Dostupné z: https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002514/

[46] Family: YvrJ (PF12841). *Pfam*. [online]. [cit. 2021-05-05]. Dostupné z: http://pfam.xfam.org/family/PF12841

[47] fliY - Flagellar motor switch phosphatase FliY. *UniProt* [online]. [cit. 2021-05-05]. Dostupné z: https://www.uniprot.org/uniprot/P24073

Seznam symbolů a zkratek

Zkratky:

RNA	…	Ribonukleová kyselina
tRNA	…	Transferová ribonukleová kyselina
mRNA	…	Mediátorová ribonukleová kyselina
rRNA	…	Ribozomální ribonukleová kyselina
sRNA	…	Malá nekódující ribonukleová kyselina
miRNA	…	Mikro ribonukleová kyselina
snRNA	…	Malá jaderná ribonukleová kyselina
DNA	…	Deoxyribonukleová kyselina
dsDNA	…	Dvouvláknová DNA
TYA	…	Tryptone yeast extract acetate agar
RCM	…	Klostridiální zesílený agar
RBS	…	Ribozomové vazebné místo
NGS	…	Sekvenování nové generace
cDNA	…	Komplementární DNA
5' UTR	…	5' nepřekládaná oblast
3' UTR	…	3' nepřekládaná oblast
ncRNA	…	Nekódující RNA
pre-mRNA	…	Primární RNA
DNáza	…	Deoxyribonukleáza
RNáza	…	Ribonukleáza
SAGE	…	Sériová analýza genové exprese
PCR	…	Polymerázová řetězová reakce
RT-qPCR	…	Reverzně transkripční kvantitativní PCR
RT	…	Reverzní transkriptáza
dPCR	…	Digitální PCR
cdPCR	…	Čipová dPCR
ddPCR	…	Kapková dPCR
TGS	…	Třetí generace sekvenování
SMRT	…	Jedna molekula v reálném čase
ZMW	…	Zero-mode waveguides
RPKM	…	Počet čtení na milion kilobází
CDS	…	Kódující sekvence

Seznam příloh

Příloha 1 – Soupis elektronických příloh

- bp_detekce_sRNA.m – skript pracující se všemi potřebnými vstupy, jehož výstupem jsou detekované úseky sRNA ve FASTA formátu. V rámci tohoto skriptu jsou volány všechny ostatní funkce.
- getindex.m – funkce, která získává potřebné informace o indexech anotovaných úseků z reference.
- getreads.m – funkce, která eliminuje nepotřebná čtení z BAM souboru.
- vysledky_B_D.mat – obsahuje uložený workspace z prostředí MATLAB. V tomto workspace jsou uloženy do proměnných získané detekované úseky z jednotlivých časových kroků pro oba replikáty B a D a veškeré informace o nich.
- B_D_primarni.xlsx – tabulka s listy, které obsahují detekované úseky z replikátů B a D na primárním vlákně v jednotlivých časových krocích T1-T6.
- B_D_sekundarni.xlsx – tabulka s listy, které obsahují detekované úseky z replikátů B a D na sekundárním vlákně v jednotlivých časových krocích T1-T6.
- readme.txt – textový soubor, popisující veškeré informace o odevzdaných elektronických přílohách.