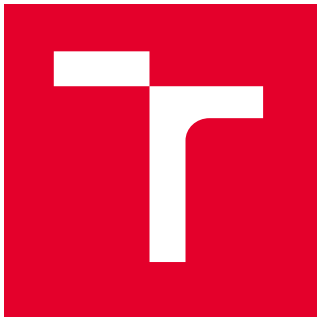
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2022 Julie Nejezchlebová



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

ODVOZENÍ OPERONOVÝCH STRUKTUR V RÁMCI CELOGENOMOVÉ ANALÝZY

OPERON STRUCTURES INFERENCE IN GENOME-WIDE ANALYSIS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — **Julie Nejezchlebová**

AUTHOR

VEDOUCÍ PRÁCE — **Ing. et Ing. Jana Schwarzerová, MSc**

SUPERVISOR

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Julie Nejezchlebová ***ID:*** 221525

Ročník: 3 ***Akademický rok:*** 2021/22

NÁZEV TÉMATU:

Odvození operonových struktur v rámci celogenomové analýzy

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s principy genové exprese vyskytující se u prokaryot. 2) Prostudujte techniky používané pro odvození transkripčních jednotek, konkrétněji se zaměřte na operony a vyberte vhodný online nástroj pro odvození operonové struktury. 3) Vytvořte si vhodný dataset obsahující predikci operonové struktury a genovou expresní informaci. 4) Práci rozšiřte o vámi navržený algoritmus implementovaný v jazyce Python, který operonovou strukturu v celém genomu upřesní o informaci získanou z genové exprese. 5) Dále se zaměřte na konkrétní vybrané operony a proveďte jejich dynamickou analýzu genové exprese. 6) Proveďte diskusi k výsledkům.

Práce bude prováděná na datech poskytnutých z laboratoře Funkční genomiky a systémové biologie, UBMI ve spolupráci VŠCHT v Praze a z veřejně dostupných databází.

DOPORUČENÁ LITERATURA:

[1] XIONG, Jin. Essential bioinformatics. Cambridge University Press, 2006.

[2] GS, STENT. THE OPERON: ON ITS THIRD ANNIVERSARY. MODULATION OF TRANSFER RNA SPECIES CAN PROVIDE A WORKABLE MODEL OF AN OPERATOR-LESS OPERON. Science (New York, NY), 1964, 144.3620: 816-820.

Termín zadání: 7.2.2022 ***Termín odevzdání:*** 27.5.2022

Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc

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

ABSTRAKT

Bakalářská práce se věnuje problematice odvození operonových struktur a vytvoření softwarového nástroje, který umožní predikci operonových struktur. Nástroj jednak predikuje operony na základě genové expresní informace, ale také upřesní již predikované operony o genovou expresní informaci. Nástroj je testován na bakteriích *Escherichia coli* BW25113 a *Clostridium beijerinckii* NRRL B-598. Teoretická část je věnována popisu struktury a funkce operonu, sekvenování genomu, analýze transkriptomu, bakteriím *Escherichii coli* BW25113, *Clostridium beijerinckii* NRRL B-598 a již dostupným online nástrojům pro odvození operonových struktur. V praktické části se práce zabývá předzpracováním surových transkriptomických dat, za účelem získání vhodného formátu pro predikci operonových struktur, testováním online nástrojů a samotné implementaci vlastního nástroje.

KLÍČOVÁ SLOVA

Operon, Genová exprese, Genom, Transkriptom, *Escherichia coli* BW25113, *Clostridium beijerinckii* NRRL B-598

ABSTRACT

The bachelor thesis is devoted to the problem of derivation of operon structures and creation of a software tool that allows prediction of operon structures. The tool both predicts operons based on gene expression information, but also refines already predicted operons with gene expression information. The tool is tested on the bacteria *Escherichia coli* BW25113 and *Clostridium beijerinckii* NRRL B-598. The theoretical part is devoted to description of operon structure and function, genome sequencing, transcriptome analysis, *Escherichia coli* BW25113, *Clostridium beijerinckii* NRRL B-598 and already available online tools for inferring operon structures. In the practical part of the thesis, the pre-processing of raw transcriptomic data to obtain a suitable format for the prediction of operon structures, testing of online tools and the actual implementation of the tool itself are discussed.

KEYWORDS

Operon, Gene expression, Genome, Transcriptome, *Escherichia coli* BW25113, *Clostridium beijerinckii* NRRL B-598

NEJEZCHLEBOVÁ, Julie. *Odvození operonových struktur v rámci celogenomové analýzy*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 55 s. Bakalářská práce. Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Julie Nejezchlebová
VUT ID autora:	221525
Typ práce:	Bakalářská práce
Akademický rok:	2021/22
Téma závěrečné práce:	Odvození operonových struktur v rámci celogenomové analýzy

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

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

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí bakalářské práce paní Ing. et Ing. Janě Schwarzerové, MSc za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

U prokaryot se na genové regulaci významně podílejí operony. Operony zajišťují efektivní hospodaření buňky a umožňují organismům reagovat na změny okolního prostředí. V současné době je většina algoritmů k odvození operonových struktur založena na algoritmech strojového učení, které fungují na základě porovnávání informací z veřejných databází, obvykle v rámci srovnání s modelovými organismy. Toto srovnání není ovšem dostatečné pro odovzení naprosté většiny nemodelových organismů. Proto se bakalářská práce zabývá vytvořením algoritmu, který predikuje operony na základě genové expresní informace pomocí korelačního koeficientu, který umožňuje přesně určit, proč jsou geny přiřazeny k danému operonu.

Bakalářská práce se v první kapitole teoreticky věnuje operonům a jejich biologickému významu. Je zde popsána problematika genové regulace prokaryot, transkriptom a jeho analýza.

Ve druhé kapitole jsou popsány sekvenační techniky Illumina, SOLiD, Ion Torent a RNA-Seq. Tyto techniky umožňují sekvenování transkriptomu a tím pádem umožňují získat informaci o genové regulaci. Tato kapitola se dále věnuje popisu vybraných bakterií, které byly použity v praktické části. Je zde popsána bakterie *Escherichia coli* BW25113. Tato bakterie byla zvolena jako zástupce reprezentující modelový organismus. Druhou zkoumanou bakterií je *Clostridium beijerinckii* NRRL B-598, která je sice nemodelovým organismem, avšak je významným butanolovým producentem, proto je vhodným organismem k bližšímu zkoumání.

Praktická část se v první řadě věnuje předzpracování surových transkriptomických dat ve formě RNA-Seq do vhodné formy pro následné odvození operonových struktur. Dále se práce zabývá testováním dostupných online nástrojů, jmenovitě, Operon-mapperem, FGENESBem a ProOpDBem. V neposlední řadě práce přináší vlastní implementaci funkce OperonIdentifier, která byla naprogramována v programovacím jazyce Python. Tato funkce umožňuje uživateli buď upřesnit již predikované operony o informaci získanou z genové exprese, nebo predikovat operony pouze na základě genové expresní informace. Závěr praktické části je věnován dynamické analýze vybraných operonů. U bakterie *Escherichia coli* BW25113 je analyzován *lac* operon, u bakterie *Clostridium beijerinckii* NRRL B-598 *sol* operon. Tyto operony jsou vhodné k ověření správnosti a bližší dynamické analýze, jelikož byly u těchto bakterií laboratorně prokázány.

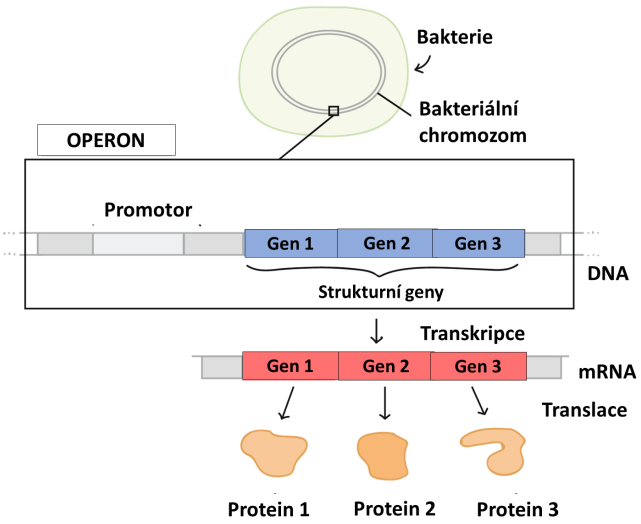
1 Operon

Operony hrají významnou roli v celkové genové regulaci bakterií. Operony jsou skupiny příbuzných genů řízené jedním operátorem [1]. Umožňují efektivně exprimovat sady genů a zároveň zajišťují, aby nebyly zbytečně vytvářeny proteiny, které buňka v daný okamžik nepotřebuje. [2]

1.1 Operon jako transkripční jednotka

Operon obsahuje jeden nebo více strukturních genů, které jsou transkribovány do jedné polycistromatické mediátorové RNA (mRNA). Jedna molekula mRNA kóduje tedy více než jeden protein. Součástí transkripční jednotky (TU) operonu je také promotor a operátor (viz 1.1). Promotor je nukleotidová sekvence, která umožňuje přepis genu, protože obsahuje vazebná místa pro RNA polymerázu. Pokud je promotor rozpoznán RNA polymerázou, dochází k navázání RNA polymerázy na vazebná místa a k iniciaci transkripce. Při syntéze RNA promotory indikují, které geny by měly být použity pro tvorbu mRNA. Tím pádem kontrolují, které proteiny buňka produkuje. Operátor je segment DNA, na který se váže regulátor. Často se nachází mezi promotorem a strukturními geny operonu. [2]

Operon může dále obsahovat regulační geny jako například represor. Represor se váže na operátor a inhibuje transkripci. Regulační geny nemusí být nutně součástí operonu, můžou být umístěny i jinde v genomu. [2]

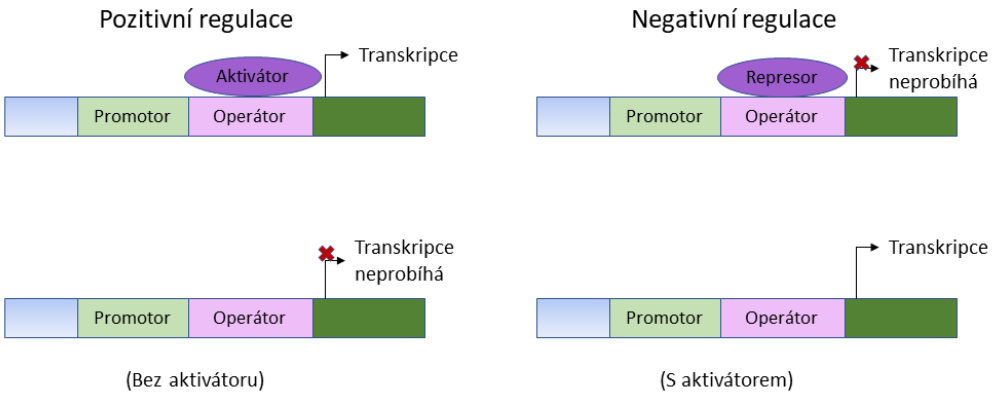


Obr. 1.1: Struktura a genová exprese operonové TU. Převzato z [2].

1.2 Význam operonu v genové regulaci

Genová regulace umožňuje organismům regulovat expresi různých genů v závislosti na okolních podmínkách [3]. Důležitou roli při expresi hraje transkripce, která se promítá do všech stupňů genové exprese [4].

Na molekulární úrovni se na regulaci podílejí molekulární regulátory. Na úrovni transkripce rozlišujeme regulátory pozitivní a negativní. Pozitivní regulátory transkripci iniciují, negativní ji zastavují. [4] V rámci genové regulace rozlišujeme indukovatelné a represivní operony. Indukovatelné operony jsou za normálních okolností deaktivovány. Aktivují se pouze pokud je přítomen konkrétní induktor. Naopak represivní operony jsou aktivovány, nicméně mohou být deaktivovány korepresorem. Regulace operonu pak může být buď pozitivní, či negativní indukcí, nebo represí. Pozitivní a negativní indukce i pozitivní a negativní represe jsou znázorněny na obrázku 1.2. [2]



Obr. 1.2: Pozitivní a negativní regulace operonu. Modifikováno z [5].

1.2.1 Negativní regulace

Negativní kontrola spočívá v navázání represoru na operátor tak, aby se zabránilo transkripci [3]. V negativních indukovatelných operonech je na operátor navázán regulační represorový protein, který brání transkripci genů na operonu. Pokud je přítomen induktor, který se naváže na represor, dochází ke změně konformace represoru tak, že se nemůže vázat na operátor. To umožňuje expresi genu nacházející se v jednom shluku operonu. [3]

U negativních represivních operonů probíhá za normálních okolností transkripce. Represorové proteiny jsou produkovány regulačním genem, ale nejsou schopny se vázat na operátor ve své normální konformaci. Pokud se na represorový protein naváže

kopresor, dochází ke změně konformace represoru. Aktivovaný represorový protein se váže na operátor a zabraňuje transkripci. [3]

1.2.2 Pozitivní regulace

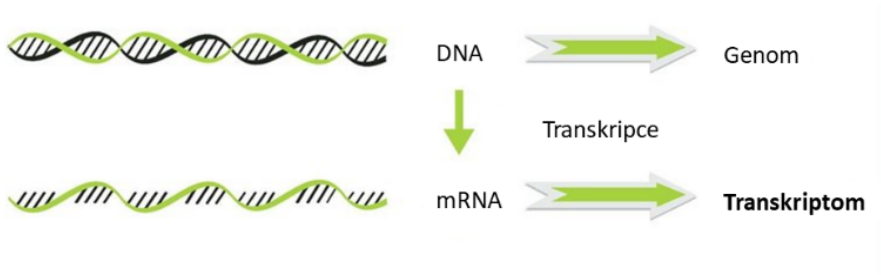
U pozitivní genové regulace je transkripce stimulována aktivátorovým proteinem, který se váže na DNA. Obvykle se neváže na promotor, ale může se navázat i jinde. [3]

V pozitivně indukovatelných operonech se aktivační proteiny nemohou vázat na příslušný segment DNA. Když dojde k navázání induktoru na aktivační protein, dochází ke změně konformace tak, že se protein může navázat na DNA a aktivovat transkripci. [3]

U pozitivních represivních operonů jsou aktivační proteiny normálně vázány na příslušný segment DNA. Pokud se inhibitor naváže na aktivátor, dochází k zastavení aktivace a transkripce systému. [3]

1.3 Transkriptom

Během genové exprese dochází k přepisu DNA do mRNA [6]. Výstupem transkripce jednoho genu je transkript. Soubour všech transkriptů se nazývá transkriptom , viz obrázek 1.3 [7].



Obr. 1.3: Schéma transkripce se zvýrazněným transkriptomem. Převzato z [8].

Z počtu transkriptů je možné stanovit míru genové exprese v určitém typu buňky nebo tkáně. Téměř každá buňka obsahuje stejné geny, ale různé buňky vykazují různé vzorce genové exprese. Shromažďování a pozorování transkriptomů různých typů buněk umožňuje pochopit, co tvoří konkrétní typ buňky, jak tento typ funguje za standartních podmínek a jak za jiných okolností. Transkriptomy navíc umožňují vytvořit komplexní genomový obraz o tom, jaké geny jsou aktivní za určitých podmínek v daných buňkách. [7]

1.3.1 Analýza transkriptomu

Genetický kód obsažený v transkriptomech nebo jeho relativní proporce můžeme určit pomocí analýzy transkriptomu [6]. Analýza transkriptomu je expresní analýza celé sady molekul RNA, které jsou produkovány buňkou za určitých fyziologických podmínek. Studiem transkriptomu se zabývá funkční genomika. Funkční genomika nám umožňuje pochopit biologické funkce celého genomu prostřednictvím automatizované exprese. Analýza transkriptomu odhaduje vzorce koexprimovaných a koregulovaných genů a umožňuje určit funkce genů, které zatím nebyly charakterizovány. [9]

Před samotnou analýzou transkriptomu je třeba extrahovat molekuly mRNA z buněk a pomocí reverzní transkripce vytvořit komplementární DNA (cDNA). Posledním krokem před samotnou analýzou je sekvenování molekul cDNA. Analýza transkriptomu se pak provádí například pomocí sériové analýzy genové exprese (SAGE) nebo pomocí analýzy založené na mikročipech. [6]

SAGE

SAGE se používá k vytvoření knihovny krátkých sekvenačních značek. Každá značka slouží k jednozačné identifikaci transkriptu. Počet detekcí každé značky koreluje s úrovní genové exprese příslušného transkriptu. [10]

Ze sekvencí cDNA jsou vyříznuty krátké fragmenty DNA, které se používají jako unikátní markery genových transkriptů. Právě tyto sekvenační fragmenty nazýváme jako značky. Značky jsou spojeny dohromady, klonovány a sekvenovány. [9]

Analýza transkriptů se provádí sériovým způsobem. Jakmile jsou značky jednoznačně identifikovány, jejich absolutní četnost udává úroveň genové exprese. SAGE umožňuje sekvenování více značek v jednom klonu, takže může odhalit i slabě exprimované geny. [9]

Nejdražším a časově nejnáročnějším krokem metody SAGE je sekvenování. Je obtížné zjistit, kolik značek je potřeba sekvenovat, aby bylo dosaženo dobrého pokrytí celého transkriptomu. Pro většinu laboratoří jsou náklady spojené s touto metodou neúnosné. Další nevýhodou této metody je chybovost. Jedna nebo dvě sekvenační chyby ve značce můžou vést k nejedoznačné nebo chybné identifikaci značky. Dalším problém je, že správně sekvenovaná značka může někdy odpovídat několika genům nebo naopak nemusí odpovídat žádnému genu. Pro zlepšení citlivosti a specifičnosti je potřeba prodloužit délku značek. Mezi komplexní softwarové nástroje pro analýzu SAGE patří například SAGEmap, SAGExProfiler nebo SAGE Genie. [9]

Analýza transkriptomu založená na mikročipech

Nejpoužívanější metodou, která se používá k profilování genové exprese, je metoda založená na DNA mikročipech. Mikročipem je podložní sklíčko s oligomery DNA, případně cDNA, které reprezentují celý genom studovaného druhu. Každý oligomer slouží jako sonda, na kterou se váže unikátní cDNA. Celá populace cDNA, označená fluorescenčními barvivy nebo radioizotopy, se nechá hybridizovat sondami na mikročipu. Množství fluorescenčních nebo radioaktivních značek odráží množství odpovídající mRNA na daném místě v buňce. Pomocí této analýzy je možné zkoumat vzorce globální genové exprese v buňce. Teoreticky lze identifikovat skupiny genů, které jsou zapojeny do stejných regulačních nebo metabolických drah. [9]

Analýza transkriptomu pomocí mikročipů zahrnuje několikastupňový postup. V první fázi je vyroben mikročip fixací správně navržených oligomerů, které reprezentují specifické geny. V druhé fázi dochází k hybridizaci cDNA na mikročipu. Ve třetí fázi je snímán a analyzován obraz z mikročipu. Ve čtvrté fázi dochází k transformaci a normalizaci dat. V poslední fázi je prováděna analýza dat za účelem identifikace exprimovaných genů i souboru genů, které jsou koregulovány. [9]

Mikročipy na rozdíl od metody SAGE udávají relativní četnost úrovně exprese mRNA. Tato metoda je také vhodnější ke zkoumání rozdílné exprese genů mezi různými tkáněmi a buňkami. Dalším cílem analýzy pomocí mikročipů je identifikace koordinovaných vzorců genové exprese, což vyžaduje shlukovou analýzu dat. Mezi nejoblíbenější metody shlukování patří hierarchické metody, samoorganizující se mapy (SOM) a k-means. [9]

Značnou nevýhodou této metody je vícestupňový postup, při kterém může docházet k chybám a zkreslením v každém kroku. Může tedy obsahovat velké množství falešně pozitivních a falešně negativních výsledků. Výsledky analýzy mikročipů poskytují pouze hypotézy o funkcích genů na základě klasifikace expresních dat. [9]

2 Laboratorní data

V úvodní části této kapitoly je stručný přehled sekvenačních procesů a technik, které předchází bioinformatickým počítačovým analýzám, mezi které patří právě i predikce a detekce operonových struktur. Konkrétně jsou zde stručně popsány technologie Illumina, SOLiD, Ion Torrent a RNA–Seq.

Druhá část kapitoly je věnována bakterii *Escherichii coli* BW25113. V kapitole je popsána jednak samotná bakterie, ale i transkriptomická data, která jsou použita v praktické části. Poslední kapitola se zabývá bakterií *Clostridium beijerinckii* NRRL B-589 a jejích replikáty reprezentující genovou expresi.

2.1 Sekvenační techniky

Sekvenační techniky se primárně používají pro sekvenování genomické informace. Využívají se k určení přesné sekvence bází v molekule DNA [11]. Informace o sekvenci DNA jsou důležité pro zkoumání funkcí genů. S rozvojem informačních technologií se sekvenování stává stále dostupnějším. Sekvenování celého genomu je však stále složitý proces. Aby bylo možné osekvenovat celý genom, je potřeba rozdělit genom na kratší úseky DNA. Následně každý kousek jednotlivě osekvenovat a všechny kousky poskládat do jednoho konsensusu. [12]

První komerčně používanou technologií byla Sangerova metoda vyvinutá v roce 1977 Fredem Sangerem a jeho kolegy[13]. V Human Genome Project bylo Sangerovo sekvenování použito k určení sekvencí mnoha relativně malých fragmentů lidské DNA. Fragmenty byly zarovnány na základě překrývajících se částí a shromážděny do sekvence větších oblastí DNA. I dnes je Sangerovo sekvenování stále využíváno [12]. Především se používá pro sekvenování jednotlivých částí DNA, jako jsou fragmenty používané při klonování DNA nebo generované pomocí polymerázové řetězcové reakce (PCR). [12] Avšak v současné době se nejvíce používají technologie sekvenování nové generace „Next Generation“. Umožňují sekvenování mnoha fragmentů DNA najednou, jsou nákladově efektivnější a rychlejší než systémy první generace. Mezi systémy „Next generation“ patří například Illumina, SOLiD nebo Ion Torrent. [14]

2.1.1 Illumina

Technologie Illumina pracuje na principu sekvenování syntézou. Pomocí této metody je produkována drtivá většina světových sekvenačních dat. Illumina podporuje

masivní paralerní sekvenování pomocí patentované metody, kterou detekuje jednotlivé báze, jakmile jsou připojeny do rostoucích řetězců DNA.[15]

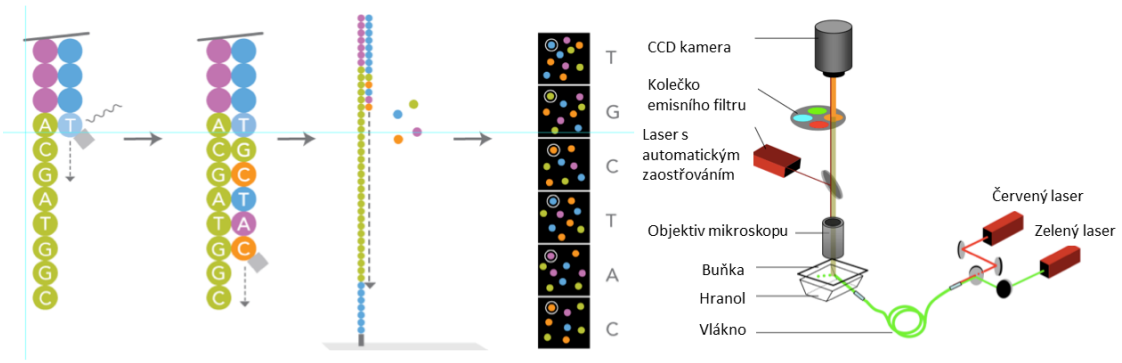
Při sekvenování je DNA nejprve fragmentována na úseky o velikosti 200 - 600 bp. Krátké úseky DNA, neboli adaptéry, jsou připojeny k fragmentům a přichyceny na malou destičku. Každá molekula DNA se opakovaně namnoží, dokud nevznikne mozaika milionů klastrů. Do rostoucího řetězce jsou zařazeny báze s navázanou fluorescenční barvou, které syntézu zastaví. Blokace této syntézy je vratná. Nově přidaná báze je přečtená vysoce citlivou kamerou. Dochází k enzymatickému odstranění fluorescenčních značení a blokující části molekuly. Následně se může přidat další báze a může proběhnout další kolo reakce. Kamera snímá signál z destičky a podle rozdílné fluorescence pozná, která báze byla přidána u každé z milionů skupin. Počítač analyzuje záznam a podle toho, jak se mění fluorescenční signál v rámci každé skupiny, zrekonstruuje přesnou sekvenci molekul DNA v příslušné skupině. Citlivost čtení dosahuje až 99,9 %. [13]

Sekvenační technologie Illumina má několik odlišných přístupů. Jelikož jsou v praktické části použita RNA–Seq data, která byla sekvenována systémem NextSeq, je níže uvedena podkapitola s podrobnějším popisem této technologie.

NextSeq

Sekvenační systémy NextSeq představují rychlé, výkonné stolní sekvenátory umožňující sekvenování exonů, celých genomů a transriptomů [16]. Sekvenování syntézou umožňuje přesné určení sekvenace milionů unikátních fragmentů DNA najednou [17].

Během každého sekvenačního cyklu jsou reverzibilně značené nukleotidy promývány po povrchu oligo klastrů, díky čemuž je jeden nukleotid navázán na rostoucí komplement každého jednotlivého oligo vlákna. Označené nukleotidy jsou zobrazeny a barvivo je následně chemicky odštěpeno, aby se mohl navázat další nukleotid v sekvenci. V každém cyklu se k zachycení bází používá více kombinací barviv. Přístroje NextSeq pořizují v každém cyklu dva snímky, jeden se zeleným filtrem, druhý s červeným filtrem (viz obrázek 2.1). [17]



Obr. 2.1: Obrázek shrnující pochody při sekvenování technologií NextSeq. Převzato z [18].

2.1.2 SOLiD

Technologie SOLiD je enzymatická metoda sekvenování, která používá enzym ligázu. Ligáza umí jednoduše připojit jednořetězcovou molekulu DNA ke stávajícímu řetězci DNA. [13]

Při sekvenování se k templátovému vláknu přidají kousky DNA sondy. Každá sonda nese jednu ze čtyř fluorescenčních značení a začíná všemi možnými dvojkombinacemi 4 základních bází, takže celkem je 16 různých typů sond. K novému rostoucímu řetězci je v každém kroku připojena pomocí ligázy sonda, která odpovídá templátové DNA. Snímačem přečte fluorescenční značení sondy, která je následně odstraněna a umožní připojení další sondy. Templátová molekula je čtena opakovaně, aby každá báze byla přečtena několikrát a došlo tak k přečtení kompletní sekvence. Výstupem SOLiD jsou krátké sekvence o velikosti maximálně 100 bp. Nevýhodou této metody je problematické čtení palindromatických sekvencí. [13]

2.1.3 Ion Torrent

Ion Torrent je technologie na polovodičovém čipu umožňující přímý převod chemicky kódovaných bází na digitální signál. Technologie Ion Torrent je jednodušší, rychlejší a nákladově efektivnější než jakákoli jiná dostupná sekvenační technologie. [19]

Po připojení nové báze do řetězce DNA je jako vedlejší produkt uvolňován vodík. Vodík způsobuje změnu pH roztoku. Podle intenzity změny pH je možné poznat, kolik bází bylo připojeno. Problém nastává při čtení delších homopolymerních řetězců. Polovodičový čip není schopen určit přesný počet nukleotidů. [13]

2.1.4 RNA–Seq

RNA–Seq je sekvenační metoda, která využívá technologií „Next Generation“. Tato vysoce účinná metoda se používá ke studiu transkriptomu a genové exprese [20]. Data použitá v praktické části semestrální práce byla sekvenována pomocí této metody.

RNA–Seq má několik kroků. V prvním kroku je extrahována RNA. Pomocí reverzní transkripce je RNA přepsána do cDNA. cDNA se fragmentuje a na oba konce fragmentů jsou připojeny adaptéry obsahující funkční prvky umožňující sekvenování. Po amplifikaci, selekci velikosti, čištění a kontrole kvality je pomocí sekvenování „Next Generation“ analyzována knihovna cDNA. Výstupem jsou krátké sekvence, které odpovídají celému fragmentu nebo jeho části, z níž byly odvozeny. [20]

Sekvenování může probíhat buď metodou sekvenování s jedním koncem nebo s párovým koncem. Dále se volí mezi protokolem specifickým pro řetězec a protokolem nespecifickým pro řetězec. V případě protokolu specifického pro řetězec je zachována informace o tom, které vlákno DNA bylo přepsáno. Na závěr lze všechny čtení zarovnat s referenčním genomem. [20]

2.2 *Escherichia coli* BW25113

Escherichia coli je gram negativní tyčinkovitá bakterie, která se vyskytuje ve střevech teplokrevných organismů. Většina kmenů *E. coli* je neškodná, avšak některé kmeny mohou u lidí nebo některých zvířat způsobit vážnou otravu. Neškodné kmeny jsou součástí střevní mikroflóry a produkují vitamín K2, který zabraňuje usazování bakterií ve střevech. *E. coli* je modelovým organismem, proto je vhodným adeptem pro návrh algoritmu pro predikci operonových struktur. [21], [22]

V práci je konkrétně použita bakterie *E. coli* BW25113. Tento laboratorní kmen byl vytvořen v laboratoři Barryho L. Wannera a byl použit v metodě využívající bakteriofágový rekombinační systém lambda red k provádění genových disrupcí pomocí PCR [23]. *E. coli* BW25113 a její deriváty se používají pro různé studie, například se používá ve fenotypových výzkumech [23], [24], [25].

Genom bakterie *E. coli* BW25113 s acession number CP009273.1 byl stažen z NCBI. Z databáze Gene Expression Omnibus (GEO) [26] byla stažena CountTable *E. coli* BW25113, která obsahuje genovou expresní informaci celkem 18 replikátů. Tyto replikáty byly použity a jsou podrobněji popsány ve studii od Roperse a kol. [27]. Všechny replikáty této bakterie byly sekvenovány sekvenační metodou RNA-Seq pomocí přístroje Ion Torrent S5. Před sekvenací byly bakterie uchovány při teplotě $80\,°C$ v roztoku, který vznikl smícháním 500 μl kultivačního bujónu

a 500 μl fenolu nebo ethanolu. Růst bakterií probíhal v bioreaktoru o objemu 2 l při teplotě 37 $°C$ a pH 6,8. Jako médium bylo použito médium M9 a glukóza. Celkem byly sekvenovány 2 genotypy - genotyp W-gly (wild-type) a R-gly (growth - regulated). Jednotlivé replikáty jsou blíže charakterizovány v tabulce 2.1.

Tab. 2.1: Popis dat, která byla použita k vytvoření CountTable pro *E. coli*. Od 600 je optická hustota vzorku při vlnové délce 600 *nm*.

Název	**Genotyp**	**Replikát**	**Čas [*h*]**	**Od 600**
W-gly-1	W-gly	1	3,5	5,22
W-gly-1	W-gly	2	3,5	6,08
W-gly-1	W-gly	3	3,7	6,4
W-gly-2	W-gly	1	5,5	15,4
W-gly-2	W-gly	2	5,5	17,7
W-gly-2	W-gly	3	5,7	18,4
R-gly-1	R-gly	1	4,1	5,36
R-gly-1	R-gly	2	4,1	5,38
R-gly-1	R-gly	3	4,2	5,94
R-gly-2	R-gly	1	5,5	7,6
R-gly-2	R-gly	2	5,5	7,58
R-gly-2	R-gly	3	5,7	7,87
R-gly-3	R-gly	1	7,5	9,62
R-gly-3	R-gly	2	7,5	9,44
R-gly-3	R-gly	3	7,7	9,78
R-gly-4	R-gly	1	23	11,36
R-gly-4	R-gly	2	23	10,92
R-gly-4	R-gly	3	23,1	12,6

2.3 *Clostridium beijerinckii* NRRL B–598

Clostridium beijerinckii patří do rodu *Clostridium*. Bakterie z tohoto rodu jsou obecně grampozitivní anaerobní organismy, z nichž některé druhy mohou být patogenní. Avšak většina druhů je nepatogenních a mohou být průmyslově zpracovávány. Některé druhy jsou schopné fixovat dusík a přeměňovat sacharidy na jiné sloučeniny a molekuly, jako je oxid uhličitý, acetát nebo butyrát. [28]

Clostridium beijerinckii je solventogenní bakterie, která se používá pro svou schopnost produkovat buthanol při fermentaci aceton–buthanol–ethanolu (ABE). V současné době může bakteriální výroba buthanolu konkurovat jeho syntetické výrobě. [29]

Právě kvůli potenciálu v produkci biobuthanolu je bakterie *C. beijerinckii* NRRL B-598 vhodným adeptem k bližšímu zkoumání. Proto je algoritmus pro predikci operonových struktur aplikován i na tuto bakterii. Acession number použité bakterie je CP011966.3. V práci jsou použity replikáty B, C, D, E, F a G, které byly sekvenovány v šesti časových bodech {T1, T2, T3, T4, T5, T6} sekvenační metodou RNA–Seq [30].

Replikáty B, C, D a E byly sekvenovány za standartních kultivačních podmínek. To znamená, že kmen *C. beijerinckii* NRRL B-598 byl kultivován při teplotě 37 °C a byl udržován ve formě suspenze spor. Replikáty B, C jsou podrobněji popsány ve studii od Sedláře a kol. [31], replikáty D, E jsou popsány ve studii Patákové a kol. [32]. Všechny tyto replikáty byly sekvenovány v šesti časových bodech tak, aby byly zachyceny všechny metabolické fáze během 23 hodin. Konkrétně byly sekvenovány v časech {3,5 *h*, 6 *h*, 8,5 *h*, 13 *h*, 18 *h*, 23 *h*}.

Replikáty F, G byly sekvenovány při butanolovém šoku. Butanol byl přidán v čase 6 *h*. Tyto replikáty byly sekvenovány v časových bodech {6 *h*, 6,5 *h*, 7 *h*, 8 *h*, 10 *h*, 12 *h*}. Oba replikáty jsou popsány ve studii od Sedláře a kol.[33].

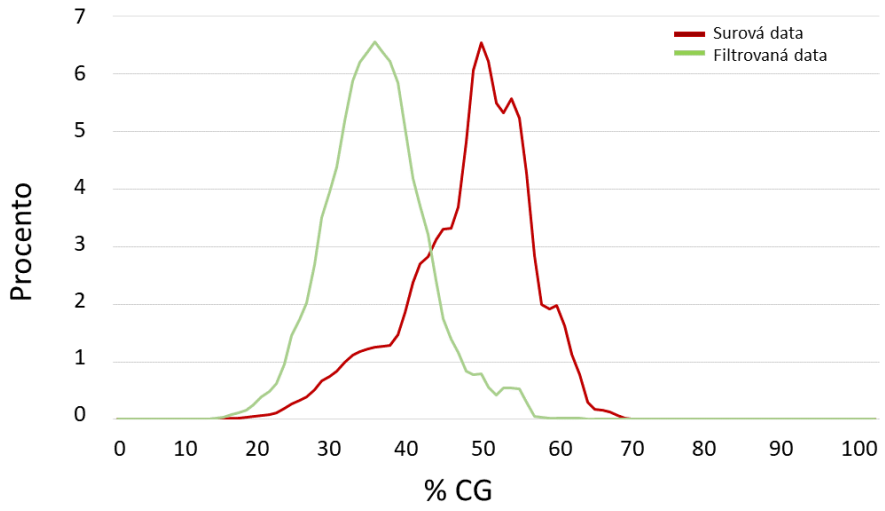
3 Metody odvození operonových struktur

Úvodní část kapitoly se věnuje předzpracování RNA-Seq dat bakterie *C. beijerinckii* NRRL B-598. Druhá část kapitoly je věnována popisu a testování dostupných online nástrojů, které se používají k predikci operonových struktur. Konkrétně jsou zde popsány nejčastěji citované online nástroje pro predikci operonů - Operon–mapper [34] , balíček FGENESB [35] a Prokaryotická operonová databáze (ProOpDB) [36] . Ve třetí části kapitoly je popsána vlastní naprogramovaná funkce OperonIdentifier. V závěru kapitoly jsou jednotlivé nástroje srovnány.

3.1 Předzpracování RNA–Seq dat

K vytvoření vhodného datasetu pro predikci operonových struktur na základě informace o genové expresi jsou použita RNA–Seq data bakterie *Clostridium beijerinckii* NRRL B–598. Předzpracování je provedeno pomocí shell-skriptů dostupných z `https://github.com/JanaSchwarzerova/Analytical-pipeline-rawRNA-Seq` [37]. Samotné předzpracování pomocí těchto shell-skriptů bylo provedeno na výpočetních zdrojích dostupných na portálu Metacentrum VO - virtuální organizace [38]. Před zahájením předzpracování RNA–Seq dat je pomocí nástrojů FASTQC [39] a MultiQC [40] zkontrolována kvalita počátečních dat. MultiQC je modulární nástroj, který slouží k agregaci výsledků bioinformatických analýz [40]. Výstupem této části je HTML dokument neboli report. FASTQC se používá ke kontrole celkové kvality sekvence. Kontroluje se například procentuální rozložení CG a přítomnost nebo absence nadměrně zastoupených sekvencí [37].

Pro maximalizaci kvality sekvenačních dat je třeba odstranit co nejvíce ribozomální RNA (rRNA). K filtraci rRNA slouží nástroj SortMeRNA [41]. Tento nástroj umožňuje analýzu čtení systémů „Next Generation“. Hlavní aplikací SortMeRNA je filtrování rRNA z metatranskriptomických dat. Vstupem je čtení ve *fasta* nebo *fastq* formátu a jeden nebo více souborů rRNA. [41] Po filtraci rRNA dat je následně provedena kontrola kvality filtrovaných dat. Na obrázku 3.1 je znázorněna změna struktury surových a filtrovaných dat. Rozložení GC u filtrovaných dat se mnohem více blíží teoretickému normálnímu rozložení GC.



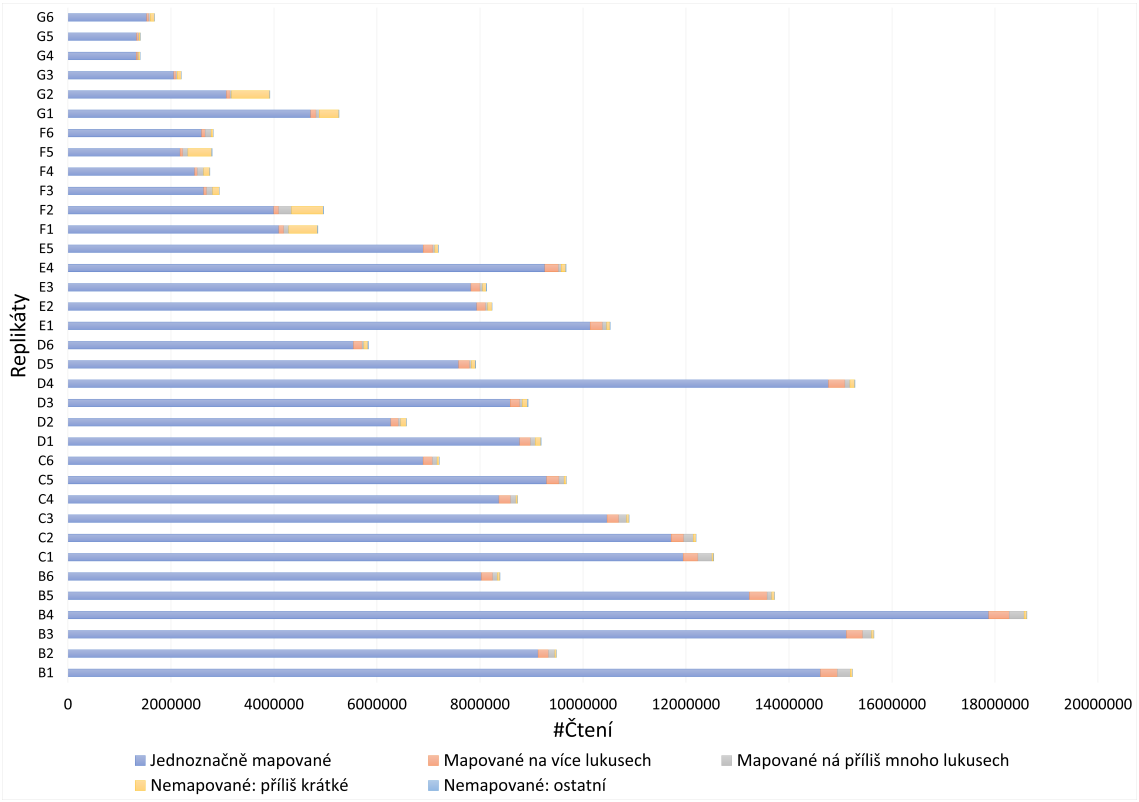
Obr. 3.1: Výskyt GC na sekvenci.

V dalším kroku je použit nástroj Trimmomatic [42]. Tento nástroj slouží k ořezávání dat získaných metodou Illumina. [42] Data jsou ořezávána od 3' konce dokud není dosaženo požadovaného Phred skóre, které udává míru spolehlivosti čtení. Za spolehlivé Phred skóre je považováno skóre nad 20 [43].

Následně je nutné zarovnat čtení k referenci. K tomu je použit nástroj STAR [44]. Před použitím tohoto nástroje je nutné vytvořit indexový genom [37]. V tomto případě byl stažen z Národního Centra pro Bioinformatické Informační databáze (NCBI) genom s acession number CP011966.3 ve formátu *.gff3*.

Pro každé čtení, které STAR zarovná, je nalezena nejdelší sekvence, která přesně odpovídá jednomu nebo více místům v referenčním genomu. Nejdelší shodné sekvence se nazývají jako maximální mapovatelné předpony (MMP). V nezmapované části čtení je následně vyhledána další nejdelší sekvence, která přesně odpovídá referenčnímu genomu. Tento postup se opakuje, dokud nejsou dohledány všechny MMP. Jednotlivé MMP jsou následně shlukovány do kompletního čtení. [44]

Na obrázku 3.2 jsou znázorněny výsledky mapování pomocí nástroje STAR [44]. Průměrná hodnota jednoznačně namapovaných čtení ze všech vzorků je 8,3 milionu. Maximum jednoznačně namapovaných vzorků čtení je u replikátu B v pátém časovém bodě. Minimum jednoznačně namapovaných čtení se nachází u replikátů F a G v pátém časovém bodě.



Obr. 3.2: Výsledky mapování pomocí nástroje STAR.

Pomocí nástroje Samtools [45] jsou seřazeny ořezané sekvence a uloženy ve formátu BAM. Následně je vytvořena výsledná tabulka počtů, neboli tzv. CountTable, pomocí funkce featureCounts [46]. Tato funkce je také zakomponována v shell-skriptech. Tabulka 3.1 je ukázkou výsledné tabulky počtů.

Tab. 3.1: Ukázka výsledné tabulky počtů.

Locus tag	**B1**	**B2**	**B3**	**B4**	**...**	**G5**	**G6**
X276_26820	3772	2278	2923	2730	...	386	435
X276_26815	2644	1570	1901	1973	...	174	236
X276_26810	599	423	453	516	...	35	62
X276_26805	5670	2403	3088	3151	...	398	482
X276_26800	1977	950	1153	1215	...	150	170
X276_26795	13022	7239	9263	7921	...	1088	1275

3.2 Online nástroje

V této kapitole jsou popsány vybrané online nástroje sloužící k predikci operonových struktur. Konkrétně jsou zde rozebrány nástroje Operon-mapper a FGENESB, které

jsou zároveň i prakticky testovány. Poslední část této kapitoly je věnována nástroji ProOpDB.

3.2.1 Operon–mapper

Webový server Operon-mapper slouží k predikci operonů v libovoných sekvencích bakteriálního nebo archeálního genomu. Předpovědi operonů jsou založeny na vzdálenosti sousedních genů a na funkčních vztazích jejich produktů, které kódují proteiny v dané nukleotidové sekvenci. Operon–mapper vyhledává všechny otevřené čtecí rámce (ORF) spolu s jejich genomickými souřadnicemi, ortologickými skupinami a funkčními vztahy. [34]

Celý proces Operon–mapperu lze rozdělit do tří fází: získávání dat, sekvenační analýza a doručení výsledků. V první fázi uživatel nahraje na webovou stránku nukleotidovou sekvenci ve formátu *fasta*. Následně je provedena sekvenační analýza. Operon–mapper predikuje operony na bázi umělé neuronové sítě. Předpovědi operonů jsou založeny na integrovaných vzdálenostech sousedních genů a na funkčních vztazích jejich produktů kódující proteiny. Po dokončení analýzy si uživatel může zobrazit výsledný soubor nebo sadu souborů na webové stránce. Výsledky analýzy jsou zároveň zaslány na uživatelem zvolený e-mail. [34]

Z NCBI byl stažen genom bakterií *E. coli* BW25113 a *C. beijerinckii* NRRL B–598 ve formátu *fasta* a *gff3* [47]. Tyto soubory reprezentují vstupní soubory v rámci online nástroje Operon–mapper. Tento nástroj predikuje 821 operonových struktur a 1447 TU v genomu *E. coli* BW25113. V genomu *C. beijerinckii* NRRL B-598 predikuje nástroj 948 operonových struktur a 2307 TU. Ukázka souboru s výslednou predikcí operonů je na obrázku 3.3.

Operon	IdGene	Type	COGgene	PosLeft	postRight	Strand	Function
1							
	dnaA	CDS	COG0593	500	1849	+	[L] ATPase involved in DNA replication initiation
	cds-ALB48607.1	CDS	COG0592	2111	3211	+	[L] DNA polymerase sliding clamp subunit (PCNA
	yaaA	CDS	COG2501	3248	3454	+	[S] Uncharacterized conserved protein
	recF	CDS	COG1195	3503	4606	+	[L] Recombinational DNA repair ATPase (RecF pathway)
	cds-ALB48605.1	CDS	ROG0244	4606	4866	+	NA
	gyrB	CDS	COG0187	4921	6828	+	[L] Type IIA topoisomerase (DNA gyrase/topo II,
	gyrA	CDS	COG0188	6852	9332	+	[L] Type IIA topoisomerase (DNA gyrase/topo II,
2							
	rna-X276_26785	rRNA	NA	9833	11344	+	NA
	rna-X276_26780	rRNA	NA	11684	14597	+	NA
	rrf	rRNA	NA	14706	14822	+	NA
	rna-X276_26770	tRNA	NA	14835	14911	+	NA
	rna-X276_26765	tRNA	NA	14913	14988	+	NA

Obr. 3.3: Ukázka souboru s predikcí operonů u bakterie *C. beijerinckii* pomocí Operon–mapperu.

3.2.2 FGENESB

FGENESB je balíček pro automatickou detekci anotací bakteriálních genomů. Algoritmus predikce genů je založen na modelech Markovova řetězce kódujících oblastí a translačních terminačních míst. FGENESB umožňuje pracovat se sadou sekvencí získanou z bakteriálních komunit. Pro anotaci sekvencí společenstev obsahuje program AB split, který odděluje archebakteriální a eubakteriální sekvence. Tento nástroj byl použit ve vůbec prvním publikovaném projektu anotace bakteriálních společenstev. [35]

Predikce operonů je prováděna na základě vzdáleností ORF a četností různých genů sousedících ve známých bakteriálních genomech. Kompletní balíček FGENESB není dostupný online, ale je možné si jej nainstalovat. Online verze obsahuje pouze predikci genů a značně zjednodušenou část programu pro predikci operonů. Výsledná anotace je ve formátu GenBank. [35]

Pomocí FGENESB jsou predikovány operony u bakterie *C. beijerinckii* NRRL B-598. Aby bylo možné operony predikovat, bylo nutné vložit genom *C. beijerinckii* NRRL B–598 ve formátu *fasta* a z nabídky organismů vybrat organismus, který je této bakterii nejbližší. Z rodu *Clostridium* jsou v nabídce celkem tři organismy - *C. acetobutylicum* ATCC 824, *C. perfringens* str 13 a *C. tetani* E88. Všechny tři uvedené organismy byly pomocí nástroje BLAST [48] porovnány s *C. beijerinckii* NRRL B-598. Z 92 % se bakterie *C. beijerinckii* NRRL B-598 shodovala s *C. acetobutylicum* ATCC 824. Z nabízených organismů se *C. acetobutylicum* ATCC 824 shodovala s bakterií *C. beijerinckii* NRRL B- 598 nejvíce, proto byla tato bakterie zvolena jako nejbližší organismus. Tento nástroj predikuje celkem 929 operonových struktur a 2626 TU. Ukázka souboru s predikcí operonů je na obrázku 3.4.

```
Prediction of potential genes in microbial genomes
Time:   Tue Jan  1 00:00:00 2005
Seq name: CP011966.3 Clostridium beijerinckii NRRL B-598 chromosome, complete genome
Length of sequence - 6186993 bp
Number of predicted genes - 5559
Number of transcription units - 3544, operons - 1095
    N      Tu/Op   Conserved  S             Start         End    Score
                  pairs(N/Pv)
    1     1 Tu  1     .       +    CDS        500 -      1849   1145
    2     2 Op  1     .       +    CDS       2111 -      3211   1042
    3     2 Op  2     .       +    CDS       3242 -      3454    239
    4     2 Op  3     .       +    CDS       3503 -      4606    990
    5     2 Op  4     .       +    CDS       4606 -      4866    275
    6     2 Op  5     .       +    CDS       4921 -      6828   1826
    7     2 Op  6     .       +    CDS       6852 -      9332   2638
    8     3 Tu  1     .       -    CDS      11684 -     12136   -237
    9     4 Op  1     .       +    CDS      15082 -     15597    307
   10     4 Op  2     .       +    CDS      15612 -     16133    481
   11     4 Op  3     .       +    CDS      16136 -     17521   1326
   12     4 Op  4     .       +    CDS      17538 -     17765    127
   13     5 Tu  1     .       +    CDS      17871 -     18053     92
   14     6 Op  1     .       +    CDS      18665 -     19180    289
   15     6 Op  2     .       +    CDS      19177 -     22656   2894
```

Obr. 3.4: Ukázka souboru s predikcí operonů pomocí FGENESB.

3.2.3 ProOpDB

Databáze ProOpDB je aktuálně jedna z nejkompletnějších databází, která slouží k predikci operonů [36]. Narozdíl od jiných operonových databází je úspěšnost algoritmu predikce operonů mimořádně úspěšná i v případech, kdy nejsou trénovací a testovací organismy stejné. [36],

ProOpDB umožňuje vyhledávání operonů podle běžně používaných kritérií, jako je název genu nebo ID genu v konkrétních genomech. Umožňuje ale i vyhledávání a vizualizaci operonů podle konkrétních metabolických drah buněčných procesů. Tento způsob vyhledávání operonů umožňuje uživateli provést analýzu operonů organismu nebo skupiny organismů z celkového hlediska v souladu s jeho buněčnými a biochemickými vlastnostmi. Pomocí ProOpDB může uživatel snadno identifikovat regulační motivy, které se vztahují nejen k určité rodině genů, ale i k určité metabolické dráze.[36]

3.3 OperonIdentifier

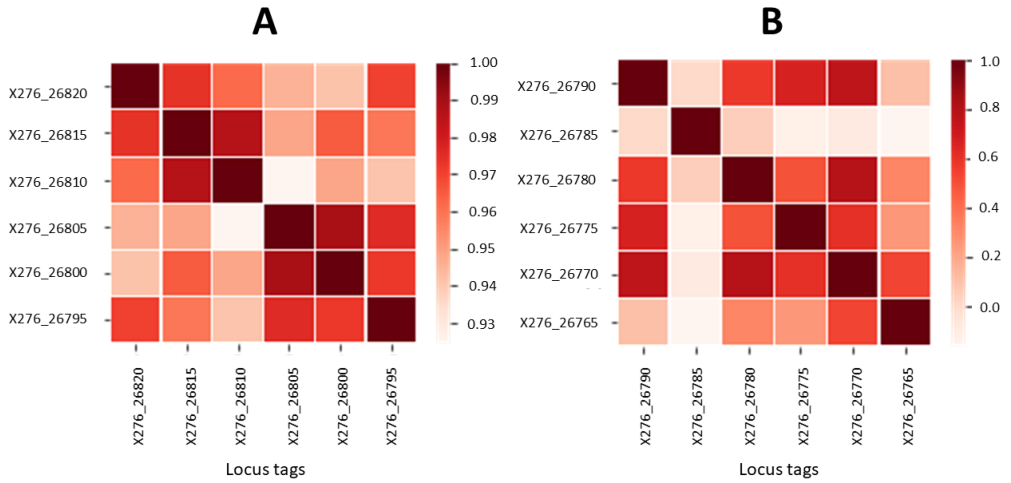
Pro predikci operonových struktur byla vytvořena funkce OperonIdentifier. Tato funkce umožňuje upřesnit již predikované operonové struktury o informaci získanou z genové exprese, ale zároveň umožňuje predikovat operonové struktury pouze na základě informace o genové expresi. Funkce byla implementována v Pythonu (verze 3.9) za pomocí balíčků NumPy [49], Pandas [50], Statistics [51] a openpyxl [52].

Funkce OperonIdentifier predikuje operonové struktury na základě Pearsonova korelačního koeficientu. Pearsonův korelační koeficient vyjadřuje v tomto případě vztah mezi dvěma geny. Pokud je korelace mezi geny vysoká, jsou tyto geny s velkou pravděpodobností součástí jednoho operonu. Pokud je korelace mezi geny nízká, pak tyto geny nejsou součástí jednoho operonu viz obrázek 3.5. Pearsonův korelační koeficient je vyjádřen rovnicí:

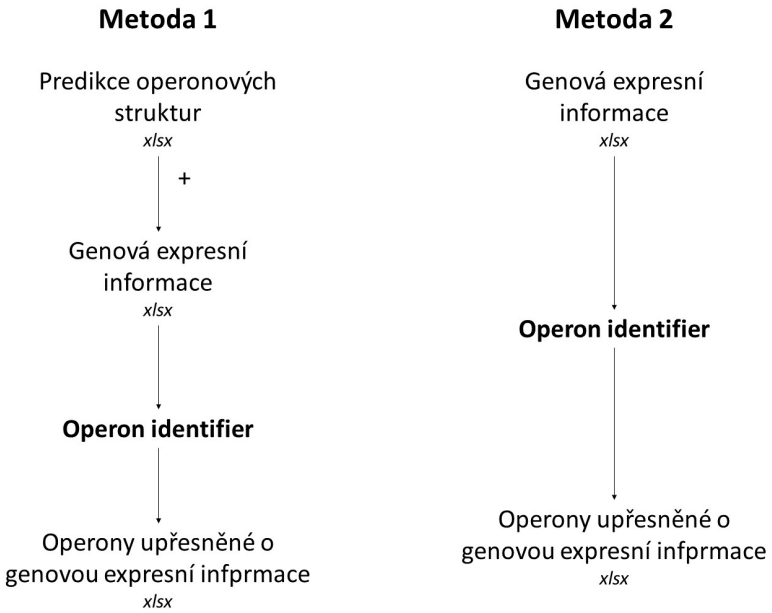
$$r_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}, \tag{3.1}$$

kde r_{xy} vyjadřuje korelaci mezi genem x a následujícím genem y.

Vstupem funkce OperonIdentifier je soubor s predikcí operonových struktur ve formátu *xlsx*, soubor s expresní genovou informací ve formátu *xlsx* a metoda, kterou uživatel zvolí zadáním čísla 1 nebo 2. Pokud uživatel zvolí 1, jsou již predikované operonové struktury upřesněny o informaci získanou z genové exprese. Jestliže je zvolena 2, pak jsou operony predikovány pouze na základě genové expresní informace, viz obrázek 3.6.



Obr. 3.5: Heat mapy, které znázorňují korelaci mezi šesti po sobě jdoucími geny. Mapa A představuje geny, které patří do jednoho operonu, mapa B znázorňuje geny, které nepatří do jednoho operonu. Tmavě červená barva znázorňuje vysokou korelaci mezi geny a naopak světlá barva znázorňuje nízkou korelaci mezi geny.



Obr. 3.6: Schéma metody 1 a 2 funkce OperonIdentifier.

Operonové predikce z online nástrojů nejsou ve formátu *xlsx*. Aby mohly být tyto predikce použity jako vstupní soubory funkce OperonIdentifier, je nutné je převést do formátu *xlsx*. Operony jsou u bakterie *E. coli* BW25113 přiřazeny na základě

ID genu. U bakterie *C. beijerinckii* NRRL B-598 jsou přiřazeny na základě označení locus tag. Výsledný soubor, který je pak vstupem funkce OperonIdentifier, obsahuje v prvním sloupci ID genu a ve druhém sloupci predikci operonových struktur pro všechny geny, pro které je známá genová expresní informace. Pořadí ID genu v souboru s operonovou predikcí musí být stejné jako v souboru s genovou expresní informací. V tabulce 3.3 je ukázka vstupního souboru s operonovou predikcí.

Tab. 3.2: Ukázka vstupního souboru s predikcí operonových struktur pro bakterii *C. beijerinckii* NRRL B-598. V prvním slupci je locus tag, ve druhém sloupci je predikce operonvých struktur pomocí nástroje FGENESB.

Locus tag	**FGENESB**
X276_26820	1
X276_26815	2
X276_26810	2
X276_26805	2
X276_26800	2

Pokud uživatel zvolí metodu 1, tak OperonIdentifier vezme veškerou genovu informaci pro všechny geny, které byly přiřazeny ke stejnému operonu, vypočítá korelační matici a zjistí průměr korelační matice. Poté funkce projde postupně všechny řádky korelační matice. Pokud je průměr řádku matice menší než stanovený limit, tak se operon rozdělí. Pro zbývající geny je přepočítána korelační matice a celý postup se opakuje. Když je OperonIdentifier u posledního řádku korelační matice, pokusí se do operonu připojit další gen. Opět je přepočítána korelační matice a celý postup se opakuje, dokud je průměrná korelace posledního řádku matice větší než předem stanovený limit. Podrobnější vývojový diagram metody 1 je znázorněn v příloze B.1.

V případě metody 2 vezme OperonIdentifier nejprve veškerou genovou expresní informaci prvních dvou genů a vypočítá jejich vzájemnou korelaci. Jestliže je korelace genů větší než předem stanovený limit, tak je do operonu přidán další gen a je vypočítána korelační matice těchto tří genů. Stejně jako u metody 1 pak OperonIdentifier zjišťuje, zda je průměr posledního řádku korelační matice větší než limit. Pokud je, tak je do operonu přidán další gen, pokud není, tak poslední gen k operonu nepatří a je iniciován nový operon se dvěma geny. Celý postup se opakuje, dokud funkce neprojde všechny geny. Podrobnější popis metody 2 je znázorněn ve vývojovém diagramu B.2, který je dostupný v příloze.

Na závěr funkce OperonIdentifier rozliší operony a TU. Postup, kterým jsou operony a TU rozlišeny, je znázorněn ve vývojovém diagramu v příloze B.3. Výsledný soubor s predikcí operonových struktur je znázorněn v tabulce 3.3.

Tab. 3.3: Ukázka souboru s predikcí operonů pomocí funkce OperonIdentifeir. Ve sloupci "IG_gene"je uloženo ID genu, ve druhém sloupci "Operon_prediction"jsou uložena čísla predikovaných operonových struktur nebo TU. Ve třetím sloupci "Operon/TU"je uložena informace, zda se jedná o TU nebo operonovou strukturu.

	ID_gene	**Operon_prediction**	**Operon/TU**
0	thrL	1	Operon
1	thrA	1	Operon
2	thrB	1	Operon
3	thrC	1	Operon
4	yaaX	1	Operon
5	yaaA	2	TU
6	yaaJ	3	TU

Pomocí funkce OperonIdentifier jsou predikovány operony jak u bakterie *E. coli* BW25113, tak u bakterie *C. beijerinckii* NRRL B-598. Pomocí metody 2, která je založena pouze na genové expresní informaci, je predikováno u *E. coli* 1111 operonových struktur a 1193 TU. U *C. beijerinckii* je touto metodou predikováno 1183 operonových struktur a 593 transkrupčních jednotek.

Pokud je u metody 1 použit Operon-mapper jako výchozí nástroj pro predikci operonových struktur, tak je pro *E. coli* predikováno 889 operonů a 1691 TU. U bakterie *C. beijerinckii* je predikováno 992 operonů a 563 TU.

Pro *C. beijerinckii* je také jako výchozí nástroj pro predikci operonových struktur použit nástroj FGENESB. V tomto případě je predikováno 988 operonů a 564 TU.

3.4 Srovnání výsledků predikce operonových struktur

V rámci bakalářské práce je prakticky provedena predikce operonových struktur pomocí online nástroje Operon-mapper a funkce OperonIdentifier u bakterií *E. coli* BW25113 a *C. beijerinckii* NRRL B-598. U bakterie *C. beijerinckii* je ještě navíc provedena operonová predikce pomocí nástroje FGENESB. Základní statistiky predikce operonových struktur jsou uvedeny v tabulkách 3.4, 3.5, 3.6.

Tab. 3.4: Základní statistiky predikce operonových struktur pro bakterii *E. coli* pomocí nástroje Operon-mapper a metody 1/2 funkce OperonIdentifier (OI_I/OI_2).

Nástroj	**Operon-mapper**	**OI_I**	**OI_II**
Počet TU	1447	1691	1193
Počet operonů	821	889	1111
Průměrná délka operonu	3	3	3
Maximální počet genů v operonu	28	28	20

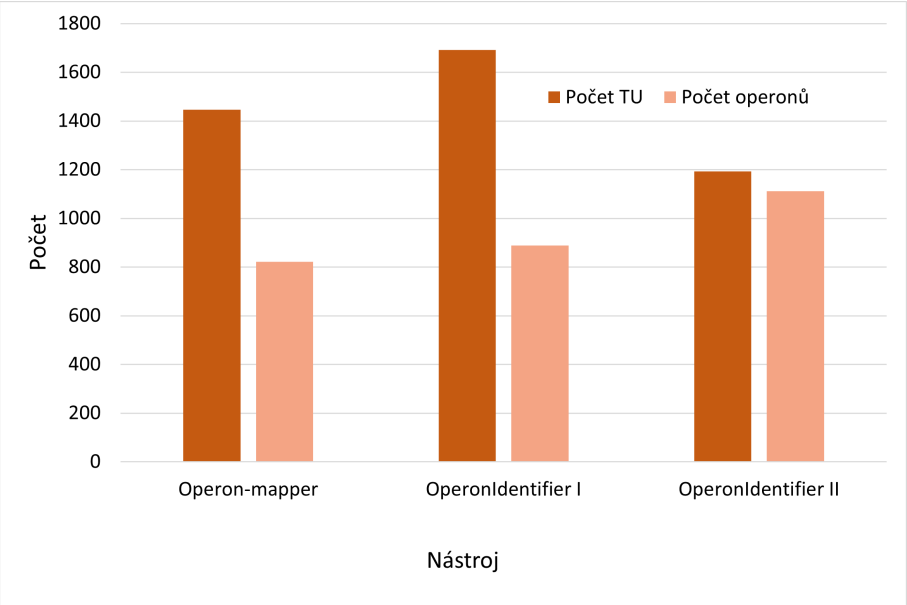
Tab. 3.5: Základní statistiky predikce operonových struktur pro *C. beijerinckii* pro online nástroje.

Nástroj	**Operon-mapper**	**FGENESB**
Počet TU	2307	2626
Počet operonů	948	929
Průměrná délka operonu	3	3
Maximální počet genů v operonu	38	31

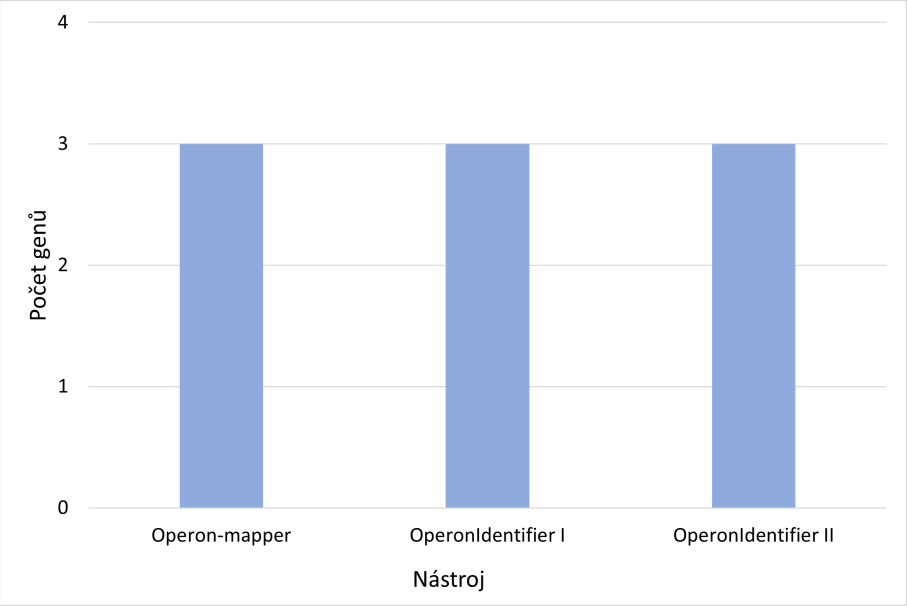
Tab. 3.6: Základní statistiky predikce operonových struktur pro *C. beijerinckii* pro OperonIdentifier (OI). (OM) a (FGENESB) znamená, že je vstupní soubor s predikcí operonů predikován pomocí nástroje Operon-mapper nebo FGENESB.

Nástroj	**OI_I (OM)**	**OI_I (FGENESB)**	**OI_II**
Počet TU	563	564	593
Počet operonů	992	988	1183
Průměrná délka operonu	5	5	4
Maximální počet genů v operonu	53	53	53

U bakterie *E. coli* BW25113 predikuje nejvíce operonů funkce OperonIdentifier za použití metody 2. Pomocí této metody je zároneň predikováno nejméně TU. Nejméně operonů predikuje nástroj Operon-mapper. Nejvíce TU predikuje metoda 1 funkce OperonIdentifier. U bakterie *E. coli* BW25113 obsahuje operon průměrně 3 geny. Průměrná délka operonu je u všech použitých nástrojů stejná. Tyto statistiky jsou znázorněny v grafech 3.7, 3.8.

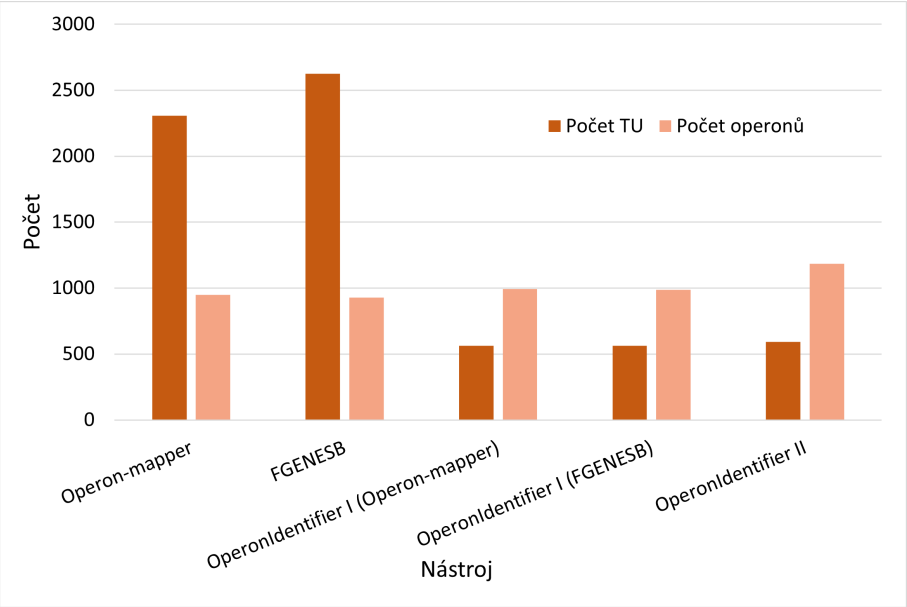


Obr. 3.7: Graf počtu predikovaných operonových struktur a TU pomocí různých nástrojů pro bakterii *E. coli* BW25113. OperonIdentifier I/II značí metodu 1/2 funkce OperonIdentifier. V závorce je uveden online nástroj, který je použit k predikci operonových struktur.

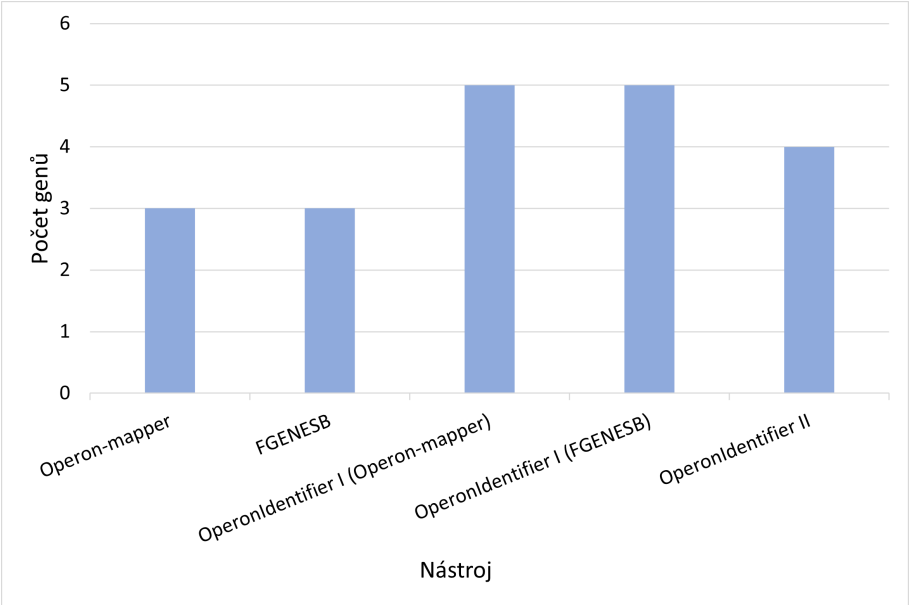


Obr. 3.8: Graf zobrazující průměrnou délku operonu u jednotlivých nástrojů pro bakterii *E. coli* BW25113. OperonIdentifier I/II značí metodu 1/2 funkce OperonIdentifier. V závorce je uveden online nástroj, který je použit k predikci operonových struktur.

Výsledné predikce TU se u bakterie *E. coli* BW2511 liší mnohem méně než u bakterie *C. beijerinckii* NRRL B-598. Online nástroje predikují mnohem více TU než funkce OperonIdentifier u *C. beijerinckii* NRRL B-598. V počtu predikovaných operonů vybočuje pouze metoda 2 funkce OperonIdentifier. Tato metoda predikuje u bakterie *C. beijerinckii* NRRL B-598 nejvíce operonových struktur. Ostatní metody se v počtu predikovaných operonů významně neliší, ale liší se v délce operonu. Funkce OperonIdentifier predikuje v průměru operony o dva geny delší než online nástroje. V grafech 3.9, 3.10 jsou znázorněny počty operonů a TU a průměrné délky operonů predikovaných u bakterie *C. beijerinckii* NRRL B-598.



Obr. 3.9: Graf počtu predikovaných operonových struktur a TU pomocí různých nástrojů pro bakterii *C. beijerinckii* NRRL B-598. OperonIdentifier I/II značí metodu 1/2 funkce OperonIdentifier. V závorce je uveden online nástroj, který je použit k predikci operonových struktur.



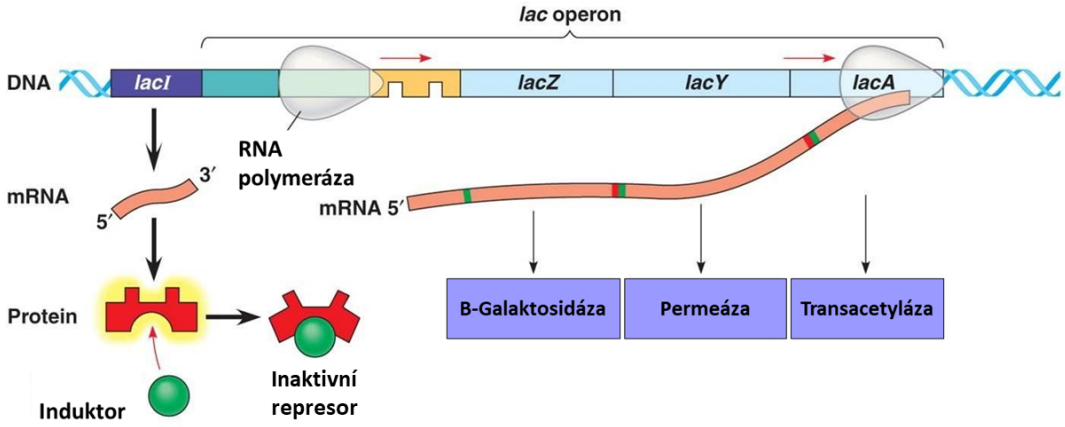
Obr. 3.10: Graf zobrazující průměrnou délku operonu u jednotlivých nástrojů pro bakterii *C. beijerinckii* NRRL B-598. OperonIdentifier I/II značí metodu 1/2 funkce OperonIdentifier. V závorce je uveden online nástroj, který je použit k predikci operonových struktur.

4 Identifikace operonových struktur

Tato kapitola se věnuje nejznámnějším a nejvíce prozkoumaným operonům bakterií *E. coli* BW25113 a *C. beijerinckii* NRRL B-598. Konkrétně je zde popsán *lac* operon, který se vyskytuje u *E. coli* BW25113 a *sol* operon, který se nachází u bakterie *C. beijerinckii* NRRL B-598. U obou operonů byla provedena dynamická analýza. Tyto operony byly u zmíněných bakterií laboratorně prokázány, proto jsou použity k ověření správnosti predikce operonových struktur. [53]

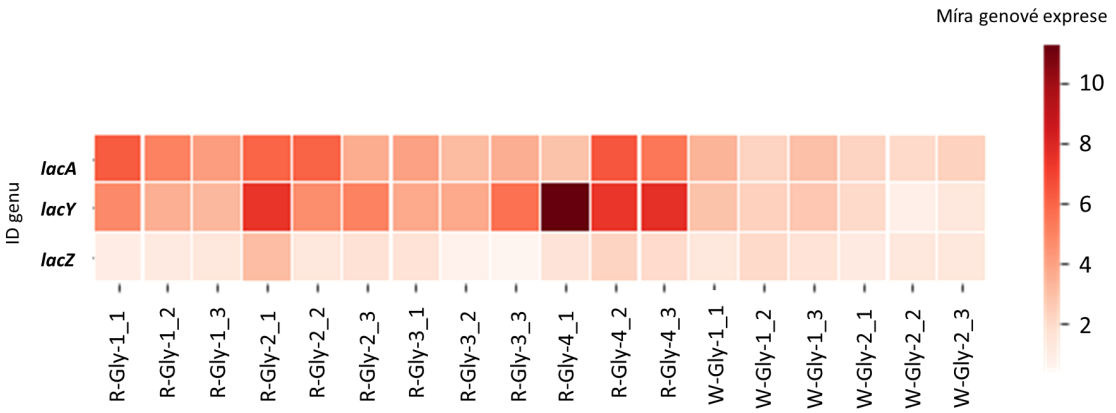
4.1 *lac* operon

Nejznámnějším operonem bakterie *E. coli* je *lac* operon (viz obrázek 4.1). Tento operon se skládá ze třech kódujících genů - *lacA*, *lacY*, *lacZ*, které jsou přepisovány do proteinu z jednoho promotoru. Tento operon je regulován *lac* represorem, který je produktem genu *lacI* přepisovaného z vlastního promotoru.



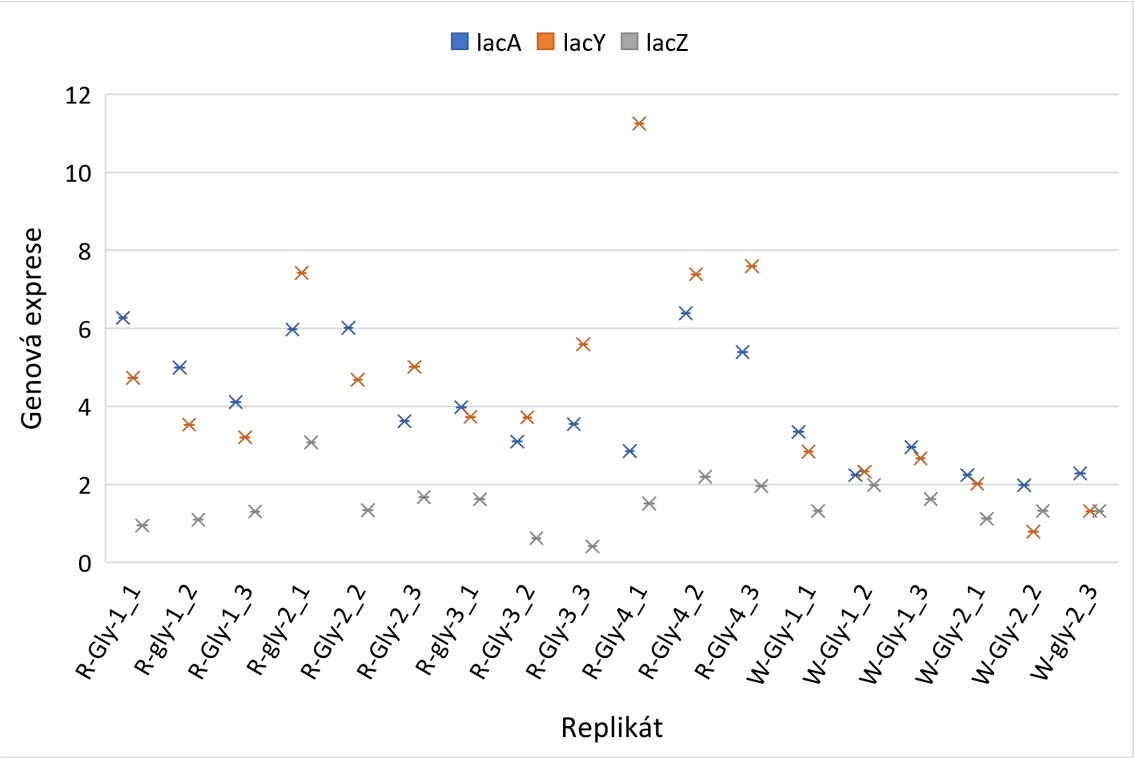
Obr. 4.1: Struktura *lac* operonu. Převzato z [54].

Pro *lac* operon je vykreslena heat mapa (viz obrázek 4.6). Z heat mapy je zřejmé, že geny *lac* operonu spolu korelují. Největší genovou expresi vykazuje growth-regulated replikát 1 R-Gly-4 genu *lacY*. Naopak nejnižší genovou expresi vykazuje growth-regulated replikát 3 R-Gly-3 genu *lacZ*.



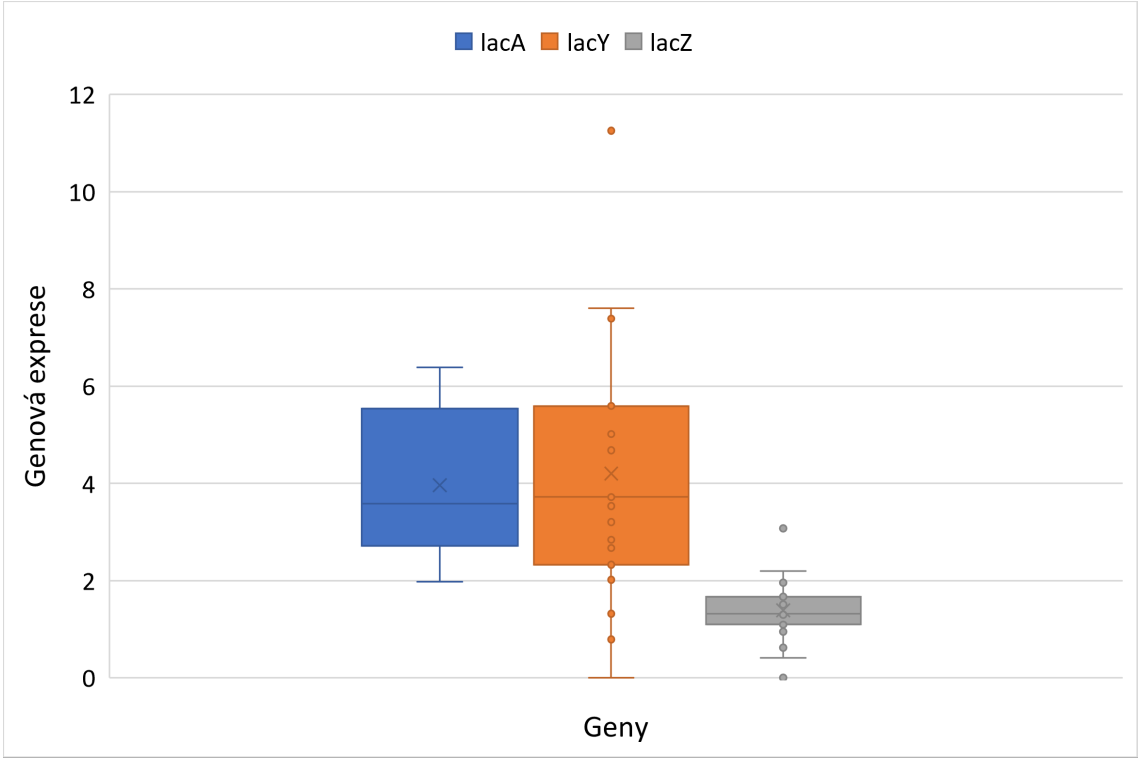
Obr. 4.2: Heat mapa znázorňující míru genové exprese v *lac* operonu. Čím větší je míra genové exprese, tím tmavší políčko heat mapy je.

Pro ještě lepší analýzu genové exprese *lac* operonu je sestrojen graf 4.3. Je značně patrné, že největší genovou epxresi skutečně vykazuje growth-regulated replikát 1 R-Gly-4 genu *lacY* a nejnižší genovou expresi vykazuje growth-regulated replikát 3 R-Gly-3 genu *lacZ*. Celkově je na první pohled zřejmé, že nejnižší genovou expresi vykazují replikáty genu *lacZ*. V tomto grafu je vidět, že geny *lacA* a *lacY* spolu značně korelují. Pokud kleasá míra genové exprese u jednoho genu, klesá míra genové exprese i u druhého genu. Pokud míra genové exprese u jednoho genu roste, tak zároveň roste míra genové exprese i u druhého genu. S těmito dvěma geny koreluje o něco méně i gen *lacZ*. *β-galaktosidáza* produkovaná genem *lacZ* štěpí laktózu a *laktosa permeasa*, která je produktem genu *lacY*, transportuje laktózu do buňky. Oproti tomu *thiogalaktosid transacetyláza*, produkovaná genem *lacA*, zbavuje buňku *thiogalaktosidů*, takže se svojí funkcí značně liší od ostatních genů. To může být důvod, proč *lacA* koreluje méně s geny *lacY* a *lacZ*. [55], [56]



Obr. 4.3: Graf znázorňující míru genové exprese pro jednotlivé replikáty v *lac* operonu.

V grafu 4.4 je zobrazeno rozložení genové exprese pro jednotlivé geny v *lac* operonu. Geny *lacA* a *lacY* vykazují podobnou míru genové exprese. Naopak u genu *lacZ* je míra genové exprese výrazně menší. Celkově největší míru genovou exprese vykazuje gen *lacY*.



Obr. 4.4: Graf rozložení genové exprese v jednotlivých genech *lac* operonu.

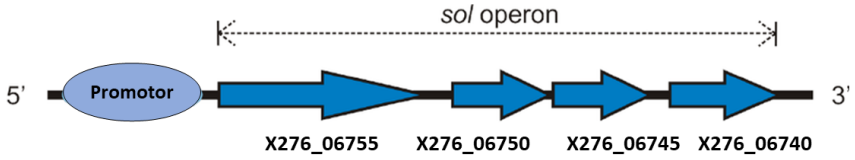
Jak ukazuje tabulka 4.1 *lac* operon je pomocí všech testovaných nástrojů predikován správně.

Tab. 4.1: Tabulka uvadí čísla, pod kterými je *lac* operon detekován u jednotlivých nástrojů. OI_I/OI_II metodu 1/2 funkce OperonIdentifier.

ID genu	**Operon-mapper**	**OI_I**	**OI_II**
lacA	179	214	182
lacY	179	214	182
lacZ	179	214	182

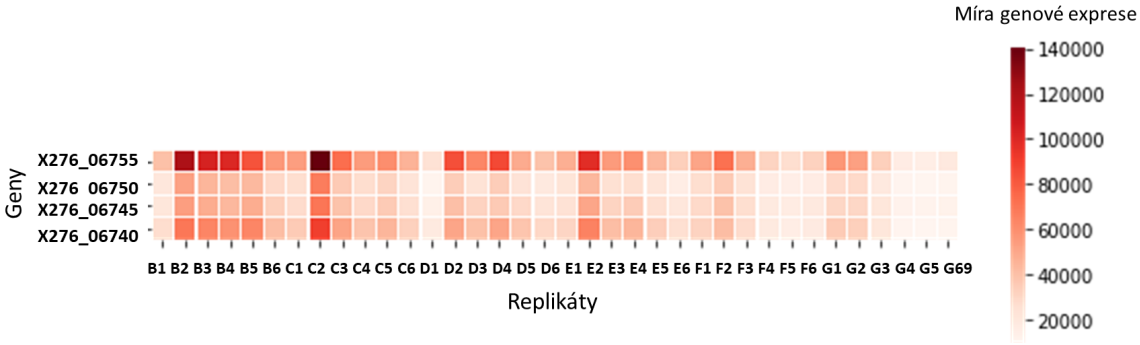
4.2 *Sol* operon

Nejznámnějším operonem bakterie *C. beijerinckii* je *sol* operon (viz obrázek 4.5). Součástí *sol* operonu této bakterie jsou čtyři geny - *X276_06755, X276_06750, X276_06745 a X276_06740*. U těchto genů je provedena dynamická analýza.



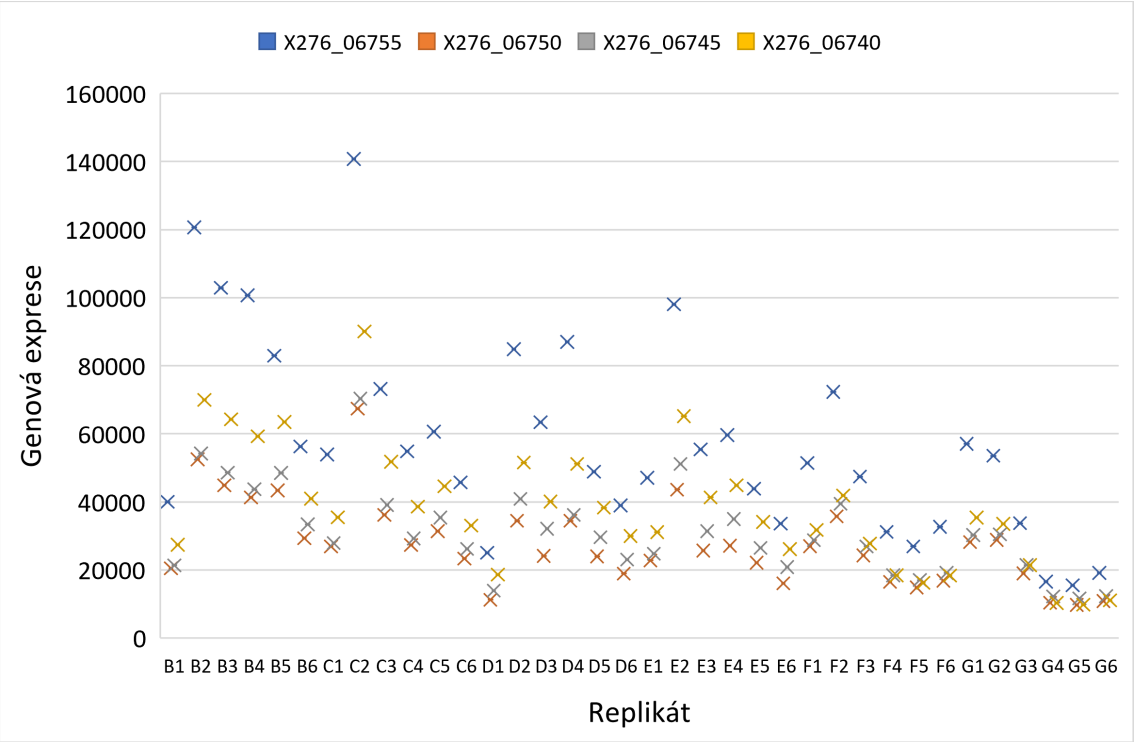
Obr. 4.5: Struktura *sol* operonu. Převzato z [57].

Na obrázku 4.6 je zobrazena heat mapa tohoto operonu. Z heat mapy je patrné, že geny tohoto operonu spolu korelují, jelikož mají v rámci jednotlivých replikátů podobnou míru genové exprese. Největší genovou expresi vykazuje replikát C2 u genu *X276_06755*. Naopak nejmenší míra genové exprese je u replikátů G4, G5 a G6 u genů *X276_06750, X276_06745, X276_06740*.



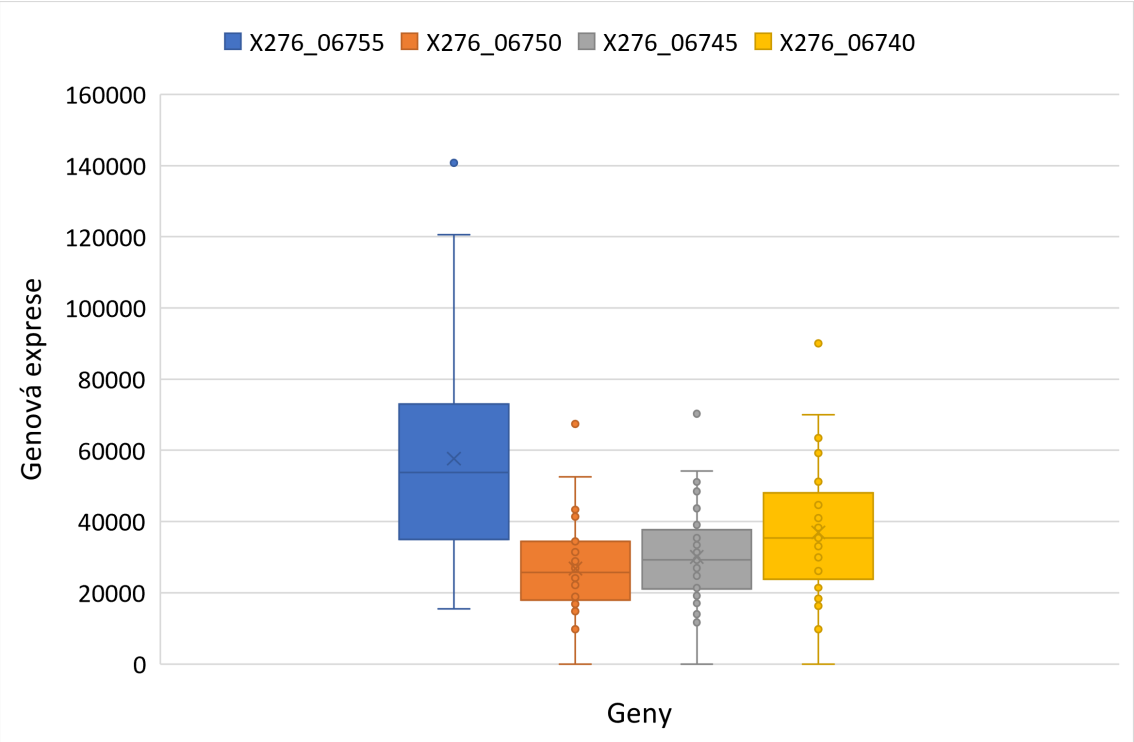
Obr. 4.6: Heat mapa znázorňující míru genové exprese v *sol* operonu. Čím větší je míra genové exprese, tím tmavší políčko heat mapy je.

Míra genové exprese pro všechny replikáty všech genů *sol* operonu je podrobněji zobrazena v grafu 4.7. Z grafu je vidět, že opravdu značně nejvyšší expresi vykazuje replikát C2 u genu *X276_06755*. Naopak vůbec nejnižší genovou expresi vykazuje replikát G5 genu *X276_06750*. Dále je patrné, že celkově největší míru genové exprese vykazuje gen *X276_06755*. Naopak nejnižší míru genové exprese má gen *X276_06750*. Tento graf potvrzuje, že geny *sol* operonu spolu opravdu korelují. Míra genové exprese má u všech genů zároveň buď klesající nebo rostoucí tendenci.



Obr. 4.7: Graf znázorňující míru genové exprese pro jednotlivé replikáty v *sol* operonu.

Rozložení genové exprese pro jednotlivé geny *sol* operonu je zobrazeno v grafu 4.8. Graf ukazuje, že geny *X276_06750* a *X276_06745* mají velmi podobné rozložení genové exprese. Naopak výrazně vybočuje gen *X276_06755*.



Obr. 4.8: Graf rozložení genové exprese v jednotlivých genech*sol* operonu.

Tab. 4.2: Tabulka uvadí čísla, pod kterými je predikován *sol* operon pomocí jednotlivých nástrojů. OI_I/OI_II představuje metodu 1/2 funkce OperonIdentifier.

Locus_tag	**OM**	**OI_1(OM)**	**FGENESB**	**OI_1(FGENESB)**	**OI_2**
X276_06755	2566	1153	2677	1151	1326
X276_06750	2567	1153	2677	1151	1326
X276_06745	2567	1153	2677	1151	1326
X276_06740	2567	1153	2677	1151	1326

Na rozdíl od *lac* operonu, který je všemi nástroji detekován správně, predikuje nástroj Operon-mapper *sol* operon špatně. Gen *X276_06755* je pomocí tohoto nástroje detekován jako TU, ostatní tři geny jsou detekovány jako jeden operon, viz tabulka 4.2. Predikce operonů pomocí nástroje Operon-mapper je založena na vzdálenosti sousedních genů a na funkčních vztazích jejich produktů, proto není vhodné použití Operon-mapperu k predikci operonů u málo prozkoumaných bakterií. V tomto ohledu se jednoznačně jeví výhodnější použití funkce OperonIdentifier, která umožňuje predikci operonů pouze na základě genové expresní informace bez hlubších znalostí organismu.

Závěr

Bakalářská práce se zabývá problematikou genové regulace prokaryot. U prokaryot se na genové regulaci významně podílejí operony. Práce se zabývá programováním vlastní funkce, která k operonové predikci využívá genovou expresní informaci.

První kapitola bakalářské práce je věnována operonům. Je zde popsána problematika genové regulace prokaryot, transkriptom a jeho analýza. Druhá kapitola popisuje vybrané sekvenační techniky, které umožňují sekvenování transkriptomu a tím pádem umožňují získat genovou expresní informaci. Konkrétně jsou zde popsány sekvenační technologie Illumina, SOLiD, Ion Torrent a RNA-Seq. Tato kapitola se dále věnuje bakteriím *E. coli* B25113 a *C. beijerinckii* NRRL B-598, které byly použity v praktické části.

Praktická část ze zabývá vytvořením vhodného datasetu k predikci operonových struktur a operonovou predikcí. V první části kapitoly je popsáno předzpracování RNA-Seq dat bakterie *C. beijerinckii* NRRL B-598, které je nutné provést, aby bylo možné získat genovou expresní informaci ve vhodné formě pro operonovou predikci. Další části kapitoly jsou věnovány odvození operonových struktur v rámci celogenomové analýzy. Jsou zde jednak popsány a testovány online nástroje Operon-mapper, FGENESB a ProOpDB, ale především se kapitola zabývá vytvořením a otestováním funkce OperonIdentifier. Tato funkce umožňuje uživateli jednak upřesnit operonové struktury o genovou expresní informaci, ale také umožňuje predikovat operony pouze na základě genové expresní informace. Poslední část práce se věnuje dynamické analýze známých operonů - *lac* operonu a *sol* operonu. K ověření správné predikce operonů je použit u bakterie *E. coli* BW25113 *lac* operon a u bakterie *C. beijerinckii* NRRL B-598 *sol* operon. Tyto operony byly u těchto organismů laboratorně potvrzeny, proto byly vybrány jako kontrolní operony.

Navržená funkce byla implementována v Pythonu a testována na *E. coli* BW25113 a *C. beijerinckii* NRRL B-598. Funkce ve všech případech správně detekovala *lac* operon i *sol* operon narozdíl od online nástrojů. *Sol* operon nebyl správně detekován pomocí nástroje Operon-mapper. U *C. beijerincki* NRRL B-598 se online nástroje lišily oproti funkci OperonIdentifier především v počtu predikovaných TU. Online nástroje predikují u této bakterie mnohem více TU než OperonIdentifier. Oproti tomu u *E. coli* se počty TU liší značně méně. U *E. coli* BW25113 se průměrná délka operonů predikovaných pomocí Operon-mapperu a funkce OperonIdentifier neliší. Z toho vyplývá, že operony predikované online nástrojem a funkcí OperonIdentifier jsou přibližně stejně dlouhé. Oproti tomu u bakterie *C. beijerinckii* NRRL B-598 se délka operonů u online nástrojů a funkce OperonIdentifier lišila přibližně o dva geny. Online nástroje predikovaly kratší operony než funce OperonIdentifier, proto bylo u online nástrojů predikováno mnohem více TU než u funkce OperonIdentifier.

Literatura

[1] Operon: Operon Definition. Biology dictionary [online]. Dostupné z: https://biologydictionary.net/operon/

[2] *Overview: Gene regulation in bacteria. Khan Academy [online]. Dostupné z: https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/regulation-of-gene-expression-and-cell-specialization/a/overview-gene-regulation-in-bacteria*

[3] *What is Operon? Role in Protein Synthesis [online]. Dostupné z: https://www.golifescience.com/operon/*

[4] *ROSYPAL, Stanislav. Nový přehled biologie. Scientia, 2003. ISBN 80-7183-268-5.*

[5] *Regulatory Proteins [online]. Dostupné z: https://www.mun.ca/biology/scarr/bio4241_regulatoryproteins.htm*

[6] *What is Transcriptome Sequence?. WARWICK [online]. Dostupné z: https://warwick.ac.uk/fac/sci/lifesci/research/vegin/geneticimprovement/transcriptome/*

[7] *Transcriptome Fact Sheet: What is a transcriptome?. National Human Genome Institute [online]. Dostupné z: https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet*

[8] *Transcriptomics [online]. Dostupné z: http://alnelsongen564s17.weebly.com/transcriptome.html*

[9] *XIONG, Jin. Essential bioinformatics. Cambridge University Press, 2006.*

[10] *Serial Analysis of Gene Expression (SAGE) by Sequencing [online]. Dostupné z: https://www.thermofisher.com/cz/en/home/life-science/sequencing/rna-sequencing/gene-expression-sequencing/serial-analysis-gene-expression-sage-sequencing.html*

[11] *DNA Sequencing. National Human Genome Research Institute [online]. Dostupné z: https://www.genome.gov/genetics-glossary/DNA-Sequencing*

[12] *DNA sequencing. Khan Academy [online]. Dostupné z: https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/dna-sequencing*

[13] *KOLÍSKO, Martin. Moderní metody sekvenování DNA. Živa, 2017, 3: 2017.*

[14] *DNA sequencing. Britannica [online].. Dostupné z:* https://www.britannica.com/science/DNA-sequencing

[15] *Explore Illumina sequencing technology [online]. Dostupné z:* https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

[16] *NextSeq Series of Sequencing Systems [online]. Dostupné z:* https://ostr.ccr.cancer.gov/wp-content/uploads/2018/06/nextseq-500_specifications.pdf

[17] *NextSeq [online]. Dostupné z:* https://cores.research.asu.edu/genomics/equipment/nextseq

[18] *NextSeq 500's new chemistry described [online]. Dostupné z:* http://enseqlopedia.com/2014/01/nextseq-500s-new-chemistry-described/

[19] *SCIENTIFIC, Thermo Fisher. Ion TorrentTM next-generation sequencing technology. Dostupné z:* https://www.thermofisher.com/za/en/home/life-science/sequencing/nextgeneration-sequencing/ion-torrent-next-generation-sequencing-technology.html,2016.

[20] *MACKENZIE, Ruairi J. RNA-seq: Basics, Applications and Protocol. Dostupné z:* https://www.technologynetworks.com/genomics/articles/rna-seqbasics-applications-and-protocol-299461,2018.

[21] *E. coli. Word Health Organization [online]. Dostupné z:* https://www.who.int/news-room/fact-sheets/detail/e-coli

[22] *REN, Lujing, et al. Microbial production of vitamin K2: current status and future prospects. Biotechnology advances, 2020, 39: 107453.*

[23] *GRENIER, Frédéric, et al. Complete genome sequence of Escherichia coli BW25113. Genome announcements, 2014, 2.5: e01038-14.*

[24] *LONG, Christopher P., et al. Fast growth phenotype of E. coli K-12 from adaptive laboratory evolution does not require intracellular flux rewiring. Metabolic engineering, 2017, 44: 100-107.*

[25] *CHANG, Vicky, et al. The effect of lipopolysaccharide core structure defects on transformation efficiency in isogenic Escherichia coli BW25113 rfaG, rfaP, and rfaC mutants. J. Exp. Microbiol. Immunol, 2010, 14: 101-107.*

[26] *BARRETT, Tanya, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research, 2012, 41.D1: D991-D995.*

[27] *ROPERS, Delphine, et al. Multiomics Study of Bacterial Growth Arrest in a Synthetic Biology Application. ACS Synthetic Biology, 2021, 10.11: 2910-2926.*

[28] *Clostridium: What is it? Morphology, Classification, Characteristics [online]. Dostupné z:* `https://www.microscopemaster.com/clostridium.html`

[29] *BRANSKA, Barbora, et al. 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.*

[30] *SEDLAR, Karel, et al. A transcriptional response of Clostridium beijerinckii NRRL B-598 to a butanol shock. Biotechnology for biofuels, 2019, 12.1: 1-16.*

[31] *SEDLAR, Karel, et al. Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. BMC genomics, 2018, 19.1: 1-13.*

[32] *PATAKOVA, Petra, et al. Acidogenesis, solventogenesis, metabolic stress response and life cycle changes in Clostridium beijerinckii NRRL B-598 at the transcriptomic level. Scientific reports, 2019, 9.1: 1-21.*

[33] *SEDLAR, Karel, et al. A transcriptional response of Clostridium beijerinckii NRRL B-598 to a butanol shock. Biotechnology for biofuels, 2019, 12.1: 1-16.*

[34] *TABOADA, Blanca, et al. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics, 2018, 34.23: 4118-4120.*

[35] *FGENESB Suite of Bacterial Operon and Gene Finding Programs [online]. [cit. 2021-12-06]. Dostupné z:* `http://www.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb`

[36] *TABOADA, Blanca, et al. ProOpDB: Pro karyotic Op eron D ata B ase. Nucleic acids research, 2012, 40.D1: D627-D631.*

[37] *SCHWARZEROVÁ, Jana. Reproducible analytical pipeline for using raw RNA-Seq data from non-model organisms.*

[38] *ŠUSTR, Z., et al. Metacentrum, the czech virtualized ngi. In: EGEE Technical Forum. 2009.*

[39] *FASTQC. Babraham Bioinformatics [online]. Dostupné z:* `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

[40] *EWELS, Philip, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 2016, 32.19: 3047-3048.*

[41] *KOPYLOVA, Evguenia; NOÉ, Laurent; TOUZET, Hélène. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 2012, 28.24: 3211-3217.*

[42] *BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, 30.15: 2114-2120.*

[43] *Phred - Quality Base Calling [online]. Dostupné z:* `https://www.phrap.com/phred/`

[44] *DOBIN, Alexander, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, 29.1: 15-21.*

[45] *LI, Heng, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16: 2078-2079.*

[46] *LIAO, Yang; SMYTH, Gordon K.; SHI, Wei. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 2014, 30.7: 923-930.*

[47] *Clostridium beijerinckii NRRL B-598. NCBI [online]. Dostupné z:* `https://www.ncbi.nlm.nih.gov/bioproject/?term=clostridium%20beijerinckii%20NRLL%20B-598`

[48] *ALTSCHUL, Stephen F., et al. Basic local alignment search tool. Journal of molecular biology, 1990, 215.3: 403-410.*

[49] *HARRIS, Charles R., et al. Array programming with NumPy. Nature, 2020, 585.7825: 357-362.*

[50] *MCKINNEY, Wes, et al. pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 2011, 14.9: 1-9.*

[51] *Statistics — Mathematical statistics functions [online]. Dostupné z:* `https://docs.python.org/3/library/statistics.html`

[52] *GAZONI, Eric; CLARK, Charlie. openpyxl-A Python library to read/write Excel 2010 xlsx/xlsm files. Dostupné z:* `http://openpyxl.readthedocs.org/en/default`

[53] *The lac operon. Khan Academy [online]. Dostupné z:* https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/regulation-of-gene-expression-and-cell-specialization/a/the-lac-operon

[54] *Epigenetic regulation of Lac Operon in E. coli [online]. Dostupné z:* https://www.onlinebiologynotes.com/epigenetic-regulation-of-lac-operon-in-e-coli/

[55] *CLARK, David P.; PAZDERNIK, Nanette J. Molecular biology. Elsevier, 2013.*

[56] *Lac Operon Concept in Bacteria [online]. Dostupné z:* https://noteshippo.com/lac-operon-definition-structure-situations-or-conditions

[57] *SEDLAR, K., et al. Identification and characterization of sol operon in clostridium pasteurianum NRRL B-598 genome. In: In: Proceedings Of The 2nd International Conference On Chemical Technology. Czech Society of Industrial Chemistry, Mikulov. 2014. p. 435-439.*

Seznam symbolů a zkratek

cDNA	komplementární DNA
DNA	deoxyribonukleová kyselina
ESTs	expressed sequence tags
FGENESB	pipeline bakteriální genomové anotace
GEO	omnibus genové exprese
HMM	hidden Markov model
MMP	maximální mapovatelné předpony
mRNA	mediátorová RNA
NCBI	Národní centrum pro bioinformatické informace
OI	OperonIdentifier
ORF	otevřený čtecí rámec
ProOpDB	Prokaryotická operonová databáze
RNA	ribonukleová kyselina
rRNA	ribosomální ribonukleová kyselina
RNA-Seq	RNA sekvenování
SAGE	sériová analýza genové exprese
SOLiD	sekvenování nukleotidů pomocí ligace a detekce
SOM	samoorganizující se mapy
TU	transkripční jednotka

A Obsah elektronické přílohy

A.1 *E. coli* BW25113

Ve složce *E_coli* je uložena funkce *OperonIdentifier* s nastavenými parametry pro bakterii *E. coli* BW25113 a všechny vstupní soubory této funkce. Součástí složky jsou také výsledné predikce operonových struktur pomocí funkce *OperonIdentifier*.

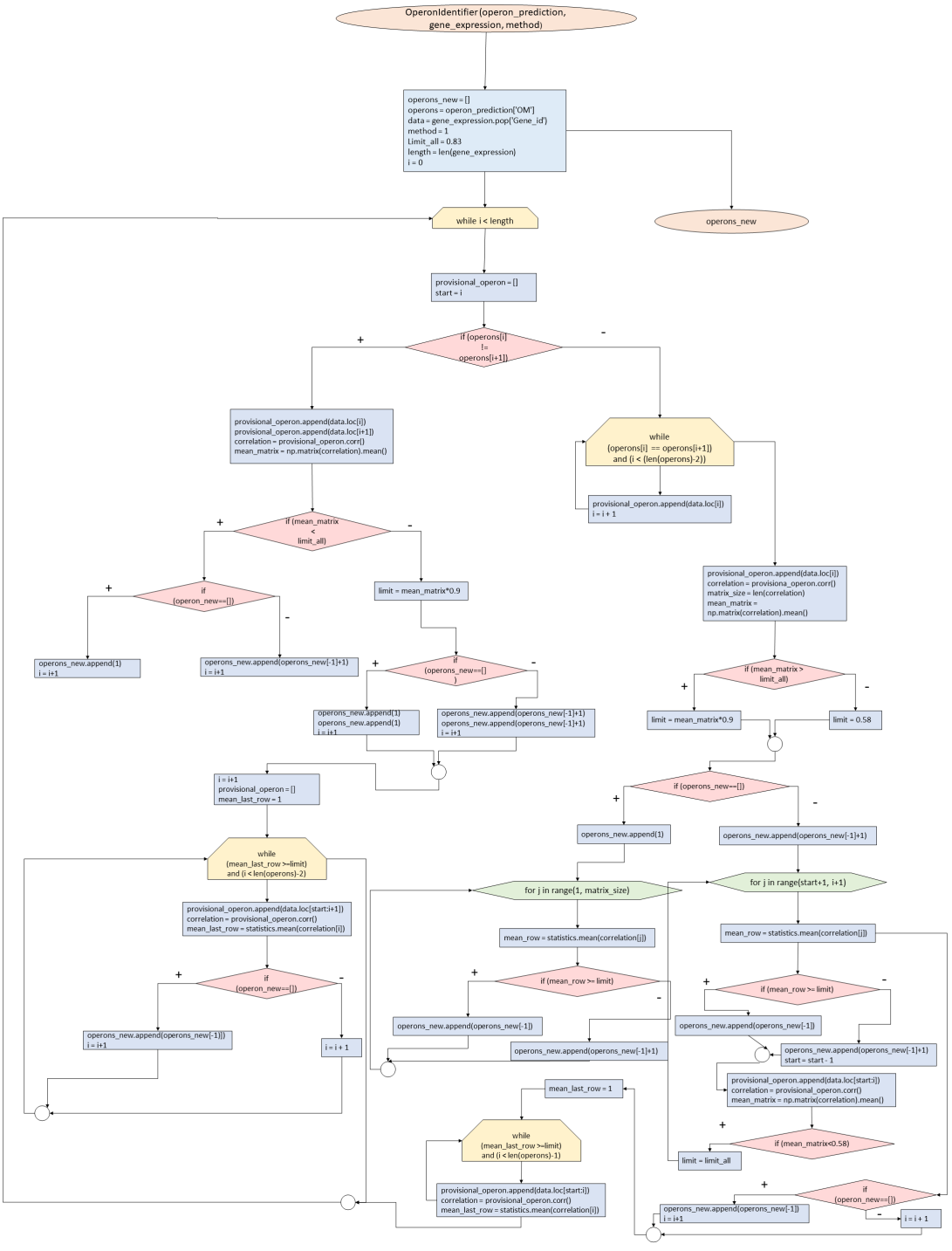
A.2 *C. beijerinckii* NRRL B-598

Tato složka obsahuje funkci *OperonIdentifier* s nastavenými paramtery pro bakterii *C. beijerinckii* NRRL B-598. Součástí slžky jsou také všechny potřebné vstupní soubory této funkce a soubory s výslednou predikcí operonových struktur pro tutu bakterii.

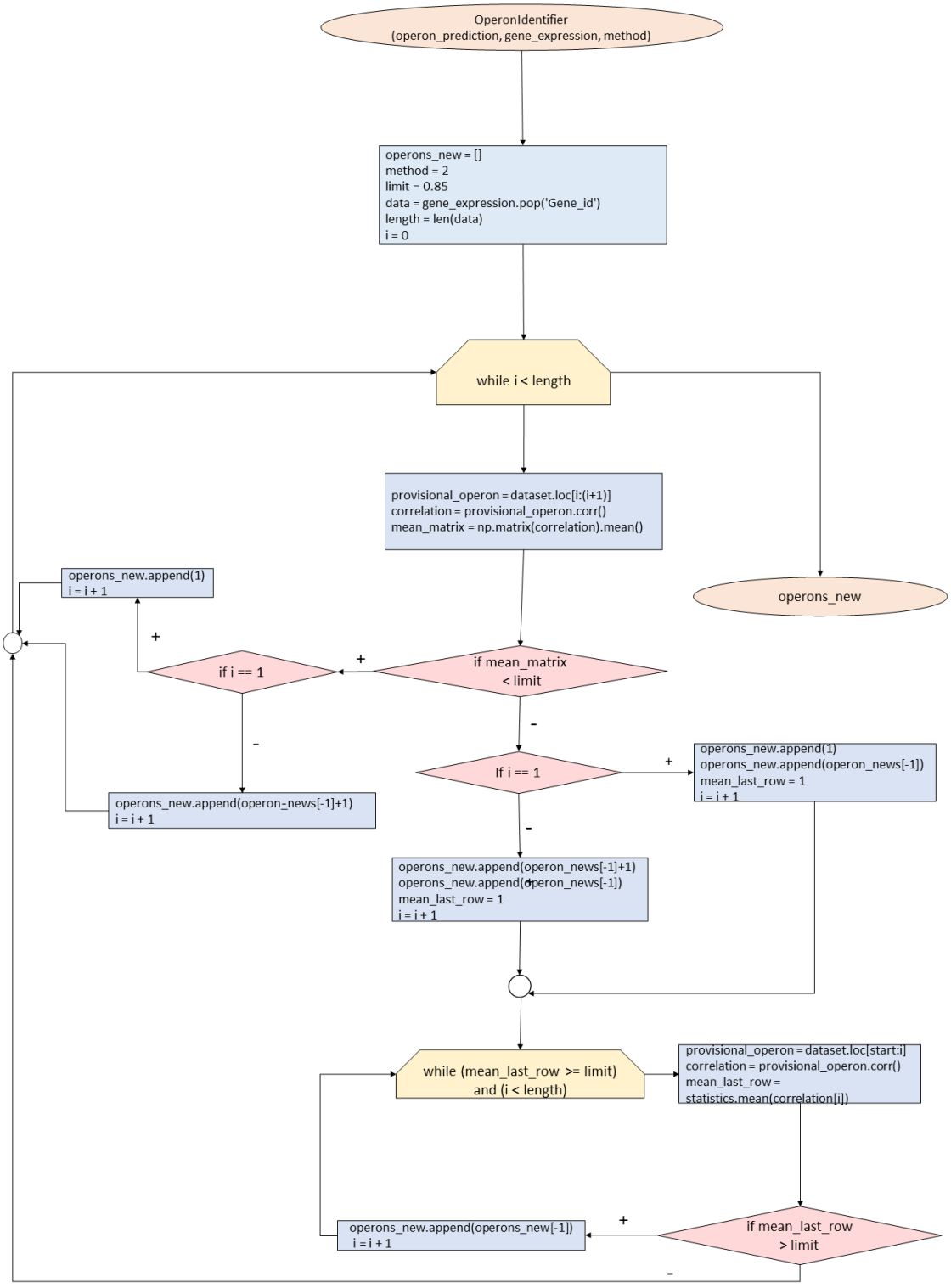
A.3 CountTables

Složka obsahuje CountTables, které byly vytvořeny předzpracováním RNA-Seq dat bakterie *C. beijerinckii* NRRL B-598.

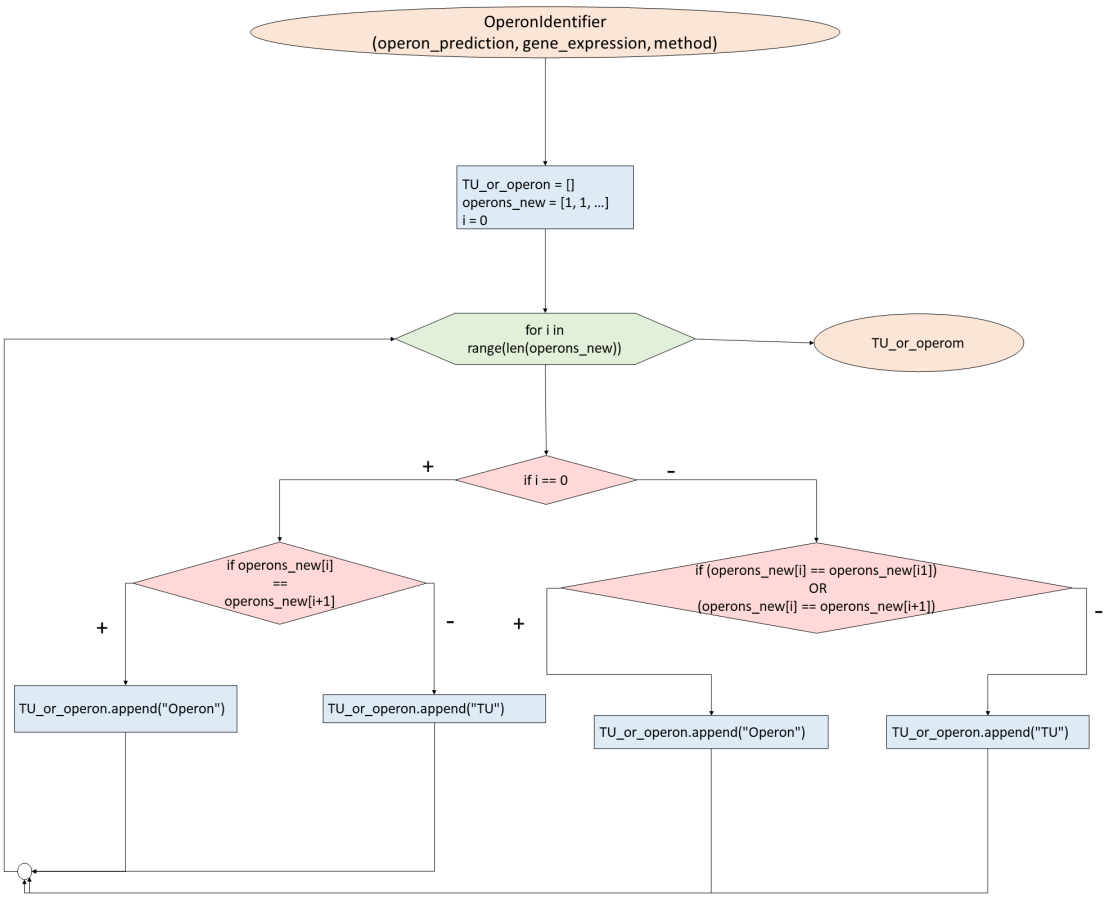
B Vývojové diagramy



Obr. B.1: Vývojový diagram metody 1 funkce *OperonIdentifier* pro bkaterii *C. beijerinckii* NRRL B-598 se vstupní predikcí operonových struktur z nástroje Operon-mapper.



Obr. B.2: Vývojový diagram predikce operonových struktur pouze z informace o genové expresi pro *C. beijerinckii* NRRL B-598 .



Obr. B.3: Vývojový diagram části funkce *OperonIdentifier*, která rozliší operonové struktury a TU.