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

KONZERVACE POZICE GENŮ V BAKTERIÁLNÍCH GENOMECH
GENE ORDER CONSERVATION IN BACTERIAL GENOMES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Tereza Martinková**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Denisa Maděránková, Ph.D.**
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Tereza Martinková ***ID:*** 183354
Ročník: 3 ***Akademický rok:*** 2017/18

NÁZEV TÉMATU:

Konzervace pozice genů v bakteriálních genomech

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma konzervace pozice genů (synteny bloky) v prokaryotních genomech a na téma komparativní genomiky založené na porovnání pozice genů. 2) Sestavte rozsáhlý dataset anotovaných bakteriálních genomů, který bude sloužit k analýze jejich fylogenetických vztahů porovnáním vzájemných pozic genů. 3) V libovolném programovém prostředí vytvořte funkci pro výpočet pozičního profilu vybraných genů anotovaného genomu vůči vybranému referenčnímu genomu. 4) Navrhněte a implementujte metodu porovnání pozičního profilu vybraných genů s cílem vyhodnocení fylogenetických vztahů mezi bakteriálními genomy. 5) Proveďte analýzu sestaveného datasetu a výsledky diskutujte a porovnejte s referenčním fylogenetickým stromem.

DOPORUČENÁ LITERATURA:

[1] MAHADEVAN, P. a D. SETO. Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder 4.0. BMC Research Notes. 2010, 3:41.

[2] SODERLUND, C., NELSON, W., SHOEMAKER, A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Research, 2006, roč. 16, s. 1159-1168.

Termín zadání: 5.2.2018 ***Termín odevzdání:*** 25.5.2018

Vedoucí práce: Ing. Denisa Maděránková, Ph.D.
Konzultant:

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

Abstrakt

Teoretická část práce se zabývá základními pojmy jako bakteriální genom, komparativní genomika a synteny bloky. Je zde vysvětleno, co synteny je a v čem spočívá její důležitost. Dále je v teoretické části zmíněn GenBank formát, jeho obsah a využití. Praktická část je zaměřena na vyhledávání podobností v sekvencích DNA referenční bakterie s vybranou bakterií, jejich setřídění pomocí hladového algoritmu a zobrazení podobnosti fylogenetickým stromem.

Klíčová slova

Gen, genom, synteny, komparativní genomika, DNA, CDS, fylogenetika, breakpointová metoda

Abstract

Theoretical part of the thesis deals with basic concepts such as bacterial genome, comparative genomics and mainly synteny blocks. Here is explained what synteny is and what is its importance. In the theoretical part, the GenBank format is also mentioned, its content and usage. The practical part is focused on searching similarities in DNA sequences of reference bacteria with selected bacteria, their sorting by means of greedy algorithm and visualization of similarities using phylogenetic tree.

Keywords

Gene, genome, synteny, comparative genomics, DNA, CDS, phylogenetics, breakpoint method

Bibliografická citace:

MARTINKOVÁ, T. *Konzervace pozice genů v bakteriálních genomech*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2018. 41s. Vedoucí práce: Ing. Denisa Maděránková, Ph.D.

Prohlášení

„Prohlašuji, že svou bakalářskou práci na téma Konzervace pozice genů v bakteriálních genomech 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 literatury na konci práce. Jako autor 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 jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestně právních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne **25. května 2018** ……

podpis autora

Poděkování

Děkuji vedoucí semestrální práce Ing. Denise Maděránkové, Ph.D. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé semestrální práce. Dále bych chtěla poděkovat své rodině za podporu při studiu.

V Brně dne **25. května 2018** ……

podpis autora(-ky)

Obsah

Seznam obrázků

Seznam tabulek

ÚVOD

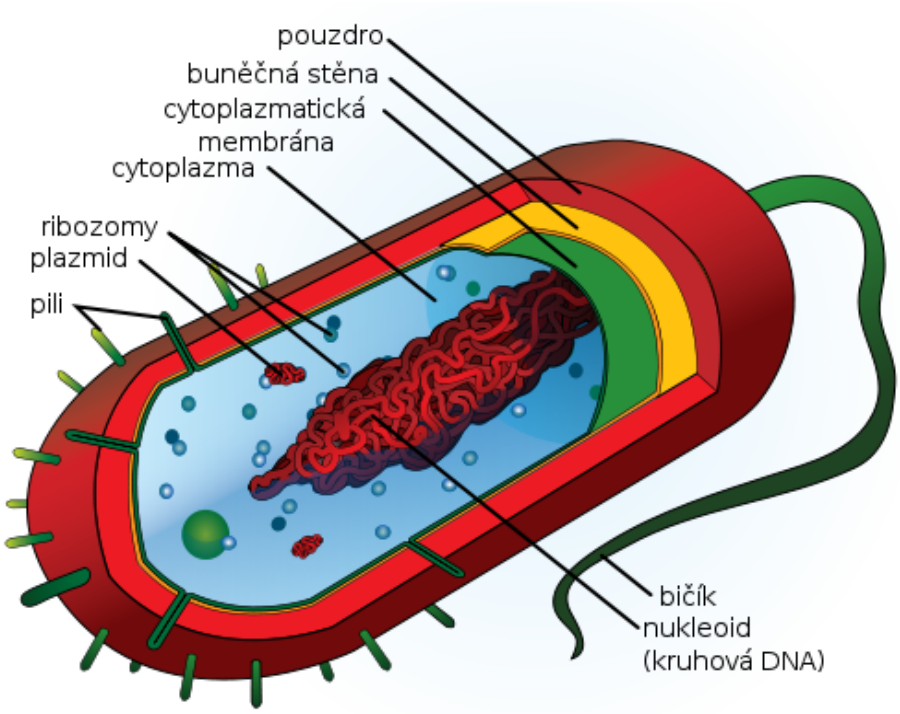
Komparativní genomika, jedno z odvětví genomiky, je obor zajímající se o podobnosti a vztahy mezi chromozomy, geny nebo genomy. Využívá se pro porovnání různých organismů, například díky komparativní genomice byla zjištěna podobnost lidského genomu s hlodavčím. Lépe se tedy zkoumají evoluční procesy, jelikož umožňuje nalézt i dosud neanotované geny. O evolučních procesech více prozradí i tzv. synteny bloky, což jsou vzájemně podobné oblasti v genomech rozdílných organismů, kde se porovnává pořadí synteny. V teoretické části práce je kromě synteny popsán i bakteriální genom, anotace genu nebo rozsáhlý formát genbank, který obsahuje velké množství informací o daném záznamu sekvence, která je v něm uložena. Teoretickou část uzavírá krátké shrnutí o použitých bakteriích, kde byla podobnost vybírána na základě jejich gramnegativního barvení.

Praktická část této práce je věnována přímému vyhledávání podobností mezi jednotlivými bakteriemi. V genbank formátu je uložena jak sekvence DNA, tak spousta dalších informací o organismu. Jelikož se nemůžeme úplně spoléhat na jednotlivé popisy genů uložené v genbank formátu, většinou nejsou kompletní, bude vyhledávání probíhat na třech úrovních. Nejprve bude porovnáván název genu referenční bakterie spolu s další bakterií, dle nalezených shod se poté zjednoduší algoritmus pro vyhledávání názvu produktu genu. Nakonec se pomocí lokálního zarovnání porovnají translace jednotlivých proteinů.

Vyhledávání shod slouží k zjištění množství změn v pořadí genů mezi referenční a vybranou bakterií. Pořadí genů je proto setříděno jedním z hladových algoritmů a zjištěn počet změn, které byly potřebné k tomu, aby byl poziční vektor seřazen. Počet kroků metody určuje evoluční vzdálenost mezi referenční a vybranou bakterií, s jeho pomocí se spočítá *p*-distance, dle ní se následně sestaví fylogenetický strom, určující podobnost bakterií.

1 BAKTERIÁLNÍ GENOM

Bakterie se zařazují mezi prokaryotní organismy. Jsou tvořeny prokaryotní buňkou s jediným chromozomem kruhovitého tvaru umístěným volně v cytoplazmě. Postrádají i složitý systém membránových organel, např. neobsahují mitochondrie a vakuoly, ovšem jejich funkce je nahrazována. Metabolické děje u bakterií probíhají volně v cytoplazmě. Molekuly DNA jsou v prokaryotických buňkách uloženy buď v hlavním chromozomu, nebo v plazmidech – malé molekuly DNA, též kružnicového tvaru (Obrázek 1.1). Bakterie se v genetice využívají hlavně kvůli rychlosti rozmnožování. Neprobíhají zde meiotické ani mitotické děje, bakterie se rozmnožují nepohlavně příčným dělením. Za krátkou dobu je tedy možné vytvořit mnohamilionové populace, kde se zvyšuje šance na pozorování velmi vzácných mutací, které by se u menších populací téměř jistě neprojevily. [1][4]



Obrázek 1.1 Stavba prokaryotické buňky

1.1 Genetická informace bakterií

Jak již bylo zmíněno, prokaryotní organismy obsahují pouze jeden hlavní chromozom, jsou tedy trvale haploidní. Na chromozomu jsou geny uspořádány v určitém pořadí. Pozice, která je určena pro daný gen, se nazývá lokus. Prokaryota obsahují jednoduché

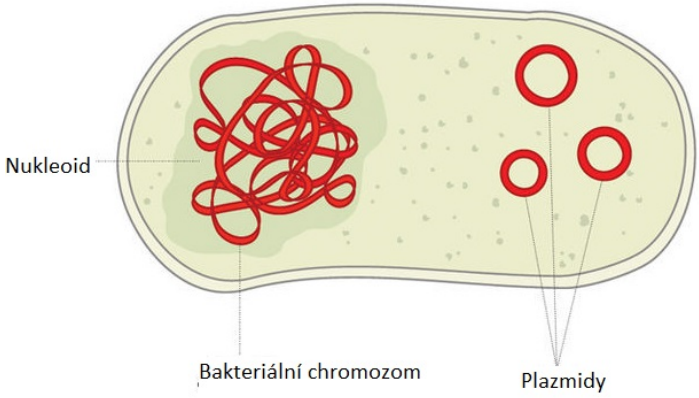
strukturní geny, které se po transkripci nemusí upravovat. Chromozom neobsahuje introny, může se tedy z funkčního hlediska celý považovat za exon. Do RNA se přepisuje jako celek a většina DNA se v něm nachází v genové formě. Jeho replikace probíhá souměrně, symetricky a z jednoho místa, kde je chromozom uchycen k cytoplazmatické membráně, tzv. OriC. Hlavním enzymem replikace je DNA polymeráza, vzniká RNA primer nutný pro iniciaci replikace a celou reakci katalyzuje DNA primáza. Úkolem DNA ligázy v průběhu replikace je spojit tvz. Okazakiho fragmenty, což jsou části již replikované DNA, které se tvoří na opožděném řetězci. Zvláštností replikace u prokaryot jsou geny, které se přepisují jako celek do jedné molekuly RNA. Tyto geny se nazývají operony a jsou společně regulovány – buď se přepisují všechny nebo žádný. [3][4][5]

Promotor je většinou počáteční úsek sekvence DNA, který zahajuje transkripci. RNA polymeráza rozezná promotor a naváže se na něj. Zahajuje se tak transkripce konkrétního genu. Na bakteriální promotor se naváže dočasně RNA polymeráza, spíše jedna z jejích podjednotek, tzv. sigma faktor. Sigma faktor umožňuje iniciaci transkripce a výběr promotoru, u bakterií je důležitým prvkem při regulaci genové exprese. Bakterie mají více sigma faktorů s rozdílnými funkcemi. Jeden je určen na provoz buňky, další umožňují bakteriím měnit transkripční program. Promotor obsahuje dvě vysoce konzervované sekvence, které sigma faktor rozeznává. Pro nejčastěji se vyskytující sigma faktor σ^{70} (RpoD), se jedná o tzv. Pribnow box v pozici -10. Pribnow box je sekvence 6 nukleotidů (TATAAT), nazýván také jako -10 sekvence, jelikož je umístěn zhruba 10 párů bazí před místem iniciování transkripce. Jeho funkce je podobná jako funkce TATA boxu u eukaryotických buněk.[3][4]

1.1.1 Plazmidy

Kromě hlavního chromozomu je genetický materiál v bakteriích uložen v menších cirkulárních DNA – plazmidech. Plazmid je genetický element, který se replikuje nezávisle na hlavním chromozomu. Není nezbytný pro přežití buňky, avšak pokud např. nese geny pro rezistenci proti antibiotiku, které bylo přidáno do růstového média, je pro ni výhodou. Plazmidy začaly vědce zajímat až ve chvíli, kdy přišli na to, že hrají důležitou roli při konjugaci, výměně genetické informace a rezistenci proti antibiotikům. Postupem času se zjistilo, že charakteristické vlastnosti bakterií využívající se v medicíně, průmyslu a životním prostředí, jsou právě díky genům, které nesou plazmidy. Díky dalšímu studiu plazmidů jako rekombinace a replikace se získaly informace o základních biologických procesech. Jsou snadno izolovatelné z bakterií, transformací mohou být znovu zavedeny do jiných buněk, proto mají zásadní význam při studiu chromozomového přeskupení v bakteriích a rostlinách. [1][2]

V buňce se vyskytují různé druhy plazmidů s rozdílnými funkcemi. R-plazmidy (rezistenční) nesou geny, které zodpovídají za rezistenci buňky k antibiotikům a jiným antibakteriálním látkám. Některé R-plazmidy podmiňují schopnost konjugace bakterií. Tato vlastnost má velký význam při šíření genů rezistentních k antibiotikům v populacích patogenních bakterií. Ovšem v medicíně způsobily R-plazmidy veliký problém. Kvůli nadměrnému používání antibiotik se mnohé bakterie staly proti nim rezistentní. [1][2][6]



Obrázek 1.2 Bakteriální chromozom s plazmidy, převzato z [6]

1.1.2 Epizomy

Epizomy mají jedinečné vlastnosti. Jedná se o genetický element postradatelný pro hostitele, replikuje se autonomně nebo může být začleněn do chromozomu bakterie. Schopnost se začleňovat záleží na přítomnosti inzerčních sekvencí (krátké úseky DNA), které jsou transponovatelné (mohou se pohybovat z jednoho chromozomu na druhý). [1]

2 ANOTACE GENOMU

Anotací genomu se rozumí snaha o nacházení genů, proteinových interakcí a dalších stejně významných oblastí v sekvenci DNA. V bioinformatice se na základě laboratorně nalezených oblastí snaží předpovídat pozice dalších oblastí pomocí výpočetních metod. U prokaryotních organismů se využívají metody založeny na znalostech vlastností promotorů. U eukaryot je použití značně složitější, jelikož strukturně je jejich genom mnohem komplikovanější.[7]

2.1 Komparativní genomika

Porovnáním kompletních genomových sekvencí pomocí různých nástrojů, včetně počítačové analýzy, se zabývá komparativní genomika. Pomocí pečlivého porovnávání charakteristik, kterými jsou definovány jednotlivé organismy, se vědcům daří určit oblasti podobností mezi jednotlivými organismy. Tyto znalosti napomáhají porozumět lidskému genomu a bojovat s lidskými onemocněními. [8][17]

Mezi hlavní úkoly komparativní genomiky se řadí identifikace sekvencí DNA, které byly konzervovány v mnoha různých organismech po dlouhou dobu. Je to totiž jeden z důležitých kroků k pochopení genomu samotnému a klíč k jeho studiu. Hledá především geny, které jsou důležité pro život buňky a organismu nebo se vztahují k různým biologickým systémům, jejichž vlastnosti se též dají využít k inovativnosti léčby různých onemocnění. [17]

Komparativní genomika je důležitým nástrojem pro studium evoluce. Analýzou evolučních vztahů mezi druhy a odpovídajících rozdílů v jejich DNA se daří porozumět, jak se časem měnil vzhled, chování a biologie organismů. V současné době se začalo komparativní genomiky využívat i v zemědělství, biotechnologiích nebo zoologii, kde se nachází rozdíly mezi druhy zvířat. [17]

3 SYNTENY

Vývoj genomu ovlivňují jak malé bodové mutace v sekvenci DNA, tak i velké události, jako je přeskupení, které přeskládá genetický materiál v buňce. Přeskupení se týká buď jen pár genů, například mutace nebo zlom v důsledku nashromáždění bodových mutací nebo tandemových duplikací. Ale může být i v mnohem větším měřítku, jako jsou dlouhé inverze nebo duplikace celého genomu. Všechny takové události jsou velice důležité v evolučních procesech a plyne z nich, že pokud porovnáme dva a více chromozomů, je velice nepravděpodobné, že pořadí genů bude totožné, dokonce i pro příbuzné druhy. Ovšem pořadí genů není náhodné a při srovnání dvou chromozomů nebo jejich částí, každý od jiného příbuzného druhu, lze nalézt shodné nebo alespoň podobné i celé genové úseky zakonzervované v určitém pořadí. Jak moc jsou si jednotlivé organismy podobné udává míra podobnosti. [12]

3.1 Mutace

Mutaci můžeme definovat jako kvalitativní nebo kvantitativní změnu v genetické informaci, považuje se za nevratný děj. Přenos genetické informace podléhá náhodným vlivům (mutagenům), např. fyzikálním, chemickým či jiným faktorům. Mezi fyzikální faktory se řadí ionizující nebo UV záření. Jako mutagen mohou působit i viry, které indukují v buňkách mutace. Mutace mohou postihnout zárodečné buňky a přenáší se tak do potomstva (gametické mutace). Pokud jsou mutace tělní a nepřenáší se do potomstva, hovoří se o mutacích somatických. Negativním dopadem mutací může dojít ke změnám v regulačních oblastech transkripce, sekvencích promotorů nebo v signálních sekvencích. [9][10]

Mutace se mohou rozdělit do tří základních skupin podle místa a organizační úrovně jejich vzniku na mutace genové, chromozomové a genomové. Mutace mohou být též slučitelné se životem (vitální) nebo se životem neslučitelné (letální). Ovšem mutace letální nemusí být vždy neslučitelné se životem, jsou to tzv. podmíněné letální mutace, kdy jsou letální pouze v jednom, restriktivním, prostředí, ale slučitelné v jiném, permisivním, prostředí. Tyto mutace jsou nejužitečnější z hlediska genetických studií, protože umožňují studie esenciálních genů. [1]

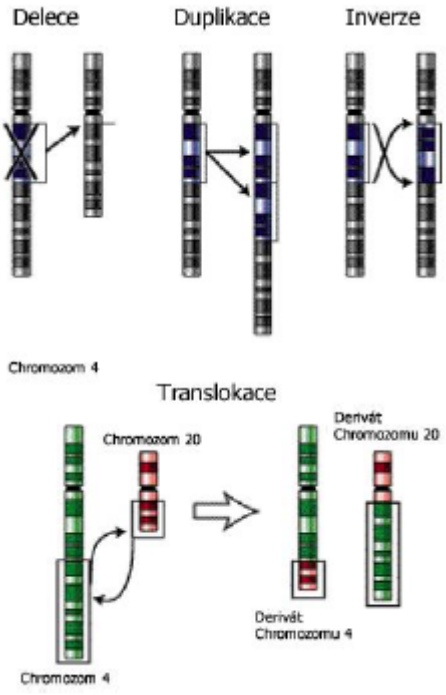
Podle okolností jejich vzniku se dělí mutace na spontánní a indukované. Spontánní mutace vznikají díky chybě při replikaci DNA. Dochází k nim bez zásahu vnějšího prostředí. DNA polymeráza je ovšem velice přesná a má samoopravnou funkci, proto je pravděpodobnost vzniku takových mutací velice malá. Z toho plyne, že naprostá většina mutací vzniká díky působení vnějších faktorů, označují se jako mutace indukované. Další definicí mutace může být změna v genotypu organismu oproti normálnímu stavu. Jsou to

náhodné změny, s cílenými mutacemi se dá setkat pouze v rámci výzkumu. Z hlediska evoluce jsou mutace velice přínosné, někdy jsou označovány jako tzv. hybná síla evoluce. [9][10][11]

3.1.1 Chromozomové mutace

Chromozomovou mutací se rozumí změna v počtu, tvaru nebo struktuře chromozomů. Obecně jsou označovány jako chromozomové aberace. Nejčastější příčinou těchto změn jsou zlomy chromozomu, kdy zlom je přerušení souvislosti DNA řetězce, jehož vinutím chromozom vzniká. Zlomy jsou způsobeny nadměrným působením mutagenů na jedince, či zhoršenou funkcí reparačních mechanismů. Následky aberací jsou závislé na tom, jestli je zachováno normální množství genetické informace, pokud ano, nazývají se změnami balancovanými. U mutací balancovaných nedochází k fenotypovým projevům. Změna fenotypu může nastat, pokud se zlom vytvoří uprostřed genu nebo mezi kódující sekvencí a jejím promotorem. K fenotypovým projevům může dojít u aberací nebalancovaných, které mají často velice závažné až letální následky.[9][11]

Chromozomové mutace se dělí na inverzi, duplikaci, translokaci a deleci. Inverze nastává, pokud se zlomený chromozomální fragment obrátí a připojí se zpět v obrácené poloze. Duplikace je opakování chromozomálního fragmentu. Při translokaci nastává transfer části chromozomu na jiné místo. Delecí se rozumí ztráta fragmentu chromozomu (Obrázek 3.1). [1]



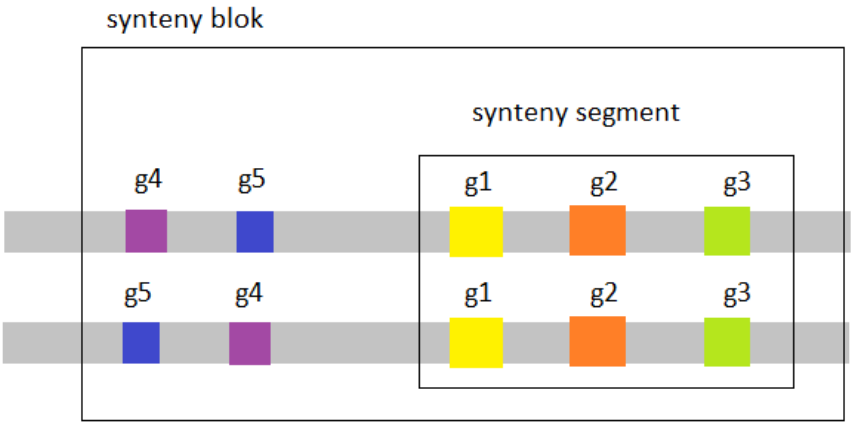
Obrázek 3.1 Příklad chromozomálních aberací

3.2 Synteny bloky

Synteny rozumíme skupinu homologních sekvencí (sekvence odvozeny od stejné původní sekvence, tzn. mají společného předka), které se ve chvíli porovnání genomů vyskytují zároveň. Pokud se porovnávají dva potomci, lze nalézt skupiny po sobě jdoucích sekvencí, ovšem jejich pořadí se může lišit.[12]

Genomové sekvenování a mapování umožnilo srovnání struktur genomů mnoha různých organismů. Zjištěním bylo, že některé organismy mají podobné genové úseky v podobných polohách v genomu. Například mnohé z genů lidí jsou syntenické s těmi jiných savců – lidoopů, myší atd. Studie synteny může ukázat vývoj genomu v průběhu evoluce. [12]

Pokud jsou místa výskytu sekvence shodné, včetně pořadí, označují se jako synteny segment. Ve chvíli, kdy je větší výskyt synteny segmentů, označuje se tento úsek jako synteny blok. V synteny bloku jsou obsaženy i inverzní a permutované synteny segmenty. V praxi je ovšem vyhledávání synteny mnohem náročnější. Jelikož segment je oblast nalezená ve dvou genomech, kde homologní geny zachovávají stejné pořadí i relativní polohu v genu v obou genomech a savčí genom je složen z 25000 genů o celkové délce $3x10^9$ párů bází. To znamená, že v průměru by segmenty, které obsahují tři geny, přesáhly 300 000 párů bází, což by pro hledání synteny bloků nebylo vhodné. Navíc je velice nepravděpodobné, že při takové délce by sekvence zůstaly naprosto shodné. [12]



Obrázek 3.2 Příklad synteny segment a synteny blok

Proto se využívá takzvaných sekvenčních kotev (anchors). Jako sekvenční kotva je definována konzervovaná neopakující se subsekvence, která se vyskytuje v porovnávaných genomech společně s určitou sekvencí. Počet sekvenčních kotev může být vyšší, než je počet genů. Předpokladem může být subsekvence f_i v genomu F shodná se subsekvencí g_i v genomu G, a zarovnání f_i ke G, dává největší shodu g_i, a naopak. Subsekvence f_i a g_i mohou být použity jako sekvenční kotvy nebo orientační značky

(landmarks), které napomáhají v porovnávání dvou genů. V zásadě může být mnohem více sekvenčních kotev, než je genů, což zpřesní rozdělení synteny bloků. Synteny bloky mohou být též definovány pomocí pozičně ukotvených sekvencí (sekvenčních kotev). Je to soubor sekvenčních kotev, které se objevují společně, ovšem ne nutně ve stejném pořadí, v genomu dvou různých organismů. [12]

3.3 Evoluční procesy a synteny bloky

Informace ohledně proběhlých evolučních procesech jsou pro lidstvo velice důležité, proto vzniká snaha o analýzu přestavby různých genomů. Součást jakékoliv analýzy je i řešení permutačních vektorů. Je to snaha o nalezení sledu permutací, jehož výsledkem je shoda v pořadí segmentů v první sekvenci s pořadím, v jakém se nalézají i v sekvenci druhé. Stanovuje se minimální počet permutací a to proto, že se nesnaží nalézt evoluční kroky, ale určit příbuznost jednotlivých druhů – čím nižší počet permutací, tím vyšší příbuznost.[12]

Permutační vektor je množina s n prvky, a každý prvek bude představovat úsek DNA (geny v podobě synteny bloku). Na příkladu, kdy $n=5$, se definuje první sekvence jako uspořádaný permutační vektor S_1 = [1 2 3 4 5] a druhá, přeházená sekvence jako S_2 = [2 4 1 5 3]. Cílem teď bude získat shodné pořadí obou vektorů pomocí transformace vektoru S_2. Využívá se metody eliminace bodů zlomu inverzí nebo transpozicí. [12]
Inverze je asi nejběžnější způsob přeskupování, kdy se část vektoru (sekvence) jednoduše převrátí. Tímto způsobem se dá též dosáhnout i uspořádání. Pokud se vezmou sousedící prvky vektoru a jejich diference je rovna 1, pak se berou jako navazující. Ovšem, pokud je diference větší, než je 1, místo se označuje jako bod zlomu. Tato metoda přidává na začátek a konec setřiďovaného vektoru záchytné body, s nimiž není možná manipulace. Počet provedených inverzí by měl být větší nebo roven polovině počtu zlomů a zároveň menší nebo roven počtu bodů zlomu ve vektoru. [12]

Eliminace bodů zlomů pomocí transpozice se zakládá na vyjmutí jednoho nebo více prvků z vektoru a jejich následné přeložení do správné pozice. Použití transpozice je výhodnější, jelikož lze odstranit až tři body zlomu najednou. Nejčastěji se ovšem využívá kombinace obou metod. [12]

4 GENBANK

GenBank je databáze všech veřejně přístupných anotovaných nukleotidových sekvencí. Databázi spravuje Národní centrum pro biotechnologické informace (NCBI), které je součástí National Institutes of Health (NIH) v USA. K databázi je umožněn přístup přes internet anebo si lze celou databázi bezplatně nainstalovat. GenBank zajišťuje nejaktuálnější obsáhlé informace ohledně nukleotidových sekvencí. Jediný problém databáze je, že do databáze mohou přispívat různí autoři a jejich příspěvky nejsou kontrolovány, proto se nelze plně spoléhat na správnost dat.
[13][14]

4.1 GenBank formát

GenBank formát slouží pro zápis biologických sekvencí z databáze NCBI, jeho přípona je *.gb. Výhodou formátu je, že může obsahovat více záznamů od různých autorů, a též spoustu doplňkových informací, což jiné, takhle využívané formáty, nedokáží. Příkladem je třeba FASTA formát. Samotný GenBank formát se skládá z hlavičky a sekvence nukleotidů. Sekvence je uvozena slovem „ORIGIN“, ukončená znakem „//“.[14][15]

V hlavičce se nachází různé užitečné informace ohledně dané sekvence. Nalezneme zde identifikační číslo sekvence, např. SCU49845, kde první tři znaky označují organismus, čtvrtý a pátý znak přesněji určuje, co sekvence kóduje. Dnes již není takové identifikační číslo dostačující pro uvedení všech informací, proto se v současné době přiřazuje identifikační číslo tak, aby bylo jedinečné pro každou sekvenci. Dále v hlavičce nalezneme délku sekvence, což je počet nukleotidových párů bází (dále jen bp). Není zde uveden žádný limit pro velikost vložené sekvence, jediné, co je limitováno, je délka záznamu na minimálně 50 bp a maximálně 350 kbp. Tento limit byl odsouhlasen pro usnadnění manipulace se sekvenčními daty pomocí softwarových programů. Po délce sekvence je uveden typ molekuly, která byla sekvenována, např. DNA, RNA, transfer RNA, ribosomal RNA. [9][10]

Následuje oddíl GenBank, v hlavičce označen třípísmennou zkratkou názvu oddílu, do které je daný organismus zařazen. Databáze GenBank má 18 oddílů, do kterých můžeme sekvence zařadit – PRI – sekvence primátů (primate sequence), MAM – ostatní savčí sekvence (other mammalian sequences), PLN – sekvence rostlin, hub a řas (plant, fungal and algal sequences), atd. Některé oddíly obsahují sekvence od různých skupin organismů, jiné obsahují data, která byla generována specifickými sekvenčními metodami z mnoha různých organismů. Datum na konci hlavičky určuje den, kdy byla provedena poslední úprava záznamu.[15]

Kromě hlavičky, jsou informace o sekvenci uloženy v metadatech, která jsou součástí formátu GenBank. Stručný popis sekvence nalezneme v části s názvem „DEFINITION“, který zahrnuje jméno organismu, název genu (popř. proteinu), pokud je sekvence nekódující, tak popis její funkce. Pokud má sekvence kódující úsek (CDS), přidává se poznámka ohledně kompletnosti, například „complete CDS“ (kompletní kódující úsek). Po popisu sekvence následuje přístupové číslo záznamu sekvencí „ACCESSION“. Obvykle je kombinací písmen a číslic, např. jedno písmeno následováno pěti číslicemi. Některá přístupová čísla mohou být delší, v závislosti na délce sekvence. V části „VERSION“ je přístupové číslo doplněno o identifikační číslo verze nukleotidové sekvence, tzn. pokud bude sekvence neupravována od jejího nahrání, bude za přístupovým číslem následovat .1, pokud bude jednou upravena, následovat bude .2, atd. Pokud se změní tohle identifikační číslo, je zároveň sekvenci přiděleno i nové identifikační číslo sekvence „GenInfo Identifier“. Každá proteinová translace má taktéž své vlastní GI číslo, které se při jakékoliv úpravě změní. [15]

V metadatech nalezneme i „KEYWORDS“ neboli klíčová slova, která popisují sekvenci. Tato klíčová slova se ovšem vyskytují pouze ve starších záznamech, v novějších pouze když o to sám autor zažádá nebo záznam obsahuje sekvenci generovanou sekvenční metodou. Dále je zde zdroj „SOURCE“, který obsahuje informace o organismu, ze kterého sekvence pochází. Nachází se tu i informace o autorech, článcích, ve kterých byly dané sekvence zmíněny autory, dále místo, kde byl článek zveřejněn. Důležitou součástí metadat jsou tzv. „FEATURES“, kde jsou informace ohledně genů, genetických produktů, významných úsecích v sekvenci, jako např. oblasti kódující proteiny, mRNA, tRNA, a řadu dalších vlastností. Ve Features se nachází shrnutí informací ohledně sekvence (její délka, zdrojový organismus a taxonomické identifikační číslo), ovšem nejdůležitější jsou informace ohledně CDS (kódující úsek), obsahuje start a stop kodon, překlad do aminokyselin, název vzniklého proteinu. [15]

5 PŘEHLED BAKTERIÍ

5.1 Escherichia coli

Escherichia coli neboli *E. coli* bude využívána jako referenční bakterie pro výpočet pozičního profilu genů vůči ostatním. *E. coli* je součást fyziologické mikroflóry tlustého střeva a nachází se též v distální části ilea. Vyskytuje se v organismu takřka od narození, nejčastěji alimentární cestou nebo přenosem od jiného jedince, který *E. coli* už osídlen. *E. coli* není schopna dlouhodobě existovat mimo hostitele.

Kmeny *E. coli* u zdravého jedince nevyvolávají onemocnění, ovšem v případě narušení poměru jednotlivých druhů mikrobů ve střevě mohou způsobit zdravotní komplikace, např. přemnožením *E. coli*. I některé patogenní kmeny mohou způsobovat závažné infekce. Kultivace *E. coli* je nenáročná, nejčastěji se využívá laktózový nebo krevní agar, na kterém roste v šedých koloniích. Enzymatické aktivity se využívá při inkubaci kolonie, jelikož umožňuje přesné zařazení druhu podle jeho biochemických vlastností. *E. coli* patří mezi gramnegativní bakterie. Databázové číslo používané *E. coli* v NCBI databázi je NC_000913. [16]

5.2 Ostatní použité bakterie

Tabulka 5.1 Příbuzné porovnávané bakterie a jejich výskyt

Jméno bakterie	Gramovo zbarvení	Místo výskytu	Databázové číslo NCBI
Klebsiella pneumoniae	G-	v ústech, fermentace laktózy	NZ_AP014950
Legionella pneumophila	G-	nemoc legionelóza	NC_018139
Salmonella enterica	G-	průjmové onemocnění	NC_003198
Salmonella bongori	G-	salmonelóza	NC_015761
Shigella flexneri	G-	shigelóza (průjmové onemocnění)	NC_004337
Yersinia pestis	G-	morová nemoc	NC_003143
Pseudomonas aeruginosa	G-	v odpadních vodách, v půdě	NC_002516
Pseudomonas syringae	G-	rostlinný parazit	NC_007005

6 VÝPOČET POZIČNÍHO PROFILU GENŮ

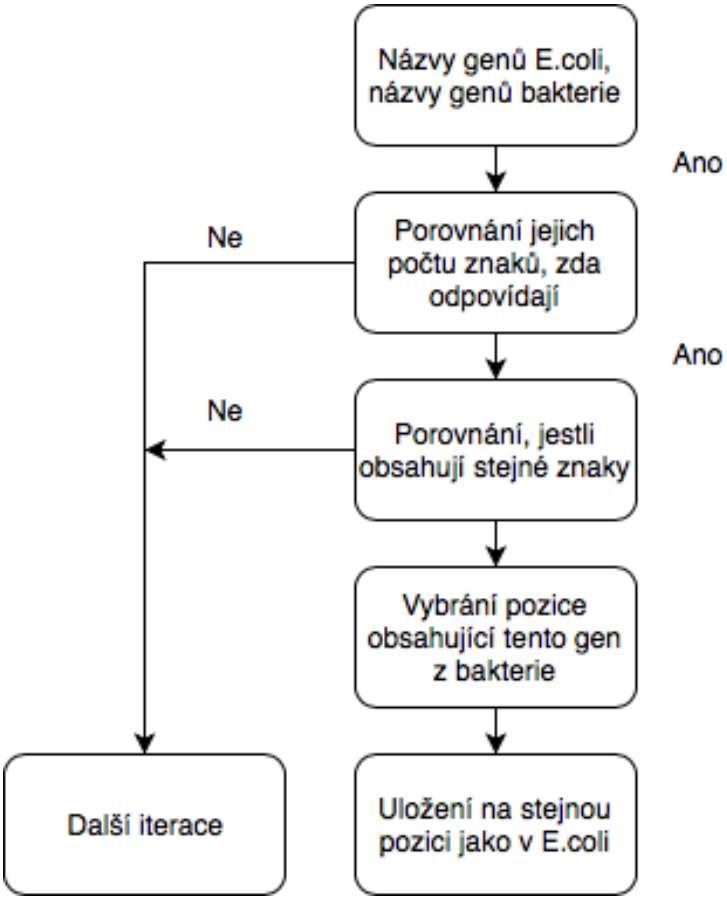
V následující části práce je popsáno porovnávání anotovaných genomů vybraných bakterií zmíněných výše, vůči referenční *E. coli*. Soubory bakteriálních genomů jsou stáhnuty z databáze NCBI ve formátu genbank. Všechny soubory s bakteriálními genomy jsou vybrány na základě označení *RefSeq*, které se nachází v části Keywords a označuje, že sekvence byla anotována pomocí anotace prokaryotického genomu NCBI, která by měla zlepšovat konzistenci celého datového souboru. Tyto sekvence jsou nadále anotovány pomocí nástrojů unikátních pro tento genom. Soubory s označením *RefSeq* by proto měly být ucelenější a kvalitnější. Některé soubory jsou jako referenční sekvence vybrány na základě dlouhodobého využívání a širokého uznání, např. referenční genom Escherichia coli str. K-12 substr. MG1655, který je využíván i v této práci. Všechny funkce budou programovány v prostředí MATLAB. [18]

6.1 Vyhledávání podobností na základě genů

Sekvence vybraných bakterií (viz Tabulka 5.1) byly stáhnuty z databáze NCBI, kde byl vybrán celý genom, ne pouze jeho části, které se dají taktéž stáhnout. Příkazem *genbankread* byly nahrány jednotlivé sekvence do prostředí Matlab, ovšem formát genbank obsahuje velké množství v tuto chvíli nepotřebných informací. Potřebné pouze informace o názvech genů. Jelikož referenční *E. coli* obsahuje přes 4000 CDS, byl program nejprve natrénován pouze na části genomu – prvních 50 genů. Genbank formát neobsahuje všechny názvy hledaných genů, proto pro lepší práci s algoritmem byla prázdná pole nahrazena pouhým znakem ‚X‘.

Po detailnějším prozkoumání načteného souboru byly nalezeny shodné názvy některých proteinů, označeny jako 'insertion element IS1 protein InsB' a 'insertion element IS1 protein InsA', zkráceně InsB a InsA. Jelikož měly tyto geny stejné označení, stejný produkt a shodné sekvence aminokyselin, byly ze souboru zcela odstraněny. Jejich podobnost by totiž zkreslovala výsledek vyhledávání.

Hlavním úkolem algoritmu bylo vyhledat geny, které jsou obsaženy jak v *E.coli*, tak v ostatních bakteriích. Ovšem spolu se shodou je důležitá i informace o pozicích genů, která pomůže určit, jak moc jsou si dané bakterie příbuzné (Obrázek 6.1). Proměnná *bakterie* označuje uživatelem vybranou bakterii, která se má spolu s *E.coli* porovnávat.



Obrázek 6.1 Blokové schéma cyklu pro vyhledávání shody v názvech genů

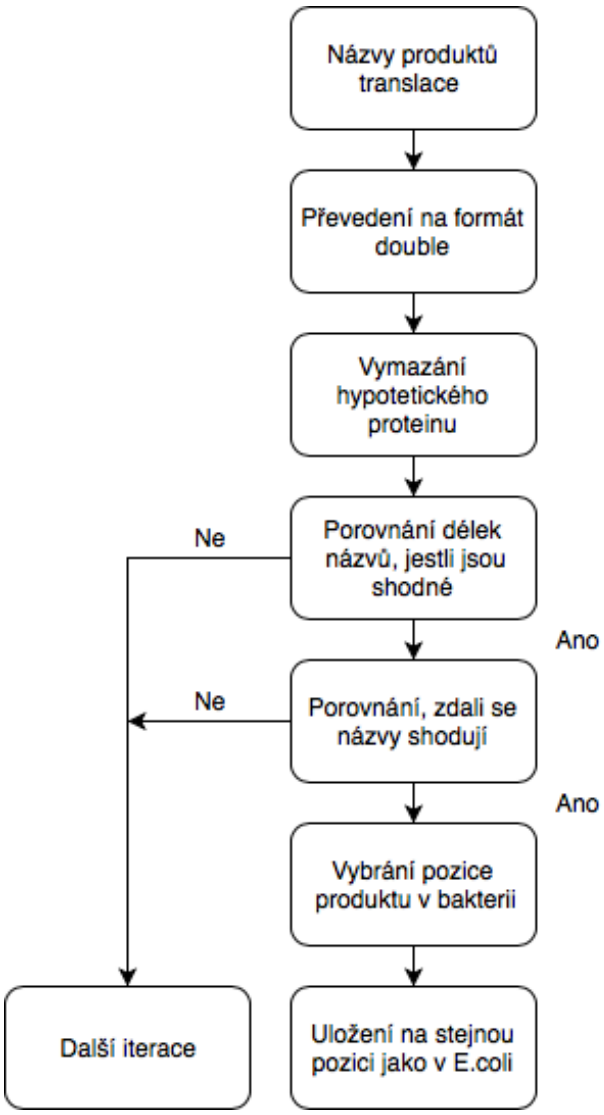
Algoritmus pro vyhledávání shody názvů genů funguje na základě porovnávání, nejprve délek daných slov – to proto, aby se zjednodušila výpočetní náročnost algoritmu, protože pokud název genu z porovnávané bakterie bude mít rozdílnou délku než název genu *E.coli*, cyklus přejde na další iteraci. Z cyklu jsou též odstraněny i ona ‚X', pro jejich nepotřebnost při porovnávání. Poté se porovnává přímo obsah daných proměnných. Výsledkem funkce je dvouřádkové buňkové pole, kde na prvním řádku jsou názvy genů dle referenční *E.coli* a na řádku druhém pozice jednotlivých genů dle vybrané bakterie.

6.2 Vyhledávání podobností na základě produktu translace

Vyhledávání podobností na základě produktu translace bylo o něco složitější než hledání shody v názvu genu. Pro zjednodušení byla hned ze začátku funkce vytvořena matice, do jejíhož prvního sloupce byly ukládány jedničky nebo nuly, dle toho, zdali se nalezla podobnost v porovnání názvu genů. Pokud se podobnost nalezla a v matici se na daném místě nachází jednička, je z dalšího zpracování tento gen vyřazen. V produktech translace souborů se často nacházel i tzv. hypothetical protein, jehož existence byla predikována,

ovšem ještě neexistuje její experimentální důkaz. Je tedy těžké přiřadit i jeho funkci. Ze souboru byl též odstraněn, ale ne zcela, pouze jen jako název produktu translace.

Názvy produktů jsou pro svou délku v Matlabu uloženy ve formátu *char* (character array). Tento formát není pro porovnávání znaků úplně vhodný, pro zjednodušení byly tedy všechny názvy převedeny na formát *double* pomocí stejnojmenné funkce. Funkce *double* nám jednotlivá písmena ve slově převede na čísla, tím pádem je pak porovnání snazší. Nadále se postupuje podobně jako v předchozím algoritmu. Nejprve jsou porovnány délky názvu u referenční a zkoumané bakterie, poté se porovnává, jestli se shodují i v obsahu (Obrázek 6.2). Výstupem funkce je dvouřádkový vektor, který obsahuje informace o názvu produktu a jeho pozici v porovnávané bakterii.



Obrázek 6.2 Blokové schéma cyklu pro vyhledávání shody v názvu produktu

6.3 Vyhledávání podobností na základě translace

Vytvořená pomocná matice jedniček a nul v této části je velmi nápomocná pro zlepšení výpočetní náročnosti. Jedničky zde reprezentují, že daný gen byl nalezen buď porovnáním jeho názvu nebo názvu jeho produktu translace. Proto nalezené shody budou vynechány z algoritmu lokálního zarovnání.

Smithův – Watermanův algoritmus provádí lokální zarovnání, což je hledání nejpodobnějších úseků různých délek mezi dvěma sekvencemi. Oblasti, které jsou od těchto úseků vzdáleny, nejsou při zarovnání brány v potaz. Lokální zarovnání povoluje mezery, je to algoritmus dynamického programování, který sestavili pánové Temple F. Smith a Michael S. Waterman v roce 1981. Základem lokálního zarovnání je Needleman-Wunschův algoritmus pro globální zarovnání, ovšem lokální zarovnání pokládá negativní hodnoty rovny nule. [21] Jako skórovací matice, která penalizuje vložení mezery, byla použita v Matlabu pro proteiny defaultně nastavená BLOSUM50. Algoritmus je ovšem výpočetně náročný, proto jakékoliv zjednodušení je pro další zpracování dat velice výhodné.

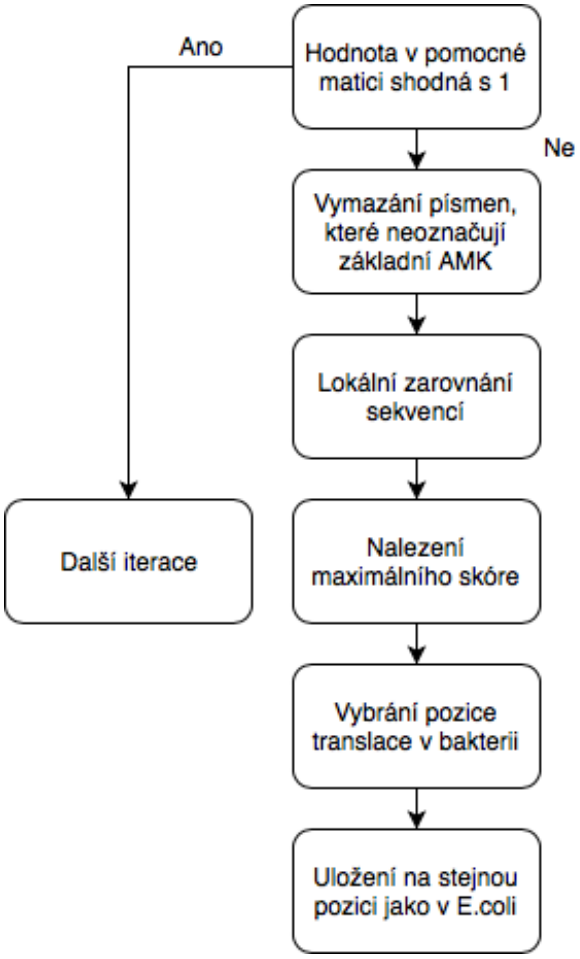
Když je shoda nalezena v názvech genů nebo názvech produktů, dále se nepoužívají. Vybírají se pouze sekvence aminokyselin, u kterých se provádí lokální zarovnání pomocí funkce *swalign*. Výstupem této funkce je zarovnání sekvencí a jejich skóre. V matici, kde se nalézá skóre se najde maximální hodnota, která určuje podobnost sekvence aminokyselin *E.coli* k sekvenci aminokyselin, která byla vzata z porovnávané bakterie. Maximální hodnoty skóre se mohou na první pohled zdát velice odlišné, např. jedna maximální hodnota se pohybuje okolo 50, druhá okolo 300. Může se vše na první pohled zdát jako náhodná shoda nebo špatné vyhodnocení algoritmu, ovšem záleží na délce sekvence, penalizaci mezer a jejich prodloužení.

Kvalita sekvencí z NCBI databáze někdy neodpovídá jejich označení RefSeq, proto mohou být výsledky vyhledávání zkreslené. Ve chvíli, kdy se najde malá shoda v názvech genů nebo produktů (protože většina sekvencí tyto informace ani neposkytuje), tak algoritmus hledá většinu shod pomocí zarovnání translací. V tuto chvíli vzniká největší šance na výskyt náhodné shody, kdy je vyhodnocováno maximální skóre, ovšem muselo by se ručně kontrolovat, zdali jsou si sekvence skutečně podobné nebo jen nastala náhodná shoda. Kvůli náhodným shodám nastane v některých záznamech bakterií situace, že pro dvě sekvence aminokyselin *E.coli*, se nalezne maximální skóre u jedné a té stejné sekvence aminokyselin bakterie.

Nejlepší výsledky porovnání byly pro bakterii *shigella flexneri*, která obsahovala dostatečné množství informací o názvech genů a o jejich produktech. Shoda se našla v 2398 názvech genů a 2047 názvech produktů. Čím více je nalezených shod v genech

a produktech, tím je algoritmus výpočetně náročnější, protože není potřeba zarovnávat tolik sekvencí.

Jelikož programovací prostředí Matlab umí rozpoznat v sekvenci aminokyselin pouze základních 20 označení pro aminokyseliny, ostatní znaky, ač též jsou rozšířením označení pro aminokyseliny, Matlab nedokáže vyhodnotit a vypíše chybovou hlášku. Proto bylo potřebné tyto znaky ze sekvence odstranit.



Obrázek 6.3 Blokové schéma cyklu pro lokální zarovnání sekvencí AMK

7 POROVNÁNÍ POZIČNÍHO PROFILU GENŮ

Pro porovnání pozičního profilu genů je třeba využití greedy algoritmů, kdy greedy se dá přeložit jako hladový. Použití hladových algoritmů může být při porovnávání sekvencí DNA, které se odlišují pouze sekvenčními chybami, mnohem rychlejší než dynamické programování. I tak je jejich výsledek relativně optimální. Hladové algoritmy na rozdíl od ostatních pracují s rozdíly mezi jednotlivými sekvencemi. Nejjednodušeji se dají popsat jako algoritmy počítající četnost rozdílů. Vzdálenost mezi sekvencemi je poté definována jako minimální počet rozdílů v jakémkoliv jejich zarovnání. [19]

Když si vezmeme příklad, na kterém by se hladová strategie dala aplikovat: pokladník uvažuje v každém kroku pouze největší denominaci menší nebo rovnu M. Jelikož cílem bylo minimalizovat počet mincí, které se vrátily zákazníkovi, zdá se to jako velice dobrá strategie. Přeci by se nevrátily nikdy čtyři pětikoruny místo jedné dvacetikoruny. Tento příklad použil to, co se zdálo jako nejlepší volba a nepovažoval ostatní možnosti. Což je to, co dělá tento algoritmus „hladovým“. Společná vlastnost hladových algoritmů je, že často přináší neoptimální výsledky, za to trvají velmi krátce. Pokladník uvažuje v každém kroku pouze největší denominaci menší nebo rovnu M. [20]

7.1 Breakpointová metoda

Breakpointová metoda se řadí mezi hladové algoritmy. Minimalizuje body zlomu mezi synteny bloky. Bod zlomu se umisťuje mezi dva bloky (geny), které na sebe nenavazují. Využití počtu zlomů vede k lepšímu algoritmu pro třídění reverzí, protože produkuje řešení bližší optimálnímu než např. metoda maximálního prefixu.

Permutace P_1 až P_n se prodlouží o hodnoty $P_0=0$ a $P_{n+1}=n+1$ na konec. Tyto hodnoty svou polohu v průběhu třídění nikdy nemění. Pokud sousední prvky splňují podmínku, že P_i a P_{i+1} (kdy $0 \leq i \geq n$) jsou po sobě jdoucí čísla, nazývají se přilehlými (adjacency), pokud ne, nazývají se body zlomu (breakpoint).

0 2 1 3 4 5 8 7 6 9

Obrázek 7.1 Příklad permutačního vektoru s přidanými pevnými hodnotami na obou koncích

Permutace na obrázku 7.1 má 5 hodnot přilehlých (2 1, 3 4, 4 5, 7 6) a čtyři body zlomu (0 2, 1 3, 5 8, 6 9). Permutace může mít maximálně $n+1$ bodů zlomu, také nemusí obsahovat žádný, a to v případě, že permutace shodná. Každý bod zlomu odpovídá dvojici prvků permutačního vektoru P_i a P_{i+1}, kde sice tyto dva prvky spolu sousedí v permutačním vektoru P, ale v setříděné matici tomu tak být nemá. Proto jsou

nenásledné prvky v třídícím procesu odděleny od sebe a setříděny. Tímto způsobem se dá pozorovat třídění reverzí jako proces odstraňování zlomových bodů. Každá reverze může vést k eliminaci nejvýše dvou breakpointů, což znamená, že:

$$d(P) \geq \frac{b(P)}{2}, \qquad (7.\ 1)$$

kde *b(P)* je počet bodů zlomu v *P*. Při setřiďování je vhodné si definovat části permutačního vektoru, které se nachází mezi dvěma po sobě jdoucími body zlomu. Například na obrázku 7.1 se nachází pět takovýchto úseků (0, 2 1, 3 4 5, 8 7 6, 9). Tyto úseky se nadále mohou hodnotit jako vzestupné (3 4 5) nebo sestupné (2 1, 8 7 6). Jednoprvkové části by se dali klasifikovat buď jako vzestupné nebo i sestupné, ovšem je vhodné je klasifikovat jako sestupnou část, s výjimkou krajních hodnot (0 a *n+1*), které budou vždy rostoucí.[22]

7.2 Breakpointová metoda v Matlabu

Breakpointová metoda byla použita pro setřídění pozic genů bakterií, kdy se z CDS bakterie vybrala položka *indices*, která představuje počáteční a konečnou pozici sekvence genu. Vybrala se pouze počáteční pozice, pro zjednodušení metody se neuvažovalo ani to, zdali se gen nachází na hlavním nebo komplementárním vlákně.

Breakpointová metoda není pouze obyčejná setřiďovací metoda, která by vektor pořadí genů vzestupně seskládala. Nejdůležitější vlastností a výstupem naprogramované breakpointové funkce je, že počítá počet potřebných kroků k seřazení vektoru. Tyto kroky v tomto případě znázorňují evoluční vzdálenost mezi *E.coli* a vybranou bakterií. Čím menší bude počet kroků, tím více jsou si bakterie příbuzné.

V první fázi metody se provedlo nalezení vzestupných částí pozičního vektoru, kdy se v pomocném vektoru ukládaly na jednotlivá místa hodnoty jedna nebo nula podle toho, zda byla část vzestupná (1) nebo sestupná (0). Pomocí další funkce se nalezla pozice první nesetříděné hodnoty. Tyto hodnoty poté vstupují do hlavní funkce, která provádí setřídění pomocí breakpointů. Výpočet probíhá ve *while* cyklu, který zajišťuje, že na výstupu funkce bude seřazený vektor. Je totiž podmíněno, že *while* cyklus bude probíhat až do chvíle, než se vektor pozic bude rovnat uměle vytvořenému vektoru o stejné délce se seřazenými čísly.

Funkce pomocí vytvořeného vektoru jedniček a nul, nalezne hodnoty, které se nachází na stejné pozici v permutačním vektoru. Tyto pozice dále vyhodnotí a dle potřeby upraví. Úpravy probíhají pomocí přetočení sestupných částí a jejich zařazení na správnou

pozici. Tento algoritmus pro základní breakpointovou metodu ovšem není úplně ideální pro data, která jsou více, náhodněji zpřeházená. Jako trénovací a testovací data byly použity vektory náhodných čísel, poté už vektory s pozicemi uměle přeházenými. Při použití těchto vektorů, algoritmus pracoval, jak měl. Na obrázku 7. 2 je uveden příklad na výše zmíněné sekvenci z obrázku 7. 1:

```
>> pocet_kroku = breakpoint_final([2 1 3 4 5 8 7 6])

pocet_kroku =

     1

     0     1     2     3     4     5     8     7     6     9


pocet_kroku =

     2

     0     1     2     3     4     5     6     7     8     9

     1     2     3     4     5     6     7     8


pocet_kroku =

     2
```

Obrázek 7.2 Příklad výstupu breakpointové metody, Command Window Matlab

Z obrázku lze vidět, že jak v prvním, tak ve druhém kroku metoda přetočila hodnoty na pozicích (1 2) a (6 7 8), jelikož úseky byly správně vyhodnoceny jako sestupující. Počet změn, které byly potřeba k setřídění permutačního vektoru je roven 2. Takové výsledky ovšem nenastaly při zpracování reálných vektorů. Reálné vzorky jsou mnohem delší a více přeházené. Bohužel jsou přeházeny až na tolik, že jednoduchý třídící algoritmus na ně nestačí. Pozice vektorů ve většině případů jsou ve vektoru uloženy osamoceně, bez jakékoliv návaznosti na pozice uloženy okolo něj. Bylo by nutné metodu optimalizovat, aby brala v potaz i tyto hodnoty a přesunovala je na jejich správné pozice. Tato optimalizace by byla časově náročná a změnila by podstatu breakpointové metody. Protože ve chvíli, kdyby byla vytvořena taková funkce optimalizace, která by nalézala tyto odlehlé hodnoty a přesunovala je zpět na místo, trvala by velice dlouho, kvůli tomu, že když by nalezla odlehlou hodnotu a přesunula ji na svůj správný index, není zaručeno, že není přesunuta opět mezi dvě další pozice, se kterými nesouvisí. Funkce by tedy měla velké množství opakování a naprosto by změnila chování a výstup breakpointové metody.

Odlehlost hodnot je způsobena hledáním shod pomocí translace, kdy, jak už bylo zmíněno, není možné říct bez zásahu a nahlédnutí, zdali jsou sekvence přiřazeny správně nebo nastala náhodná shoda. Protože právě tyto náhodné shody mohou způsobit to, že geny nemají vzestupné a sestupné části. V práci se místo těchto hodnot používaly vektory náhodných čísel se stejnou délkou jako je délka bakterie.

Tabulka 7.1 Náhodně vytvořené počty kroků pro jednotlivé bakterie

Název bakterie	Počet kroků
Shigella flexneri	4045
Klebsiella pneumoniae	5319
Salmonela bongori	4256
Salmonella enterica	4106
Pseudomonas syringae	5085
Pseudomonas aeruginosa	5567
Yersinia pestis	3796
Legionella pneumopila	3101

V tabulce 7. 1 se nachází počty kroků jednotlivých bakterií využitých dále v této práci. Poziční vzdálenosti je vytvořena tak, že je vytvořen vektor již seřazeny a pomocí příkazu *randperm* roztřízen. Z toho plyne, že pro každé spuštění cyklu bude počet kroků rozdílný, proto je přiložena tabulka, která obsahuje hodnoty dále zpracovávaných počtů kroků.

7.3 Proporcionální vzdálenost

Proporcionální vzdálenost neboli *p*-distance, udává počet mutací v sekvenci úměrně k délce sekvence. Vzorec pro její výpočet je následující:

$$p = \frac{n_p}{n} \qquad (7.\ 1)$$

kde n_p je počet mutovaných pozic dvou sekvencí a n je počet všech pozic. Čím menší je *p*-distance, tím více jsou si sekvence podobné. V případě této práce se jako n_p bere počet kroků provedených v breakpointové metodě a n je délka sekvence bakterie. V následující tabulce lze vidět počet kroků metody, délku sekvencí a jejich proporcionální vzdálenost.

Tabulka 7.2 Seznam hodnot pro jednotlivé bakterie

Název bakterie	Počet kroků	Délka sekvence	*p*-distance
Shigella flexneri	4045	4051	0.9985
Klebsiella pneumoniae	5319	5327	0.9985
Salmonela bongori	4256	4267	0.9974
Salmonella enterica	4106	4110	0.9990
Pseudomonas syringae	5085	5089	0.9992
Pseudomonas aeruginosa	5567	5572	0.9991
Yersinia pestis	3796	3798	0.9995
Legionella pneumopila	3101	3115	0.9955

Proporcionální vzdálenost u náhodných dat je téměř shodná pro všechny bakterie, ovšem na reálných datech by byla mnohem rozmanitější, jelikož si sekvence nejsou tolik podobné.

8 SESTAVENÍ FYLOGENETICKÉHO STROMU

Fylogenetika je vědním oborem, který se zabývá studiem příbuzenských vztahů mezi organismy (zkoumá evoluci). Odmítá hierarchii a místo toho se snaží naleznout skutečnou příbuznost. Tyto příbuzenské vztahy analyzuje tak, že zkoumá vznik a vývoj jednotlivých vývojových linií (taxonů). Pomocí fylogenetických stromů zobrazuje pořadí a způsob větvení taxonů v průběhu evoluce. Základním předpokladem fylogenetiky je, že všechny organismy pochází z jednoho společného předka. Fylogeneze je vývoj druhu v evolučním procesu. Je to historický proces, který nelze přímo pozorovat, ale rekonstruuje se na základě evolučních teorií. Fylogeneze je předmětem studia fylogenetiky. Grafickým znázorněním vzájemných vztahů mezi organismy jsou fylogenetické stromy. [23]

8.1 Fylogenetické stromy

Fylogenetickým stromem se rozumí grafické znázornění taxonů a jejich příbuzenských vztahů. Místo taxonů mohou být vyobrazeny i jednotlivé geny či biologické skupiny a jejich podobnosti se posuzují na základě morfologické či genetické podobnosti. Vrcholy stromu, které jsou spojeny hranami se dvěma a více dalšími vrcholy se označují jako vnitřní vrcholy, ty, které jsou spojeny jen s jedním dalším vrcholem, se nazývají listy. Každý vrchol představuje taxon, hrana vztah mezi nimi. Podle typu fylogenetického stromu může délka hrany udávat dobu vývoje nebo míru podobnosti mezi taxony.

Fylogenetické stromy se dají rozdělit na nezakořeněné a zakořeněné. Nezakořeněné stromy vykreslují vztahy mezi taxony, ale nezabývají se specifikací společného předka. Zakořeněné stromy mají jeden vrchol, který se označí jako kořen a je brán jako společný předek zobrazených taxonů. Každý z vrcholů označuje předchůdce, listy zase reálné taxony.

Na základě znalostí molekulární biologie je možno porovnávat taxonomické jednotky, např. pomocí informací o příslušných sekvencích DNA a aminokyselin. Díky nim se dají zjistit genetické vzdálenosti mezi jednotlivými dvojicemi taxonů. Pomocí zarovnání sekvencí se zjistí, jak moc se od sebe dané sekvence liší – procentuální odlišnost bází mezi sekvencemi. Existují různé metody konstrukce fylogenetických stromů, použitá UPGMA metoda je popsána v následující podkapitole. [25]

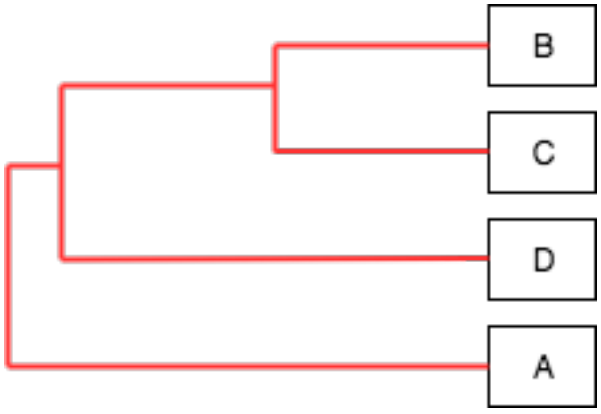
8.2 UPGMA

UPGMA (Unweighter Pair Group Method with Arithmetic mean) je nejjednodušší metoda konstrukce fylogenetických stromů, nazývá se též distanční shlukovací analýza. UPGMA pracuje s distanční maticí, která spojuje taxony na základě minimální vzdálenosti v distanční matici. Mějme matici distancí (Tabulka 8.1), která se skládá z taxonů A, B, C a D. Hodnoty v jednotlivých polích udávají genetické vzdálenosti. [24]

Tabulka 8.1 Matice distancí – příklad

	A	**B**	**C**	**D**
A	-			
B	0.5	-		
C	0.45	0.15	-	
D	0.55	0.4	0.35	-

Nalezne se nejmenší hodnota distance pro dva taxony a spojí se dohromady. V tomto případě je to B a C. Pomocí aritmetického průměru se distanční matice přepočítá (vzdálenost A ku BC a vzdálenost D ku BC) a B a C se spojí dohromady. Další krok je obdobný, kdy se opět najde minimální hodnota distance, která bude mezi BC a D a opět se spojí dohromady. Jelikož zbyl už jen jediný taxon A, připojíme ho jako poslední. Výsledný fylogenetický strom bude vypadat následovně:



Obrázek 8.1 Fylogenetický strom příklad

Délky větví se zjistí tak, že se hodnota prvního minima vydělí dvěma, a to je vzdálenost každé větve k uzlu (v příkladu BC = 0.15, délka B k uzlu je 0.075). Dále se pokračuje podobně, Délka větve D(BC) je spočítána tak, že se vzdálenost D(BC)/2 odečte

od vzdálenosti BC/2. Větev D k uzlu má poté délku D(BC)/2. Stejným principem je pak dopočítán i zbytek.

UPGMA je úsporným výpočetním algoritmem, který se dá aplikovat i na větší množství sekvencí. Má i spoustu nevýhod, například minimální evoluce nemusí odpovídat minimální délce větví, může také vytvářet stromy s nesprávnou topologií. [24]

9 DISKUZE VÝSLEDKŮ

Výsledkem práce měla být zjištěna podobnost jednotlivých bakterií k referenční *E. coli*. Vyhledávání pomocí názvu genů je relativně rychlá výpočetní metoda, ovšem většina bakterií neobsahovala informace o názvech genů. Nejlepší výsledky byly pro bakterii *shigellu flexneri*, kde se nalezlo přes 2000 shod z celkových 4051. Bohužel u ostatních bakterií shoda nebyla tak vysoká a pohybovala se v rozmezí 200–1000.

Vyhledávání shody pomocí názvu produktu translace bylo úspěšnější, na výstupu algoritmu bylo nalezeno více shod než u algoritmu pro vyhledávání názvu genů, ovšem i tak největší část nalezení shodných genů spočívala v lokálním zarovnání sekvencí aminokyselin, která byla časově náročná. Průměrná doba spuštění algoritmu pro všechny tři úrovně vyhledávání byla 5 hodin. Nalezené poziční vektory byly dány na vstup funkce *breakpoint*, ovšem ani po delší době funkce nebyla schopna vektor setřídit, kvůli špatné kvalitě pozičních vektorů. Tyto vektory obsahovaly velmi malé množství sestupných a vzestupných částí, spíše připomínaly náhodně vytvořený vektor. Proto je výstup metody zobrazen na náhodných vektorech pro ilustraci principu této funkce.

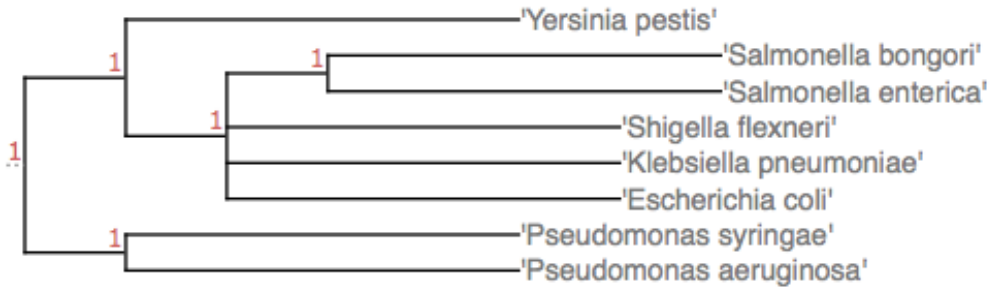
Podobnost bakterií k *E.coli* se ale dá odhadnout i z pozičních vektorů, kdy se dá pozorovat, že některé bakterie mají přece jen méně změn v pořadí genů než jiné. Na obrázku níže (Obrázek 9.1) je zobrazeno prvních 20 hodnot vektoru pozic pro tři různé bakterie, ze kterého se dá podoba s *E. coli* odhadnout.

```
salmonella_enterica =

  Columns 1 through 8

         1           2           3           4        2353           5           6           8

  Columns 9 through 16

         9          10          11          12          13          15          32          33

  Columns 17 through 20

      3272          36          37        3869


shigella =

  Columns 1 through 8

         1           2           3           4           6           7        4565           8

  Columns 9 through 16

         9          10          11          12          13          15        3273        1967

  Columns 17 through 20

      2743          17          18        4437
```

```
yersinia =

  Columns 1 through 8

       607         609         610        3502         613        2206         615         617

  Columns 9 through 16

       625        4240        4269         626         628        3707        1009        4190

  Columns 17 through 20

       631         633         858        2048
```

Obrázek 9.1 Prvních 20 hodnot pozičních vektorů bakterií *samonella enterica, shigella flexneri a yersinia pestis*

Z těchto hodnot je patrné, že bakterie *salmonella enterica* a *shigella flexneri* jsou *E. coli* mnohem více podobné než *yersinia pestis*. Tato podobnost značí, že jsou si evolučně bližší (jsou příbuznější), i přes to, že všechny bakterie pochází z jednoho řádu *Enterobacteriales.* Ze stejného řádu pochází skoro všechny vybrané bakterie, až na *pseudomonas aeruginosa* a *pseudomonas syringae*, s nimi patří pouze do stejné třídy, řád už se liší. Na následujícím obrázku (Obrázek 9.2) je vyobrazen fylogenetický strom, vytvořen pomocí NCBI funkce *Common taxonomy tree*, která po zadání jednotlivých organismů ukáže strom jejich příbuznosti. Tato funkce ovšem poskytne soubor ve formátu *.phylip, lze ovšem využít různé internetové nástroje pro zobrazení fylogenetických stromů.



Obrázek 9.2 Referenční fylogenetický strom vytvořený pomocí funkce *Common taxonomy tree* na internetových stránkách NCBI

Dle referenčního stromu lze posoudit, že předchozí odhad příbuznosti pomocí pozičního vektoru byl správný. K *E. coli* je bakterie *shigella flexneri* nejpodobnější, zatím co *yersinia pestis* a *pseudomonas syringae* a *aeruginosa* nejméně.

Metoda UPGMA byla v textu zmíněna proto, že pomocí ní měl dělat strom s výslednými *p*-distancemi. Hodnoty *p*-distancí z reálných pozičních vektorů bez řádné

optimalizace breakpointové metody nejsou dostupné, tudíž nebylo možné tento strom vytvořit.

ZÁVĚR

V této bakalářské práci bylo hlavním cílem vytvoření algoritmu pro setřídění pozičních vektorů bakterií a zjištění jejich podobností. V teoretické části práce je popsán bakteriální genom, synteny bloky, mutace a obsah formátu genbank. V další kapitole se nachází informace o referenční bakterii a je popsán vytvořený data set vybraných bakterií, které měly být nějakým způsobem příbuzné. V tomto případě jsou všechny bakterie gramnegativní. Jako referenční bakterie byla zvolena *Escherichia coli*, jedna z nejvíce prozkoumaných bakterií, vyskytující se ve střevní mikroflóře. Dále byly vybrány bakterie náhodně, jen s podmínkou gramnegativního zbarvení. Byly vybrány bakterie způsobující různá onemocnění (*shigella, salmonella, yersinia*) nebo i bakterie vyskytující se v odpadních vodách nebo parazitující na rostlinách.

Další část práce se zabývá vytvořením algoritmu pro výpočet pozičního profilu genů vybraných genomů bakterií vůči referenčnímu genomu. Zvoleny byly tři úrovně vyhodnocování podobností, na základě informací obsažených v genbank záznamu sekvencí vybraných bakterií. První úroveň představovala vyhledávání na základě shody v názvu genu, druhá pak shodu na základě názvu produktu translace. Třetí úroveň hledání shody byla pomocí lokálního zarovnání, ze kterého byla poté vybrána maximální hodnota skóre a určena podobná sekvence.

Na výstupu funkce pro vyhledávání shody je neseřazený poziční vektor, který se pomocí hladového algoritmu, konkrétně breakpointové metody, dále setřídí. Setřídění reálných dat neproběhlo, jelikož breakpointová metoda měla pouze základní vlastnosti, které nepředpokládaly složitost nesetříděného pozičního vektoru. Proto je třídění ilustrováno alespoň na příkladu.

Následující kapitola se věnuje fylogenetice, fylogenetickým stromům a jejich konstrukci, více je zde rozepsána metoda UPGMA. Pomocí fylogenetického stromu se měla zjistit podobnost bakterií vůči referenční *E.coli*. Jelikož výsledky metody pro reálné vektory nejsou dostupné, je v práci zařazen alespoň referenční fylogenetický strom, vytvořený pomocí nástrojů dostupných v databázi NCBI.

Práce by mohla být rozšířena o vytvoření optimalizace breakpointové metody a vylepšení algoritmu hledání náhodné shody, která by minimalizovalo výskyt náhodné shody, především u translace. Také dataset bakterií by mohl být ještě více rozšířen.

Literatura

[1] SNUSTAD, D. Peter a Michael J. SIMMONS, RELICHOVÁ, Jiřina, ed. *Genetika*. Přeložil Anna MATALOVÁ. Brno: Masarykova univerzita, 2009. ISBN 978-80-210-4852-2.

[2] EDITED BY DONALD R. HELINSKI, STANLEY N. COHEN, DON B. CLEWELL, DAVID A. JACKSON a Alexander HOLLAENDER. *Plasmids in Bacteria.* Boston, MA: Springer US, 1985. ISBN 9781461324478.

[3] GEORGE P. RÉDEI. *Encyclopedia of genetics, genomics, proteomics, and informatics*. 3rd ed. New York: Springer, 2008. ISBN 9781402067532.

[4] *Genetika – Biologie* [online]. [cit. 2018-01-27]. Dostupné z: http://www.genetika-biologie.cz/prokaryota

[5] KAGUNI, Jon M. DnaA: Controlling the Initiation of Bacterial DNA Replication and More. *Annual Review of Microbiology* [online]. 2006, **60**(1), 351-371 [cit. 2018-01-30]. DOI: 10.1146/annurev.micro.60.080805.142111. ISSN 0066-4227. Dostupné z: http://www.annualreviews.org/doi/10.1146/annurev.micro.60.080805.142111

[6] *Bacterial DNA – the role of plasmids* [online]. [cit. 2018-01-03]. Dostupné z: https://www.sciencelearn.org.nz/resources/1900-bacterial-dna-the-role-of-plasmids

[7] CVRČKOVÁ, Fatima. *Úvod do praktické bioinformatiky*. Praha: Academia, 2006. ISBN 80-200-1360-1

[8] *Aktuální genetika - Multimediální učebnice lékařské biologie, genetiky a genomiky* [online]. 2006 [cit. 2018-01-03]. Dostupné z: http://biol.lf1.cuni.cz/ucebnice/komparativni_genomika.htm

[9] KOLEKTIV, Oldřich Nečas a. *Obecná biologie pro lékařské fakulty*. 3., přeprac. vyd., V nakl. H. Jinočany: H, 2000. ISBN 80-86022-46-3.

[10] *Genotoxicita a karcinogeneze* [online]. [cit. 2018-01-30]. Dostupné z: https://is.muni.cz/do/rect/el/estud/prif/ps13/genotox/web/pages/02_mutace.html

[11] *Genetika - Biologie* [online]. [cit. 2018-01-30]. Dostupné z: http://www.genetika-biologie.cz/mutace

[12] RICHARD C. DEONIER, MICHAEL S. WATERMAN a Simon TAVARÉ. *Computational Genome Analysis An Introduction*. New York, NY: Springer Science+Business Media, 2005. ISBN 0387288074.

[13] *GenBank Overview* [online]. [cit. 2018-01-03]. Dostupné z: https://www.ncbi.nlm.nih.gov/genbank/

[14] BENSON, D. A., I. KARSCH-MIZRACHI, D. J. LIPMAN, J. OSTELL a D. L. WHEELER. GenBank. *Nucleic Acids Research* [online]. 2007, **35**(Database), D21-D25 [cit. 2018-01-03]. DOI: 10.1093/nar/gkl986. ISSN 0305-1048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl986

[15] *Sample GenBank Record* [online]. [cit. 2018-01-03]. Dostupné z: https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

[16] J. Horáček, *Základy lékařské mikrobiologie*. Karolinum, 2000.

[17] *National Human Genome Research Institute* [online]. [cit. 2018-01-31]. Dostupné z: https://www.genome.gov/11509542/

[18] *Prokaryotic RefSeq Genomes* [online]. [cit. 2018-05-14]. Dostupné z: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

[19] ZHANG, Zheng a Scott SCHWARTZ. *A Greedy Algorithm for Aligning DNA Sequences* [online]. [cit. 2018-05-18]. Dostupné z: https://doi.org/10.1089/10665270050081478

[20] JONES, Neil C. a Pavel. PEVZNER. *An introduction to bioinformatics algorithms*. Cambridge, MA: MIT Press, c2004. ISBN 0-262-10106-8.

[21] SMITH, T. F. a M. S. WATERMAN. *Identification of Common Molecular Subsequences* [online]. 1981, , 195-197 [cit. 2018-05-21]. Dostupné z: https://dornsife.usc.edu/assets/sites/516/docs/papers/msw_papers/msw-042.pdf

[22] JONES, Neil C. a Pavel. PEVZNER. *An introduction to bioinformatics algorithms*. Cambridge, MA: MIT Press, c2004. ISBN 0-262-10106-8

[23] ROSYPAL, Stanislav. *Přehled biologie*. 3. upr. vyd., v nakl. Scientia 2. vyd. Praha: Scientia, 1998. ISBN 80-7183-110-7.

[24] HAMPL, Vladimír. *Molekulární taxonomie: Výpočet genetických distancí a tvorba stromu distančními metodami* [online]. 10.11.2017 [cit. 2018-05-22]. Dostupné z: http://web.natur.cuni.cz/~vlada/moltax/

[25] MOUNT, David W. *Bioinformatics: sequence and genome analysis*. 2nd ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press, c2004. ISBN 0-87969-712-1.

SEZNAM PŘÍLOH

Příloha 1: Obsah přiloženého CD

1. Bakalářská práce, formát .pdf
2. Složka „Algoritmy a bakterie"
    - soubory bakterií ve formátu .gb
    - Skript_Martinkova.m
    - vyber_genu.m
    - vyber_proteinu.m
    - breakpoint.m
    - serazene.m
    - vzestupne.m
    - nahrada_pozic.m