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

DETEKCE MOBILNÍCH GENETICKÝCH ELEMENTŮ POMOCÍ ČÍSLICOVÉHO ZPRACOVÁNÍ GENOMICKÝCH SIGNÁLŮ
MOBILE GENETIC ELEMENTS DETECTION BY GENOMIC SIGNAL PROCESSING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR
Bc. Jarmila Nováková

VEDOUCÍ PRÁCE
SUPERVISOR
Ing. Helena Škutková, Ph.D.

BRNO 2017

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
TECHNICKÉ V BRNĚ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské inženýrství a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Bc. Jarmila Nováková ***ID:*** 146199
Ročník: 2 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Detekce mobilních genetických elementů pomocí číslicového zpracování genomických signálů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši typů mobilních genetických elementů (MGE) vyskytujících se v bakteriálních genomech a nástrojů využívaných k jejich detekci. 2) Navrhněte metodu detekce konkrétního vybraného typu MGE v bakteriální DNA využívající číslicové zpracování genomických signálů a dílčí kroky otestujte v programovacím prostředí Matlab. 3) Realizujte detektor vybraných typů MGE v bakteriální DNA využívající číslicové zpracování genomických signálů. 4) Vytvořený detektor otestujte na vhodně sestavených datasetech kompletních genomů různých bakteriálních kmenů. 5) Proveďte statistické srovnání navrženého detektoru s běžně využívanými nástroji využívajícími znakový zápis DNA.

DOPORUČENÁ LITERATURA:

[1] LIM, Kian Guan, Chee Keong KWOH, Li Yang HSU a Adrianto WIRAWAN. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. Briefings in Bioinformatics, Jan 2013, 14(1), 67-81.

[2] HAWKEY, Jane, Mohammad HAMIDIAN, Ryan R. WICK, David J. EDWARDS, et al. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genomics, 2015, 16(1), 1-11.

Termín zadání: 6.2.2017 ***Termín odevzdání:*** 19.5.2017

Vedoucí práce: Ing. Helena Škutková, Ph.D.
Konzultant:

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

ABSTRAKT

Práce se zabývá mobilními genetickými elementy. Je zaměřena na jejich vlastnosti využitelné pro jejich detekci. Zároveň se věnuje problematice transformace symbolické sekvence do numerické formy.

Jsou vysvětleny klasifikace mobilních genetických sekvencí, popsány základní typy mobilních genetických sekvencí, principy numerických map a detekce v symbolické reprezentaci

Je zpracována konverze symbolických genetických sekvencí dle zvolené numerické mapy a provedena analýza vlastností mobilních genetických elementů v numerické reprezentaci pro návrh detektoru. Na závěr je vytvořena knihovna motivů používaná navrženým detektorem.

KLÍČOVÁ SLOVA

Mobilní genetické elementy, bakterie, inzerční sekvence, transpozon, numerické reprezentace, detektor

ABSTRACT

Mobile genetic elements are occupied by this project. It is aimed at their features, which can be used for their detection. It also deals with issue of conversion of symbolic sequence into numerical form.

Classifications of mobile genetic elements are explained, basic types of mobile genetic sequences are described, and principles of numerical maps and detection in symbolic represetation are also clarified.

Conversion of symbolic genetical sequences by chosen numerical map and calculation of normalized correlation values for set of mobile genetic elements are compiled. Analysis of the mobile genetic elements properties is performed for design of detector. The library of themes is created at the end for usage by designed detector.

KEYWORDS

Mobile genetic elements, bacteria, insertion sequences, transposon, numerical representation, detector

NOVÁKOVÁ, J. Detekce mobilních genetických elementů pomocí číslicového zpracování genomických signálů. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2017. 81 s. Vedoucí diplomové práce Ing. Helena Škutková, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma Detekce mobilních genetických elementů pomocí číslicového zpracování genomických signálů jsem vypracovala samostatně pod vedením vedoucího diplomové 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 diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové 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 následujícího 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.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji vedoucí diplomové práce Ing. Heleně Škutkové, Ph.D. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne ....

(podpis autora)

Obsah

Úvod

Základní dogma bylo a je jedním z hlavní principů genetiky. Podle něj je v pořadí nukleotidů uchována genetická informace. Odtud je přepisem do molekuly mRNA následně tato informace přeložena do pořadí aminokyselin na ribozomu. Dnes již víme, že dogma má několik odboček. Jednou z nich je i neměnnost genetické informace. Genetická informace totiž může být změněna nebo dokonce poškozena (mutována). Kromě mutací mění knihovnu genů organismů i takzvané mobilní genetické elementy.

Mobilní genetické elementy jsou specifické sekvence genomu schopné se množit a vkládat do jiných oblastí genomu. Jsou dnes hodně prozkoumávanou oblastí bioinformatiky a genomiky vzhledem k jejich velkému vlivu na variabilitu genomu. Během posledních pěti desetiletí byly objevovány postupně různé typy mobilních genetických elementů a některé z nich velice nedávno. Lze tak předpokládat, že se jejich řady budou dále rozšiřovat.

V sekvenci mobilních genetických elementů mohou být zahrnuty i geny pro různé vedlejší vlastnosti. Například u bakterií je velice často přítomen gen umožňující rezistenci k antibiotikům. A díky jejich schopnosti rychlého přenosu a kopírování představují riziko pro zdraví člověka. Navíc jejich množení a přenos je častější v bakteriálních genomech než v eukaryotických.

Doposud bylo implementováno plno odlišných algoritmů pro detekci různých typů mobilních genetických algoritmů. [1] Každý s různou mírou spolehlivosti a různými parametry pro detekci. Tyto detektory jsou velice citlivé i na nepatrnou změnu parametrů. Až na vzácné vyjímky operují všechny v symbolickém zápisu genetické sekvence, což velice omezuje možnosti analýzy a množství použitelných algoritmů. Číslicové zpracování signálů nabízí mnohem více odlišných metod a algoritmů pro analýzu, klasifikaci nebo detekci nejen v biologických signálech. Ačkoliv stále převládá analýza v symbolické reprezentaci gentické sekvence, začínají se zkoušet i nástroje číslicového zpracování signálů.

Tato práce se zabývá analýzou alternativní možnosti detekce mobilních genetických elementů v genomických signálech. Díky dnes dostupným sofistikovaným numerickým mapám lze transformovat symbolickou sekvence do numerické za zisku genomického signálu. Zde již lze uplatnit prostředky číslicového zpracování signálů. Vzhledem k nižší úspěšnosti detekování mobilních genetických elementů v symbolickém zápisu genetické sekvence [1] se nabízí možnost prozkoumat i použití metod číslicového zpracování signálů, kterým se tato práce věnuje. Na základě poznatků získaných analýzou genomických signálů a jejich částí odpovídající mobilním genetickým elementům představuje i nový nástroj pro jejich detekci.

1 Bakterie

Prokaryotické jednobuněčné organismy můžeme rozdělit do dvou domén – bakterie a archea. Bakterie jsou velice jednoduché mikroskopické buňky s primitivní stavbou žijící v různých prostředích. Patří mezi nejrozšířenější organismy na planetě, dokáží žít v půdě, v oceánu, ve vřídlech (85°C) i v lidském organismu. Některé jsou prospěšné pro člověka, například Escherichia coli v tlustém střevě tráví cukry s tvorbou kyseliny mléčné a vitamínu B, jiné jsou oproti tomu škodlivé a způsobují různá onemocnění. Mezi nejvíce diskutovány dnes patří MRSA (methicillin-resistant Staphylococcus aureus) a VRSA (vancomycin-resistant Staphylococcus aureus), jelikož disponují rezistencí na nejvíce používaná antibiotika.

Na povrchu je bakterie ohraničená buněčnou stěnou a cytoplazmatickou membránou, avšak vnitřním organelám ohraničení biomembránou chybí. Většina bakterií má zahrnutý peptidoglykan do buněčné stěny. Dovoluje to rozlišení bakterií na grampozitivní a gramnegativní, kdy grampozitivní mají více peptidoglykanu na nějž se naváže více krystalové violeti.

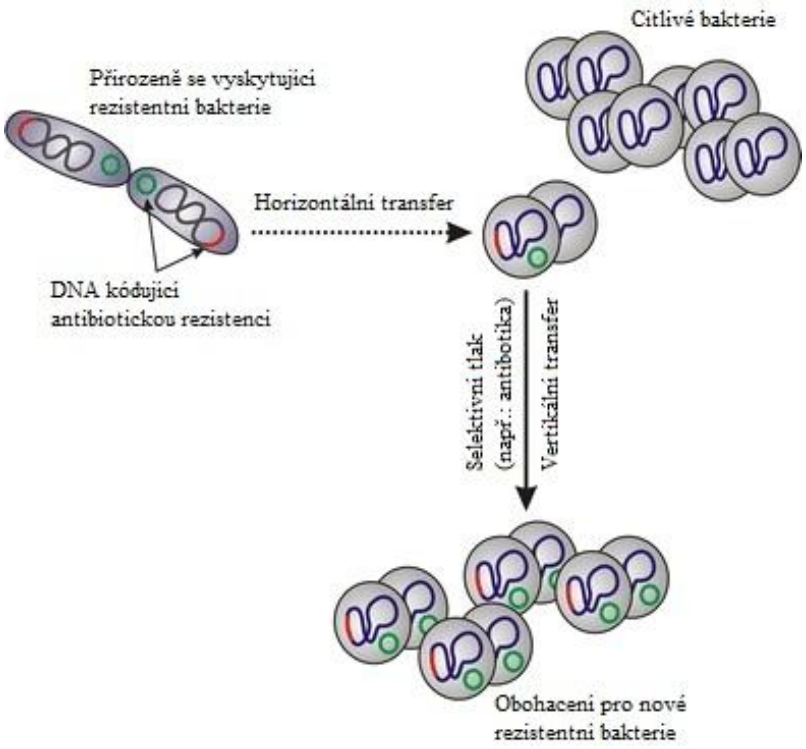
Na rozdíl od eukaryotické jaderné DNA, která je uložena v jádře s dvojitou jadernou membránou, bakteriální jaderná DNA je uložena volně v cytoplazmě vyplňující vnitřní prostor bakterie. Tento nositel genetické informace je tvořen jedinou dvoušroubovicí stočené do kruhu – nukleoid. Je asi tisíckrát delší než buňka samotná a proto je dále poskládána do smyček až vznikají pentlicové struktury.

Tvorba esenciálních bílkovin probíhá ve sférických ribozomech. U některých buněk můžeme nalézt i plazmidy, což jsou další malé kruhové nebo lineární molekuly DNA schopné autoreplikace. Pro buňky jsou postradatelné, jelikož nesou doplňkovou informaci. V mnoha případech se jedná o výhodu ve formě rezistence k antibiotikům, schopnosti vytvářet vlastní antibiotika nebo určení patogenity bakterií. [2] Oproti tomu buňce bakterie chybí mitochondrie, plastidy, endoplazmatické retikulum a mnoho dalších organel, jež nalezneme v eukaryotické buňce.

Dle výše uvedených informací lze vydedukovat, že u bakterií se genom převážně skládá z kruhové DNA a plazmidů. Bakteriální genom vykazuje silný pozitivní vztah s množstvím genů, tudíž bakterální genom obsahuje velmi malé množství nekódující DNA - intronů. Funkčně podobné geny jsou sdružovány do operonů, které reguluje společný promoter. Bakteriální genom je velice proměnlivý ve smyslu genového repertoáru, přesto pozoruhodně stabilní ve smyslu chromozomové organizace. Nukleoid nabývá různých rozměrů, od 10 kb až k 1Mb, ale u volně žijících bakterií byla objevena i velikost téměř 1,4 Mb (půdní bakterie Ktedonobacter racemifer). Můžeme v něm najít i několik determinantů pro míru jeho stočení. [3]

Protože bakterie má pouze jeden chromozom, nenajdou se různé alely jednoho genu v jedné buňce, přesto existují a tudíž DNA s alelu vzniklou v jiné buňce musí být nějakým způsobem inkorporována do DNA dané bakterie.

K přenosu nových genů nebo jejich alel může dojít několika způsoby. Transfer MGE mezi buňkami je znám jako laterální nebo horizontální přenos genů (HGT). Zatímco vertikální přenos genů je klasická dědičnost genetické informace z mateřské na dceřinou buňku, HGT probíhá z prokaryota na prokaryota, z prokaryota na eukaryota nebo z eukaryota na eukaryota. [4] Bakterie získává genetickou informaci z okolního prostředí nebo buněk třemi způsoby HGT – transformací, transdukcí nebo konjugací. [4]



Obr. 1: Princip vertikální a horizontálního přenosu genů [4]

Bakteriální genom lze rozdělit na dvě části, na základní a doplňkový genom. Základní obsahuje všechny esenciální geny potřebné pro správný chod metabolismu, syntézu DNA a RNA a replikaci. Doplňkový genom reprezentuje diverzitu mezi jednotlivými druhy bakterií a zodpovídá za adaptaci bakterie na prostředí. Doplňkové geny obvykle mají jiný poměr guaninu a cytosinu oproti základní části genomu, jelikož se často jedná o geny inkorporované z jiných druhů bakterií. [4] Tento jev je obvyklý u gramnegativních bakterií avšak není společný všem. [2]

2 Mobilní genetické elementy

Fenomén evoluce všech organizmů je spojen se změnami v genetickém kódu, ať už způsobené vnějšími či vnitřními faktory. Tyto změny označujeme souhrnně jako mutace. Mezi nepřeberným množstvím typů mutací se vyskytují i mobilní genetické elementy (někdy nazývané jako předatelné elementy nebo transpozony) reprezentující rozptýlené repetice v genetické informaci.

Mobilní genetické elementy jsou specifické sekvence genomu, které se umí množit a následně vkládat do nových oblastí genomu. V některých případech se umí stěhovat celé, a to v rámci daného genomu nebo mezi genomy. [5] Definice je označuje jako „specifické DNA segmenty, které se mohou opakovaně vkládat do jednoho nebo mnoha míst v jednom nebo několika genomech“. [6]

Dříve se genetika řídila takzvaným základním dogmatem, dle něhož je genetická informace uchovávána v pořadí nukleotidů v molekule DNA a že je neměnná. Avšak dogma má ve skutečnosti plno odboček – vyjímek. Jednou z nich je neklidnost a dynamičnost struktury genetického kódu vedoucí ke změnám knihovny genů. Za touto vlastností stojí také mobilní genetické elementy. [5]

Za objevením transpozonů stojí Barbara McClintock, která ukázala ve 40. a 50.letech minulého století při studiu chromozomových zlomů u kukuřice, že její genom obsahuje mnoho mobilních elementů způsobujících somatické mutace a za tento objev byla odměněna Nobelovou cenou. [7] Od toho momentu se genetici setkali s mnoha různými genetickými jevy ve kterých genetické elementy byly schopné pohybu v genomu aktivně i pasivně. Tyto genetické elementy, nyní prezentované jako mobilní genetické elementy, vyvolávají zájem ve výzkumných kruzích pro jejich bizardní chování v porovnání s konvenčními genetickými jevy jako jsou transkripce, translace, replikace , rekombinace a podobně.

Základní skupiny mobilních elementů jsou velmi staré a najdeme je téměř u všech eukaryotických druhů. U rostlin často zabírají i přes 80% celkové genomové DNA, u hub a metazoí reprezentují mnohem menší, ale přesto podstatnou část jejich genomu. U člověka tvoří mobilní elementy až polovinu genetické informace. V menší míře (5 – 10%) se transpozony vyskytují dokonce i u prokaryot. [8] U zlatého stafylokoka bylo objeveno dokonce až 20%. [2]

Mobilní genetické elementy jsou hlavními hráči v pohybu a reorganizaci genů, ať už v rámci daného genomu (intracelulární mobilita) nebo mezi bakteriálními buňkami (mezibuněčná mobilita). Mobilní genetické elementy jsou definované jako DNA sekvence o různorodé délce (od 1 báze až k několika stovkám kb) a které často nesou funkce ovlivňující jejich přenos a rekombinaci v hostitelském genomu. Dneska převažuje názor, že jsou

klíčovými hráči v přeskupování genetického materiálu, jenž v kombinaci s mutacemi a „přírodním výběrem“ tvoří esenciální prvky evoluce. [9]

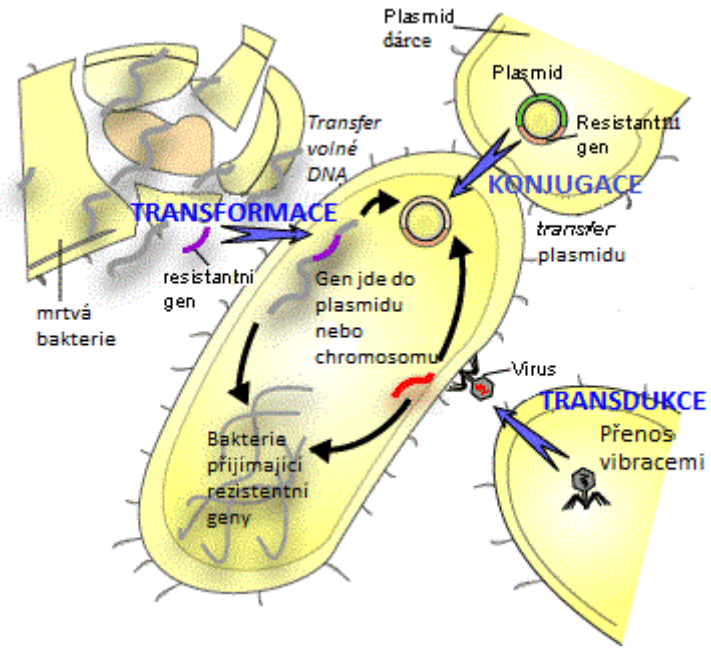
2.1 Podstata přenosu MGE

Jak je zmíněno výše, existují 3 typy HGT pro zisk a začlenění nové genetické sekvence do genomu bakterie - transformace, transdukce a konjugace. [4]

Při procesu transformace vstoupí volná DNA do nové buňky z okolního prostředí penetrací přes buněčnou stěnu a membránu. [10] Transdukce probíhá infekcí určitými virusy (bakterofágové) schopnými infikace bakterie se současným dopravením fragmentu jiné bakteriální DNA. [11] Konjugace se týká přenosu plazmidů při přímém fyzickém kontaktu mezi bakteriemi. [10]

Samotná transpozice prokaryotních mobilních genetických elementů se týká pouze přenosu v rámci jedné buňky, neprobíhá mezibuněčně ani se nejedná o přijetí DNA z okolního prostředí. Tento přenos může probíhat replikativním a konzervativním způsobem. [12]

V replikativním způsobu dochází k vytvoření kopie elementu a ta se integruje na nové místo v genomu. Na originálním místě zůstává původní element. U druhého způsobu, konzervativního, dochází k excizi (vystřižení) elementu z původního místa a přenosu na nové místo v genomu. [13]



Obr. 2: Tři typy HGT [11]

2.2 Klasifikace mobilních genetických elementů

Doposud neexistuje jednoznačné rozdělení a klasifikace typů mobilních genetických elementů. Tradičně byly MGE rozděleny na bakteriofágy, plastidy a transpozony, ale s identifikací nových typů elementů jako například genomických ostrovů, konjugativní transpozony apod. se stalo tohle tradiční rozdělení zastaralým a přežitým. [14],[6] Dnes je snahou najít a identifikovat všechny typy mobilních genetických elementů. Tahle idea vedla k založení databáze ACLAME (A CLAssification of Mobile genetic Elements) a UCL webové stránky. [6]

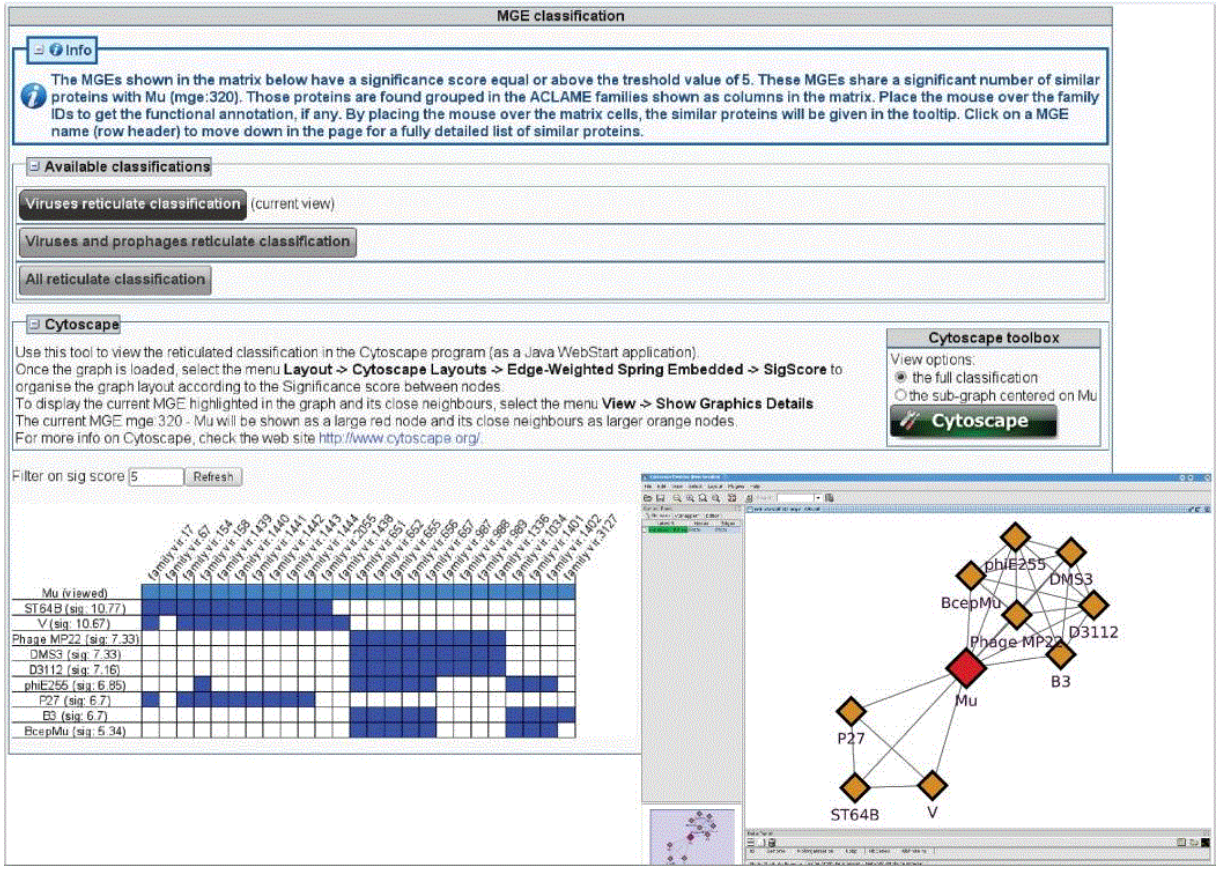
ACLAME databáze je zasvěcena sbírání, analýze a klasifikaci sekvencovaných mobilních genetických elementů získaných z rozličných zdrojů do skupin s podobnými znaky. [9],[14] Poskytuje informace o obsahu MGE a klasifikace jsou přístupné z různých pohledů na jejich organizaci. Genový/proteinový pohled, genomový pohled a populační pohled na MGE. Poslední verze databáze ACLAME 0.4 nyní jímá informaci o 457 genomech bakteriofágů, o 1109 plasmidech a o 760 profázích. [14]

Vůči každému MGE jsou posbírány sekvence proteinů, které MGE kóduje. Z nich jsou dále vytvořeny proteinové rodiny. Sekvence proteinů se posbírají z různých databází a tudíž musí být podrobeny porovnání za pomocí SSERCH programu s pradvěpodobnostní hranicí 0.001. Proteiny, jež mají největší míru similarity, jsou dále podrobeny Markovovu algoritmu shlukování (cluster). [14] Zároveň je tato databáze přístupná dobrovolníkům pro dodatečné přidání MGE nebo úpravu již existujícího MGE i s jeho zařazením. Při založení této databáze se tímto způsobem získalo 437 shluků minimálně o 3 členech vytvořených z pouhých 50% poskytnutých proteinových sekvencí. [9]

ACLAME navrhuje také odlišné zobrazení pro genové transfery, mozaicismus a genomová přeskupení pozorovaná v MGE populacích. Zavedli zobrazení v grafu (Obr. 3) oproti původnímu tradičnímu stromovému.

UCL webová stránka oproti tomu slouží k přiřazování sekvenčních čísel k jednotlivým objeveným a zaregistrovaným MGE. Komise během schůzky věnované DNA Inzercím v Cold Spring Harbor roku 1976 navrhla sadu pravidel pro pojmenovávání MGE. Byl navrhnut systém, kdy jakýkoliv element inzerční sekvence (IS element) byl pojmenován IS1, IS2, atd. a stejný postup aplikován i na traspozony (Tn1, Tn2, atd.), neboli přiřazení sekvenčního čísla za zkratku názvu MGE. Avšak postupem času došlo ke konfliktům, jelikož po přídělení Tn4685 nebyl žádný seznam publikován a čísla byla přidělována náhodně nebo jedno číslo bylo přiděleno několika elementům. Avšak dnes je snaha v tomto systému pokračovat na nově objevené MGE prostřednictvím této webové stránky určené k registraci nově objevených sekvencí (spolu s popisem, referencemi a přístupovým číslem) identifikovaných jako MGE. [6]

Existují různá rozdělení MGE, dle úhlu pohledu. V některých případech se jejich vlastnosti dokonce překrývají nebo proplétají. [15] Vzhledem k dříve zmíněnému výsledku shlukovacího Markovova algoritmu nejsou ještě dokonce všechny typy MGE identifikovány. [9],[14] Mnoho genomových sekvenovacích projektů stále probíhá anebo jsou plánovány, což vede k dobrému důvodu věřit, že dojde k objevení i dalších MGE. [6] Naše znalosti nejsou prozatím dostatečné natolik, abychom si mohli dovolit úspěšnou generalizaci. Proto práce dále uvádí základní typy mobilních genetických elemetů s nimiž se lze setkat, avšak nemůže postihnout všechny již nalezené nebo nenalezené.



Obr. 3: Klasifikační sekce na webu ACLAME se zobrazením v podobě grafu
V matici na levé straně obrázku představují sloupce proteinové rodiny, kde daný MGE kóduje proteiny. První řádek v matici koresponduje s hledaným MGE (zde fág MU) a další řádek nejbližší další MGE. Číslo vedle jména MGE označuje podobnosti s hledaným MGE. Modře vybarvené čtverce v matici říkají, že zde protein kódovaný hledaným MGE je i v proteinové rodině daného sloupce. Graf vpravo ukazuje vztahy mezi 7 mutačními fágy. **[14]**

2.3 Základní typy mobilních genetických elementů

Základní typy genetických elementů nebyly objeveny záměrně při výzkumu a hledání každého z nich. Ve skutečnosti byly nalezeny pouhou náhodou při studiu jiných genetických jevů, problémů a částí genetického kódu. Tyto výzkumy probíhaly odděleně v rozmezí 30 let, mluvíme-li o bakteriích. [16]

První takovou náhodou byl výše zmíněný objev Barbary McClintock při studiu chromozomových zlomů v kukuřici za použití markerů. To byl však jen objev existence mobilních genetických elementů. Markery měly v určitých oblastech větší míru výskytu a dokonce poloha tohoto zvýšeného výskytu markerů nebyla stabilní, protože se u některých liní přesouvala. To vedlo k myšlence existence aktivátoru. [7]

Nacházení typů prokaryotického transposonového elementu začalo až o 20 let později. Fág Mu je prvním nalezeným typem objeveného prokaryotického mobilního elementu. V Denverských odpadních vodách jej zaznamenal A.L.Taylor. Při zájmu o regulaci genové exprese skupina genetiků izolovala spontánní pleiotropní mutace. Tyto následně identifikované segmenty pojmenovaly jako inzerční sekvence (Insertion sequence, IS), což je druhý typ MGE. Poslední objev se týkal transpozonu. Studium lékům odolných plasmidů pomohlo najít i tento poslední typ. Nicméně samotná analýza těchto elementů mohla být ve skutečnosti prováděna až od konce 70.let minulého století. [16]

Další typy elementů jsou bakteriofágové, integrovaný profág, přenositelný profág, integrovaný satelitní profág, jednotlivé transpozony, konjugativní transpozony, integrativní konjugativní transpozony, mobilizovatelné transpozony, integrativní mobilizitovatelné transpozony, kompozitní transpozony, integrony, inzerční sekvence, plazmidy, genomické ostrovy, ostrovy patogenity, chromozomové kazety, skupina I intornů, skupina II intronů, IStron intein. [4],[6],[17]

2.3.1. Fág Mu

Fág Mu je normální bakteriofág nalezený v Denverských odpadních vodách. Je pojmenovaný podle slova mutace. Disponuje fascinující schopností vkládání svého profága na mnoho různých míst v chromosomu *Escherichie Coli*, které občas způsobují mutace. Obecně se dokáže integrovat do bakteriálního i plasmidového genomu. Mutace se může projevit prakticky na jakémkoliv lokusu na chromosomu a mají nízkou specificitu. Dále se zvládá integrovat v jakékoliv orientaci. Umožňuje mobilizaci jakékoliv DNA a její začlenění kamkoliv v genomu.

2.3.2. Inzerční sekvence

Spontánní pleiotropní mutace operonů lac a gal nalezené při výzkumu regulace genové exprese ukázaly zajímavé vlastnosti jako například extrémní polaritu v expresi cistronů směrem od promotoru, od místa mutace. Tyto nalezené mutace bylo možné revertovat, což vyloučilo možnost mutace typu delece. A protože reakce těchto mutací na mutageny v reverzních testech byly odlišné od již známých typů mutací, dospěli genetici k hypotéze, že jde o vložení dodatečné DNA. Hypotéza byla s pomocí transdukčních fágů brzy potvrzena. Bylo objeveno, že několik vzdálených segmentů DNA může být jednotlivě vloženo na množství různých míst. To vedlo k onomu názvu „Inzerční sekvence“, jinak “IS

elementy“. IS elementy hrají významnou roli k restrukturalizaci bakteriálního genomu za použití mutace a DNA přenosu. [8] V některých druzích bakterií tvoří až 10% genomu. [8]

Jejich velikost se pohybuje v řádu stovek bází. [8] Obvykle 0.7 – 2.5 kb. [15] Strukturálně mají invertovaná opakování na koncích daného segmentu o velikosti desítek bází (viz Obr. 4). [8] Mnoho IS, ale ne všechny, nosí tato krátká nedokonalá invertovaná opakování o délce menší než 40 bp a generují malou duplikaci o 2 až 14 bázích s hostitelskou DNA lemující místo inzerce. [17] Disponují jedním nebo dvěma ORF. V této sekvenci nalezneme geny kódující proteiny odpovědné za funkce zahrnuté v pohyblivosti IS elementu. [18] V nejzákladnější formě se jedná o pouhý jeden gen kódující místně specifickou rekombinázu (nazývaný transpozáza). [18] Žádné jiné geny neobsahuje. [17] Vloží-li se IS do kódující sekvence genu, způsobí jeho inaktivaci. Avšak IS může obsahovat dokonce transkripční a translační terminační signály, čímž dokážou blokovat expresi dalších genů ležících za místem vložení IS směrem od promotoru. [9] Jejich hlavním účinkem jsou změny v expresi genů, hlavně aktivace nebo deaktivace. [4] Tento projev označujeme jako polární mutace (ovlivnění všech genů transkripčně směrem od promotoru). [8]

Obr. 4: Struktura inzerční sekvence [9]

2.3.3. Transpozon

Nejlépe zkoumaným a dokumentovaným současným příkladem genetických změn v evoluci je vznik přenosné antibiotické rezistence jako bakteriální odpověď na antibakteriální chemoterapii. Proběhl velmi intenzivní výzkum týkající se genetiky determinantů antibiotické rezistence. [8],[16] Mnoho determinantů projevovalo anomální rekombinaci a nakonec bylo zjištěno, že se transponují z jednoho replikonu na další. Díky tomu si získali označení transpozony. Později se vyjasnilo, že plno dalších fenotypových znaků, jako schopnost degradovat specifické substráty nebo produkce toxinů, mohou být také nošeny na traspozonech. [16]

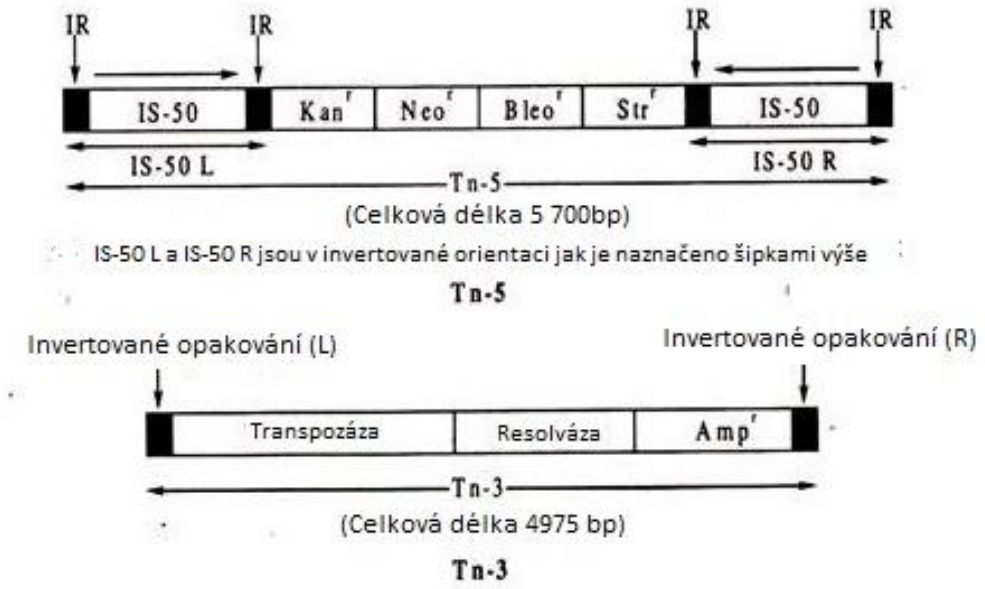
Vyskytují se v bakteriálních plazmidech, které se označují jako R faktory (R=rezistence). Oblast transpozonu zodpovědná za přenos zmíněné rezistence se nazývá RTF oblast.

Molekulární analýza v 70. letech minulého století identifikovala dvě odlišné třídy transpozonů. První je reprezentovaná amphicilinem Tn3 a strukturou se podobá IS elementům. Oproti nim se však odlišuje větším množství kódujících sekvencí pro specifický

fenotyp mezi invertovanými opakováními na koncích daného elementu. [8] Obsahuje také invertovaná opakovaní jako IS element. Nazývají se nekompozitní transpozony. 0

Druhá třída, kompozitní transpozony, se skládá ze sloučených elementů. Každá z nich obsahuje kódujicí sekvenci obklopenou z každé strany IS elemetem. IS elementy se zde starají o mobilitu sloučeniny elementů a jakýkoliv úsek bakteriálního genomu může být vložen mezi ně a vytvořit sloučený transpozon. [8]

Transpozony se obecně vyznačují velikostí od 2.5 do 60kb, dlouhými invertovanými opakováními na koncích a jedním nebo několika doplňkovými geny, které propůjčují hostitelskému genomu dodatečný fenotyp. [15] Převážně kódují antibiotickou rezistenci v úseku mezi dvěma IS elementy na koncích. [4] A zároveň si nesou gen pro vlastní enzym transpozázu. Mnoho z nich není místně specifických, tudíž mnoho kopií toho samého transpozonu se může objevit v bakteriální buňce, na chromozomu nebo v plazmidu. [2]



Obr. 5: Struktura kompozitního (nahoře) a nekompozitního (dole) transpozonu. [9]
Invertovaná opakování (IR) se rozdělují na levé (Left, L) a pravé (Right, R) dle své pozice.

3 Symbolická reprezentace

Genetické sekvence jsou vždy posloupnosti písmen zastupujících deoxyribonukleotidové báze v DNA řetězcích (A,T,G,C), nebo RNA řetězcích (A,U,G,C). Jedná se o standardizovaný formát využívající IUPAC kódování. Taková posloupnost písmen tvoří esenciální informaci pro organismus prostřednictvím překladu do sekvence aminokyselin tvořících bílkoviny. Navzdory písemné reprezentaci sekvence existuje mnoho nástrojů nebo sofistikovaných algoritmů pro jejich analýzu v tomto formátu. Pro porovnávání dvou nebo více sekvencí znaků se pro genetické sekvence používá procedura zarovnávání sekvencí, kde lze kvantifikovat míru podobnosti. K dispozici je mnoho metod a algoritmů, například BLAST, Needlemanův-Wunschův algoritmus, Smithův-Watermanův algoritmus. Dále lze zkoumat počet bodových mutací pomáhající při výpočtu evoluční vzdálenosti, variabilitu sekvencí a tak dále.

3.1 Detekce mobilních genetických elementů v symbolické reprezentaci

Vysoký zájem o mobilní genetické elementy spolu s přístupem ke genomické DNA vedl k vytvoření nové generace metod a nástrojů pro jejich naleznutí v sekvenci a analýzu v symbolické reprezentaci. Převažuje přístup vyhledávání repetic, protože konkrétní MGE se může v sekvenci nebo genomu objevit více než jednou díky jejich schopnosti replikativní transpozice. Navíc mnoho MGE obsahuje repetici již ve své struktuře a to v invertovaných koncích. [28], [34]

Tandemové repetice jsou opakující se sekvence, které se neobjevují pouze náhodně, ale i přímo příléhající jedna ke druhé. Většina dnes přístupných vyhledávacích nástrojů využívá dvoufázový přístup detekce tandemových repetic: vyhledat a vyfiltrovat. [34] V první fázi je obvykle implementován jeden ze dvou typů algoritmů. Buď časově velice náročný vyčerpávající kombinatoriální vyhledávací algoritmus porovnávání každého úseku s každým anebo algoritmus založený na statistickém či heuristickém přístupu (např.: malé segmentové okno skenuje po sekvenci dokud nenajde první repetici a následně tyto malé perfektní repetice spojuje dohromady pro naleznutí delších repetic).

V druhé fázi se přistupuje k jedné z mnoha různých metod filtrování pro zajištění identifikace a výtah MGE. [34] Tandemové repetice nemusí být pouze MGE, ale i segmentované duplikace, satelity nebo jen náhodné repetice. Odlišení MGE od ostatních je náročné díky biologické komplexnosti zahrnující:

- Fragmentovaná povaha MGE
- Promíchání MGE do jiných typů repetic
- Rozlišení mezi odlišnými MGE, které sdílí úseky různých délek

- Konkrétní MGE může mít v genomu degenerovanou kopii či kopie [28]
- MGE mohou být navzájem vnořené

Jako filtraci lze použít jednoduchý, avšak nejméně účinný, filtr délky anebo komplexnější metody jako shluková analýza K-means, statistické modely anebo zarovnávání sekvencí. [34]

Právě zarovnávání je jednou z nejvíce používaných filtrovacích metod anebo její částí v nástrojích pro vyhledávání MGE nebo repetic. Zarovnávání sekvencí je definováno jako procedura porovnávání dvou nebo více sekvencí vyhledáváním série individuálních znaků nebo charakteristických znaků přítomných v sekvencích ve stejném pořadí. K tomu je potřeba kvantifikovat podobnost sekvencí za použití párové vzdálenosti např.: proporcionální vzdálenosti *p*.

$$p = počet\ rozdílných\ pozic / počet\ pozic \quad \textbf{(1)}$$

Používanější ze dvou metod zarovnávání sekvencí je dynamické programovaní, které umožňuje globální (Needleman-Wunschův algoritmus) i lokální zarovnání (Smith-Watermanův algoritmus). Logicky je výhodnější pro detekci MGE a repetic využít lokálního zarovnání. Přistupuje se k použití metody BLAST (Basic Local Alignment Search Tool) anebo Smith-Watermanovu algoritmu:

$$F(i,j) = max \begin{cases} 0 \\ F(i-1,j-1) + s(x_i, y_i) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \quad \textbf{(2)}$$

Kde *i,j* označují pozice v sekvencích, *d* zastupuje hodnotu mezery a *s* hodnotu shody nebo neshody. [38]

3.1.1. Metody založené na homologii

Zarovnávání sekvencí se využívá v několika filtrovacích metodách. BLAST se používá v metodách založených na homologii, jež vyhledávají MGE za použití její protein kódující části sekvence. Zaměřují se na detekci již identifikovaných repetic. Také nejsou aplikovatelné na některé MGE, které jsou kompletně tvořené nekódujícími sekvencemi. Navíc detekce založená na homologii s protein kódujícími úseky vyžaduje další analýzu strukturních znaků k získání celé délky referenční sekvence.

V metodách založených na struktuře se pouze zarovnávají lokální repetice nalezené za pomocí různých čtecích rámců pro geny *gag*, *pol* a *env*, míst vázaní primerů,

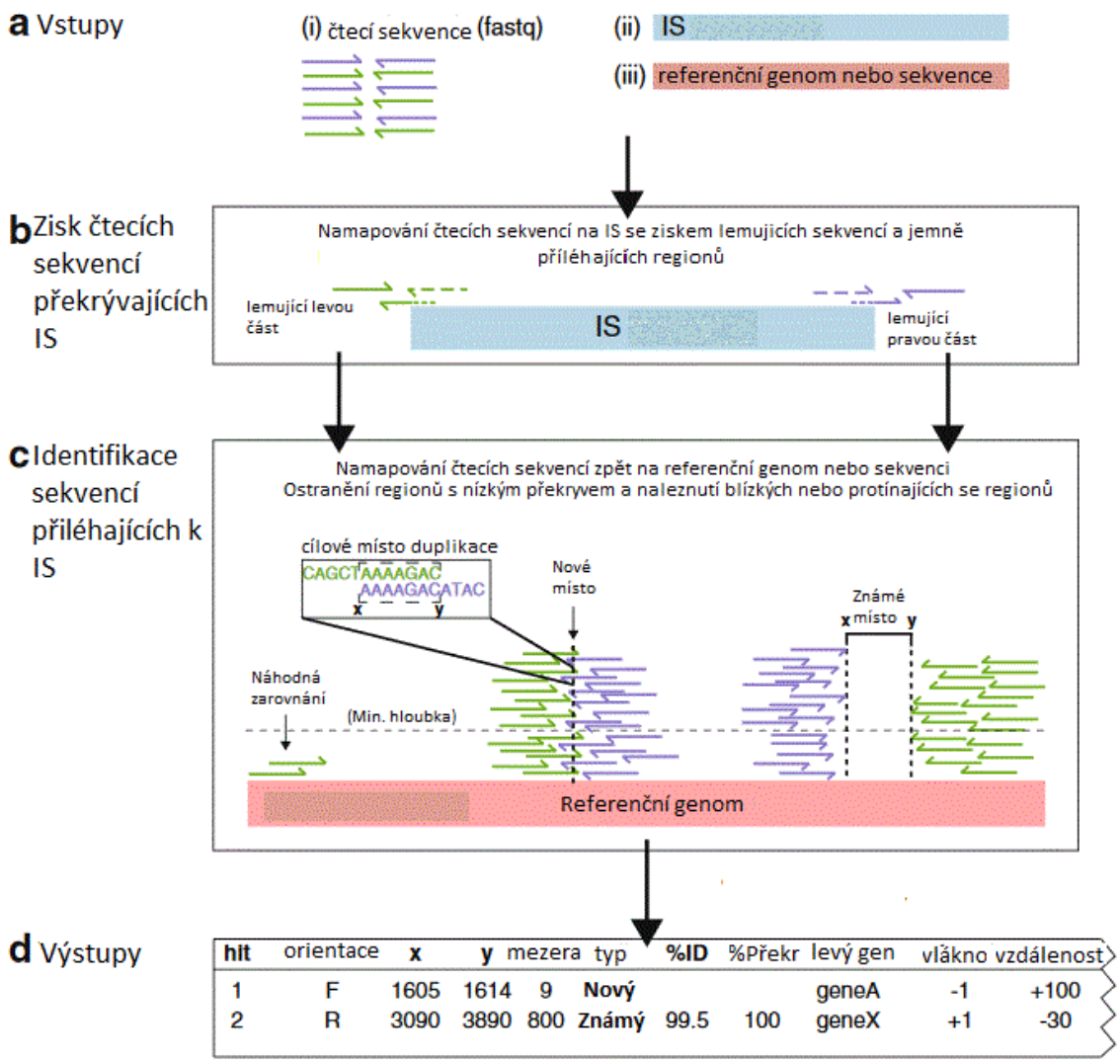
polypurinových traktů atd. Párové zarovnání takovýchto repetic je dále využíváno pro odhad hranic repetic zpětně na sekvenci. Algoritmus je aplikován v nástroji LTR_STRUC. Naneštěstí LTR_STRUC je neschopný identifikovat neúplné repetice.

3.1.2. Metody založené na porovnávání genomů

Metody srovnání genomů spoléhají na schopnost detekovat velké inzerce v mnohonásobném zarovnání způsobené transpozicí. MGE je detekováno, pokud je zarovnání narušeno inzerčním regionem v jednom nebo více druzích organismů (délka nad 200bp). Po odfiltrování náhodných a jednoduchých repetic jsou inzerční regiony lokálně zarovnány pro zjištění repetic v inzerčních regionech. Tento proces je omezen na shromáždění pouze inzerčních regionů, které jsou odvozeny od stejného listu fylogenetického stromu. Tato metoda řeší problém s určování hranic repetice, ale je závislá na kvalitě celogenomového zarovnání. To bude velice chudé v místech bohatých na MGE. Jestli všechny inzerce MGE budou zděděné vzhledem k dalším analyzovaným druhům, žádný MGE nebude detekován.

3.2 Detekce mobilního genetického elementu

Konkrétní MGE sice může mít degeneraci ve svých kopiích, ale pokud jsou známy, lze vytvořit konsensuální sekvencí konkrétního MGE (např.: IS1012) a za použití zarovnání sekvencí nalézt v jiné delší sekvenci nebo genomu jeho umístění. Tento přístup se Smith-Watermanovým zarovnáním je využíván v nástrojích Censor a RepeatMasker. [28] Ty používají referenční set sekvencí MGE získaných z metod zmíněných dříve. Z těchto dat jsou pak schopny určit konsensuální sekvenci MGE, její typ a vyhledávat ji v genomech nebo dlouhých sekvencích. Konsensuální sekvence reprezentuje nejlepší aproximaci realizací konkrétního genetického elementu lišících se degeneracemi. Evoluční vzdálenost mezi konsensuální sekvencí MGE a konkrétními genetickými sekvencemi MGE je menší než mezi dvěma konkrétními genetickými sekvencemi MGE. Detekce konsensuální sekvencí vykazuje vyšší sensitivitu než přímé použití sekvence MGE.



Obr. 6: Postup algoritmu ISMapperu [37]
a. Potřebné vstupy – párové čtecí sekvence, sekvence MGE (zde IS) a referenční genom nebo sekvence
b. Zisk souboru párových čtecích sekvencí, které se překrývají se začátkem nebo koncem IS za použití metod zarovnání
c. Namapování párových čtecích sekvencí získaných z předchozího kroku na referenční genom nebo sekvenci. Místa s množstvím namapovaných párových čtecích sekvencí větším než nastavený práh jsou vyhodnoceny jako místa začátků nebo konců MGE. Podprahové detekce jsou odstraněny.
d. Výstupem je matice informací o místech detekce namapovaných párových čtecích sekvencí.

3.2.1. Metoda mapování dílčích čtecích sekvencí

Výše zmíněné metody a nástroje jsou velice citlivé na počáteční nastavení parametrů vyhledávání jako minimální délka repetice či MGE, váhy zarovnání, minimální skóre a další. Proto jsou hledány jiné přístupy v detekci MGE. Jedním z nich je i ISMapper využívající k detekci krátké párové čtecí sekvence. Dále vyžaduje jako vstup sekvenci IS a referenční genom nebo sekvenci. Využijí se pouze krátké čtecí sekvence, jež se úspěšně namapovaly (zarovnaly) na začátek nebo konec IS. Ty jsou dále namapovány (zarovnány) každý zvlášť

na referenční genom nebo na sekvenci ze vstupu. Inzerční místo je identifikováno v pozici nebo jejím bezprostředním okolí, kde se objeví více než šest krátkých čtecích sekvencí (neboli filtrace čtecí hloubkou). [37] Přehled a podrobnější popis algoritmu viz (Obr. 6).

Tab. 1: Porovnání vlastností nástrojů pro vyhledávání repetic [28],[34]

Nástroj	**iMEx [51]**	**RepeatMasker [52]**	**ISMapper [37]**
Vyhledávací algoritmus	Kombinatoriální	Heuristický	Heuristic
Rozsah velikosti překrytí	1-5 bp	0-100%	5-100bp
Min. délka repetice [bp]	Bez omezení	2	Bez omezení
Detekce nedokonalé repetice	Ano	Ne	Ano
Poznámky	Web	Web	Python v2.7.5

Výše zmíněné vyhledávací nástroje používají symbolické reprezentace genetických sekvencí a na základě jejich algoritmů lze předpokládat vysoké nároky na výpočetní paměť. Některé sice používají kompresní techniky pro snížení zátěže paměti, ale u kombinatoriálních vyhledávacích algoritmů stoupá časová náročnost exponenciálně, obzvláště pokud potřebujeme vyhledávat i překrývající se nebo neúplné MGE. Základní vlastnosti některých nástrojů jsou zmíněny v Tab. 1.

4 Numerická reprezentace

Nevýhodou symbolického zápisu DNA řetězců je špatná aplikovatelnost matematických modelů a vztahů. Avšak existuje i možnost převedení symbolické sekvence do numerické a provedení analýzy v číselné reprezentaci genetických sekvencí za použití standardních metod číslicového zpracování signálů. Tato konverze se využívá pro digitální analýzu signálu, jehož použití se v posledních letech stále zvyšuje pro výzkum genomické DNA. Umožňuje identifikovat skryté znaky a pravidelnosti, které nemohou být odhaleny za pomoci konvenčních DNA symbolických a grafických reprezentačních technik. [20]

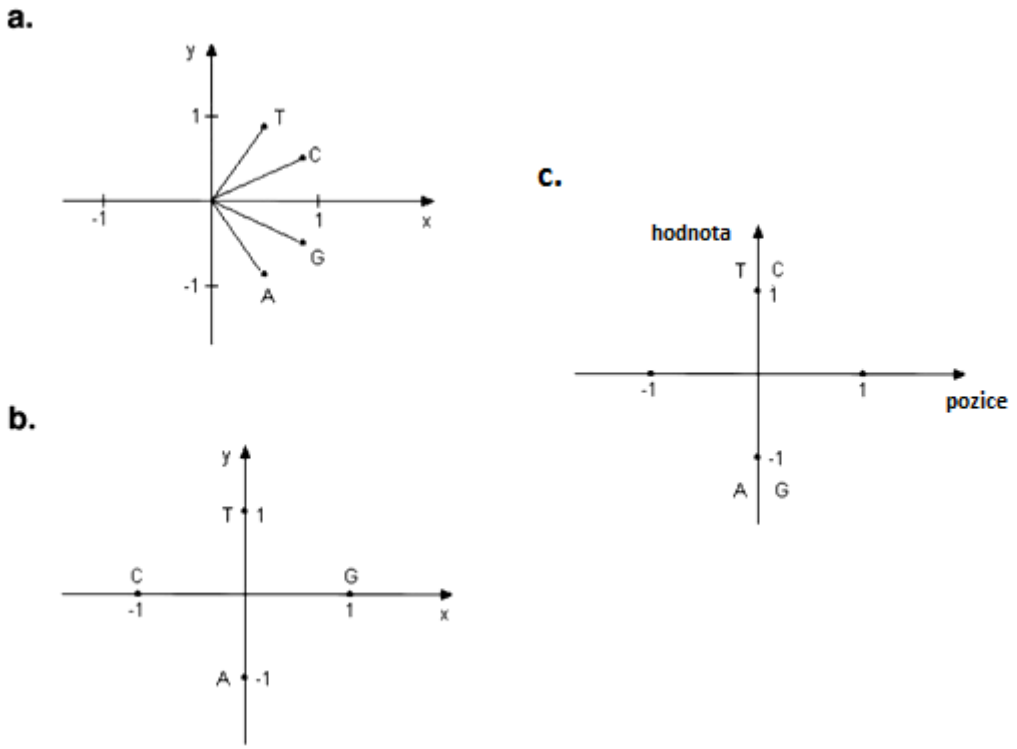
Pro konverzi diskrétních bází DNA sekvence do diskrétní numerické sekvence bylo představeno již mnoho metod. Základní dnes používané metody numerické reprezentace DNA sekvencí společně se základními vztahy konverze (numerickými mapami) a popisem převodu lze nalézt v tabulce (Tab. 10) v příloze. [20] Zároveň jsou některé metody ze zmíněné tabulky rozebrány níže. Grafy jednotlivých numerických reprezentací dále zobrazené v této kapitole vznikly za použití jedné sekvence. Jedná se o mobilní genetický element, v tomto případě inzerční sekvenci IS186A organismu *Escherichia Coli* o délce 1341 bp. Má stejnou strukturu jako na Obr. 4, tudíž se skládá z invertovaných opakování na koncích (zeleně) a navíc má gen pro transpozázu mezi nimi (žlutě). Genetický kód s barevně odlišenými částmi je:

cccataagcg ctaacttaag ggttgtggta ttacgcctga tatgatttaa cgtgccgatg
aattactctc acgataactg gtcagcaatt ctggcccata ttggtaagcc cgaagaactg
gatacttcgg cacgtaatgc cggggctcta acccgccgcc gcgaaattcg tgatgctgca
actctgctac gtctggggct ggcttacggc cccgggggga tgtcattacg tgaagtcact
gcatgggctc agctccatga cgttgcaaca ttatctgacg tggctctcct gaagcggctg
cggaatgccg ccgactggtt tggcatactt gccgcacaaa cacttgctgt acgcgccgca
gttacgggtt gtacaagcgg aaagagattg cgtcttgtcg atggaacagc aatcagtggc
cccgggggcg gcaccgctga atggcgacta catatgggat atgatcctca tacctgtcag
ttcactgatt ttgagctaac cgacagcaga gacgctgaac ggctggaccg atttgcgcaa
acggcagacg agatacgcat tgctgaccgg ggattcggtt cgcgtcccga atgtatccgc
tcacttgctt ttggagaagc tgattatatc gtccgggttc actggcgagg attgcgctgg
ttaactgcag aaggaatgcg ctttgacatg atgggttttc tgcgcgggct ggattgcggt
aagaacggtg aaaccactgt aatgataggc aattcaggta ataaaaaagc cggagctccc
tttccggcac gtctcattgc cgtatcactt cctcccgaaa aagcattaat cagtaaaacc
cgactgctca gcgagaatcg tcgaaaagga cgagtagttc aggcggaaac gctggaagca
gcgggccatg tgctattgct aacatcatta ccggaagatg aatattcagc agagcaagtg
gctgattgtt accgtctgcg atggcaaatt gaactggctt ttaagcggct caaaagtttg
ctgcacctgg atgctttgcg tgcaaaggaa cctgaactcg cgaaagcgtg gatatttgct
aatctactcg ccgcattttt aattgacgac ataatcagcc atcgctggat ttcccccca
gaagtgcgga tccgaaaaga agaactaact cgttgtggag aataacaaaa atggtcatct
ggagcttaca ggtggccatt cgtgggacag tatccctgac agcctacaaa acgcaattga
agaacgcgag gcatcgtctt aacgaggcac cgaggcgtcg cattcttcag atggttcaac
ccttaagtta gcgcttatgg g

Ideální numerická mapa, která slouží pro tranformaci symbolické sekvence do numerické formy, by měla nést stejné množství informace jako symbolická sekvence,

tudíž by měla být schopna postihnout všechny její vlastnosti, neměla by přidávat další vlatsnosti nad rámec již existujících v symbolické sekvenci a měla by být dostatečně jednoduchá, aby umožňovala rychlost a efektivnost analýzy nebo zpracování.

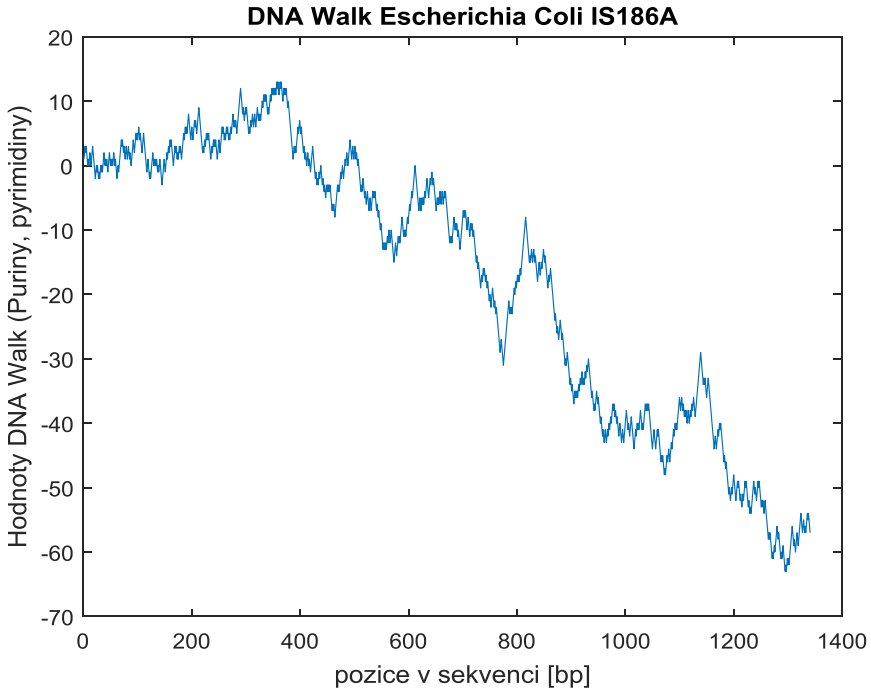
Mezi techniky fixního mapování patří Voss, tetrahedron, celá čísla, reálná čísla a komplexní čísla. Do technik uvažujících fyzikálně-chemické vlastnosti DNA biomolekul patří EIIP (elektron-iontový interakční potenciál ≈ electron-ion interaction potential), atomové číslo, párová čísla, DNA Walk a Z křivka. Každá metoda má své klady a zápory a je vhodná pro jiný typ analýzy. Reálná čísla a celá čísla zavádějí nový matematický vztah mezi bázemi, novou parazitní vlastnost, která není přítomna v původní symbolické sekvenci. Jedná se o různou váhu (hodnotu) každé báze (např.: T<C<A<G) dle přiřazeného čísla, což neodpovídá žádné vlastnosti symbolické sekvence. Takové reprezentace se musí používat velice opatrně. [20]



Obr. 7: Schéma pro reprezentace DNA Walk
a. Yauova reprezentace, b. Gateova reprezentace, c. Krok [21]

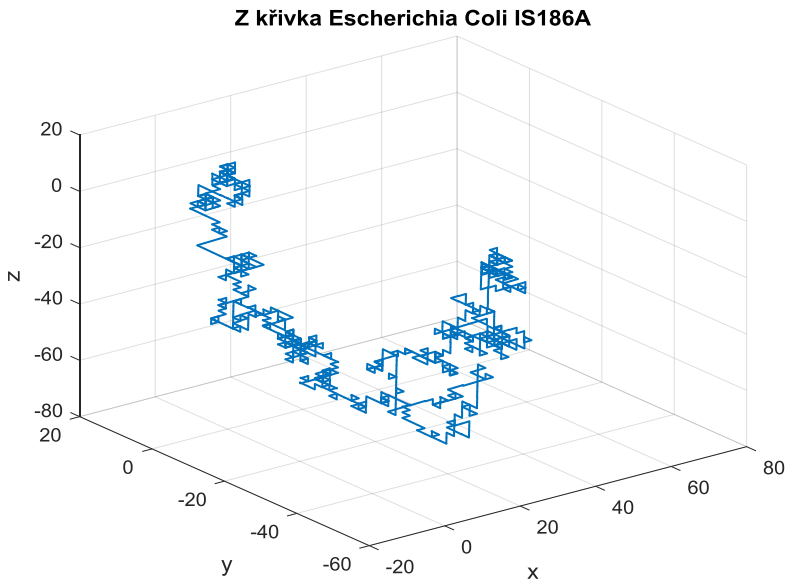
DNA Walk představuje celý soubor technik a tudíž nemá jednoznačně definovanou jednu numerickou mapu. Avšak všechny jeho metody spojuje stejný postup transformace, kdy bereme jednotkové kroky po symbolické sekvenci a tím za sebe klademe jednotkové vektory v číselné reprezentaci jejichž směr (resp. hodnoty) je určen typem nukleotidu (pyrimidin nebo purin, silná nebo slabá vazba, amino nebo keto skupina) na daném kroku. Obvykle se hodnota sekvence zvedá nahoru po krocích kumulativním způsobem. DNA Walk lze použít jako nástroj pro vizualizaci změn v nukleotidové kompozici, vzorů párových bází a evoluci podél DNA sekvence. [20] DNA Walk je výborný pro analýzu při grafickém

zobrazení. Poskytuje jednoduchý způsob jak prohlédnout, zařadit, rozřadit a porovnávat různé genové struktury. Mezi DNA Walk spadá i reprezentace jednotkovými vektory ve všech kvadrantech (Gateova reprezentace) nebo pouze v prvním a čtvrtém kvadrantu (Yauova reprezentace), viz Obr. 7a,b. [21] Yau vytvořil pyrimidonovo-purinový graf na dvou kvadrantech kartézského souřadnicového systému. [22] Bohužel při převodu z vyššího počtu dimenzí do nižšího může docházet k degeneraci sekvence. Například při použití Gateovy reprezentace mají sekvence AGTC, AGTCA a AGTCAG stejnou grafickou reprezentaci. [21] Dalším způsobem DNA Walk je pouhé krokování o -1 níž na pozici purinové báze a o1 výš na pozici pyrimidinové báze jak je ukázáno v Obr. 7c a také zmíněno v Tab. 10 v příloze. [20] Tento poslední způsob nedovoluje získat zpět původní sekvenci oproti prvním dvoum zmíněným metodám DNA Walk a budeme jej nazývat DNA Walk (Krok).

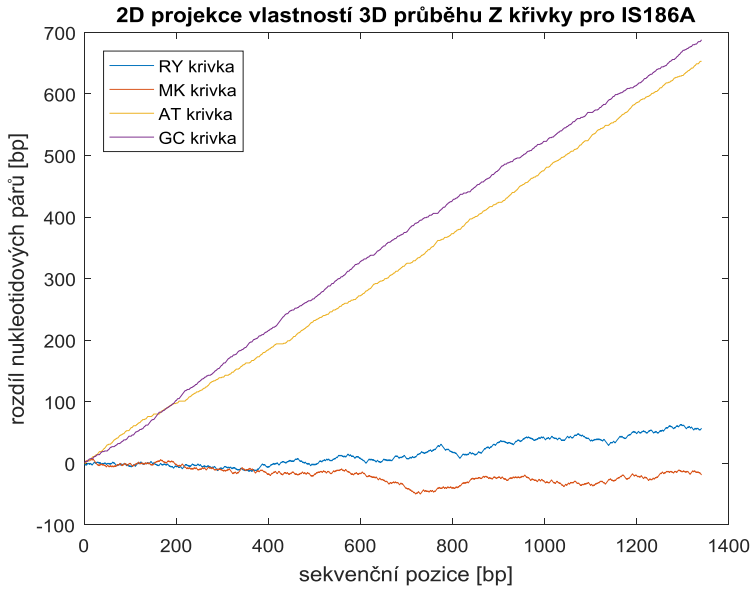


Obr. 8: Numerická reprezentace IS dle numerické mapy DNA Walk (Krok)

Z křivka je unikatní reprezentace pro vizualizaci a analýzu DNA sekvence. Její 3 komponenty odpovídají 3 nezávislým nukleotidovým distribucím, které kompletně popisují DNA sekvenci a jedná se o zatím jednu z nejdokonalejších numerických map, protože nabízí nedegenerativní transformaci. Její komponenty x_n, y_n, z_n zobrazují distribuci purinů proti pyrimidinům (R versus Y), aminů oproti ketonům (M versus K) a silných vazeb oproti slabým vazbám (S versus W) dle vzorců v Tab. [20] Ukázka 3D průběhu Z křivky je na Obr. 9 a její 2D projekce vybraných vlastností jsou používány pro analýzu obsahu sekvence, například AT křivka a GC křivka (dvě části SW křivky) v Obr. 10 ukazující rozdíl zastoupení nukleotidů podle síly vazby. Inverzí lze získat přesný obsah původní sekvence, avšak pro velmi dlouhé sekvence ztrácí Z křivka dobrou rozlišitelnost. [36]



Obr. 9: Numerická reprezentace IS186A dle 3D numerické mapy Z křivky



Obr. 10: 2D projekce křivek RY. MK, AT a GC pro IS186A

5 Analýza numerických sekvencí

Existující detektory MGE pracují se symbolickou reprezentací zabírající větší množství paměti a jedná se o náročné algoritmy. Dohromady tudíž představují výpočetně náročné detektory s dlouhým výpočetním časem. Nabízí se testování detekce v numerické reprezentaci, kdy se s posloupností čísel lépe pracuje a umožňuje větší množství matematických operací. Jelikož tento způsob nebyl doposud prozkoumán, je potřeba provést detailní analýzu různých přístupů k numerické reprezentaci sekvence a možnosti detekce. Metody detekce k testování lze založit na základě již vytvořených pro symbolickou reprezentaci anebo experimentálně testovat jiné pohledy na sekvenci za použití matematických přístupů detekce při číslicovém zpracování signálů.

5.1 Základní data

Data, sekvence, používaná dále ve funkcích a skriptech bylo třeba nejprve stáhnout z internetu a upravit. Z domény bakteriálních genomů bylo vybráno zaměření na čeleď enterobakterií. Jedná se o gramnegativní nesporulující bakterie nacházející se v trávicím ústrojí. [39] Tato čeleď obsahuje více jak 130 druhů a mnoho z nich disponuje velkým množstvím MGE různých typů. [40] Tím tvoří bohatý zdroj na analýzu a realizaci detektoru i otestování. Po analýze několika organismů vzhledem k množství MGE v genomu byly vybrány k dalšímu použití genomy organismů *Klebsiella pneumoniae* druh 12839 (Accession number EU780013.1), *Salmonella typhimurium* LT2 (Accession number NC_003197.2), *Salmonella enterica* plasmid pHCM1 (Accession number NC_003384.1), *Yersinia pestis* CO92 chromozom (Accession number NC_003143.1), *Escherichia Coli* CFT073 (Accession number AE014075.1). Jako inspirace pro výběr těchto sekvencí byly použity [28], [34],[37], kde jsou konkrétně zmíněny přístupová čísla (Accession number) pro nalezení sekvence v NCBI (National Center for Biotechnology Information). Zároveň genomy těchto sekvencí již byly detailně prozkoumány pro MGE i pro použití na detektorech MGE v symbolické reprezentaci. Jednotlivé genomy byly staženy z webových stránek NCBI ve formátu FASTA. Ze stejných stránek byly zároveň získány informace o MGE konkrétního organismu uvedené v Tab. 12, Tab. 13, Tab. 14, Tab. 15 a Tab. 16 v přílohách. Pro zachování integrity jsou sekvence MGE po celou dobu jejich používání pojmenovány podle názvu korespondujícího MGE a zkratkou organismu (např.: KP_IS26 pro IS26 v organismu *Klebsiella pneumoniae*). Genomy jsou pojmenovány dle organismu.

Z genomů bylo třeba vyjmout úseky zájmu neboli MGE. Sekvence byly po stažení načteny do prostředí MatLab použitím funkce `fastaread` a následně z nich vyjmuta nejprve celá sekvence (`FASTAData.Sequence`) do proměnné typu char. Dále z nich byly vystřiženy požadované úseky podle informací v Tab. 12, Tab. 13, Tab. 14, Tab. 15 a Tab. 16 a opět uloženy do proměnné typu `char`.

Dále byly vyjmuty ze stažených genomů i kódující sekvence neobsahující MGE tak, aby měla podobnou délku jako MGE. Tyto sekvence slouží pro porovnání s MGE a nalezení odlišností, které by mohly pomoci nalézt motiv pro budoucí detektor. Od každého organismu byl vybrán jeden až dva geny. Jejich přehled je k vidění v Tab. 2. Pro pokrytí všech možností a variací byl vytvořen i soubor 5 náhodných sekvencí pro analýzu a porovnávání se sekvencemi MGE. Náhodné sekvence byly vytvořeny v symbolické reprezentaci a pro další použití převedeny do numerické společně se sekvencemi MGE. Nejdůležitější pro začátek analýzy však je vytvořit si modelovou krátkou sekvenci MGE podle Obr. 4

```
tcatacgtat gtctttagcg gattgtgcga gacaatgcta ccttaccggt cggaactcga
tcggttgaac tctatcacgc cgctaaagac atacgtatga
```

Všechny použité sekvence a genomy lze najít ve složce *Data* na přiloženém DVD ve formátu umožňujícím jednoduchou manipulaci a použití pro skripty a funkce, jež jsou taktéž k naleznutí na přiloženém DVD ve složce *Funkce*.

Tab. 2: Geny pro porovnání [49]

Organismus	**Accession number organismu**	**Gen**	**Pozice začátku**	**Pozice konce**	**Délka**
Klebsiella pneumoniae	EU780013.1	Gentamicin-(3)-N-acetyltransferáza	23216	24076	861
Salmonella typhimurium	NC_003197.2	Homoserin kináza	2789	3730	942
Salmonella enterica	NC_003384.1	Methyláza DNA modifikace	37651	38475	825
Yersinia pestis	NC_003143.1	Asparagin synthetáza AsnA	1435	2427	993
Escherichia Coli	AE014075.1	Protein chaperonu DnaK	12946	14862	1917

5.2 Konverze dat do numerické reprezentace

Pro aplikaci standardních metod číslicového zpracování signálů je třeba převést připravené symbolické sekvence a genomy do numerických. Byly tedy vytvořeny funkce podle matematických vztahů v numerických mapách v Tab. 10, avšak ne pro všechny zmíněné v této tabulce. Byly vybrány ty reprezentace, které mají nízkou nebo žádnou redundanci a jsou vhodnější pro další zpracování. Mezi nimi jsou Voss (10), tetrahedron (11), komplexní čísla (14), EIIP (16), DNA Walk (19), (20), (21), Z křivka (22). Funkce jsou pojmenovány adekvátně `voss.m`, `Tetrahedron.m`, `complexRep.m`, `eiip.m`, `DNA_WalkPP.m`, `DNA_WalkGate.m`, `DNA_WalkYau.m` a `Zkrivka.m`. Všechny tyto funkce mají stejný požadovaný vstup a to sekvenci v symbolickém zápisu ve formátu `char`. Počet výstupních

proměnných z každé této funkce odpovídá počtu sekvencí numerické mapy dle vzoru výstupu v posledním sloupci v Tab. 10 a jejich výpočet taktéž. Výstup je již převedená sekvence ze symbolického zápisu do numerického. Typ numerické sekvence je daný použitou funkcí.

Tab. 3: Výstupní proměnné funkcí pro konverzi do numerické reprezentace

Funkce	**Výstup**	**Formát výstupních proměnných**
`voss.m`	`[C, G, A, T]`	`logical`
`Tetrahedron.m`	`[xr,xg,xb]`	`double`
`complexRep.m`	`[SB]`	`complex double`
`eiip.m`	`[SB]`	`double`
`DNA_WalkPP.m`	`[SB]`	`double`
`DNA_WalkGate.m`	`[x,y,z]`	`double`
`DNA_WalkYau.m`	`[Cx, Cy]`	`double`
`Zkrivka.m`	`[Cx, Cy]`	`double`

Každá ze zmíněných funkcí pro konverzi genetické sekvence převede pouze jednu sekvenci. Pro konverzi skupiny genetických sekvencí uložených v proměnné `sekvence` byla vytvořena nadstavbová funkce `reprezentace.m`, avšak lze použít na konverzi i jedné sekvence. Prvním vstupem této funkce je proměnná typu cell obsahující názvy sekvencí v prvním sloupci a samotné sekvence v druhém sloupci, oboje typu char. Druhým vstupem je textový řetězec `typ_repre` identifikující jaká numerická mapa se má použít pro konverzi. Výstupem je opět proměnná typu cell, v rámci funkce nazvaná `ElementyRepre`. První sloupec obsahuje stejnou informaci jako vstupní proměnná se sekvencemi, tedy jejich názvy. Celkově však může mít dva až pět sloupců. Počet je daný množstvím výstupů jednotlivých funkcí pro konverzi. Například funkce `voss.m` dává 4 výstupní vektory a tudíž výstupní proměnná funkce reprezentace bude mít 5 sloupců, jeden s názvy a další čtyři s vektory. Jsou uloženy ve výstupní proměnné ve stejném pořadí jako výstup použité funkce pro konverzi. Vstupní proměnná `typ_repre` řídí podmíněný příkaz `switch-case-end`, čímž je zajištěn výběr určité numerické mapy. Každý případ příkazu `switch` zajistí v rámci `for` cyklu postupné vybrání všech sekvencí ze vstupní proměnné se sekvencemi, jejich převod do numerické reprezentace a uložení vektoru/ů numerické reprezentace do výstupní proměnné `ElementyRepre`. Zároveň zobrazí v oddělených oknech grafy vektorů pro jednotlivé MGE i s nadpisem udávajícím jeho název a popisky os.

5.3 Analýza podobnosti mezi mobilními genetickými elementy

Jako první analýza numerické reprezentace MGE se nabízí nalezení podobnosti mezi takovými signály. A protože máme k dispozici sekvence v diskrétní numerické reprezentaci, můžeme se přiklonit k použití prostředků digitální analýzy signálů. Pro hledání podobnosti mezi dvěma signály je nejvhodnější použití korelace. [24]

5.3.1. Teorie o korelaci

Účelem analýzy signálů v časové oblasti je nacházet vazby mezi hodnotami signálu v různých zvolených časových okamžicích a vhodně tyto vazby kvantitativně vyjádřit (nebo popřípadě zjistit neexistenci takových vazeb). Pravděpodobnostní vztahy mezi hodnotami signálů, např. f_m a f_n, lze pak vystihnout pomocí korelace, jejíž vztah je určen jako

$$R_{fg}(m,n) = E\{f_w(m)g_w(n)\} \approx \frac{1}{M}\sum_{w=w_1}^{w_M} f_{w_i}(m)g_{w_i}(n) \qquad \textbf{(3)}$$

Kde E představuje souborovou střední hodnotu, $f_w(m)$ je hodnota funkce f_w v okamžiku m, $g_w(n)$ je hodnota funkce g_w v okamžiku n, a M je počet diskrétních vzorků shodný pro obě funkce. [24] Tato funkce je však dvourozměrná, závisí totiž na časových okamžicích m a n. Pokud zavedeme nezávislost pravděpodobnostních charakteristik na čase, tedy při jejich definování nezáleží na stanovení počátku osy, můžeme zjednodušit vztah na jednorozměrný. Závislost se pak bude týkat rozdílu času a ne na absolutních dvou časech, tedy vztah se změní na

$$R_{fg}(m,n) = R_{fg}(m+d,n+d) = R_{fg}(m-n) = R_{fg}(\tau) \qquad \textbf{(4)}$$

Rozdíl τ (posun) 2 funkcí vůči sobě, neboli vzdálenost mezi m a n, a d je hodnota posunu na každé funkce. [24] Takto upravený vztah lze jednoduše aplikovat na porovnávání signálů ve formátu

$$r_{xy}(\tau) = \frac{1}{M}\sum_{i=0}^{M-1} x(i)y(i-\tau) \qquad \textbf{(5)}$$

Definovanou pro $x(n)$ a $y(n)$, kde $n = 0, \ldots, M-1$. $r_{xy}(\tau)$ představuje hodnotu korelace mezi dvěma signály x a y pro hodnotu posunu τ, a M je počet diskrétních vzorků signálu. [24]

5.3.2. Korelace v Matlabu

Tento vztah (5) lze najít v prostředí MatLab realizovaný pod funkcí `xcorr` integrovaný v Signal Processing Toolboxu. Při jeho nejjednodušším použití jako `r = xcorr(x,y)` nemusí mít signály stejnou délku. Tato funkce si kratší signál doplní nulami na jeho konec tak, aby měly signály stejnou délku. [25] Výstupem je vektor neváhovaných hodnot korelace mezi dvěma signály pro každý vzájemný posun. Tato funkce umožňuje definovat dodatečné parametry ovlivňující jak vstupní signály, tak výstupní vektor. Umožňuje volbu normalizace výstupního vektoru a tudíž zisk váhovaných hodnot korelace. Například `'coeff'` normalizuje hodnoty korelace tak aby autokorelace signálu při nulovém posunu byla rovna 1.

$$\hat{R}_{xy,coeff}(m) = \frac{1}{\sqrt{\hat{R}_{xx}(0)\hat{R}_{yy}(0)}}\hat{R}_{xy}(m) \qquad \textbf{(6)}$$

Ve tomto vztahu představuje $\hat{R}_{xx}(0)$ autokorelaci signálu x při nulovém posunu a $\hat{R}_{yy}(0)$ má stejný význam. $\hat{R}_{xy}(m)$ je hodnota korelace mezi signály x a y při posunu m. Tímto způsobem je dosaženo zmíněné normalizace, kdy při $\hat{R}_{xy,coeff}(m) = 1$ jsou signály x a y naprosto shodné na vájemném posunu m. [25]

5.3.3. Výsledky korelace mobilních genetických sekvencí

Funkce `korelace_vysledky.m` byla vytvořena za účelem zisku hodnot nejlepších korelací mezi numerickými sekvencemi i s pozicí vzájemného posunu pro tuto hodnotu. Na základě těchto výsledků lze dále stanovit, zda se MGE navzájem podobají.

Vstupem zůstává proměnná typu buňkového pole (cell). Funkce je koncipovaná na práci s výstupem `ElementyRepre` z funkce `reprezentace.m`.

Základem funkce je výpočet korelaci dvou numerických sekvencí ze vstupní proměnné `ElementyRepre`. Pokud je jedna ze sekvencí kratší nebo delší, posouvá se okno po delší sekvenci a korelace se počítá v okně. Délka okna je rovná délce kratšího signálu. Tímto způsobem se eliminuje problém s požadavkem stejné délky vstupních sekvencí do funkce `xcorr` pokud potřebujeme získat jako výstup normalizované hodnoty korelace.

Výstup funkce jsou dvě proměnné typu cell, kde nalezneme v první výstupní proměnné `podobnostiXcorr` je matice hodnot korelací mezi všemi možnými dvojicemi signálů a v druhé výstupní proměnné `posunyXcorr` pozici vzájemného posuvu pro korespondující hodnoty v proměnné `podobnostiXcorr`. Každá výstupní proměnná má tolik buněk, kolik je vektorů numerické reprezentace a jsou ve stejném pořadí jako výstupy funkce počítající konverzi symbolické sekvence dle numerické mapy. Na zvoleném souboru dat, na 10 náhodně

vybraných sekvencích MGE (Tab. 4), 5 sekvencích genů (Tab. 2) a 5 náhodných sekvencích, byla provedena korelace v plovoucích oknech pro 8 různých numerických reprezentací genetické sekvence. Pro zajištění validity analýzy byly zahrnuty sekvence neobsahující žádný MGE (sekvence genů a náhodné sekvence) a vůči nim byly testovány korelace zmíněných MGE. Tento soubor dat symbolických sekvencí lze nalézt v souboru `korelaceMGEdata.mat` na přiloženém DVD.

Tab. 4: Náhodně vybrané sekvence MGE pro korelaci

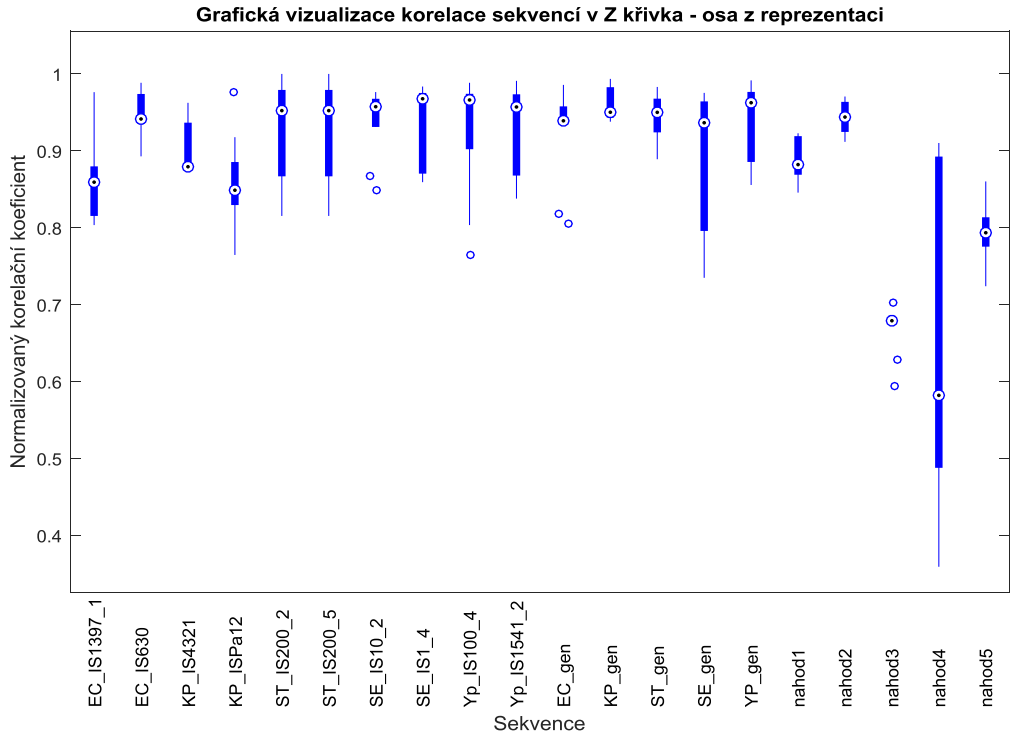
Organismus	**Délka genomu [bp]**	**Typ MGE**	**Délka [bp]**
Klebsiella pneumoniae	37606	IS4321	1327
		ISPa12	1387
Salmonella typhimurium	4857450	IS200_2	709
		IS200_5	709
Salmonella enterica	218160	IS10	1315
		IS1	748
Yersinia pestis	4653728	IS100	1954
		IS1541	712
Escherichia Coli	5231428	IS1397	1427
		IS630	1110

Každá matice korelací všech možných dvojic mezi sekvencemi byla prozkoumána a byla zjišťována průměrná hodnota korelace, maximální hodnota korelace a minimální hodnota korelace pro každý převod dle numerické mapy zvlášť. Zároveň bylo zkoumáno, zda je signifikantní rozdíl mezi vzájemnými hodnotami korelace MGE oproti hodnotám korelace mezi MGE a sekvencemi bez MGE (sekvence genů a náhodné sekvence). Také se pohlíželo na minimální hodnotu korelace a její výskyt. Obzvlášť jestli minimální hodnota korelace mezi MGE je nižší nebo vyšší než minimální hodnota korelace mezi kódujícími sekvencemi a ostatními MGE. K zobrazení hodnot korelací a nalezení rozdílu nebo podobnosti nejlépe posloužilo zobrazení v boxplotech (viz Obr. 11)

Mnoho matic korelací různých numerických reprezentací neudává žádný rozdíl, tedy se nenašel žádný rozdíl mezi numerickou reprezentaci sekvencí bez MGE a sekvencemi s MGE, například v Tab. 19 a všechny sekvence v souboru dat si byly velice podobné. Zaznamenatelný rozdíl se vyskytuje při použití numerických map DNA Walk (puriny, pyrimidny) a u Z křivky pro převod symbolické genetické sekvence ještě větší.

Při konverzích Z křivky jsou v hodnotách korelací signálů z-ové osy viditelné rozdíly mezi korelacemi signálů MGE versus náhodné sekvence, ale i signály genů versus inzerční

sekvence, viz Tab. 18. Z-ová osa hovoří o distribuci silných vazeb oproti slabým vazbám v sekvenci. Při pohledu na statistiky minimálních a maximálních hodnot korelací v Tab. 18 lze říci, že rozsah hodnot korelací mezi geny a MGE je <0,73; 0,99> se značně překrývá s rozsahem pro korelace mezi MGEnavzájem <0,76; 1,00>. Nelze na základě takto velkému překryvu vyvodit žádný rozdíl v numerické reprezentaci MGE oproti numerické reprezentaci genů. Významnější rozdíl je mezi MGE a signály náhodných sekvencí. Normalizované korelační hodnoty mezi nimi se pohybují v rozsahu <0,36; 0,97>. Přesto je mezi nimi velká míra podobnosti znázorněná i překryvem boxplotů v Obr. 11.



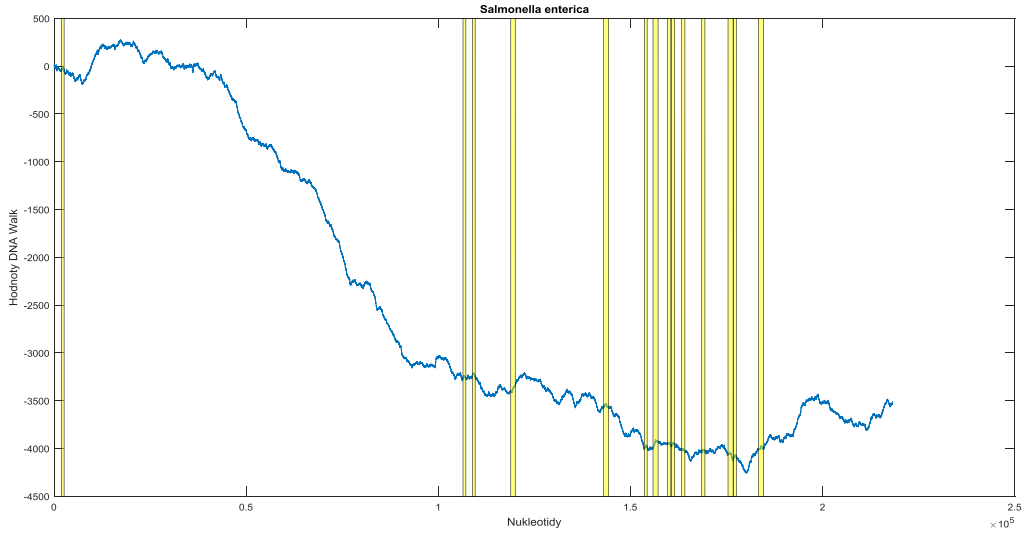
Obr. 11: Boxploty hodnot korelací mezi sekvencemi pro z-ovou osu Z křivky

5.4 Analýza charakteru genomického signálu

Po prozkoumání a hledání strukturální podobnosti mezi číslicovými signály reprezentace MGE bylo přistoupeno k hledání charakteristického znaku nebo znaků v rámci celého genomu v organismu. Protože dnes již jsou k dispozici prostředky pro zisk celého genomu organismu, můžeme tak převést s pomocí numerické mapy celý genom a zkoumat jeho charakteristiky jako celku nebo procházet genom po částech a sledovat jeho vývoj.

Pro zkoumání MGE v rámci celého genomu je třeba vyznačit části genomu, kde se MGE nachází. K tomu byla na základě dat z Tab. 12, Tab. 13, Tab. 14, Tab. 15 a Tab. 16 vytvořena buňková pole obsahující posloupnosti pozic začátku a konce každého typu MGE zvlášť po genomech (např.: `EC_pozice`). Po převedení symbolické reprezentace genomu

do numerické a jeho vykreslení, byly zvýrazněny za použití těchto vektorů části sekvence obsahující MGE. (Obr. 12) Pro barevné zvýraznění těchto částí byla použita volně dostupná funkce `ShadePlotForEmpahsis` na webu Mathworks. [41]



Obr. 12: Genom *Salmonella enterica* s vyznačenými MGE
Použitá numerická mapa DNA Walk (Krok)

Jednotlivé numerické reprezentace prozkoumávaných organismů byly postupně procházeny a vizuálně studovány jakékoliv strukturální změny v průběhu signálu. Byly hledány znaky, jež se objevují v průběhu jiných biologických signálů a charakterizují určitý hledaný jev používány dále v diagnostice. Záznam časové změny elektrického potenciálu způsobeného aktivitou žaludeční stěny, elektrogastrogram, se při kontrakcích svalstva vyznačuje zvýšením kmitočtu stahů viditelného na signálu. [42] Změna frekvence v signálu je stěžejní i ve studiu elektroencefalogramu. Kmitočtový obsah záznamu elektrické aktivity mozku je členěn do skupin dle frekvence. Tato frekvenční pásma odpovídají různým stavům (spánek, bdění, …). Nehodnotí se pouze jednotlivé rytmy, ale i amplitudy, tvary, netypické grafoelementy a další. To vše má diagnostický význam např.: u posuzování epilepsie a jeho ložiska nebo stanovení mozkové smrti. Grafoelementy jsou obvykle vázané na patologické jevy a mohou to být komplexy hrotů a vln, vřetena, ostré vlnky apod. [43],[44]

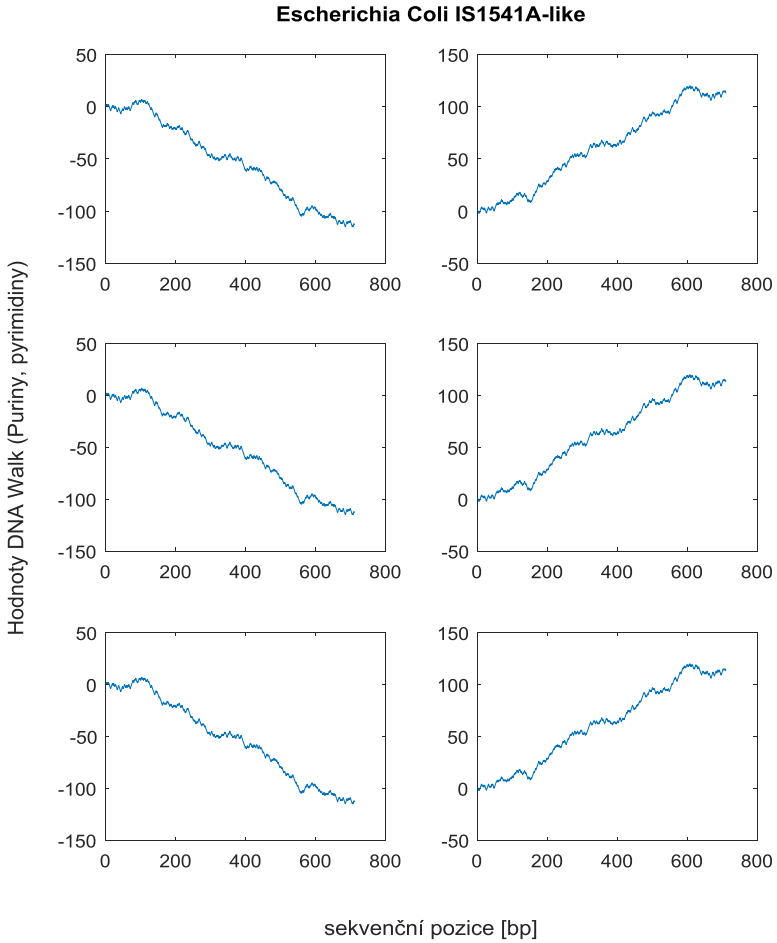
Procházení signálů genomů s vyznačenými úseky MGE slouží k vizuální analýze průběhu a studiu případných strukturálních odlišností charakterizujících přítomnost MGE. Zaměření bylo na strukturální odlišnosti nebo změny v průběhu signálu typické pro jiné biologické signály:

- Změna frekvence
- Změna amplitudy
- Tvary (změna nebo jeden tvar či struktura objevující se napříč MGE)

- Grafoelementy (píky, hroty, vlny, vřetena, průchody izolinií atd)

Pro vizuální analýzu celého průběhu genomu nejsou některé numerické mapy vhodné. Hlavními důvody jsou postrádání signálového charakteru například u Voss reprezentace, délka genomu nebo nepřehledný charakter numerické reprezentace. Ve vhodných reprezentacích genomů nebyly průzkumem nalezeny žádné odlišnosti mezi částmi MGE a zbytkem signálu. Nebyl zpozorován ani nějaký druh přechodového jevu v okolí počátku nebo konce MGE. Naopak bylo zjištěno že konkrétní typ MGE má téměř stejný tvar ať už se nachází kdekoliv v numerické sekvenci genomu. (Obr. 13) Tudíž opačný efekt, vliv pozice v genomickém signálu, úseků CDS, genů, intronů, exonů apod nemá vliv na tvar MGE také. Proto byly vyše zmíněné odlišnosti dále hledány jak v části sekvence signálu odpovídající pouze MGE tak před začátkem části signálu MGE i za koncem části signálu s MGE.

Z genomického signálu byly vyjmuty jak části MGE s úseky před a za nimi tak samostatné začátky a konce jednotlivých MGE – invertovaná opakování. K získání zmíněných úseků byly vytvořeny další vektory s hodnotami pozic vycházejících z původních vektorů (např.: `EC_pozice.mat`).



Obr. 13: DNA Walk (Krok) elementu IS1541A-like z různých míst genomu *Escherichie Coli*
Grafy na pravé straně patří reverzně komplementárním elementům IS1541A-like

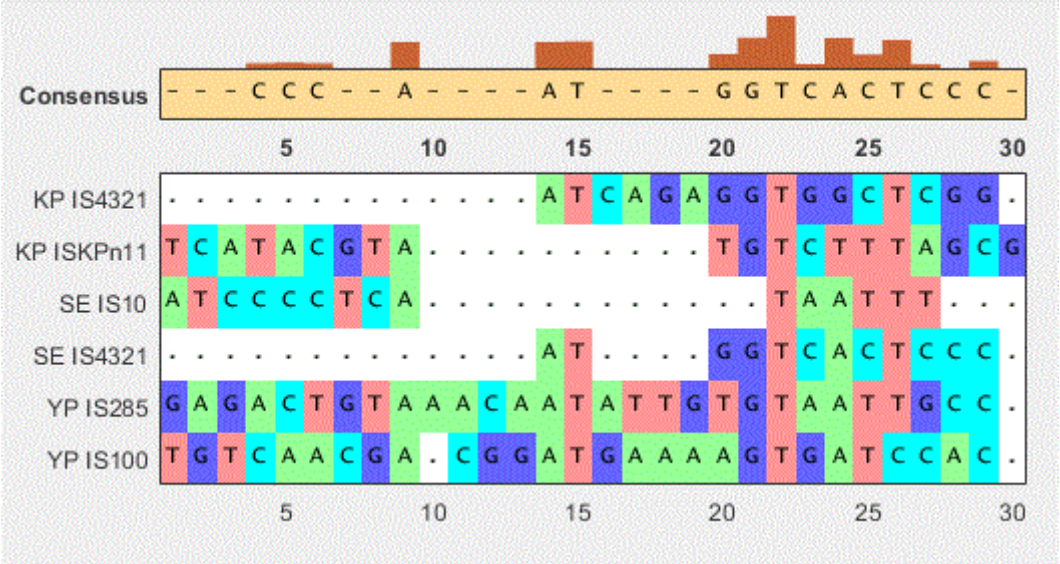
Úsekům před a za MGE v signálu byla věnována pozornost navíc. Byly zobrazeny jak samostatná invertovaná opakovaní tak i s přesahem do oblastí před a za invertovanými opakování. Délka prodloužených úseků byla zvolena o velikosti 200 vzorků (bp), protože se průměrná délka MGE ve zvolených organismech pro tuto práci pohybuje kolem 1000 bp a úsek o 200 vzorcích představuje dvacet procent této délky. Opět některé numerické reprezentace nedávaly dostatečně vhodné zobrazení použitelné pro další detekce nebo naleznutí charakteristického znaku nebo znaků. Při vykreslení těchto prodloužených úseků MGE ve vhodných numerických reprezentacích byl potvrzen předchozí poznatek ohledně stejného průběhu daného MGE nezávisle na pozici v sekvenci a není ovlivněm žádným jiným parametrem ze signálu, jak je možné vidět na IS1541A-like na Obr. 13, protože navzdory výseku z genomu, všechny vypadají stejně. Ani před MGE ani za MGE nebyl zpozorován žádný jev, který by všechny MGE doprovázel a byl využit jako znak přítomnosti MGE v signálu.

Tab. 5: Invertovaná opakování (IRL, IRR) mobilních genetických elementů [49]

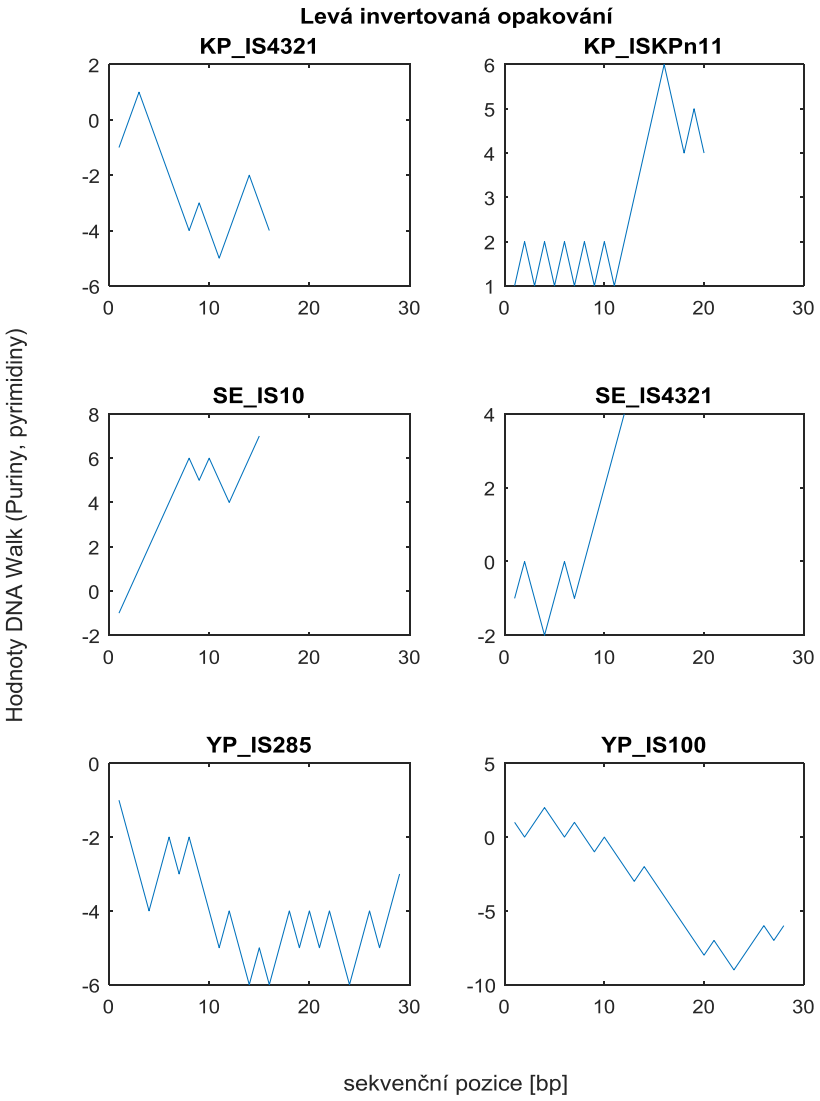
Organismus	**MGE**	**Začátek IRL [bp]**	**Konec IRL [bp]**	**Délka IRL [bp]**	**Začátek IRR [bp]**	**Konec IRR [bp]**	**Délka IRR [bp]**
Klebsiella pneumoniae	ISKpn11	1	16	16	1275	1290	16
	ISPa12	1	20	20	1368	1387	20
Salmonella enterica	IS10	1	15	15	1301	1315	15
	IS4321	1	12	12	1306	1317	12
Yersinia pestis	IS285	1	29	29	1287	1315	29
	IS100	1	28	28	1927	1954	28
Escherichia Coli	IS1397	1	25	25	1402	1426	25

Invertovaná opakování mají různé délky pro různé MGE. Pohybují se od 7 bp až po 150 bp. [17] Pro zobrazení byly vybrány MGE s invertovanými opakováními o délce 12 až 29 bp pro snadnější vizuální analýzu. [17] Zvlášť se zkoumala levá invertovaná opakování (počátek MGE, IRL) a pravá invertovaná opakování (konec MGE, IRR). Grafy byly tvořeny daty různými MGE ze všech organismů (viz Tab. 5). Tyto úseky však nebyly zkoumány pouze na úrovni numerické reprezentace, ale byly vyjmuty i ze symbolické reprezentace genomů, zarovnány a zobrazeny pro porovnání s numerickou reprezentací. Tímto by měl být získán přehled o obou reprezentacích a umožněna snadná kontrola úseku nebo tvaru ze signálu s jejím protějškem v symbolické genetické sekvenci. I zde nebylo možné pro zkoumání použít některé numerické reprezentace. Na grafech znázorňujícím několik IRL MGE (Obr. 15) lze vidět, že rozdíly v délce mohou být signifikantní. Tento fakt by ztěžoval naleznutí společného motivu IRL nebo IRR. Avšak v žádné z os Z křivky (např.: Obr. 27, Obr. 28), v EIIP nebo

v DNA Walk (např.: Obr. 15, Obr. 26) reprezentaci není viditelná společná charakteristika mezi IRL nebo IRR navzájem. Při průzkumu zarovnaných symbolických sekvencí (Obr. 14, Obr. 29) je potvrzena odlišná skladba sekvencí způsobujících různé signály v reprezentaci jak pro IRL tak pro IRR. K zarovnání viditelnému v Obr. 14, Obr. 29 byla využita funkce matlabu `multialign` a pro zobrazeni funkce `seqalignviewer`. Multialign byl použit s defaultními parametry, metoda váhování sekvencí (`WeightsValue`) nastavená na `THG`, skórovací matice (`ScoringMatrixValue`) na `NUC44`.



Obr. 14: Zarovnání symbolických sekvencí IRL zobrazených v Obr. 15



Obr. 15: Porovnání IRL po konverzi numerickou mapou DNA Walk (Krok)

5.5 Spektrální analýza

Předchozí analýzy genetického signálu v numerické podobě se odehrávaly v časové oblasti. Stále zbývá probádat frekvenční oblast, jinak časově-frekvenční. Frekvenční analýza poskytuje alternativní formu zobrazení, která rozloží signál do komponent, jež od sebe oddělí jednotlivé jevy promíchané v časové oblasti. [45] Je potřeba převést časové signály do frekvenční oblasti. Signály získané převody dle numerických map nejsou ze své povahy časové, ale pro použití matematických vztahů umožňujících frekvenční analýzu se bude považovat genetický signál za časový, přičemž jeden nukleotid bude odpovídat jedné sekundě.

5.5.1. Fourierova transformace

Rozložení komponent signálu může být dosáhnuto několika způsoby, jedním z nich je Fourierova transformace. [45] Podle Fourierovy transformace je signál aditivní směs nekonečného počtu harmonických složek, to je amplitudy a počáteční fáze všech jeho složek. Spektrum diskrétního signálu (to je vzorkovaného signálu), představující rozložení lineární kombinace diskrétních harmonických složek, je dáno jeho diskrétní Fourierovou transformací

$$F(k\Omega) = \sum_{n=0}^{N-1} f(nT)\mathrm{e}^{-jk\Omega nT}, \qquad (7)$$

kde n a k jsou celočíselné indexy a N je počet prvků transformované posloupnosti. V rovnici se signál f s časovou posloupnosti vzorků nT (T je vzorkovací perioda) rozloží na harmonické komponenty o frekvencích $k\Omega$ (Ω je N-tina vzorkovacího kmitočtu). Pak $F(k\Omega)$ představuje komplexní spektrální koeficienty. [24], [46]

Absolutní hodnota komplexního spektrálního koeficientu udává příslušné harmonické složky o kmitočtu $k\Omega$ a argument (fáze) tohoto komplexního čísla udává počáteční fázi této složky. Soubor všech N spektrálních koeficientů vytváří diskrétní spektrum daného signálu, soubor absolutních hodnot těchto koeficientů je **amplitudové spektrum** a soubor fází koeficientů tvoří **fázové spektrum**. [24]

5.5.2. Výkonové spektrum

Stochastický signál je výsledkem stochastického (náhodného) procesu a jeho hodnoty není možné předem znát, nanejvýš můžeme předpokládat, že budou v nějakém intervalu s nějakou pravděpodobností. Jsou tedy neurčité. To samé platí i o genetické sekvenci a tudíž její numerickou reprezentaci taktéž budeme považovat za stochastický signál. Je třeba mít na paměti, že jakmile proběhne realizace náhodného signálu, v našem případě skenování sekvence nebo genomu, stává se signál deterministickým, jelikož již známe celý jeho průběh a již je neměnný.

Spektrální analýzu stochastického signálu umožňuje výkonové spektrum, jež je souborovou střední hodnotou individuálních spekter přes všechny realizace procesu. V praxi lze však výkonová spektra pouze odhadovat na základě konečného počtu realizací náhodného signálu. K dispozici jsou dvě základní metody odhadu, metoda periodogramu a metoda korelogramu.

Metoda periodogramu aproximuje souborovou střední hodnotu průměrem individuálních výkonových spekter z M realizací a lze odhadovat pouze na základě konečného počtu realizací signálu. V praxi se přistupuje k váhování oknem $w(n)$ a pro každé spočítáme

spektrum díky DFT , následně se složky spektra umocní a v posledním kroku sečtou a zprůměrují.

Jelikož se pohybujeme stále v oblasti používání metod číslicového zpracování signálu na genomická data, je třeba i zde osvětlit použití některých pojmů. Například vzorkovací frekvence f_{vz} udává při snímání signálů počet vzorků zaznamenaných za jednotku času, její měřitelnou jednotkou je pak $Hz = s^{-1}$. V genomickém signálu však odpovídá jeden vzorek jedné pozici v genetické sekvenci a nijak nesouvisí s časem. Pojem vzorkovací frekvence tak je třeba upravit na pouhé relativní měřítko s bezrozměrnou jednotkou. Relativní vzorkovací frekvence $f_{vz} = 1$ je jednotná pro všechny genomické signály.

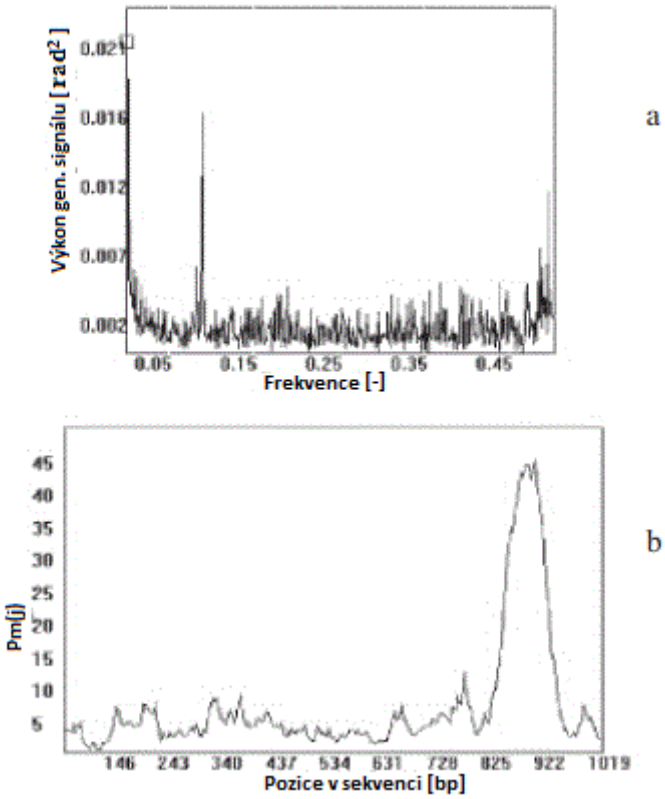
Je známým faktem, že v DNA sekvencích jsou přítomny periodicity, zvláště v určitých částech genomu jako exony. [47] V DNA sekvencích se vyskytuje tří bázová periodicita, která vytváří velice ostrý hrot ve výkonovém spektru na frekvenci 1/3. [35] Častější výskyt jednoho MGE v genomu lze taktéž označit jako peridocitu anebo repetici. To vede k myšlence detekce MGE ve výkonovém spektru.

Podle [35] lze získat kroky algoritmu detekce repetic při spektrální analýze:

- Zisk sekvence v numerické reprezentaci.
- Spočítat výkonové spektrum a průměr spektra přes celou sekvenci. Každou hodnotu výkonového spektra podělit hodnotou průměru.
- Identifikovat ve spektru všechny píky vyšší než hranice. Každá takto identifikovaná frekvence f_i je potenciální repetice o délce $N_i = 1/f$ (Obr. 16a). Na základě simulací byla v algoritmu dle [35] zvolena hranice o hodnotě 4.
- V klouzavém okně délky m vycentrovaný na pozici j v sekvenci se opět počítá výkonové spektrum, jeho průměr, podělí tyto hodnoty a následně zjistí, zda se v tomto posunu okna nachází vyšší hodnota frekvence f_i než je stanovená hranice. Tímto způsobem zjistíme, v které části sekvence se repetice nachází. (Obr. 16b)
- Nyní je známá délka repetice i úsek sekvence, v němž se nachází jedna nebo více.
- V úseku se uvažují všechny možnosti repetic o délce N_i a zarovnáním se identifikují ty vyskytující se nejčastěji.

Vysvětlení algoritmu v [35] však trpí vážnými nedostatky. Prvním je chybějící specifikace použité numerické mapy, dále vzorec uvedený pro výpočet výkonového spektra nemá vysvětleny všechny prvky a chybí specifikace typu zarovnání sekvencí. Tudíž tyto chybějící nebo nedostatečně osvětlené části byly upraveny. Pro výpočet výkonového spektra se použila metoda periodogramu, která se aplikovala na všechny numerické mapy. Provedly

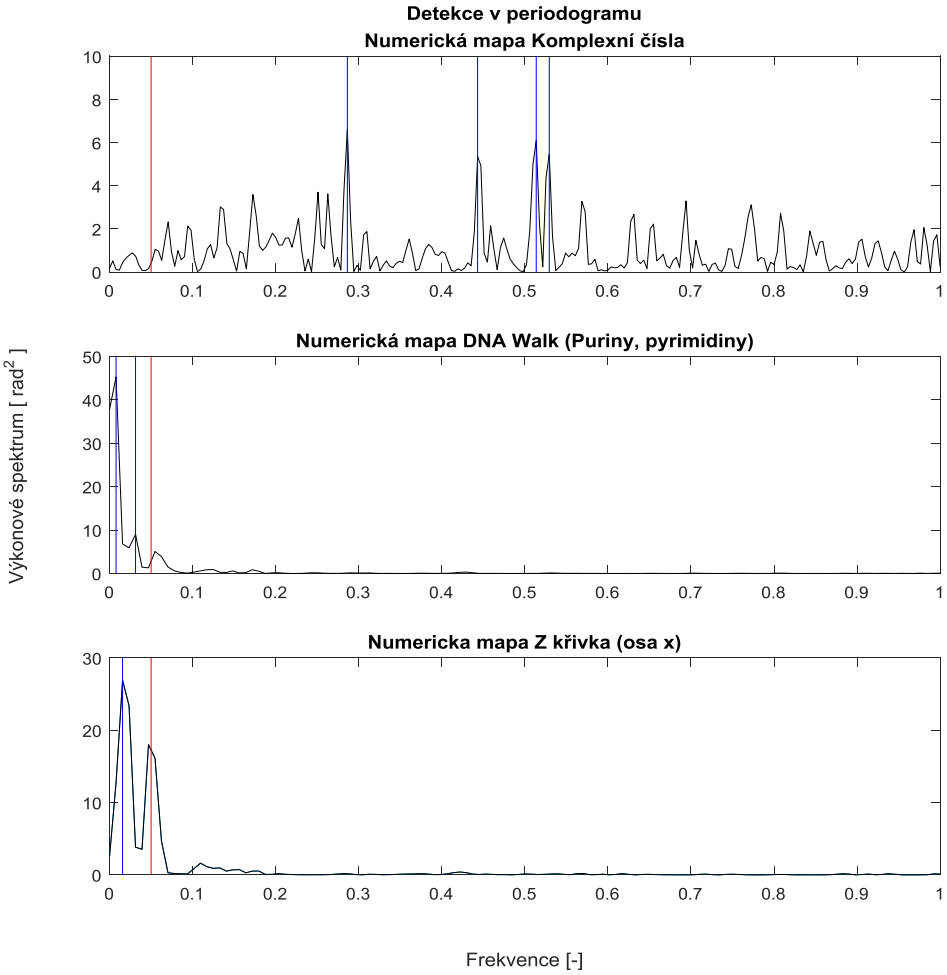
se tedy pouze první tři kroky algoritmu, aby prvotním úkolem zůstalo zjistit, zda při těchto modifikacích lze na modelové sekvenci stále detekovat MGE jako repetici, jinými slovy, zda nalezneme ojedinělé píky vyšší než hranice. Analýza byla provedena na modelové sekvenci zmíněné v kapitole 5.1 s použitím numerických map pro komplexní čísla, EIIP, všechny postupy DNA Walk a osy Z křivky. Vzhledem k repetici o délce 20 bp v modelové sekvenci byl hledán pík vyšší než hodnota 4 na frekvenci 0.05 (1/20).



Obr. 16: Detekce repetic ve výkonovém spektru – teorie [35]
a. Výkonové spektrum podělené jeho průměrem s píky na f=1/11 a f=1/2
b. Výsledek analýzy klouzavým oknem se skenovací frekvencí f=1/2, repetice se nachází mezi pozicemi 800 a 900bp

Po prozkoumání všech periodogramů modelové sekvence byly identifikovány 3 numerické mapy, kde se pík na požadované frekvenci objevil. Jedná se o DNA Walk (Krok), x-ová osa DNA Walk metody Gate a x-ová osa Z křivky. Na Obr. 17 lze vidět v prvním grafu nevyhovující upravený periodogram (podělený průměrem výkonového spektra) při použití numerické mapy Komplexní čísla. Jelikož modelová sekvence byla vytvořena jako náhodná sekvence a upraveny pouze jejich prvních a posledních 20bp, aby představovaly inverzní repetice není možné, aby bylo detekováno tolik repetic označených modrými liniemi. Červená linie ukazuje ve všech grafech správné místo píku pro modelovou sekvenci. Nicméně další dva grafy ukazují píky v ostatních upravených periodogramech. Ačkoliv již je přítomen pík na frekvenci 1/20, ukázaly se ve výkonových spektrech i další píky, které jsou nežadoucí a zavádějící. Opět by tyto píky, vyznačené modrou čárou, neměly

být přítomny. Jediný pík, který by mohl být objasnitelný, souvisí s frekvencí 1/3 spojený s tří bázovou periodicitou v kódujících sekvencích. Ten se však vyskytuje pouze v prvním grafu.



Obr. 17: Detekce repetic ve výkonovém spektru - test
Červená vyznačuje místo, kde musí být pík vyšší než hraniční hodnota 4 rad^2 kvůli repetici, modrá vyznačuje další místa vyšší než hraniční hodnota

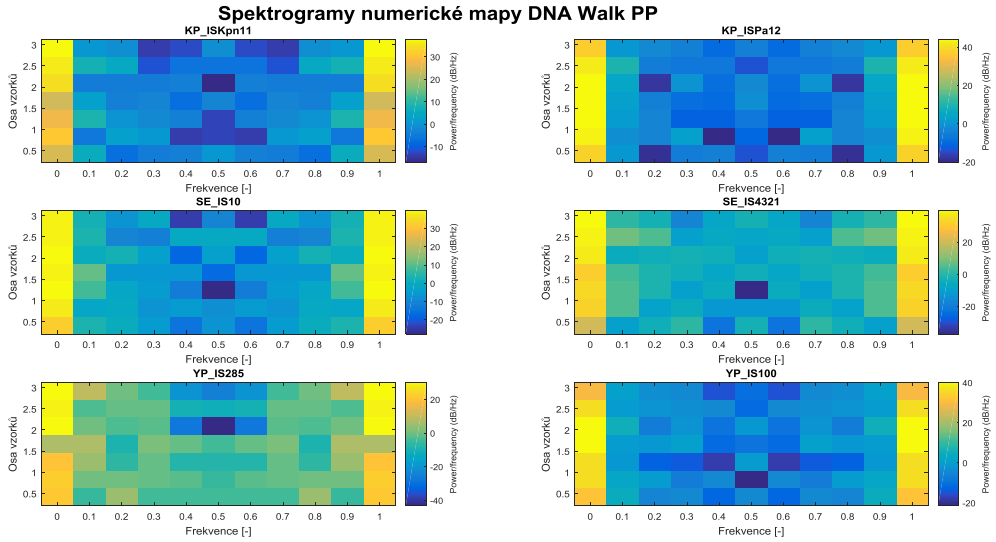
5.5.3. Spektrogram

Další časově frekvenční analýzou je vývoj spekter v čase, spektrogram. Jejím účelem je zachytit vývoj krátkodobých spekter, jejichž charakter se v čase mění. Každé dílčí spektrum je odvozené z krátkého úseku signálu, vymezeného posuvným oknem zvolené délky (počet vzorků *N*). Dostáváme tak spektra s úměrně malým frekvenčním rozlišením, ale odpovídající jednoznačně definovanému úseku časové osy. Kvůli snazšímu sledování vývoje ve spektrech se dílčí krátkodobá spektra obvykle sestavují do souborů.

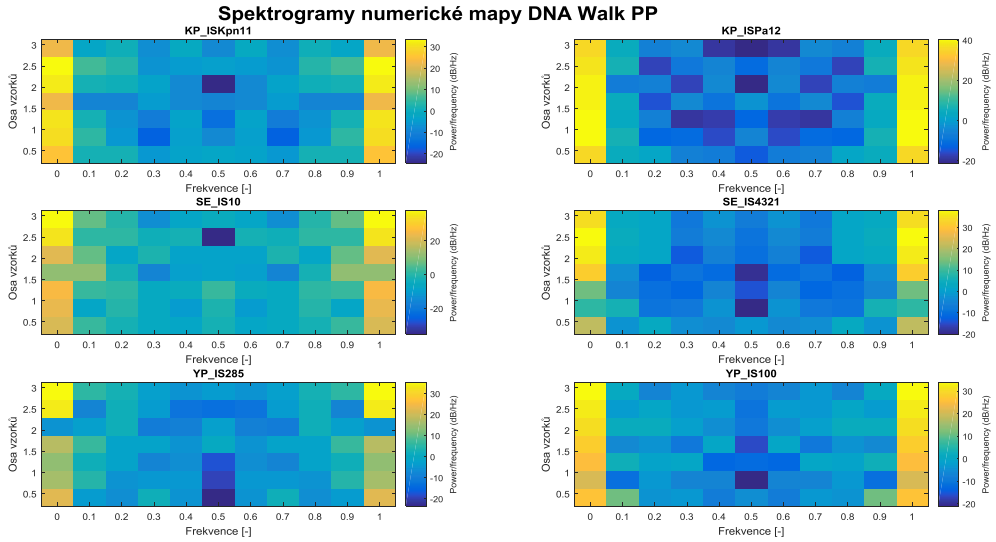
Vývoj výkonových spekter v čase se využívá pro analýzu různých signálů. V analýze variability srdečního rytmu s ní lze odhalit ischemickou epizodu na základě chybějícího výkonu na určité frekvenci přes několik dílčích krátkodobých spekter, spektrogram

fonokardiogramu může odkrýt aortové šelesty a v elektroenecefalogramu s jeho pomocí lze zjistit přesný okamžik nástupu spánku.

Jelikož spektrogram dokáže odhalit změnu vzniklou ve spektru v určité části signálu, byla provedena tato analýza i u genomického signálu. Výpočet dílčích spekter bude v posuvném okně po vzorcích, ne po časové ose.



Obr. 18: Spektrogramy IRL (numerické mapa DNA Walk PP)

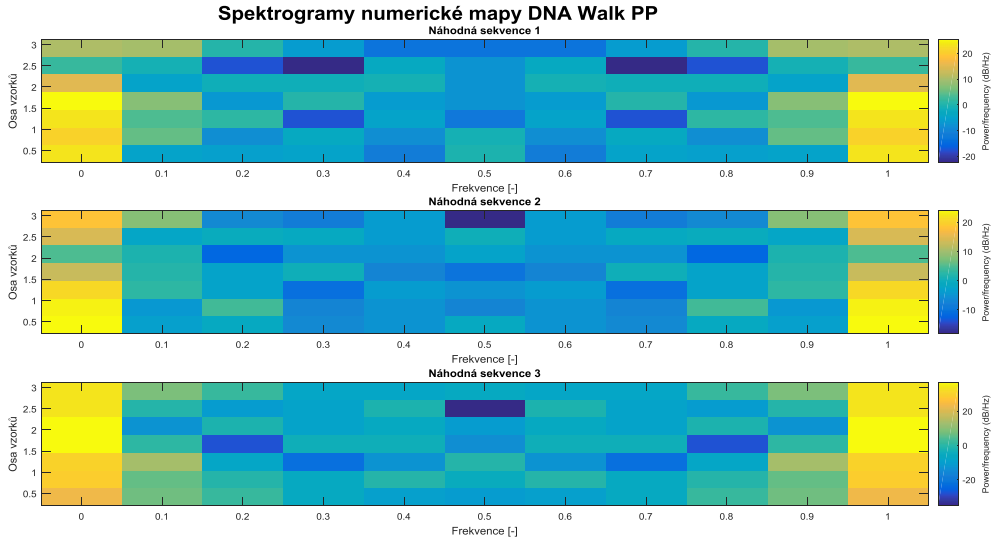


Obr. 19: Spektrogramy IRR (numerické mapa DNA Walk PP)

Spektrogram celého genomického signálu by byl příliš dlouhý a bylo by obtížné v něm najít úsek odpovídající MGE. Vhodnější se jeví analýza kratších úseků, v níž se zaměříme na začátky (IRL) a konce (IRR) MGE. K tomu lze opětovně využít data použitá v rámci analýzy zmíněné v kapitole 5.4. Ve spektrogramech zobrazíme i náhodné sekvence a kódující sekvence bez repetic pro porovnání umožňující identifikaci rozdílů charakterizujících přítomnost MGE v signálu. Pro výpočet byla použita funkce Matlabu `spectrogram`

za použití jejich dodatečných vstupních parametrů `window` (délka okna), `noverlap` (počet vzorků překryvu mezi okny), `f` (vektor frekvencí, pro něž se spočítá dílčí spektrum), `fs` (vzorkovací frekvence). Tato funkce používá k výpočtu výkonového spektra Short Time Fourier Transform (STFT), založenou na rychlé Fourierově transformaci (7).

Výběr ze spektrogramů různých numerických map začátků a konců MGE s přesahem do úseků před začátkem a za koncem o celkové délce 200 bp navíc je možné vidět na Obr. 18, Obr. 20 a pak v přílohách v Obr. 30, Obr. 31, Obr. 33, Obr. 34. Opět byly použity MGE z Tab. 5 nastavením výpočtu spektrogramu na okno o velikosti 50 vzorků a překryvem 25 vzorků. Přechod do nebo ze signálu MGE se nachází přesně uprostřed osy vzorků. Na první pohled lze vidět sníženou hodnotu spektra v okolí poloviny frekvenční osy (frekvence 0.5), avšak při porovnání se spektrogramy náhodných sekvencí (Obr. 20) je tato úvaha vyloučena jelikož i v těchto spektrogramech se nachází snížená hodnota spektra. Jiný společný jev nebyl viděn v žádném z ostatních spektrogramů ostatních numerických konverzí.

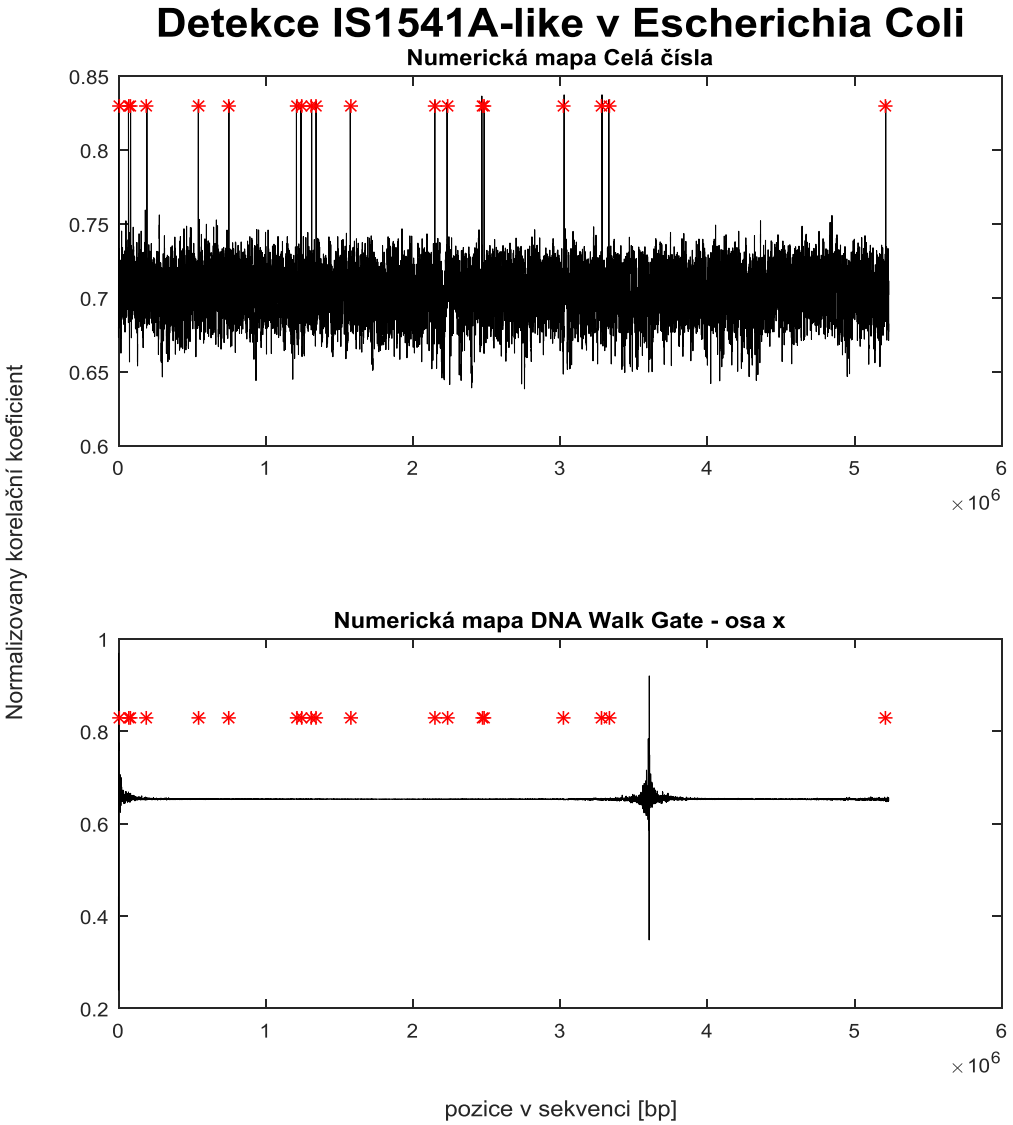


Obr. 20: Spektrogramy náhodných sekvencí (numerická map DNA Walk PP)

5.6 Korelace elementu vůči signálu

Na základě principu detekce dílčích čtecích sekvencí v ISMapperu [37] a detekce konsensuální MGE [28] byla postavena další analýza. Jak je zmíněno dříve v kapitole 5.4, numerická reprezentace konkrétního MGE má naprosto stejný tvar kdekoliv v genomu. Tudíž převedeme pouze sekvenci MGE do numerické reprezentace a tu budeme na základě korelace hledat v plovoucím okně po genomickém signálu. Jako zkušební MGE byl zvolen IS1415A-like, který se nachází i na komplementárním vláknu. Doposud se problematice detekce MGE na komplementárním vlákně nevěnovala pozornost. Místa výskytu IS1541A-like v genomu Escherichie Coli dle NCBI jsou shrnuté v Tab. 16 a jeho stejný průběh signálu kdekoliv v genomu lze vidět na Obr. 13. Tento MGE je v genomu Escherichie Coli přítomen devatenáctkrát, a sedm z nich je na komplementárním vlákně. Korelace v plovoucím okně

po genomickém signálu jasně vyznačuje váhovanými korelačními hodnotami výskyt MGE viz Obr. 21. Pokud funkci pro korelaci v plovoucím okně upravíme tak, aby výstupem byly i pozice maximálních korelací, získáme přehled o správnosti detekce viditelné v Tab. 6. Použitá numerická mapa ovlivňuje schopnost přesnosti detekce pozice začátku MGE.



Obr. 21: Detekce IS1541A-like v genomu *Escherichie Coli*
Nahoře: Detekce v signále po konverzi dle numerické mapy Celá čísla
Dole: Detekce v signále po konverzi dle numerické mapy DNA Walk Gate (osa x)

Jediná vhodná numerická mapa pro detekci elementů korelací v genomu nahrazuje bázi číslem, tedy mapa Celá čísla. Pro zisk lepších výsledků detekce bylo třeba tuto mapu upravit a místo T = 0, C = 1, A = 2, G = 3 byly hodnoty určeny jako A = 1, C = 1, G = 1, T = 1. A použitá numerická mapa ovlivňuje schopnost přesnosti detekce pozice začátku MGE také. Ostatní numerické mapy se projevily jako naprosto nevhodné identifikovat pozici jak lze vidět na Obr. 21 dole. Ačkoliv pro sledování vlastností genetické sekvence je numerická mapa Celá čísla naprosto nepřijatelná, jelikož zavádí parazitní vlastnost, pro detekci MGE se zde

projevila jako nejvhodnější. Výsledky detekce jsou pro porovnání shrnuty ve zmíněné Tab. 6 a lze vidět že v případě detekce IS1541A-like mobilního genetického elementu je maximální odchylka od správné pozice 12bp. Maximální rozdíl v detekci o 12bp je s ohledem na délku IS1541A-like 711bp více než přijatelný. Průměrně se detekce odchýlila o 3,1 bp.

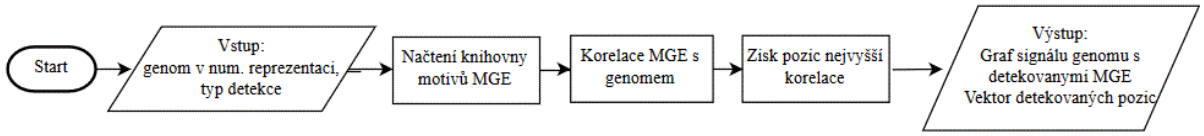
Tab. 6: Porovnání výsledků detekce se skutečnými začátky IS1541A-like

Pozice začátku dle NCBI [bp]	**Nadetekovaná pozice začátku [bp]**	**Rozdíl detekce oproti správné pozici [bp]**
308	309	1
67629	67627	-2
81932	81933	1
192688	192685	-3
542208	542209	1
749662	749663	1
1208542	1208543	1
1239332	1239330	-2
1311709	1311710	1
1341555	1341555	0
1574016	1574014	-2
2148219	2148229	12
2231360	2231361	1
2467364	2467376	12
2483124	2483125	1
3026281	3026293	12
3283628	3283640	12
3329874	3329886	12
5210335	5210335	0

Detekce MGE musela proběhnout ve dvou fázích. Výskyt MGE i na komplementárním vlákně vyžaduje druhou fázi detekce. Na komplementárním vlákně se MGE objevuje navíc v reverzní podobě. Důsledek v otočení průběhu numerické reprezentace vertikálně lze zaznamenat již na Obr. 13. Jelikož korelace funguje na principu hledání podobností mezi časově posunutými signály zvládne nanejvýš odhalit similaritu v průbězích otočených horizontálně proti sobě. Díky tomu je třeba mít každou sekvenci MGE v obou svých podobách a korelovat obě s genomickým signálem.

6 Detektor mobilních genetických elementů

Na základě informací získaných z výsledků předchozích analýz možností identifikace přítomnosti MGE v genomickém signálu lze přistoupit k návrhu detektoru MGE. Korelace numerické reprezentace MGE v genomickém signálu ukázala nejlepší určení místa výskytu MGE. Jedná se pouze o konkrétní MGE, tudíž pro detekci dalšího MGE je třeba mít jeho numerickou reprezentaci a tu vyhledávat za použití korelace v genomickém signálu. Bohužel není možné vytvořit společnou vzorovou sekvenci pro všechny MGE a tu vyhledávat v celých genomech, protože výsledky analýzy v kapitole 5.3 neukázaly žádné odlišnosti signálu kódující sekvence nebo náhodného signálu od signálů různých numerických reprezentací různých MGE. Taktéž neodhalily žádný společný rys nebo znak v reprezentacích MGE použitelný pro generalizaci na všechny typy MGE a následnou detekci.



Obr. 22: Diagram detektoru MGE v num. reprezentaci genomu

Pro sestavení detektoru MGE je nejprve potřeba vytvořit knihovnu numerických reprezentací jednotlivých mobilních genetických elementů. Dále je třeba upravit funkci pro počítání korelace v plovoucím okně po delší sekvenci, aby načítala tuto knihovnu a jejím vstupem byla numerická reprezentace genomu organismu nebo sekvence, v níž bude hledán MGE. Jako výstup se určí okno s grafem sekvence s vyznačenými detekovanými elementy a buňkové pole obsahující názvy a pozice detekovaných MGE.

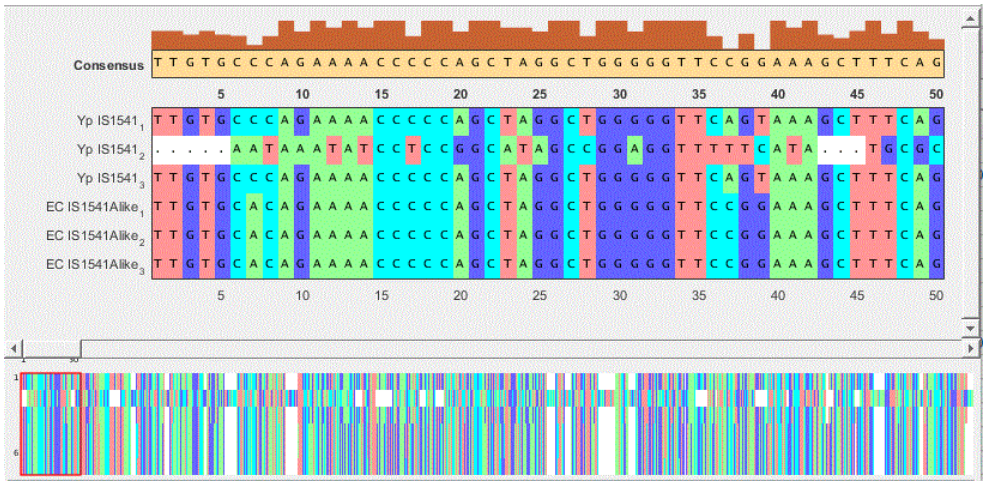
V genomech organismů využívaných v rámci této práce lze najít MGE uvedené v příloze (Tab. 12, Tab. 13, Tab. 14, Tab. 15, Tab. 16) . Ze symbolické reprezentace genomu byly již dříve vyňaty všechny úseky obsahující jakýkoliv MGE. Následně bylo potřeba zkontrolovat míru degenerace mezi jednotlivými sekvencemi konkrétního MGE. K tomu posloužilo spočítání vzdálenostní (distanční) matice mezi realizacemi daného MGE v symbolické reprezentaci (Tab. 7). Distanční matice byla získána funkcí seqpdist s defaultními parametry, což znamená že jako metoda výpočtu proporconální vzdálenosti byla nastavena na Jukes-Cantor. Pro kontrolu lze pohlédnout na zarovnání za použití příkazu `multialign` a `seqalignviewer` (Obr. 23). V distanční matici bylo ověřeno, že se sekvence daného MGE z různých míst v genomu nebo napříč genomy podobají a byla z nich vytvořena konsensuální sekvence použitím funkce `seqconsensus`. Konsensuální sekvence, představující aproximaci ze souhrnu shod mezi sadou zarovnaných sekvencí, pak byla zařazena do knihovny. [50] Pokud byla k dispozici pouze jedna sekvence konkrétního MGE

nebyla distanční matice ani výpočet konsensuální sekvence zapotřebí a rovnou se tento MGE zařadil do knihovny. Celá knihovna numerických reprezentací MGE konvertovaných dle mapy Celá čísla je uložena pod názvem `knihovna.mat`. Proměnnou typu cell `knihovna` lze kdykoliv rozšířit o další řadek, kdy se do jeho první buňky umístí název mobilního genetického elementu, druhé buňky numerická reprezentace jemu odpovídající nekomplementární sekvence a jeho komplementární podoba se umísti do třetí buňky. Knihovna v této chvíli čítá 32 mobilních genetických elementů, přičemž čtyři ukázky lze vidět na Obr. 25. Pokud byla k dispozici pouze jedna sekvence konkrétního MGE nebyla distanční matice zapotřebí a rovnou se tento MGE zařadil do knihovny.

Tab. 7: Ukázka distanční matice mezi MGE

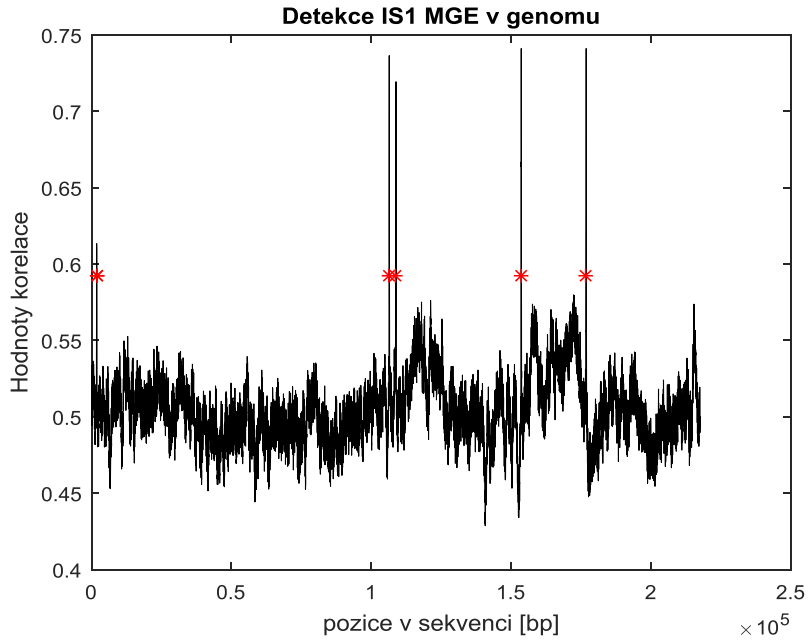
		Yersinia pestis			*Escherichia Coli*		
	MGE	IS1541_1	IS1541_2	IS1541_3	IS1541A-like_1	IS1541A-like_1	IS1541A-like_1
Yersinia pestis	IS1541_1	0					
	IS1541_2	0.0014	0				
	IS1541_3	0	0.0014	0			
Escherichia Coli	IS1541A-like_1	0.1381	0.1397	0.1381	0		
	IS1541A-like_2	0.1381	0.1397	0.1381	0.1381	0	
	IS1541A-like_3	0.1332	0.1348	0.1332	0.1332	0.1338	0

Vzhledem k velkému počtu různých MGE, jež se v organismech *Enterobacteriae* vyskytují, byl do funkce hledání MGE v genomu zařazen ještě druhý vstupní parametr, `TypDetekce`. Druhý parametr umožní definovat jeden konkrétní motiv MGE z knihovny jenže chceme detekovat (hodnota parametru je konkrétní název). Pokud bude zadán název mobilního genetického elementu, jenž není obsažen v knihovně, Matlab vypíše chybovou hlášku.

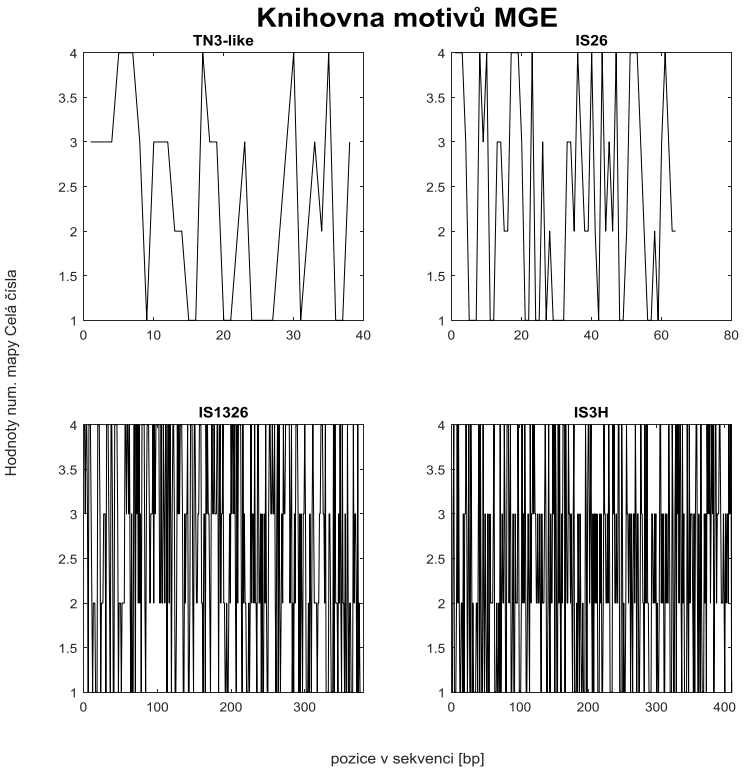


Obr. 23: Ukázka zarovnání motivů MGE

Po průchodu algoritmem je možné pozorovat v grafu nového okna numerickou reprezentaci genomu a v něm vyznačené části s mobilními genetickými elementy stejnou jako v horním grafu v Obr. 21 anebo ve výstupu detektoru Obr. 24. V detektoru je signál umocněn na druhou pro zvýraznění rozdílu mezi nízkými a vysokými hodnotami korelace. Pokud detekce proběhla pouze pro jeden typ MGE, bude jeho název obsažen i v titulku grafu. Protože se z grafu genomického signálu s detekovanými nesnadno získávají pozice, lze je dohledat ve výstupní proměnné `detekce`. Tento detektor umožňuje detekovat mobilní genetické elementy i na komplementárním vlákně v reverzní podobě, protože obsahuje i podobu tohoto motivu v knihovně.



Obr. 24: Detekovaný IS1 v genomickém signálu *Salmonelly entericy*



Obr. 25: Ukázka z knihovny motivů MGE

Takto postavený detektor je ovšem potřeba i zevaluovat v rámci jeho schopnosti správné detekce. Základními hodnotícími parametry kvality detektoru jsou statistické diagnostické testy sensitivity a specificity.

Sensitivita je množství pravdivých pozitiv, které jsou korektně identifikovány. [48]

$$Sensitivita = \frac{pravdivě\ pozitivní\ (\mathrm{PP})}{(pravdivě\ pozitivní\ (\mathrm{PP}) + falešně\ negativní\ (\mathrm{FN}))} \quad \textbf{(8)}$$

Specificita je množství pravdivých negativ, které jsou korektně identifikovány. [48]

$$Specificita = \frac{pravdivě\ negativní\ (\mathrm{PN})}{(pravdivě\ negativní\ (\mathrm{PN}) + falešně\ pozitvní\ (\mathrm{FP}))} \quad \textbf{(9)}$$

Převedení prvků sensitivity a specificity na detekci mobilních genetických elementů lze vysvětlit takto:

- pravdivě pozitivní jsou správně detekované MGE
- falešně negativní jsou ty MGE, které v sekvenci jsou, ale detektor je nenašel

- pravdivě negativní jsou místa signálu, kde nedošlo k detekci MGE a zároveň tam opravdu není MGE
- falešně pozitivní jsou detekované MGE tam, kde nejsou.

Jelikož máme k dispozici databázi NCBI s informacemi o pozicích a typu MGE a byly zvoleny pro tuto práci takové organismy, jejichž genom byl v tomto ohledu již pečlivě prozkoumán, můžeme spočítat specificitu a sensitivitu vzhledem k těmto informacím. Výsledky testování vzhledem k těmto referenčním informacím lze nalézt v Tab. 8 a Tab. 9. Všechny detektory byly spuštěny ve výchozím nastavení. Pokud byla detekce začátku nebo konce MGE mírně posunuta oproti referenčním hodnotám, byla počíítána tolerance ±20 bp. Lze tak uvést specificitu a sensitivitu jako vlastnosti detektoru. V tabulce Tab. 8 jsou zahrnuty i data z testování dalších přístupných detektorů.

Vzhledem k velikostem genomických sekvencí je dalším důležitým parametrem pro ohodnocení detektoru i rychlost detekce. Pro hodnocení je třeba znát rychlosti detekce dalších volně přístupných detektorů. Pro porovnání se zde navrženým detektorem byly vybrány detektory zmíněné v Tab. 1. Principy těchto detektorů lze nalézt v kapitole 3.1 a všechny jsou založeny na detekci v symbolické reprezentaci genomu. Každý z nich pracuje v jiném programovacím jazyce a jejich funkce nebo sety funkcí byly staženy a testovány na počítači s procesorem Intel® Core™ i5-4460 3.20 GHz o 4 jádrech a 4 vláknech a s operační pamětí dosahující 15.9 GB. Všem byla pro detekci poskytnutá stejná data - stejný genom. Byl měřen čas, který detektor strávil detekcí od rozběhnutí programu až po jeho ukončení s vydáním výsledků. Tento čas naleznete také v Tab. 8. Výsledky byly porovnány se záznamy získanými z NCBI a s těmito výstupy lze spočítat specifictu a senzitivitu i u těchto detektorů a porovnat ji se zjištěnými hodnotami pro detektor MGE v numerické reprezentaci.

Tab. 8: Senzitivita detektorů – pravdivě pozitivní (PP) a falešně negativní (FN)

MGE	**IS1**	**IS10**	**IS15/26**	**IS4321**	**IS6100**	**Jakýkoliv MGE**	**Čas detekce [s]**
	PP	PP	PP	PP	PP	FN	
NCBI	5	2	4	2	1	-	-
iMEx [51]	0	0	0	0	0	14	19
RepeatMasker	1	0	0	0	0	13	150
Navrhnutý detektor	5	2	4	2	1	0	170

Tab. 9: Specificita detektorů – falešně pozitivní detekce (FP)

MGE	IS1FP	IS10FP	IS15/26FP	IS4321FP	IS6100FP
iMEx	4321				
RepeatMasker	10				
Navrhnutý detektor	2	0	0	1	0

7 Závěr

Mnoho numerických reprezentací se ukázalo nevhodných pro hledání společného rysu mezi mobilními genetickými sekvencemi. Byly provedeny analýzy jak v časové, tak ve frekvenční oblasti, dokonce i v časově-frekvenční, na vhodných numerických reprezentacích genomů a genetických sekvencích. Nebyl nalezen žádný význačný znak společný všem typům mobilních genetických elemetů. Pro potrvzení byly prohlédnuty i samotné symbolické sekvence mobilních genetických elemetů, nebo jen jejich začátků či konců. Ty by měly být navzájem pro jeden mobilní genetický element inverzními repeticemi. Avšak ani dodatečným zkoumáním symbolické reprezentace nebyly nalezeny společné znaky. Na základě těchto analýz bylo přistoupeno k jedinému možnému způsobu navrhnutí detektoru, který se prokázal v analýze úspěšný.

Návrh detektoru je postaven na detekci numerické reprezentace konkrétního mobilního genetického elementu v numerické reprezentaci genomu. Tento přístup vedl k vytvoření knihovny motivů, neboli knihovny numerických reprezentací mobilních genetických elementů. Podobné knihovny existují pro symbolické reprezentace [28], avšak nebyla žádná taková vytvořena v numerické reprezentaci. Zde navržený detektor pracuje pouze s numerickými reprezentacemi genomu na vstupu a díky tomu pracuje s čísly a ne se symboly C, G, A, T. Práce s čísly umožňuje použití matematických postupů a nástrojů, jež není možné použít v symbolické reprezentaci. Navíc práce se symbolickou reprezentací vyžaduje naprosto jiný přístup a metody, obvykle pomalejší než práce s čísly a vyžadující více kroků. Díky využití číselného signálu reprezentujícícho genetickou sekvenci bylo možné v detektoru použít matematického vztahu pro výpočet korelace jako nástroje pro nalezení motivu mobilního genetického elementu v genomu. V symbolické reprezentaci se k tomu využívá náročnějšího lokálního zarovnání.

Detekce mobilních genetických elementů je v dnešní době jeden z důležitějších zájmů bioinformatiky. Zde představený přístup, detekce v numerické reprezentaci, je jedním z novátorských přístupů. Jedním z nich je navrhnutý detektor repetic ve spektru. [35] Ten pracuje pouze s krátkými sekvencemi okolo 100kb, oproti detektoru představeném v této práci, jež si poradí s celým genomem větším než 10Mb. [35] Detekce 5 určených MGE v genomu *Salmonella enterica* o velikosti 218160 bp trvá přibližně 170 sekund, v průměru 34 sekund na jeden motiv. Detekce všech motivů v knihovně je logicky časově velice náročná. Senzitivita a specificita je vysoko nad ostatním testovanými detektory (Tab. 8, Tab. 9). V neposlední řadě se nedetekují opakující se geny v genomu, jak je to u detektorů repetic v symbolickém zápise obvyklé, ani jiné náhodné repetice. [28] Ty tvořily víceméně veškerý výstup zkoušených volně přístupných detektorů. Testované webové detektory jsou velice citlivé na prvotní nastavení a algoritmy jsou postaveny pro detekci v eukaryotech. [34] Eukaryotní organismy mají úplně jiné typy a struktury mobilních genetických elementů. [5]

Tyto detektory nejsou tedy nejvhodnější pro detekci v prokaryotech, což je jejich další nedostatek. Nedostatkem navrženého detektoru je neschopnost detekovat nekompletní MGE s délkou nižší než 85% původní délky. Náhradním řešením je poskytnout detektoru navrženému v této práci zkrácenou sekvenci MGE, jež bude schopen detekovat.

Tato práce poskytuje nový přístup k detekci MGE v prokaryotických genomických signálech. Sestavení a využití knihovny motivů numerických reprezentací mobilních genetických elementů doposud nebylo k dispozici. Vším výše zmíněným představuje silný nástroj k vyhledání mobilních genetických elementů v genomu a pomoc při jejich výzkumu.

Reference

[1] HAWKEY, Jane, Mohammad HAMIDIAN, Ryan R. WICK, David J. EDWARDS, Helen BILLMAN-JACOBE, Ruth M. HALL a Kathryn E. HOLT. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genomics [online]. 2015, 16(1), - [cit. 2016-12-4]. DOI: 10.1186/s12864-015-1860-2. ISSN 14712164. Dostupné z: http://www.biomedcentral.com/1471-2164/16/667

[2] LINDSAY, Jodi A. Staphylococcus: molecular genetics. 2008. Norfolk, UK: Caister Academic Press, c2008. ISBN 9781904455295.

[3] ROCHA, Eduardo P.C. The Organization of the Bacterial Genome. Annual Review of Genetics [online]. 2008, 42(1), 211-233 [cit. 2016-12-4]. DOI: 10.1146/annurev.genet.42.110807.091653. ISSN 00664197. Dostupné z: http://www.annualreviews.org/doi/10.1146/annurev.genet.42.110807.091653

[4] MALACHOWA, Natalia a Frank R. DELEO. Mobile genetic elements of Staphylococcus aureus. Cellular and Molecular Life Sciences [online]. 2010, 67(18), 3057-3071 [cit. 2016-12-04]. DOI: 10.1007/s00018-010-0389-4. ISSN 1420682x. Dostupné z: http://link.springer.com/10.1007/s00018-010-0389-4

[5] Mobilní genetické elementy. Gate2Biotech [online]. České Budějovice: CREOS CZ, 2015 [cit. 2016-11-18]. Dostupné z: www.gate2biotech.cz/mobilni-geneticke-elementy

[6] ROBERTS, Adam P., Michael CHANDLER, Patrice COURVALIN, et al. Revised nomenclature for transposable genetic elements. Plasmid [online]. 2008, 60(3), 167-173 [cit. 2016-11-18]. DOI: 10.1016/j.plasmid.2008.08.001. ISSN 0147619x. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0147619X08000784

[7] RAVINDRAN, S. Barbara McClintock and the discovery of jumping genes. Proceedings of the National Academy of Sciences [online]. 2012, 109(50), 20198-20199 [cit. 2016-11-21]. DOI: 10.1073/pnas.1219372109. ISSN 00278424. Dostupné z: http://www.pnas.org/cgi/doi/10.1073/pnas.1219372109

[8] MILLER, Wolfgang J. a Pierre. CAPY. Mobile genetic elements: protocols and genomic applications. Totowa, N.J.: Humana Press, c2004. Methods in molecular biology (Clifton, N.J.), v. 260. ISBN 1588290077.

[9] LEPLAE, R. ACLAME: A CLAssification of Mobile genetic Elements. Nucleic Acids Research [online]. 2004, 32(90001), 45D-49 [cit. 2016-11-14]. DOI: 10.1093/nar/gkh084. ISSN 13624962. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh084

[10] COHEN, Stanley N. a James A. SHAPIRO. Transposable Genetic Elements. Scientific American [online]. 1980, 242(2), 40-49 [cit. 2016-12-21]. DOI: 10.1038/scientificamerican0280-40. ISSN 00368733. Dostupné z: http://www.nature.com/doifinder/10.1038/scientificamerican0280-40

[11] Bacterial Resistance to Antibiotics. TODAR, Kenneth. Online Textbook of Bacteriology [online]. Wisconsin: Todar, c2008-2011 [cit. 2016-12-21]. Dostupné z: http://textbookofbacteriology.net/resantimicrobial_3.html

[12] Aktuální genetika - Repetitivní DNA. Aktuální genetika [online]. Praha: Ústav biologie a lékařské genetiky 1.LF UK a VFN, 2006 [cit. 2016-11-12]. Dostupné z: biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm

[13] Mechanism of transposition in prokaryotes. GRIFFITHS, Anthony J. F. An introduction to genetic analysis. 7th ed. New York: W.H. Freeman, c2000. ISBN 0716735202.

[14] LEPLAE, Raphaël, Gipsi LIMA-MENDEZ a Ariane TOUSSAINT. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Research [online]. 2010, 38(suppl_1), D57-D61 [cit. 2016-11-24]. DOI: 10.1093/nar/gkp938. ISSN 03051048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp938

[15] DARMON, Elise a David R. F. LEACH. Bacterial Genome Instability. Microbiology and Molecular Biology Reviews [online]. 2014, 78(1), 1-39 [cit. 2016-11-24]. DOI: 10.1128/MMBR.00035-13. ISSN 1098-5557. Dostupné z: http://mmbr.asm.org/content/78/1/1.full.pdf

[16] SHAPIRO, James A. The Discovery and Significance of Mobile Genetic Elements. SHERRATT, David J., ed. Mobile genetic elements. 8. New York: Academic Press, 1983. ISBN 0-12-638680-3.

[17] SIGUIER, P. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Research [online]. 2006, 34(90001), D32-D36 [cit. 2016-12-04]. DOI: 10.1093/nar/gkj014. ISSN 03051048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkj014

[18] JACKSON, Robert W., Boris VINATZER, Dawn L. ARNOLD, Steve DORUS a Jesús MURILLO. The influence of the accessory genome on bacterial pathogen evolution. Mobile Genetic Elements [online]. 2011, 1(1), 55-65 [cit. 2016-12-04]. DOI: 10.4161/mge.1.1.16 432. ISSN 2159-256X. Dostupné z: http://www.tandfonline.com/doi/pdf/10.4161/mge.1.1.16432

[19] What are Transposable Elements ? | Genetics. Share Your Knowledge on Biology [online]. BiologyDiscussion.com, c2016 [cit. 2016-12-05]. Dostupné z: http://www.biologydiscussion.com/eukaryotic-cell/transposable-elements/what-are-transposable-elements-genetics/67879

[20] KWAN, Hon Keung a Swarna Bai ARNIKER. Numerical representation of DNA sequences. In: 2009 IEEE International Conference on Electro/Information Technology [online]. IEEE, 2009, s. 307-310 [cit. 2016-12-27]. DOI: 10.1109/EIT.2009.5189632. ISBN 9781424433544. Dostupné z: http://ieeexplore.ieee.org/document/5189632/

[21] YAU, S. S. -T. DNA sequence representation without degeneracy. Nucleic Acids Research [online]. 2003, 31(12), 3078-3080 [cit. 2017-01-02]. DOI: 10.1093/nar/gkg432. ISSN 13624962. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg432

[22] LIAO, Bo. A 2D graphical representation of DNA sequence. Chemical Physics Letters [online]. 2005, 401(1-3), 196-199 [cit. 2017-01-02]. DOI: 10.1016/j.cplett.2004.11.059. ISSN 00092614. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0009261404018378

[23] DEERGHA RAO, K. a M.N.S. SWAMY. Analysis of Genomics and Proteomics Using DSP Techniques. IEEE Transactions on Circuits and Systems I: Regular Papers [online]. 2008, 55(1), 370-378 [cit. 2016-12-27]. DOI: 10.1109/TCSI.2007.910541. ISSN 15498328. Dostupné z: http://ieeexplore.ieee.org/document/4378220/

[24] JAN, Jiří. Číslicové zpracování a analýza signálů: stručné skriptum. Brno: Vysoké učení technické v Brně, 2010. ISBN 9788021440180.

[25] Cross-correlation - MATLAB xcorr. MathWorks - Makers of MATLAB and Simulink [online]. Natick (Massachusetts, USA): The MathWorks, c1994-2017 [cit. 2016-12-29]. Dostupné z: www.mathworks.com/help/signal/ref/xcorr.html

[26] SHAPIRO, James Alan. Mobile genetic elements. New York: Academic Press, 1983. ISBN 0126386803.

[27] SHAPIRO, James A. Genetica [online]. 107(1/3), 171-179 [cit. 2016-11-10]. DOI: 10.1023/A:1003977827511. ISSN 00166707. Dostupné z: http://link.springer.com/10.1023/A:1003977827511

[28] BERGMAN, C. M. a H. QUESNEVILLE. Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics [online]. 2007, 8(6), 382-392 [cit. 2016-12-10]. DOI: 10.1093/bib/bbm048. ISSN 14675463. Dostupné z: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbm048

[29] MARK OSBORN, A a Dietmar BÖLTNER. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid [online]. 2002, 48(3), 202-212 [cit. 2016-11-24]. DOI: 10.1016/S0147-619X(02)00117-8. ISSN 0147619x. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0147619X02001178

[30] The Largest Prokaryotic Genomes. Sandwalk [online]. Toronto: Moran, 2013 [cit. 2016-12-17]. Dostupné z: sandwalk.blogspot.cz/2013/07/the-largest-prokaryotic-genomes.html

[31] CRISTEA, P. D. Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine [online]. 2002, 6(2), 279-303 [cit. 2016-12-27]. DOI: 10.1111/j.1582-4934.2002.tb00196.x. ISSN 15821838. Dostupné z: http://doi.wiley.com/10.1111/j.1582-4934.2002.tb00196.x

[32] CRISTEA, P. Genetic signal analysis. In: Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat.No.01EX467) [online]. IEEE, 2001, s. 703-706 [cit. 2016-12-27]. DOI: 10.1109/ISSPA.2001.950245. ISBN 0780367030. Dostupné z: http://ieeexplore.ieee.org/document/950245/

[33] Principles and Methods of Sequence Analysis. KOONIN, Eugene V a Michael Y GALPERIN. Sequence - evolution - function: computational approaches in comparative genomics. 1. Boston: Kluwer Academic, c2003, s. 1091-1102. ISBN 1-40207-274-0

[34] LIM, K. G., C. K. KWOH, L. Y. HSU a A. WIRAWAN. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. Briefings in Bioinformatics [online]. 2013, 14(1), 67-81 [cit. 2016-12-27]. DOI: 10.1093/bib/bbs023. ISSN 14675463. Dostupné z: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs023

[35] SHARMA, D., B. ISSAC, G. P. S. RAGHAVA a R. RAMASWAMY. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics [online]. 2004, 20(9), 1405-1412 [cit. 2017-05-01]. DOI: 10.1093/bioinformatics/bth103. ISSN 13674803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth103

[36] ZHANG, Ren a Chun-Ting ZHANG. Z Curves, An Intutive Tool for Visualizing and Analyzing the DNA Sequences. *Journal of Biomolecular Structure and Dynamics* [online]. 1994, **11**(4), 767-782 [cit. 2017-05-12]. DOI: 10.1080/07391102.1994.10508031. ISSN 07391102. Dostupné z: http://www.tandfonline.com/doi/abs/10.1080/07391102.1994.10508031

[37] HAWKEY, Jane, Mohammad HAMIDIAN, Ryan R. WICK, David J. EDWARDS, Helen BILLMAN-JACOBE, Ruth M. HALL a Kathryn E. HOLT. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* [online]. 2015, **16**(1), - [cit. 2017-05-13]. DOI: 10.1186/s12864-015-1860-2. ISSN 14712164. Dostupné z: http://www.biomedcentral.com/1471-2164/16/667

[38] SMITH, T.F. a M.S. WATERMAN. Identification of common molecular subsequences. *Journal of Molecular Biology* [online]. 1981, **147**(1), 195-197 [cit. 2017-05-13]. DOI: 10.1016/0022-2836(81)90087-5. ISSN 00222836. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/0022283681900875

[39] Enterobacteriaceae | Velký lékařský slovník On-Line. *Výrazy od a | Velký lékařský slovník On-Line* [online]. Praha: Maxdorf, c1998-2017 [cit. 2017-05-1]. Dostupné z: http://lekarske.slovniky.cz/lexikon-pojem/enterobacteriaceae-3

[40] Enterobacteriaceae - enterobakterie - Bakterie. *Bakterie* [online]. bakterie.eu, c2010-2013 [cit. 2017-05-13]. Dostupné z: http://www.bakterie.eu/druhy-bakterii/enterobacteriaceae

[41] AustraliaShadePlotForEmphasis - File Exchange - MATLAB Central. *MathWorks - Makers of MATLAB and Simulink* [online]. Natick (Massachusetts, USA): The MathWorks, c1994-2017 [cit. 2017-05-2]. Dostupné z: http://www.mathworks.com/matlabcentral/fileexchange/3550-shadeplotforemphasis

[42] THE ELECTROGASTROGRAM AND WHAT IT SHOWS. *JAMA: The Journal of the American Medical Association* [online]. 1922, **78**(15), 1116- [cit. 2017-05-14]. DOI: 10.1001/jama.1922.02640680020008. ISSN 00987484. Dostupné z: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.1922.02640680020008

[43] MORÁŇ, Miroslav. *Praktická elektroencefalografie*. Brno: Institut pro další vzdělávání pracovníků ve zdravotnictví, 1995. ISBN 8070132035.

[44] HOVORKA, Jiří. Klinická elektroencefalografie: základy klasifikace a interpretace. Praha: Maxdorf, 2003. ISBN 8073450011.

[45] ABBATE, Agostino, Casimer M. DECUSATIS a Pankaj K. DAS. Time-Frequency Analysis of Signals. *Wavelets and Subbands* [online]. Boston, MA: Birkhäuser Boston, 2002, s. 103 [cit. 2017-05-14]. DOI: 10.1007/978-1-4612-0113-7_3. ISBN 9781461266181. Dostupné z: http://link.springer.com/10.1007/978-1-4612-0113-7_3

[46] JAN, Jiří. *Číslicová filtrace, analýza a restaurace signálů*. 2. upr. a rozš. vyd. Brno: VUTIUM, 2002. ISBN 8021429119.

[47] BERGER, J.A., S.K. MITRA a J. ASTOLA. Power spectrum analysis for DNA sequences. In: *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings* [online]. IEEE, 2003, 29-32 vol.2 [cit. 2017-05-15]. DOI: 10.1109/ISSPA.2003.1224807. ISBN 0780379462. Dostupné z: http://ieeexplore.ieee.org/document/1224807/

[48] ALTMAN, D G a J M BLAND. Statistics Notes: Diagnostic tests 1. *BMJ* [online]. 1994, **308**(6943), 1552-1552 [cit. 2017-05-16]. DOI: 10.1136/bmj.308.6943.1552. ISSN 09598138. Dostupné z: http://www.bmj.com/cgi/doi/10.1136/bmj.308.6943.1552

[49] *National Center for Biotechnology Information* [online]. Bethesda MD, USA: U.S. National Library of Medicine, 2017 [cit. 2017-05-16]. Dostupné z: www.ncbi.nlm.nih.gov

[50] SPITZER, M., G. FUELLEN, P. CULLEN a S. LORKOWSKI. VisCoSe: visualization and comparison of consensus sequences. *Bioinformatics* [online]. 2004, **20**(3), 433-435 [cit. 2017-05-16]. DOI: 10.1093/bioinformatics/btg444. ISSN 13674803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btg444

[51] *IMEx: Imperfect Microsatellite Extractor HomePage* [online]. India: Mudunuri, c2006 [cit. 2017-05-16]. Dostupné z: http://www.mcr.org.in/IMEX/index.html

[52] RepeatMasker. *Institue for Systems Biology* [online]. [cit. 2017-05-16]. Dostupné z: http://repeatmasker.org/

Seznam zkratek a symbolů

MGE	mobilní genetické elementy
kb	kilobáze, tisíc bází DNA
Mb	megabáze, milion bází DNA
IS	Iinzeční sekvence
ORF	otevřený čtecí rámec (open reading frame)
DNA	deoxyribonukleová kyselina (deoxyribonucleic acid), nositelka genetické informace
BLAST	Basic Local Alignment Search Tool, metoda lokálního zarovnání symbolických sekvencí
IRL	levé invertované opakování (inverted repeat left)
IRR	pravé invertované opakování (inverted repeat right)

Seznam obrázků

Přílohy

Tabulky

Tab. 10: Základní metody numerické reprezentace DNA [20]

Metoda konverze	**Numerická mapa pro konverzi**	**Příklad konverze pro S(n) = [CGAT]**
Voss	$X_n = 1\ pro\ S(n) = X$ $X_n = 0\ pro\ S(n) \neq X$ **(10)**	C_n = [1, 0, 0, 0] G_n = [0, 1, 0, 0] A_n = [0, 1, 0, 0] Tn = [0, 1, 0, 0]
Tetrahedron	$x_r(n) = \frac{\sqrt{2}}{3}[2T_n - C_n - G_n]$ $x_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$ **(11)** $x_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$	$x_r(n) = \frac{\sqrt{2}}{3}[-1,-1,0,2]$ $x_g(n) = \frac{\sqrt{6}}{3}[1,-1,0,0]$ $x_b(n) = \frac{1}{3}[-1,-1,3,-1]$
Celá čísla	$T = 0, C = 1, A = 2, G = 3$ **(12)**	[1,3,2,0]
Reálná čísla	$A = -1.5, C = 0.5,$ $G = -0.5, T = 1.5$ **(13)**	$[0.5, -0.5, -1.5, 1.5]$
Komplexní čísla	$A = 1 + j, C = -1 + j$ $G = -1 - j, T = 1 - j$ **(14)**	$[-1 + j, -1 - j, 1 + j, 1 - j]$
Čtveřice (Quaternion)	$A = i + j + k, C = i - j - k$ $G = -i - j + k, T = -i + j - k$ **(15)**	$[i - j - k, -i - j + k,$ $i + j + k, -i + j - k]$
EIIP	$A = 0.1260, C = 0.1340$ $G = 0.0806, T = 0.1335$ **(16)**	$[0.1340, 0.0806, 0.1260, 0.1335]$
Atomové číslo	$A = 70, C = 58$ $G = 78, T = 66$ **(17)**	$[58, 78, 70, 66]$
Párová čísla (Paired numeric)	A nebo T $= 1$ $C\ nebo\ G = -1$ **(18)**	$P_{1n} = [-1, -1, 1, 1]$ $P_{2n} = [-1, -1, 0, 0]$& $[0, 0, 1, 1]$
DNA Walk (Krok)	$C\ nebo\ T = 1$ $A\ nebo\ G = -1$ **(19)**	$[1, 0, -1, 0]$
DNA Walk	$A = (0, -1), C = (-1,0)$ **(20)**	$x = [0, -1, 0,0,0]$

Metoda konverze	**Numerická mapa pro konverzi**	**Příklad konverze pro S(n) = [CGAT]**
(Gate)	$G = (1,0), T = (0,1)$	$y = [0,0,0,-1,0]$
DNA Walk (Yau)	$\mathrm{A} = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$ $\mathrm{C} = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$ $\mathrm{G} = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$ $\mathrm{T} = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$ **(21)**	$x = [0, 0.86, 1.73, 2.23, 2.73]$ $y = [0, 0.5, 0, -0.86, 0]$
Z křivka	$x_n = (A_n + G_n) - (C_n + \mathrm{T_n}) \equiv \mathrm{R_n} - \mathrm{Y_n}$ $y_n = (A_n + \mathrm{C}_n) - (G_n + \mathrm{T_n}) \equiv \mathrm{M_n} - \mathrm{K_n}$ $z_n = (A_n + T_n) - (C_n + \mathrm{G_n}) \equiv \mathrm{W_n} - \mathrm{S_n}$ **(22)**	$x_n = [-1, 0, 1, 0]$ $y_n = [1, 0, 1, 0]$ $z_n = [-1, -2, -1, 0]$

Tab. 11: Výhody a nevýhody reprezentací [20],[21],[22],[23]

Metoda konverze	**Výhody**	**Nevýhody**
Voss	Efektivní spektrální detektor distribuce bází a znaků periodicity, nabízí číselnou i grafickou vizualizaci.	Přidává nadbytečnou vlastnost, lineárně závislý množina reprezentace.
Tetrahedron	Detekce periodicity.	Snížená nadbytečná informace.
Celá čísla	Jednoduchá celočíselná reprezentace.	(A, G) > (C, T); nadbytečná matematická vlastnost nepřítomná v symbolické DNA sekvenci.
Reálná čísla	A-T a C-G jsou komplementární.	Nadbytečná matematická vlastnost nepřítomná v symbolické DNA sekvenci.
Komplexní čísla	A-T a C-G jsou komplexně sdružené, reflektuje komplementární charakter nukleotidů.	Zkreslení základny při analýze v časové doméně.
Čtveřice (Quaternion)	Řeší zkreslení základny v časové doméně.	Pracuje pouze s DQFT (Discrete quaternion Fourier transformation)
EIIP	Uvažuje fyzikálně-chemické vlastnosti DNA, nízká výpočetní náročnost pro analýzu, lepší schopnosti rozlišení genů.	Neschopnost detekovat kódující úseky v některých genomech.
Atomové číslo	Uvažuje fyzikálně-chemické vlastnosti DNA.	Potřeba hlubšího průzkumu.
Párová čísla (Paired numeric)	Reflektuje strukturální vlastnosti DNA, snížená komplexnost, snížená doba procesu DFT (Diskrétní Fourierova transformace), lepší idetifikace kódujících úseků než u jiných metod.	Potřeba hlubšího průzkumu.
Z křivka	Čistá biologická interpretace, nezávislé komponenty x_n, y_n, z_n, snížený výpočetní čas analýz, nadřazený technikám posuvného okna, nabízí numerickou i grafickou vizualizaci.	
DNA Walk (Krok)	Poskytuje korelační informace v dlouhém rozsahu, detekce periodicit v sekvenci, změn ve skladbě nukleotidů, nabízí numerickou i grafickou vizualizaci.	Není vhodné pro delší sekvence (víc než 1000 bází).
DNA Walk (Gate)	Jednoduchost.	Vysoké množství nadbytečných vlastností (tvorba repetitivních uzavřených smyček nebo překřížení)
DNA Walk (Yau)	Žádné uzavřené smyčky nebo překřížení, symetrie reprezentace.	Nízká míra degenerace.

Tab. 12: Pozice MGE v organismu *Klebsiella pneumoniae* druh 12836 [48]

Název MGE	**Pozice 1.báze MGE v genomu [bp]**	**Pozice poslední báze MGE v genomu [bp]**	**Délka MGE [bp]**	**Kompletní MGE?**	**Na komplementárním vlákně?**
IS4321	2182	3508	1327	Ano	Ne
integron:class 1	7172	36264	29093	Ano	Ne
ISCR1	11597	13750	2154	Ano	Ne
IS1326	17893	18268	376	Ano	Ne
Tn3-like	18269	18306	38	Ano	Ne
IS26	23109	23172	64	Ano	Ne
ISKpn11	24348	25637	1290	Ano	Ne
ISKpn12	26699	27540	842	Ano	Ne
ISKpn13	28231	29378	1148	Ano	Ne
ISPa12	30407	31793	1387	Ano	Ne
ISKpn13	32160	33307	1148	Ano	Ne
IS1326	33311	33586	276	Ano	Ne

Tab. 13: Pozice MGE v organismu *Salmonella typhimurium* LT2 pro úpravu [48]

Název MGE	**Pozice 1.báze MGE v genomu [bp]**	**Pozice poslední báze MGE v genomu [bp]**	**Délka MGE [bp]**	**Kompletní MGE?**	**Na komplementárním vlákně?**
IS200_1	1025026	1025734	709	Ano	Ne
IS200_2	2045625	2046333	709	Ano	Ano
IS200_3	2579223	2579931	709	Ano	Ano
IS200_4	3194217	3194925	709	Ano	Ano
IS200_5	3635061	3635769	709	Ano	Ano
IS200_6	4559289	4559947	659	Ano	Ne

Tab. 14: Pozice MGE v organismu *Salmonella enterica* plasmid pHCM1 pro úpravu [48]

Název MGE	Pozice 1.báze MGE v genomu [bp]	Pozice poslední báze MGE v genomu [bp]	Délka MGE [bp]	Kompletní MGE?	Na komplementárním vlákně?
IS1	1871	2585	715	Ano	Ano
IS1	106385	107132	748	Ano	Ne
IS1	108797	109544	748	Ano	Ne
IS6100	118817	119971	1155	Ano	Ano
IS10	142985	144299	1315	Ano	Ne
IS1	153546	154293	748	Ano	Ano
IS4321	155823	157139	1317	Ano	Ne
IS15/26	159639	160460	822	Ano	Ano
IS15/26	160655	161474	820	Ano	Ano
IS15/26	163306	164125	820	Ano	Ne
IS15/26	168511	169332	822	Ano	Ne
IS4321	175343	176659	1317	Ano	Ne
IS1	176752	177500	749	Ano	Ano
IS10	183224	184538	1315	Ano	Ano

Tab. 15: Pozice MGE v organismu *Yersinia pestis* CO92 chomozom pro úpravu [48]

Název MGE	Pozice 1.báze MGE v genomu [bp]	Pozice poslední báze MGE v genomu [bp]	Délka MGE [bp]	Kompletní MGE?	Na komplementárním vlákně?
IS285	17835	19149	1315	Ano	Ne
IS1541	25983	26694	712	Ano	Ne
IS100	39270	41223	1954	Ano	Ano
IS1541	103347	104058	712	Ano	Ano
IS100	105579	107532	1954	Ano	Ano
IS100	190051	192004	1954	Ano	Ne
IS1541	219704	220415	712	Ano	Ne
IS100	245726	247679	1954	Ano	Ano
IS1541	300409	301120	712	Ano	Ne
IS1541	366838	367549	712	Ano	Ano
IS100	401048	403001	1954	Ano	Ano
IS1541	509757	510468	712	Ano	Ne
IS1541	519648	520359	712	Ano	Ne
IS100	571185	573138	1954	Ano	Ano
IS1541	578376	579087	712	Ano	Ano
IS1541	586295	587006	712	Ano	Ne
IS1541	612841	612966	126	Ne	Ne
IS100	612967	614920	1954	Ano	Ano
IS1541	628060	628771	712	Ano	Ano

Tab. 16: Pozice MGE v organismu *Escherichia Coli* CFT073 pro úpravu [48]

Název MGE	Pozice 1.báze MGE v genomu [bp]	Pozice poslední báze MGE v genomu [bp]	Délka MGE [bp]	Kompletní MGE?	Na komplementárním vlákně?
IS1541A-like	308	1018	711	Ano	Ne
IS1541A-like	67629	68339	711	Ano	Ne
IS1541A-like	81932	82642	711	Ano	Ne
IS1397	115321	116747	1427	Ano	Ne
IS629	130990	132299	1310	Ano	Ne
IS1541A-like	192688	193398	711	Ano	Ne
IS1397	226129	227555	1427	Ano	Ano
IS630	253682	254791	1110	Ano	Ne
ISEc10	255335	257747	2413	Ano	Ne
IS285	323928	325231	1304	Ano	Ne
IS629	331375	332684	1310	Ano	Ano
IS1-like	376708	377302	595	Ano	Ne
IS1541A-like	542208	542918	711	Ano	Ne
IS1541A-like	749662	750372	711	Ano	Ne
IS1H	1131351	1131552	202	Ne	Ne
ISEc10	1166470	1167129	660	Ne	Ano
IS3E	1174451	1175044	594	Ne	Ano
IS3H	1175204	1175614	411	Ne	Ano
IS1541A-like	1208542	1209252	711	Ano	Ne
IS4	1209253	1209615	363	Ne	Ano
IS629	1212462	1213771	1310	Ano	Ano
ISEc10	1234406	1236815	2410	Ano	Ne
IS1541A-like	1239332	1240042	711	Ano	Ne
IS1541A-like	1311709	1312419	711	Ano	Ne
IS1541A-like	1341555	1342265	711	Ano	Ne
IS629	1398602	1399911	1310	Ano	Ne

Název MGE	Pozice 1.báze MGE v genomu [bp]	Pozice poslední báze MGE v genomu [bp]	Délka MGE [bp]	Kompletní MGE?	Na komplementárním vlákně?
ISEc10	1402805	1405214	2410	Ano	Ano
IS629	1415813	1417122	1310	Ano	Ne
IS4	1539548	1540530	983	Ne	Ano
IS1541A-like	1574016	1574726	711	Ano	Ne
ISEc2	1715771	1715914	144	Ne	Ne
IS1541A-like	2148217	2148927	711	Ano	Ano
IS1541A-like	2231360	2232070	711	Ano	Ano
IS1351	2315325	2315759	435	Ne	Ne
IS1351	2315748	2316270	523	Ne	Ne
ISSfl4	2341185	2342635	1451	Ano	Ne
ISEc10	2343328	2345737	2410	Ano	Ne
IS100X-like	2347085	2349034	1950	Ano	Ne
IS100kyp	2347085	2349034	1950	Ano	Ano
IS1541A-like	2467364	2468074	711	Ano	Ano
IS1541A-like	2483124	2483834	711	Ano	Ano
IS1541A-like	3026281	3026991	711	Ano	Ano
IS1541A-like	3283628	3284338	711	Ano	Ano
IS1541A-like	3329874	3330584	711	Ano	Ano
IS629	3439212	3440520	1309	Ano	Ano
ISEc10	3445765	3448174	2410	Ano	Ne
IS2	3451355	3452684	1330	Ano	Ano
IS4521R	3451371	3451890	520	Ne	Ne
IS4521R	3452324	3452684	361	Ne	Ne
ISEc8	3482089	3483665	1577	Ne	Ne
IS4	3532762	3534186	1425	Ano	Ano
IS1F	4283351	4284110	760	Ano	Ano

Název MGE	Pozice 1.báze MGE v genomu [bp]	Pozice poslední báze MGE v genomu [bp]	Délka MGE [bp]	Kompletní MGE?	Na komplementárním vlákně?
IS30D	4291959	4293179	1221	Ano	Ano
ISEhe3	4293180	4294096	917	Ne	Ne
IS2	4328988	4329338	351	Ne	Ne
ISEc10	4919911	4922320	2410	Ano	Ne
IS629	4931881	4933192	1312	Ano	Ano
IS629	4939378	4940687	1310	Ano	Ne
IS629	4966615	4967924	1310	Ano	Ano
ISEc2	5130002	5130107	106	Ne	Ano
IS1541A-like	5210335	5211045	711	Ano	Ne

Tab. 17: Hodnoty korelací pro numerickou reprezentaci DNA Walk (Krok)

	EC_IS1397_1	EC_IS630	KP_IS4321	KP_ISPa12	ST_IS200_2	ST_IS200_5	SE_IS10_2	SE_IS1_4	Yp_IS100_4	Yp_IS1541_2	EC_gen	KP_gen	ST_gen	SE_gen	YP_gen	Náhodná 1	Náhodná 2	Náhodná 3	Náhodná 4	Náhodná 5
EC_IS1397_1	1,00																			
EC_IS630	0,96	1,00																		
KP_IS4321	0,82	0,90	1,00																	
KP_ISPa12	0,88	0,87	0,94	1,00																
ST_IS200_2	0,99	0,99	0,99	0,96	1,00															
ST_IS200_5	0,99	0,99	0,99	0,96	1,00	1,00														
SE_IS10_2	0,85	0,89	0,92	0,94	0,95	0,95	1,00													
SE_IS1_4	0,94	0,95	0,98	0,99	0,95	0,95	0,96	1,00												
Yp_IS100_4	0,83	0,96	0,99	0,92	0,96	0,96	0,93	0,97	1,00											
Yp_IS1541_2	0,99	0,99	0,99	0,95	1,00	1,00	0,95	0,94	0,96	1,00										
EC_gen	0,91	0,96	0,98	0,93	0,99	0,99	0,92	0,97	0,91	0,99	1,00									
KP_gen	0,96	0,93	0,95	0,91	0,92	0,92	0,95	0,90	0,98	0,91	0,98	1,00								
ST_gen	0,93	0,89	0,95	0,90	0,95	0,95	0,90	0,96	0,95	0,95	0,94	0,89	1,00							
SE_gen	0,99	0,97	0,98	0,93	0,98	0,98	0,96	0,94	0,97	0,98	0,98	0,95	0,91	1,00						
YP_gen	0,98	0,94	0,98	0,94	0,99	0,99	0,92	0,97	0,97	0,98	0,98	0,95	0,96	0,97	1,00					
Náhodná 1	0,98	0,98	0,92	0,87	0,99	0,99	0,94	0,93	0,95	0,99	0,98	0,93	0,87	0,98	0,94	1,00				
Náhodná 2	0,51	0,57	0,48	0,43	0,63	0,63	0,52	0,57	0,52	0,63	0,54	0,59	0,45	0,59	0,44	0,50	1,00			
Náhodná 3	0,82	0,84	0,84	0,83	0,89	0,89	0,85	0,91	0,81	0,88	0,83	0,77	0,77	0,85	0,81	0,84	0,57	1,00		
Náhodná 4	0,90	0,87	0,92	0,95	0,94	0,94	0,90	0,96	0,91	0,93	0,92	0,90	0,87	0,92	0,92	0,87	0,38	0,78	1,00	
Náhodná 5	0,91	0,90	0,97	0,93	0,94	0,94	0,96	0,97	0,96	0,94	0,96	0,96	0,89	0,94	0,93	0,87	0,48	0,82	0,87	1,00
PRŮMĚRY pro sloupec (sekvenci)	0,90	0,91	0,92	0,90	0,95	0,95	0,90	0,93	0,92	0,94	0,93	0,91	0,89	0,94	0,92	0,91	0,53	0,82	0,88	0,90

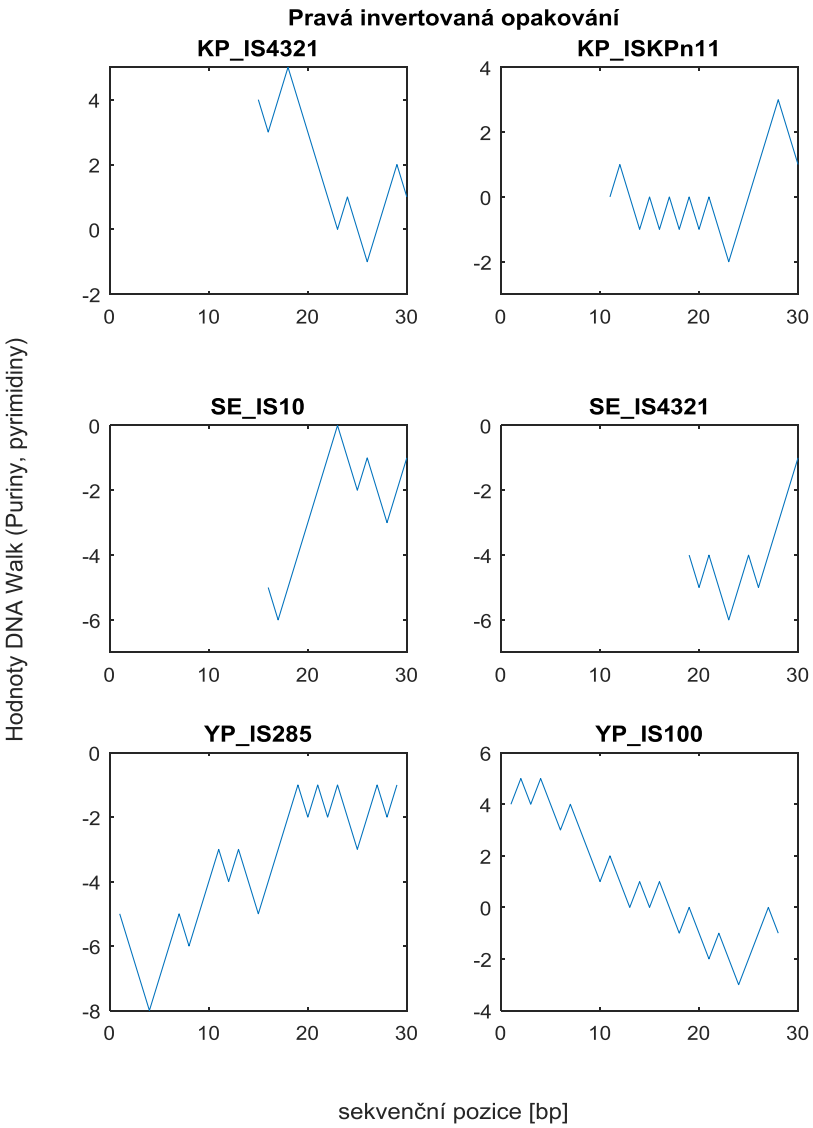
Tab. 18: Hodnoty korelací pro numerickou mapu Z křivka z-ová osa

	EC_IS1397_1	EC_IS630	KP_IS4321	KP_ISPa12	ST_IS200_2	ST_IS200_5	SE_IS10_2	SE_IS1_4	Yp_IS100_4	Yp_IS1541_2	EC_gen	KP_gen	ST_gen	SE_gen	YP_gen	Náhodná 1	Náhodná 2	Náhodná 3	Náhodná 4	Náhodná 5
EC_IS1397_1	1,00																			
EC_IS630	0,89	1,00																		
KP_IS4321	0,88	0,94	1,00																	
KP_ISPa12	0,98	0,92	0,87	1,00																
ST_IS200_2	0,82	0,94	0,88	0,83	1,00															
ST_IS200_5	0,82	0,94	0,88	0,83	1,00	1,00														
SE_IS10_2	0,87	0,98	0,96	0,85	0,95	0,95	1,00													
SE_IS1_4	0,86	0,97	0,87	0,86	0,98	0,98	0,96	1,00												
Yp_IS100_4	0,80	0,99	0,93	0,76	0,97	0,97	0,97	0,98	1,00											
Yp_IS1541_2	0,84	0,95	0,87	0,85	0,99	0,99	0,97	0,97	0,96	1,00										
EC_gen	0,81	0,96	0,96	0,82	0,94	0,94	0,99	0,94	0,94	0,95	1,00									
KP_gen	0,94	0,98	0,95	0,94	0,94	0,94	0,99	0,97	0,99	0,95	0,98	1,00								
ST_gen	0,89	0,97	0,96	0,89	0,92	0,92	0,98	0,96	0,98	0,94	0,97	0,97	1,00							
SE_gen	0,73	0,90	0,80	0,75	0,98	0,98	0,92	0,95	0,95	0,96	0,91	0,85	0,88	1,00						
YP_gen	0,98	0,97	0,97	0,96	0,86	0,86	0,99	0,90	0,99	0,89	0,98	0,97	0,95	0,78	1,00					
Náhodná 1	0,86	0,92	0,87	0,88	0,88	0,88	0,91	0,92	0,92	0,85	0,90	0,92	0,89	0,83	0,87	1,00				
Náhodná 2	0,94	0,96	0,91	0,95	0,92	0,92	0,95	0,97	0,97	0,93	0,93	0,98	0,91	0,83	0,95	0,91	1,00			
Náhodná 3	0,59	0,68	0,67	0,63	0,69	0,69	0,68	0,68	0,68	0,70	0,74	0,69	0,68	0,61	0,63	0,64	0,60	1,00		
Náhodná 4	0,36	0,52	0,49	0,36	0,89	0,89	0,55	0,90	0,62	0,91	0,56	0,77	0,63	0,84	0,44	0,49	0,44	0,39	1,00	
Náhodná 5	0,78	0,80	0,86	0,81	0,78	0,78	0,83	0,72	0,81	0,78	0,81	0,74	0,82	0,71	0,82	0,68	0,73	0,64	0,33	1,00
PRŮMĚRY pro sloupec (sekvenci)	0,82	0,90	0,87	0,83	0,90	0,90	0,91	0,91	0,90	0,91	0,90	0,92	0,90	0,85	0,88	0,84	0,88	0,65	0,60	0,75

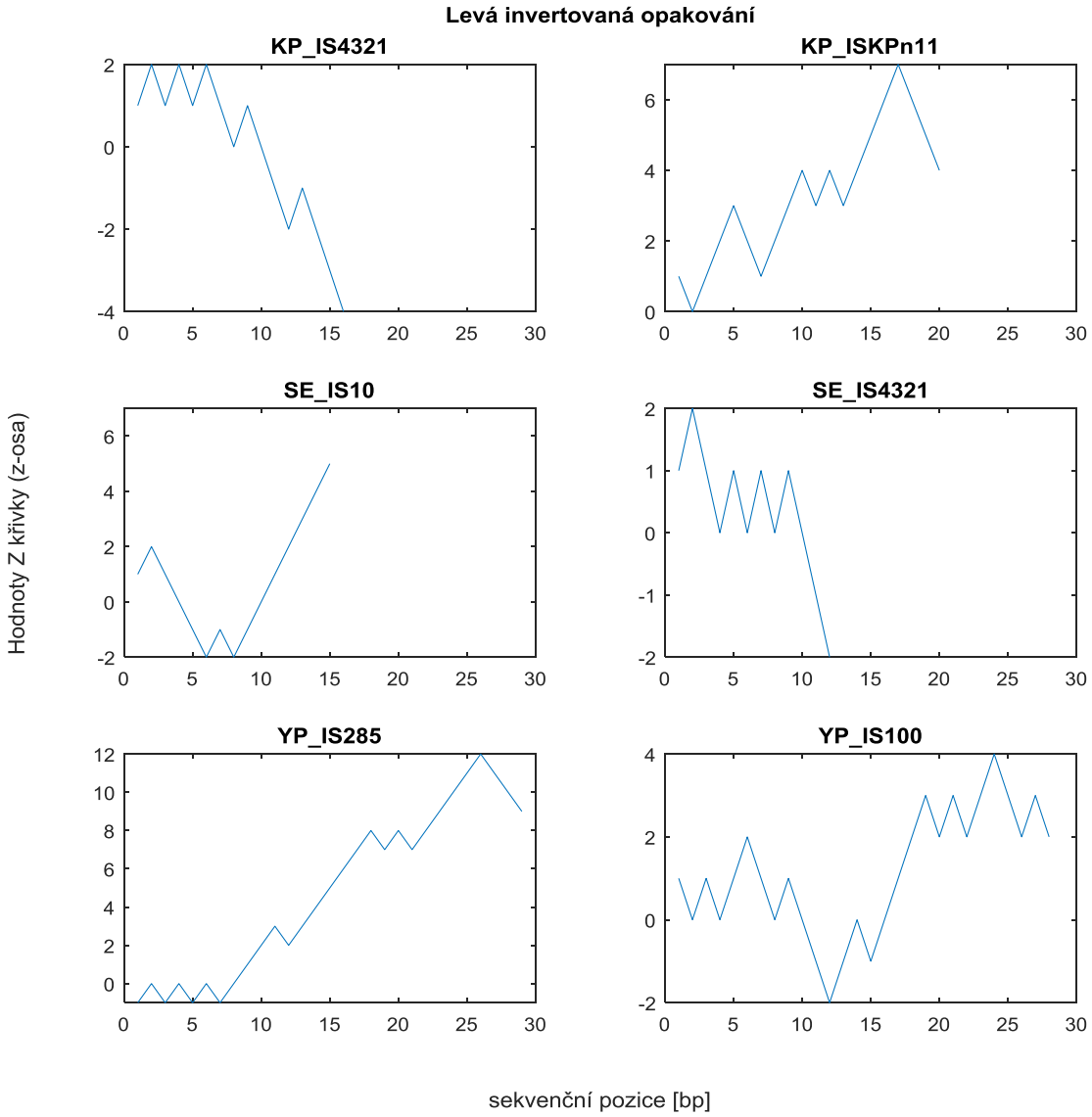
Tab. 19: Hodnoty korelací pro numerickou mapu EIIP

	EC_IS1397_1	EC_IS630	KP_IS4321	KP_ISPa12	ST_IS200_2	ST_IS200_5	SE_IS10_2	SE_IS1_4	Yp_IS100_4	Yp_IS1541_2	EC_gen	KP_gen	ST_gen	SE_gen	YP_gen	Náhodná 1	Náhodná 2	Náhodná 3	Náhodná 4	Náhodná 5
EC_IS1397_1	1,00																			
EC_IS630	0,97	1,00																		
KP_IS4321	0,97	0,97	1,00																	
KP_ISPa12	0,97	0,97	0,97	1,00																
ST_IS200_2	0,97	0,97	0,97	0,98	1,00															
ST_IS200_5	0,97	0,97	0,97	0,98	1,00	1,00														
SE_IS10_2	0,97	0,97	0,97	0,97	0,97	0,97	1,00													
SE_IS1_4	0,97	0,97	0,97	0,97	0,97	0,97	0,97	1,00												
Yp_IS100_4	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	1,00											
Yp_IS1541_2	0,97	0,97	0,98	0,98	0,99	0,99	0,97	0,97	0,97	1,00										
EC_gen	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	1,00									
KP_gen	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	1,00								
ST_gen	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	1,00							
SE_gen	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	1,00						
YP_gen	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,97	1,00					
Náhodná 1	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,97	0,97	1,00				
Náhodná 2	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,97	0,97	0,96	1,00			
Náhodná 3	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,97	0,97	0,96	0,97	1,00		
Náhodná 4	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,97	0,97	0,97	0,97	0,97	1,00	
Náhodná 5	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	1,00
PRŮMĚRY pro sloupec (sekvenci)	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97

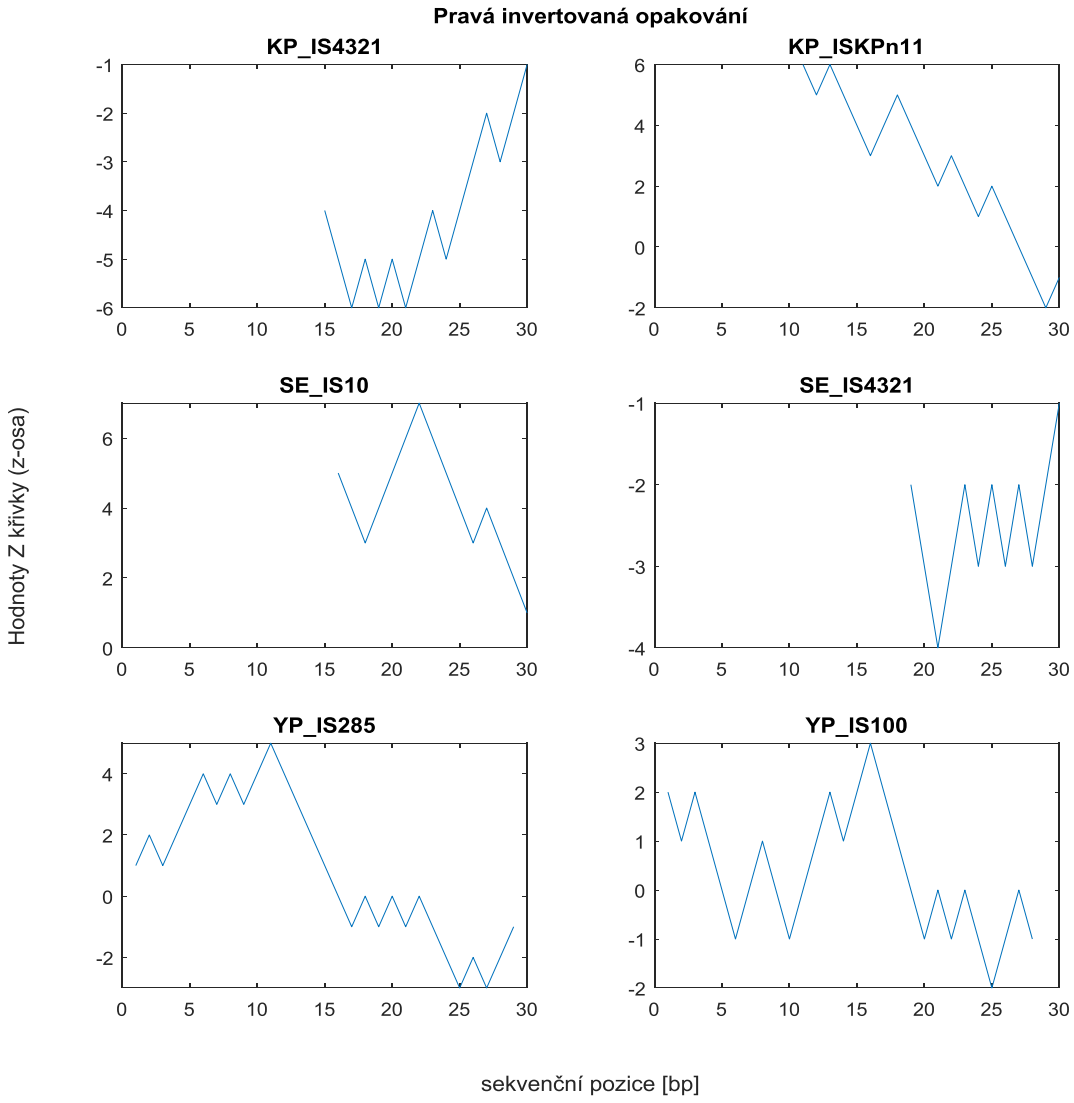
Obrázky



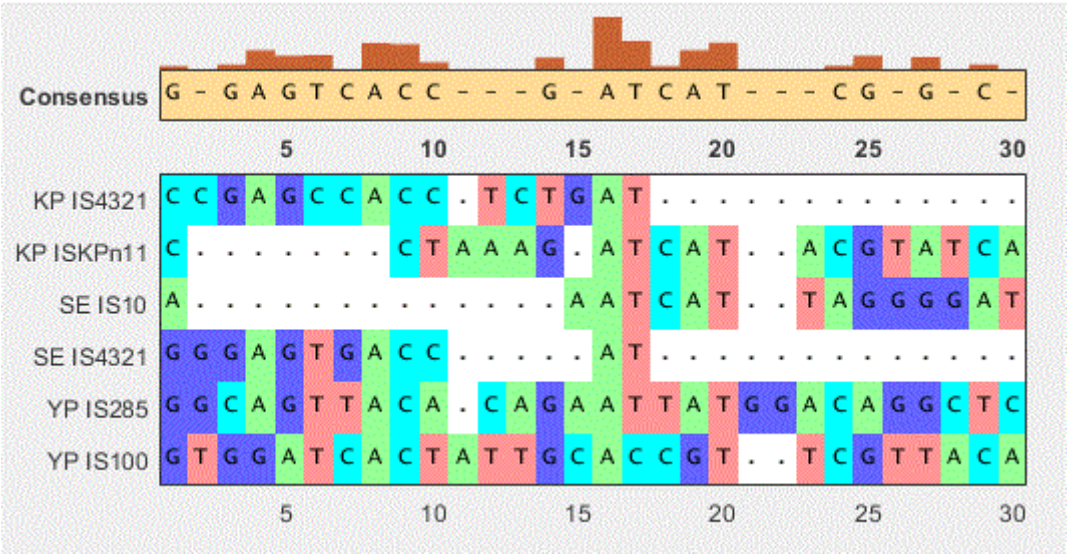
Obr. 26: Porovnání IRR po konverzi numerickou mapou DNA Walk (Puriny pirimidiny)



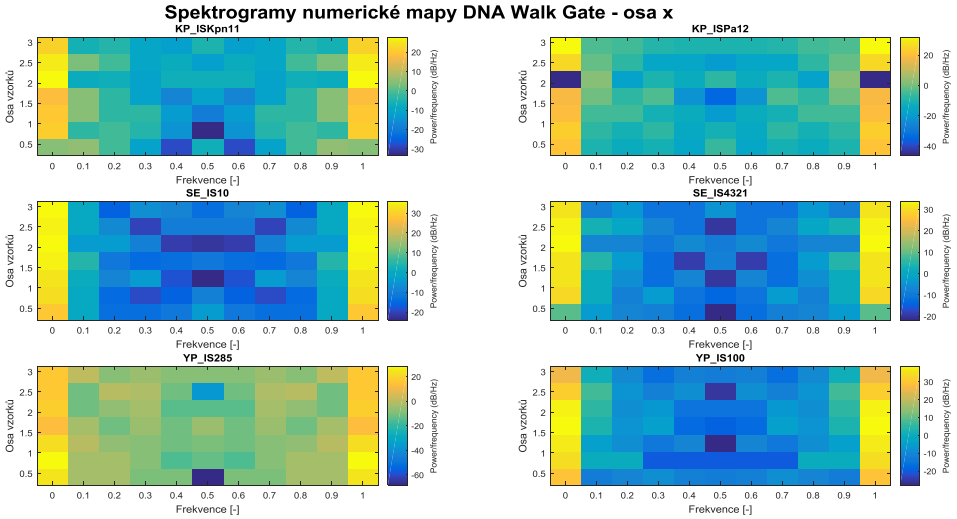
Obr. 27: Porovnání IRL po konverzi numerickou mapou Z křivka (z-osa)



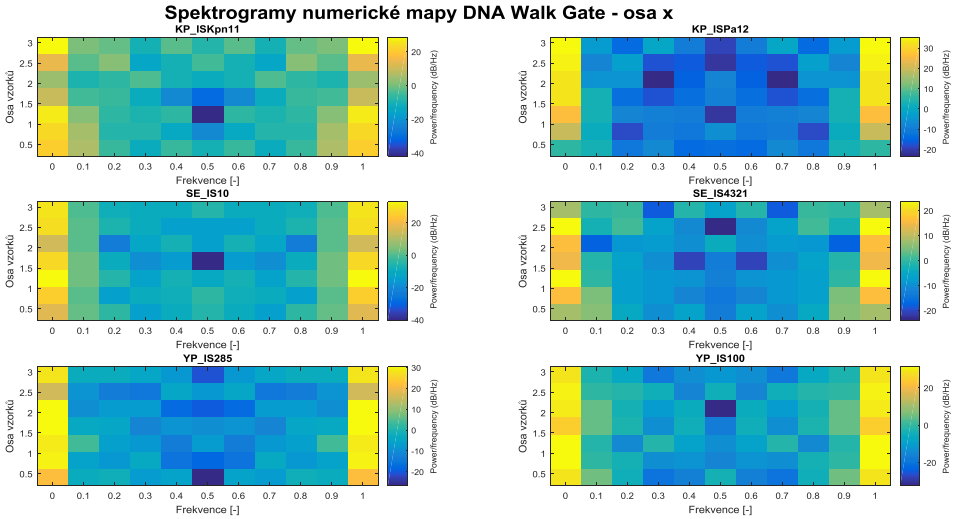
Obr. 28: Porovnání IRR po konverzi numerickou mapou Z křivka (z-osa)



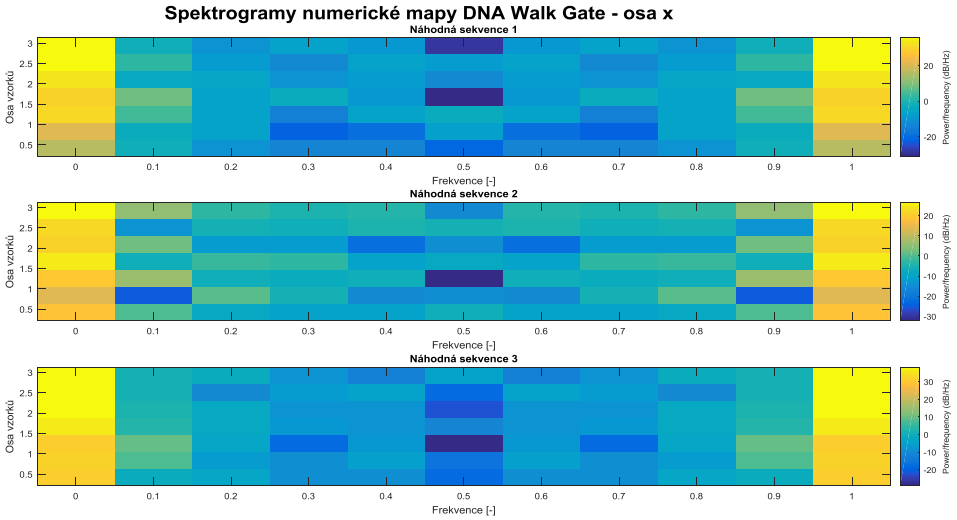
Obr. 29: Zarovnání symbolických sekvencí IRR zobrazených v Obr. 15



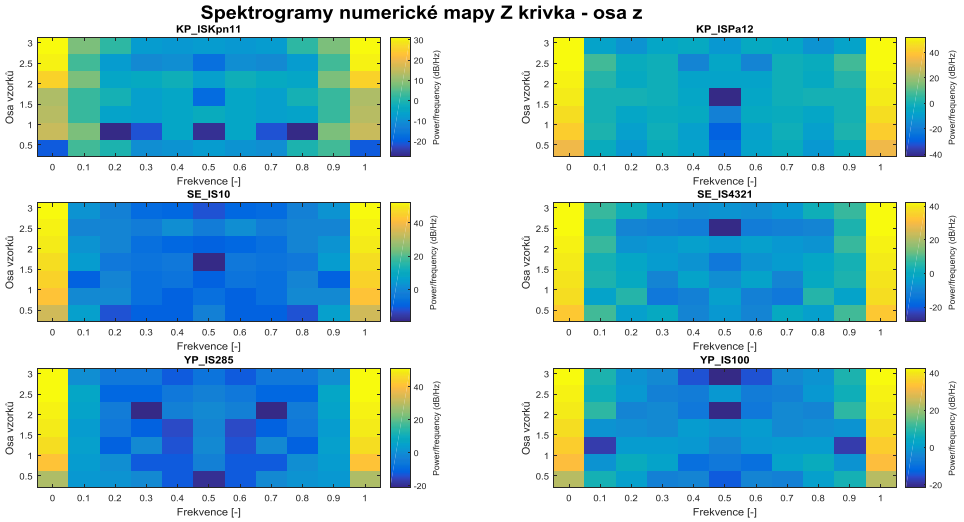
Obr. 30: Spektrogramy IRL (numerická mapa DNA Walk Gate – osa x)



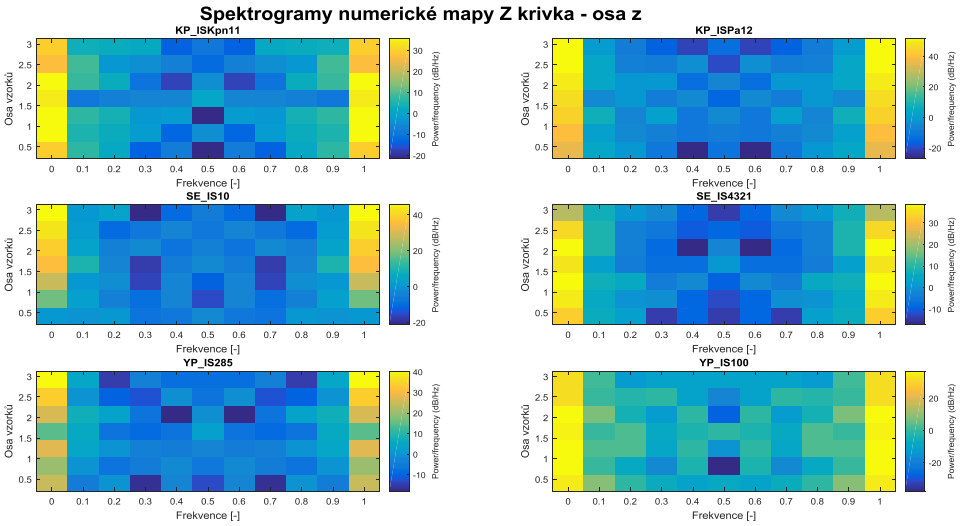
Obr. 31: Spektrogramy IRR (numerická mapa DNA Walk Gate – osa x)



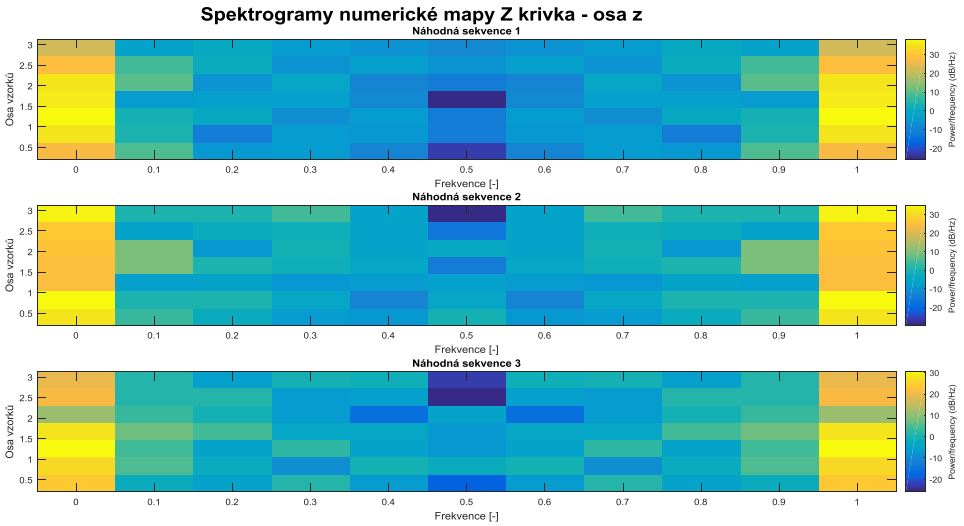
Obr. 32: Spektrogramy náhodných sekvencí (numerická mapa DNA Walk Gate – osa x)



Obr. 33: Spektrogramy IRL (numerická mapa Z křivka – osa z)



Obr. 34: Spektrogramy IRR (numerická mapa Z křivka – osa z)



Obr. 35: Spektrogramy náhodných sekvencí (numerická mapa Z křivka – osa z)