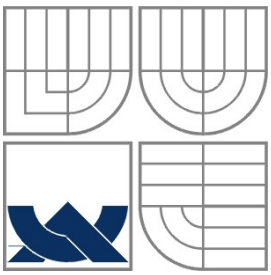
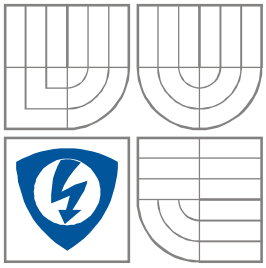


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

KOMPRESE GENOMICKÝCH SIGNÁLŮ PRO KLASIFIKACI A IDENTIFIKACI ORGANISMŮ

THE USE OF GENOMIC SIGNAL COMPRESSION FOR CLASSIFICATION AND IDENTIFICATION OF ORGANISMS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. KAREL SEDLÁŘ

AUTHOR

VEDOUCÍ PRÁCE Ing. HELENA ŠKUTKOVÁ

SUPERVISOR

BRNO, 2013

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Student: Bc. Karel Sedlář ***ID:*** 115114
Ročník: 2 ***Akademický rok:*** 2012/2013

NÁZEV TÉMATU:

Komprese genomických signálů pro klasifikaci a identifikaci organismů

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši metod konverze symbolického zápisu DNA sekvence na genomický signál. 2) Zhodnoťte možnosti využití kompresních algoritmů pro číslicové signály ke kompresi a klasifikaci genomických dat. 3) Navrhněte algoritmus pro vyhledání strukturálních podobností a redundantních složek v genomickém signálu umožňující následnou kompresi DNA sekvencí. 4) Vytvořte banku komprimovaných signálových reprezentací vhodných pro klasifikaci a taxonomické třídění organismů na základě sekvencí celých chromosomů. 5) Klasifikaci otestujte na reálných sekvencích z veřejných databází. Zvlášť vyhodnoťte prokaryotické a eukaryotické organismy. 6) Proveďte statistické vyhodnocení a diskuzi výsledků.

DOPORUČENÁ LITERATURA:

[1] CRISTEA, Paul Dan. Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine. 2002, roč. 6, č. 2, s. 279-303.
[2] CHEN, Xin, Sam KWONG a Ming LI. A Compression Algorithm for DNA Sequences. IEEE ENGINEERING IN MEDICINE AND BIOLOGY. 2001, 7-8, s. 61-66.

Termín zadání: 11.2.2013 ***Termín odevzdání:*** 24.5.2013

Vedoucí práce: Ing. Helena Škutková
Konzultanti diplomové práce:

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

UPOZORNĚNÍ:

Abstrakt

Moderní klasifikace organismů je založena na molekulárních datech. Přitom tyto metody spoléhají na vícenásobné zarovnání znakových sekvencí, což je činí výpočetně velmi náročnými. Proto je také možné zpracovávat pouze krátké úseky genomu. Tato práce ukazuje zcela nový algoritmus postavený na konverzi celogenomových sekvencí do signálů kumulované fáze. Ty jsou podrobeny ztrátové kompresi spočívající v odstranění nadbytečných frekvenčních pásem pomocí vlnkové transformace. Klasifikace organismů je následně provedena jako shluková analýza s využitím euklidovské vzdálenosti dvojic komprimovaných signálů, kde je zarovnání provedeno algoritmem pro dynamické borcení časové osy.

Klíčová slova

genomický signál, kumulovaná fáze, komprese, klasifikace, DWT, DTW

Abstract

Modern classification of organisms is performed on molecular data. These methods rely on multiple alignment of sequences of characters which make them computationally demanding. Only small parts of genomes can be compared in reasonable time. In this paper, the novel algorithm based on conversion of the whole genome sequences to cumulative phase signals is presented. Dyadic wavelet transform is used for lossy compression of signals by redundant frequency bands elimination. Signal classification is then performed as a cluster analysis using Euclidian metrics where multiple alignment is replaced by dynamic time warping.

Key words

genomic signal, cumulated phase, compression, classification, DWT, DTW

SEDLÁŘ, K. *Komprese genomických signálů pro klasifikaci a identifikaci organismů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 91 s. Vedoucí diplomové práce Ing. Helena Škutková

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma Komprese genomických signálů pro klasifikaci a identifikaci organismů jsem vypracoval 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 autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne 17.5.2013 .....

podpis autora

Poděkování

Rád bych zde poděkoval několika lidem, bez kterých by tato práce nemohla vzniknout. Především Ing. Heleně Škutkové, vedoucí diplomové práce a své mentorce, za velkou motivaci, rady a zkušenosti, které mi neustále předává.

Dále chci poděkovat svým rodičům, jejichž podpora v mém studiu je obrovská a ničím nenahraditelná. A také své babičce, jakožto první čtenářce a korektorce jazyka.

V Brně dne 17.5.2013 .....

podpis autora

Obsah

Soupis tabulek

Soupis obrázků

Úvod

Využití číslicového zpracování signálů v genomice může na první pohled působit nestandardně, protože genomická data jsme v první řadě zvyklí vnímat ve formě znakových sekvencí. Ačkoliv existuje celá řada technik na jejich zpracování, tedy i na jejich kompresi, začínají tyto metody dosahovat svých možností, co se rychlosti a efektivnosti užití týče. Toto omezení vychází právě z formy zápisu. Symbolickou sekvenci nelze zjednodušit, podvzorkovat, vhodně graficky vyobrazit, vždy musí v každém místě nabývat hodnoty jednoho z povolených znaků. Možnosti komprese jsou tak omezeny na použití pouze bezztrátových technik. Navíc jejich komprese spočívá ve změně kódování a komprimovaná data tak nelze přímo zpracovávat bioinformatickými nástroji.

Poslední desetiletí přineslo velký rozvoj v oblasti sekvenace DNA. Nové techniky jsou několikanásobně rychlejší a levnější než předchozí postupy. Standardem dnešní doby je tak sekvenování celých genomů. Současné nástroje pro identifikaci a klasifikaci organismů ale neumožňují pracovat s celými genomy v důsledku přílišné výpočetní náročnosti. Analýzu tak staví pouze na určitých genech, či jiných krátkých úsecích sekvence. Tyto úseky přitom nemusí věrně reprezentovat fylogenetický vývoj celého genomu.

V následujícím textu chci ukázat možnost porovnání celých genomů pomocí numerické reprezentace genomických dat. Technik převodu sekvence na signál existuje větší množství a poskytují rozdílné signály. Jejich vlastnosti však zůstávají do dnešní doby jen velmi málo popsané. Podrobnou analýzou chci získat techniku, která bude výpočetně nenáročná a bude poskytovat signály, které jsou specifické pro různé organismy a jsou snadno zpracovatelné.

Časově-frekvenční, respektive u genomických signálů pozičně-frekvenční analýzou, chci odhalit možnosti ztrátové komprese při zachování významných charakteristik signálu. Míra komprese pak bude volena jako kompromis mezi množstvím ztracené informace a ušetřením výpočetního času při další analýze a bude zvlášť hodnocena pro prokaryotní a eukaryotní organismy. Nejpodstatnější podmínkou komprese pak bude zachování kódování, které bude přímo zpracovatelné dalšími nástroji.

Před samotnou klasifikací se zaměřím na zarovnání signálů, což je nedílná součást bioinformatických postupů při zpracování více sekvencí. Následně chci vytvořit banku komprimovaných signálů, podle kterých by mohly být nově získané sekvence identifikovány.

1. Genomická data

V první kapitole této práce si objasníme základní pojmy, se kterými genomika pracuje. Pro numerickou reprezentaci a následnou kompresi genomických dat je podstatné, kromě pochopení principu zápisu jednotlivých nukleotidů, také pochopení jejich chemického a fyzikálního základu.

1.1 Genomika

Genomika je věda zabývající se studiem genomů organismů. Je tedy jedním z podoborů genetiky. Spíše než zkoumáním funkcí jednotlivých genů, což je úkol genetiky jako takové, se zabývá genomy jako celky. Genomem rozumíme veškerou DNA, kterou organismus obsahuje. V té geny tvoří pouze určitou část. Hlavně u vývojově vyšších organismů mnohanásobně převažují především nekódující úseky. I když struktura DNA je známa již od roku 1953, kdy J. Watson a F. Crick publikovali svoji práci o pravotočivé dvoušroubovici [1], vznik genomiky je spjat až s objevením techniky umožňující odhalení posloupnosti dusíkatých bází ve vláknech DNA, tzv. sekvenací. První masivněji využívanou metodou byla Sangerova dideoxy metoda [2], za níž autor dostal v roce 1980 Nobelovu cenu. Prvními kompletně osekvenovanými byly genomy virů a mitochondrií, které jsou v porovnáním s těmi vyšších organismů velice malé. Snad proto, že rychlost Sangerovy sekvenace není vysoká a z počátku nebyla běžně dostupná, stalo se standardem nasekvenovaná data sdílet prostřednictvím veřejných databází. V celosvětovém měřítku dnes funguje trojice databází, které jsou vzájemně propojené. Jedná se o americkou GenBank, evropskou EMBL a japonskou DDBJ [3]. Ty si každý den vyměňují data a jejich obsah je tak až na přírůstky z posledního dne totožný.

Přelom v genomice znamenal projekt The Human Genome Project (HGP) [4], jenž započal v roce 1990 a jeho hlavním cílem bylo odhalení 20 000 až 25 000 genů a zhruba 3 miliard párů bází lidského genomu. A následné uchování těchto dat právě ve veřejných databázích. Projekt byl završen s dvouletým předstihem oproti plánu v roce 2003. V rámci přidružených cílů byly sekvenovány i další organismy. Objem dat ve veřejných databázích tak začal růst velkým tempem, jelikož sekvenace se stala díky vylepšením, které projekt přinesl, dostupnější. Každý člověk je ovšem jiný, protože geny se vyskytují ve více formách. I po ukončení HGP jsou tak znovu sekvenovány již známé genomy, včetně toho lidského. V dnešní době je toto otázkou spíše sekvenátorů nové generace [5]. Ty jsou uzpůsobeny na sekvenaci více vzorků najednou a odhalování různých haplotypů jednotlivých genů. Veřejná prezentace nově objevených genotypů má pak za následek další růst databází. Ten je v dnešní době exponenciální. Zatímco v roce 1982 obsahovala databáze GenBank pouze 606 sekvencí, v dnešních dnech je jich v ní více než 100 miliónů.

1.2 Genom

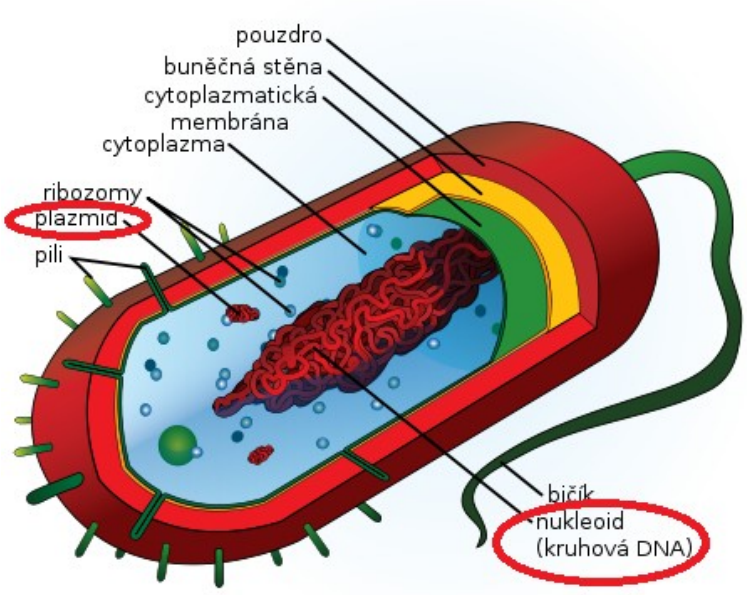
Důkladné studium genomů, jejich rozbor a popis charakteristik jakými se vyznačují, odhalí některé teoretické možnosti použitelné právě pro jejich kompresi. Genom můžeme rozdělit do dvou základních částí. Jsou to úseky kódující a úseky nekódující. Úseky kódující v sobě nesou informaci pro vytvoření příslušného proteinu. Při genové expresi jsou přepisovány (transkribovány) do mediátorové RNA (mRNA) a na základě jejich posloupnosti tripletů (kodónů) je vytvořena posloupnost aminokyselin v proteinu (proces translace mRNA do proteinu), což je funkční projev genu. Úseky nekódující tuto funkci nemají. Gen se ovšem nemusí nutně vyznačovat expresí do proteinu, ale pouze transkripcí do RNA, která dále ovlivňuje transkripci dalších genů. Důležité je brát v potaz, že genomy obsahují velké množství úseků, jenž se opakují. Tyto opakující se úseky mohou být navíc pohyblivé. K jejich vmezeření do kódující oblasti dochází pouze zřídkakdy, tedy může se zdát, že do klasifikace na základě kratších kódujících úseků nezasahují. Musíme si ovšem uvědomit, že geny se skládající z více exonů, neboli z více kódujících částí, mezi které jsou vloženy introny, které kódující funkci nemají. Pokud se do intronu vmezeří pohyblivý element (tj. transpozon), změní jeho délku. Pozice následujících exonů v rámci celé sekvence se tak změní. To může vést k zcela neoprávněné klasifikaci takového genu, či úseku DNA, pokud jsou geny získány pouze výběrem z osekvenované DNA, neboť identifikace exonů znakovými metodami v současné době ještě nedosahuje zcela uspokojivých výsledků [6]. I zde tak nalézá uplatnění signálová reprezentace a nástroje číslicového zpracování, neboť dosahují slibnějších výsledků [7]. Stále je ovšem nutné počítat s tím, že data z veřejných databází jsou zatížena různými chybami, neboť databáze velké trojky nejsou moderované. Při klasifikaci sekvencí získaných z těchto databází je velmi těžké tyto chyby odhalit, neboť vznikají již na úrovni de novo assembly [8]. Tedy na úrovni bioinformatických postupů, při kterých jsou jednotlivá krátká čtení DNA, získaná sekvenováním, skládána ve výslednou dlouhou sekvenci. Riziko takových chyb je možné snížit využitím sekvencí získaných pomocí pair-end read, tj. postupu, při kterém jsou krátké úseky DNA čteny z obou konců. Na rozdíl od single-end read, kdy se vlákno DNA čte pouze z jednoho konce, mají tato čtení lepší předpoklad pro správnou assembly [9], neboli výslednou konsezuální sekvenci poskládanou z přečtených krátkých úseků. Jednoznačně nejspolehlivější metodou je pak použít sekvence získané sekvenováním mRNA po splicingu, kdy máme jistotu, že daná sekvence je kódující. Sekvenování mRNA je ale kvůli nízké stabilitě molekuly složitější a je možné až po jejím převedení na stabilní komplementární DNA (cDNA). Přestože jsou takové úseky o mnoho kratší než sekvenování celé DNA i s introny, stále se ve většině případů jedná o úseky složené spojením více čtení. Chyba způsobená špatnou assembly je tak nižší, ale ne nemožná. Podle nepublikovaných odhadů obsahuje až 25% assemblies ve veřejných databázích chyby.

Vhodnější tak může být rozdělení genomu na repetitivní a nerepetitivní DNA [10]. Pod pojmem neopakující se DNA pak zobecníme sekvence genů, ať již protein, či pouze RNA kódující. Zbylá DNA je potom ta opakující se. Rozlišujeme v ní tandemové a rozptýlené repetice.

Zcela základní rozdíl na úrovni genomu však nalezneme mezi organismy prokaryotními a eukaryotními.

Prokaryotní genom

Prokaryotní organismy mají jednoduchou skladbu jádra buňky, které je tvořeno jedním kruhovým chromosomem, tedy jediným dvojvláknem DNA. Přesněji je tvořen z DNA, RNA a proteinů. Podíl nerepetitivní DNA (vypočítaný jako celková délka nerepetitivní DNA ku velikosti genomu) je vetší než 50%, tedy tyto sekvence mají v genomu převahu. U většiny genomů zcela zásadní, protože podíl dosahuje 85 až 90%. To teoreticky znevýhodňuje tyto genomy při kompresi, protože z obecné teorie o kompresních algoritmech je jasné, že opakující se informace může být s výhodou komprimována například algoritmy RLE [11]. Tento předpoklad platí pouze pro znakovou sekvenci a v signálové reprezentaci nemusí být překážkou, jak si později ukážeme. Prokaryotní buňky jsou jednodušší celkově, neobsahují organely, ale pouze struktury. Tedy kromě jaderného genomu už žádný další nemají. Mohou sice obsahovat plazmidy, krátké cirkulární molekuly DNA, ty ovšem pro buňku nejsou životně důležité. Samotná klasifikace a identifikace je otázkou mnohem jednodušší než u eukaryot, právě díky přítomnosti jediného chromosomu. Schéma prokaryotní buňky se zvýrazněným kruhovým jaderným chromosomem a plazmidem ukazuje Obr. 1.1.



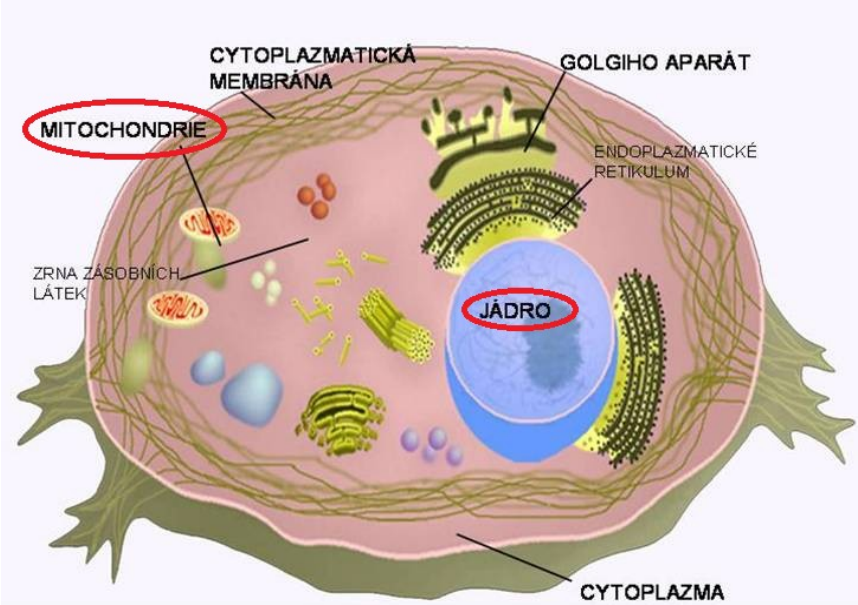
Obr. 1.1: Prokaryotní buňka [12]

Klasifikace a identifikace prokaryot má dvě základní užití. Prvním z nich je klasifikace a identifikace za účelem objevení rozdílů, či podobností mezi jednotlivými druhy. Tedy klasifikace za účelem zjištění fylogenetického vývoje. Taková klasifikace je velmi obtížná právě kvůli velkému podílu pohyblivých elementů, jenž kruhový chromosom obsahuje. Je velmi těžké vybrat pro analýzu takový úsek genomu, který by byl stabilní u všech testovaných organismů. Navíc u prokaryot dochází velmi často k horizontálnímu přenosu genetické informace, což je základní problém, se kterým se znakové metody založené na analýze pouze krátkých úseků genomu nemohou vyrovnat. Často tak není klasifikace jednoznačná, vyjádřitelná pouze stromovým grafem s binárním větvením, ale je nutné použít fylogenetické sítě [13]. Použitím celého genomu pro takovou analýzu bychom mohli klasifikaci velmi zjednodušit a především oprostit od chyb způsobených výše zmíněnými úskalími.

Druhým užitím je klasifikace a identifikace prokaryot ze vzorku odebraného na určitém místě. Tedy metagenomická analýza. Jedná se o velmi moderní přístup analýzy prokaryotních společenstev, například v půdě, ve střevě apod. V tomto typu identifikace není místo pro celogenomové signálové zpracování, protože hlavním požadavkem takové analýzy je rychlost klasifikace. Jsou tedy porovnávány pouze úseky, zpravidla 16S rRNA, jenž mohou být pokryty jedním čtením. Takové vzorky tak nejsou sekvenovány shotgun technikou, ale metodou amplikonového sekvenování. Kromě kvalitativní analýzy je u nich důležitá i analýza kvantitativní založená na počtu čtení přiřazených jednotlivým identifikovaným organizmům ze vzorku [14]. Nadále tak klasifikaci a identifikaci prokaryot budeme chápat pouze ve smyslu předcházejícího odstavce.

Eukaryotní genom

Eukaryota mají jadernou DNA dělenou do chromosomů, které se v jádře buňky navíc mohou objevovat ve více kopiích. Počet kopií je dán ploidií organismu. Například u diploidní lidské buňky je tedy každý somatický chromosom zastoupen dvakrát. Celková velikost genomu se však počítá vztažená pouze na jednu sadu chromosomů, jinak bychom zbytečně pracovali s velkým množstvím redundantní informace. Pro zpracování dat tak pro nás tento rozdíl oproti prokaryotům není podstatný. Problémem však zůstává samotné rozdělení DNA do více chromosomů, jejichž počet se může u různých organismů lišit. Klasifikaci by tedy bylo možné provádět například na základě vybraných chromozomů. Eukaryotní buňka je ovšem složitější nejen co se buněčného jádra týče, obsahuje různé organely. Velmi významnou organelou je mitochondrie se svou vlastní DNA. Schéma eukaryotní buňky se zvýrazněným jádrem a mitochondrií ukazuje Obr. 1.2 na následující straně.



Obr. 1.2: Eukaryotní buňka [15]

Mitochondriální DNA je kruhová a velmi se podobá kruhovému chromosomu prokaryotních organismů. I dnešní algoritmy na klasifikaci organismů často pracují právě s úseky mitochondriální DNA. Nabízí se tak otázka využití celé mitochondriální DNA (mtDNA) na klasifikaci organismů, neboť jeví dobrou intradruhovou stabilitu a interdruhovou variabilitu. Vhodnost jejího využití podtrhuje i fakt, že při sekvenaci genomu určitého organismu bývá zpravidla prvním krokem sekvenace jeho mitochondriální DNA. I ve veřejných databázích najdeme spoustu eukaryotních organismů, které již mají anotovaný mitochondriální genom, zatím co ten jaderný ještě ne. Proto nadále budeme chápat klasifikaci a identifikaci eukaryot na základě celého genomu, jako na základě mitochondriálního, nikoliv jaderného genomu.

1.3 Repetitivní a nerepetitivní DNA

Nerepetitivní DNA

Nerepetitivní DNA se v genomu vyskytuje pouze v jedné kopii. Tudíž jsou tyto sekvence neredundantní a jejich komprese ve znakovém zápisu je tím velice omezená. Jedná se především o kódující DNA, tedy geny. Nerepetitivní úseky jsou proto často používané při klasifikaci organismů. Jednou ze snah při klasifikaci na základě znakových metod je tak jejich odlišení. To může být založeno na GC obsahu. Pod tímto pojmem rozumíme podíl guaninu a cytosinu v dané sekvenci. GC obsah celého genomu je různý pro různé organismy. Ovšem kódující sekvence jsou GC bohatší než pozadí genomu [16]. Problém může následně nastat v rozlišení funkčních genů a pseudogenů. Pseudogeny jsou bývalé geny, které ztratili svoji kódující funkci a nejsou nadále exprimovány [17]. Jsou obecně považovány za odpadní

nekódující DNA. I přes své funkce si ovšem nadále drží rysy kódující DNA, včetně vyššího podílu GC.

Rozdíl mezi prokaryoty s vyšším podílem kódující DNA a eukaryoty, jejichž celkový podíl kódující DNA je velice malý, jsme potlačili použitím genomu mitochondriálního. Například geny člověka jsou v genomu opravdu hodně rozptýlené. Na 1 milion bází připadá pouze 12 genů. Exony mohou být velice malé, některé pouze 19 bp dlouhé. Mnoho intronů je pouze 87 bp dlouhých, délka některých však přerůstá hranici 10 000 bp [18]. Nicméně lidský mitochondriální genom obsahuje především úseky kódující.

Repetitivní DNA

Repetitivní DNA se v genomu opakuje. Kvůli opakování jsou takovéto sekvence redundantní a vhodné ke kompresi. Právě vyhledání strukturních podobností a opakujících se vzorů je klíčem pro zvýšení účinnosti kompresního algoritmu znakových dat. Opakující se DNA je zpravidla ta nekódující. V lidském genomu tedy ta bohatší na AT obsah (podíl adeninu a thyminu ku délce sekvence). Rozlišujeme přitom dva typy těchto opakujících se motivů. Pokud se vyskytují v řadě za sebou hovoříme o tandemových repeticích, motivy rozptýlené v genomu označujeme jako transpozony.

Tandemové repetice, také nazývané satelity, jsou otázkou především oblasti centromer chromosomů, což jsou oblasti genomu, jejichž sekvenace je problematická. I u lidského genomu, který považujeme za kompletně osekvenovaný je oblast centromery zapsána znaky N (cokoliv). Nicméně se vyskytují i v prokaryotních genomech a tak se s nimi běžně setkáváme. Transpozony mají svůj název spjatý se svým vznikem, tzv. transpozicí. Tu lze chápat jako přeskoky segmentu DNA z jednoho místa v genomu na jiné. Rozlišujeme přitom 2 typy těchto motivů. Prvním jsou DNA transpozony, které se však v lidském genomu vyskytují již pouze v inaktivní formě. Druhým typem jsou retrotranspozony. Ty jsou naopak velmi aktivní a v prokaryotickém genomu hojně zastoupené. [19]

Požadavky na genomické signály

Z toho co víme o nerepetitivní a repetitivní DNA, můžeme vyvodit požadavky pro ideální signálovou reprezentaci genomu pro klasifikaci a identifikaci organismů. Cílem nebude odlišení těchto úseků, jak je tomu u současných algoritmů. Ale naopak zakomponování těchto dvou částí takové, aby jejich vzájemná pozice, způsobená pohyblivými elementy, měla co nejmenší vliv na celkový tvar signálu. Samozřejmě při zachování odlišných tvarů signálů pro jednotlivé organismy.

1.4 Genomická data

Závěrem této kapitoly ještě rozebereme zápis genomických dat, jak v konvenční znakové formě, tak v signálové reprezentaci.

Genomická sekvence

Zvyklostí je zapisovat genomické sekvence jako posloupnosti IUPAC (International Union of Pure and Apllied Chemistry) znaků. A to ve směru od 5' ke 3' konci nukleové kyseliny. Standardem je použití 4 znaků: A,C,G,T. Ty reprezentují 4 nukleotidy nacházející se v DNA: adenin, cytosin, guanin, thymin. Pro RNA je pak znak T nahrazen U, protože místo thyminu se v RNA nachází uracil.

Při hledání ideální numerické mapy, tj. konvence, podle které připočteme každému nukleotidu odpovídající číselnou hodnotu, je vhodné uvažovat ještě další vlastnosti nukleotidů. A to jejich molekulární strukturu, zda se jedná o purin (A, G) nebo pyrimidin (C, T). Sílu vazby jakou se vážou ke komplementární bázi. Za vazbu silnou je považována vazba 3 vodíkovými můstky mezi C a G, slabá vazba mezi A a T je dána pouze 2 vodíkovými můstky. Nebo obsahem volných radikálů, kdy rozlišujeme amino (A, C) a keto (G, T) skupinu. I pro tyto vlastnosti má IUPAC kód svoje znaky, jsou uvedené v tabulce níže.

Tab. 1.1: IUPAC kódy pro nukleové kyseliny

IUPAC kód	význam
A	adenin
C	cytosin
G	guanin
T	thymin
U	uracil
N	cokoliv
R	A, G (purin)
Y	C, T (pyrimidin)
W	A, T (slabá vazba)
S	C, G (silná vazba)
M	A, C (amino skupina)
K	T, G (keto skupina)
.	mezera
-	mezera

Genomický signál

Základem převodu genomické sekvence na genomický signál je numerická mapa. Na ideální numerickou mapu klademe několik požadavků. Prvním požadavkem je, aby signál nesl stejnou informaci jako sekvence a nezaváděl žádnou informaci nad rámec symbolického

zápisu. Signálová reprezentace musí umožnit rychlé a efektivní zpracování. Velkou výhodou pak může být dobrá čitelnost pouze lidským okem. [18]

Numerických map a metod konverze existuje větší množství a poskytují signály s různými charakteristikami. Jednotlivým postupům tak budou věnovány speciální následující kapitoly. Zde si ovšem rozebereme základní rozdíly signálů a sekvencí a různé pohledy na tyto dvě odlišné formy zpracování.

Sekvence má přesně dané rozlišení na úrovni jednotlivých nukleotidů, na které se není možné podívat "z větší dálky". To však není vždy zcela žádoucí. Vizualizace delších úseků najednou je tak zcela vyloučena. Oproti tomu je signálové zpracování ve formě grafu jednoduše přiblížitelné a oddálitelné a umožňuje tak lidskému operátorovi vizualizaci například celého chromosomu v jenom okně. Nemluvě o změně rozlišení, například na základě podvzorkování. V takových případech je možné vypustit část informace, aniž by se změnil charakter sekvence jako celku, což při znakovém zpracování také není možné. Při tomto postupu však dochází ke ztrátě dat a takový signál již není zpětně převeditelný na sekvenci. Některé metody konverze poskytují rovnou signály, které již nelze zpětně konvertovat.

Poslední zmíněná vlastnost signálů však otevírá cestu pro využití ztrátové komprese na genomická data. Ta je pro znaková data vyloučená, vypuštění některých znaků by odporovalo dalšímu korektnímu zpracování. Signál ale můžeme zpracovat korektně i podvzorkovaný, jelikož své vlastnosti si ponechává i tak. Samozřejmě v rámci užití podvzorkování v rozumné míře.

2. Teorie komprese dat

V této kapitole probereme kompresní postupy a pojmy, které se v souvislosti s kompresí dat vyskytují. Komprese dat je založena na teorii informace, jíž věnujeme jednu z podkapitol. Dále probereme důležité pojmy potřebné pro hodnocení parametrů komprese. V závěru kapitoly si zmíníme současné kompresní algoritmy.

2.1 Teorie informace

Základem komprese je snižování redundance kódu, ve kterém jsou komprimovaná data zapsána. Těmito pojmy se zabývá teorie informace, což je abstraktní teorie zaměřená na přenos dat. Vznikla již ve 40. letech minulého století a za jejího zakladatele je považován C. E. Shannon [20]. Teorie je založená na měření množství informace, kterou nám zdroj dat, v našem případě sekvenace DNA, poskytuje. Informaci máme k dispozici ve formě zprávy složené ze slov. V případě sekvencí jsou slovy znaky, u signálu pak vzorky. Celou zprávu chápeme jako stochastický proces, deterministická zpráva totiž nenese žádnou informaci. Zdrojová slova pro počítačové zpracování kódujeme do kódových slov. Kódová slova se pak skládají z n symbolů (délka slova) a mohou obsahovat M symbolů kódové abecedy. Počet všech možných slov je pak

$$N = M^n. \tag{2.1}$$

Dále se budeme výlučně bavit o binárním kódování, jež je přirozené pro výpočetní techniku. V tom je počet symbolů kódové abecedy roven dvěma, zpravidla se jedná o znaky $\{0, 1\}$. Délka slova ve zdrojové abecedě tak nemusí být shodná s délkou slova v kódové abecedě. Možné kódování nukleotidů je vidět v následující tabulce. Toto kódování sice není ideální, počet slov, která dokážeme takto kódovat, je roven pouze 4 viz rovnice (2.1), nicméně i to stačí na pokrytí množiny 4 základních znaků používaných v sekvencích {A,C,G,T}.

Tab. 2.1: Kódování nukleotidů

zdrojové slovo	kódové slovo
A	00
C	01
G	10
T	11

Množství informace, jenž je slovem neseno přímo souvisí s pravděpodobností výskytu slova. Mějme slova A a B s pravděpodobnostmi výskytu *P(A)* a *P(B)*, potom pro množství informace *I(A)* a *I(B)* co nesou platí [28]:

- musí být nezáporné: $I(A) \geq 0$ pro $P(A) \in \langle 0,1 \rangle$
- musí být nulové pro apriorní znalost: $I(A) = 0$ pro $P(A) = 1$
- musí být větší u slov méně pravděpodobných: $I(A) > I(B)$ pro $P(A) < P(B)$
- musí být zajištěna možnost sčítání pro nezávislá slova: $I(AB) = I(A) + I(B)$

Celkové množství informace binárního kódového slova pak vypočítáme jako záporně vzatý logaritmus (při základu 2) pravděpodobnosti výskytu slova

$$I(A) = -log_2 P(A) \text{ [bit]}. \tag{2.2}$$

Množství informace na slovo však není vhodný ukazatel, jelikož tato hodnota může mezi jednotlivými slovy razantně kolísat. Řešením je vyjádření průměrného množství informace na slovo. Tuto veličinu označujeme jako *entropii*

$$H = -\sum_{i=1}^{N} P_i log_2 P_i \text{ [bit]}, \tag{2.3}$$

kde P_i je pravděpodobnost i-tého slova a N počet všech možných slov (viz rovnice). Což odpovídá fyzikální definici entropii, jako míře neuspořádanosti systému, který může nabývat N možných stavů a pravděpodobnostmi P_i. Maximální hodnota entropie, jaké pro danou zprávu může nabývat, je rovna entropii pro shodné pravděpodobnosti všech slov

$$H_{max} = log_2 N, \text{ pro } P_i = \frac{1}{N}, \ \forall i. \tag{2.4}$$

Samotná entropie pak nabývá hodnot z intervalu $\langle 0, H_{max} \rangle$. Poměr entropie ku maximální entropii určuje relativní entropii

$$H_r = \frac{H}{H_{max}}. \tag{2.5}$$

A z té již snadno dopočítáme redundanci kódu

$$\rho = 1 - H_r \tag{2.6}$$

2.2 Základní pojmy

Účinnost komprese

Základním požadavkem komprese dat je zmenšení jejich objemu. Při volbě nevhodného kompresního algoritmu však může dojít k negativní kompresi, tedy situaci, kdy mají data po kompresi větší objem. Výkon komprese můžeme měřit několika způsoby. Prvním z nich je *kompresní poměr*, který je definován jako

$$\text{kompresní poměr} = \frac{\text{objem výstupních dat}}{\text{objem vstupních dat}}.$$

Hodnota 0,5 znamená, že zkomprimovaná data mají 50% velikosti dat původních. Tedy čím menší je kompresní poměr, tím lepší komprese je. Hodnoty větší než 1 pak znamenají kompresi negativní, tedy expanzi objemu vstupních dat.

Převrácením hodnoty kompresního poměru získáme další způsob pro hodnocení kvality komprese označovaný jako kompresní faktor

$$\text{kompresní faktor} = \frac{\text{objem vstupních dat}}{\text{objem výstupních dat}}.$$

Pro mnoho lidí může být toto vyjádření přirozenější, protože čím větší hodnoty faktor nabývá, tím je komprese účinnější. Naopak čísla menší než 1 značí expanzi [11]. Jiným způsobem hodnocení kvality komprese je průměrná délka slova

$$\text{průměrná délka slova} = \frac{\text{objem dat po kompresi}}{\text{počet vzorků signálu}} \text{ [bit/vzorek]}.$$

Jedná se o způsob hodnocení spjatý se signály, jak napovídá i jednotka bit na vzorek definovaná pro průměrnou délku slova. Cílem je samozřejmě dosáhnout co nejmenšího počtu bitů na vzorek.

Rychlost komprese

Protiváhou účinnosti komprese je její rychlost. Obecně mezi nimi platí nelineární vztah, s větší mírou komprese často rapidně narůstá náročnost algoritmu a tím klesá jeho rychlost. Kompresní algoritmus musí volit vhodný kompromis mezi těmito dvěma faktory podle svého primárního určení.

Hodnocení rychlosti komprese je stejné jako hodnocení rychlosti jakéhokoliv jiného algoritmu. Tedy nehodnotí se přímo rychlost (časová náročnost), jež je podmíněna procesorem počítače, nýbrž asymptotická náročnost. Ta je definována pomocí Omikron notace (velké O notace) [21]. Je zbytečné hodnotit přesnou hodnotu náročnosti. Bereme tedy v úvahu pouze nejrychleji rostoucí člen závislosti a ostatní vypustíme. Například složitost algoritmu řízená vztahem $6n^3 + n + 5$ znamená náročnost $O(n^3)$, tedy kubickou náročnost. Kubická náročnost je typická například pro dva vnořené cykly, kdy zdvojnásobení n, berme jako zpracování dvojnásobně dlouhé sekvence, způsobí osminásobné prodloužení délky běhu kompresoru.

Ztrátovost informace

Z hlediska ztrátovosti informace při kompresi dělíme algoritmy na bezztrátové a ztrátové. Pro genomiku je typické použití bezztrátové komprese dat, právě kvůli faktu, který jsme si již jednou zmínili v předchozí kapitole. Vypuštění určité části nukleotidů ze sekvence by vedlo k nesprávné klasifikaci, protože nukleotidům nemůžeme přiřazovat důležitost. Vypuštění byť jediného vede k nesprávnému zpracování. U signálů je však ztrátová komprese možná. A pro

využití při klasifikaci i vhodná, neboť neklademe požadavek, aby komprimovaná data bylo možné navrátit do původního stavu.

2.3 Schéma komprese

Komprese genomických dat pomocí nástrojů číslicového zpracování signálů přidává do schématu jeden blok zpracování navíc. A to blok, ve kterém je potřeba sekvenční data převést na genomický signál. Prvním krokem je tedy konverze dat, která jsou v druhém kroku komprimována.

Obr. 2.1: Schéma komprese

Pro sekvenční znaková data je typická bezztrátová komprese. Pokud tedy budeme chtít, abychom po dekompresi získali zase data znaková, musí být postupy konverze a komprese reverzibilní bez ztráty informace. Tedy v žádném z bloků nesmí dojít ke ztrátě původní informace. Schéma dekomprese, je pak stejné jako schéma komprese, pouze probíhá v opačném směru.

Jak již bylo řečeno, zpracování genomických signálů otevírá cestu i pro využití ztrátové komprese. V takovém případě už tedy nebudeme požadovat po dekompresi genomickou sekvenci, kde i záměna jednoho znaku znamená následné nekorektní zpracování, jelikož sekvenční metody pracují vždy na úrovni rozlišení jednotlivých nukleotidů. Číslicové zpracování však může pracovat i s „horším" rozlišením. Na rozdíl od klasické komprese zde však máme dva bloky, kde může nastat ztráta informace. Může nastat již při konverzi, kdy signál není znovu převeditelný na sekvenci nebo až při kompresi. Navíc může nastat ztráta informace v obou dvou blocích současně.

2.4 Současné kompresní algoritmy DNA

Současné algoritmy pro kompresi DNA sekvencí jsou založeny na znakových metodách. Proto převažují algoritmy bezztrátové komprese. Existují ovšem i algoritmy pro kompresi ztrátovou, nejedná se ovšem o algoritmy uzpůsobené přímo pro kompresi DNA, spíše o obecné textové komprimační algoritmy, které fungují i při použití pouze 4 písmenné abecedy. Protože se v nukleotidové sekvenci vyskytují pouze 4 znaky, není možné použít standardní textové kompresní algoritmy. Samotná nekomprimovaná sekvence totiž používá pouze dvou bitové kódování. Použití nástrojů jako je Unix „compress" nebo MS-DOS „arj"

způsobí expanzi souboru se sekvencí [22]. V podstatě se současné algoritmy dají rozdělit do dvou skupin.

První skupinou jsou statistické algoritmy. Jejich kódování je založeno na přiřazování kratšího kódu populárnějším znakům. Popularita je zde hodnocena četností výskytu znaku v sekvenci. Statistické algoritmy mohou být postaveny na aritmetickém nebo Huffmanově kódování. Přitom se ukázalo, že pro kompresi DNA není Huffmanův kodér příliš účinný a v praxi se tak používají spíše aritmetické kodéry [23]. Dobrých výsledků mohou tyto algoritmy dosahovat při kódování sekvencí, které jsou převážně kódující nebo převážně nekódující, tedy v takových, kde bude nadbytek GC nebo AT.

Druhou skupinou jsou algoritmy založené na slovníkovém kódování. Komprese je v takovém případě prováděna náhradou delších úseků sekvence určitým kódem. Ten je získán právě porovnáním daného úseku se slovníkem. Velkého úspěchu dosahují tyto metody při kompresi opakujících se motivů. Jsou založeny na myšlence nenáhodné DNA sekvence, kdy je redundantní informace redukována použitím slovníku.

Biocompress, biocompress-2

Jedná se o první algoritmus vytvořený speciálně pro použití na DNA sekvencích [24]. Pracuje na myšlence vyhledávání opakujících se motivů. Pokud je v sekvenci nalezena subsekvence, která se již vyskytla dříve, je zakódována pouze svojí délkou a odkazem na svůj první výskyt. Metoda se tak řadí k metodám slovníkového kódování a její kompresní účinnost velmi závislá na výskytu opakujících se subsekvencí, neboť neopakující se vzory nejsou žádným způsobem komprimovány. Pokud se navíc v sekvenci objevují pouze krátké opakující se úseky, může být výsledná komprese negativní.

Novější verze algoritmu, nazvaná biocompress-2, byla publikována stejnými autory [25]. Tento algoritmus zvyšuje účinnost komprese přidáním druhé podmínky spočívající v použití aritmetického kodéru. Kdykoliv není nalezena opakující se sekvence, je dané místo kódováno na základě četnosti výskytu jednotlivých nukleoidů. Jedná se tedy o kodér nerepetitivních i repetitivních sekvencí.

Cfact

Cfact je slovníkový kodér s garantovanou kompresí [26]. V podstatě je založen na stejné myšlence jako výše zmíněný biocompress, provedení algoritmu je ale jiné. Jedná se o dvoukrokový algoritmus. V prvním kroku je projetím sekvence vytvořen suffix tree. Ten je datovou strukturou, ve které jsou uloženy všechny kombinace možných posloupností nukleotidů v komprimované sekvenci. Teprve v druhém kroku začíná samotné kódování prohlížením stromu. V každém uzlu je totiž kódována informace o maximálním možné počtu repetic v sekvenci na základě počtu větví, které z uzlu vychází. Zkoumáním pozic příslušných repetic je následně zkoumána účinnost komprese. V případě, že by slovníkové kódování pro příslušnou subsekvenci znamenalo expanzi datového souboru je použito standardní 2 bitové

kódování. Díky tomu je garantována vždy pozitivní účinnost komprese. Pro málo repetitivní sekvence je však účinnost tohoto postupu velmi malá.

Genbit

Příkladem čistě statistického algoritmu bez využití dynamického programování je Genbit [23]. I tento algoritmus je založen na myšlence redundance způsobené obsahem repetitivních subsekvencí. Místo zpracování po jednom nukleotidu rozděluje sekvenci do fragmentů o délce 4 nukleotidů. Typů těchto fragmentů může být při výskytu čtyř znaků 4^4, tedy celkově 256. Na kódování tedy stačí 8 bitů (2^8 dává právě 256 kombinací). K těmto 8 bitům se přidává ještě jeden bit, ve kterém je kódována informace o dvou sousedních fragmentech. Pokud jsou fragmenty stejné, je uložen devátý bit jako 1, v jiném případě je kódován 0. Tento algoritmus je poměrně nenáročný díky nepoužití dynamického programování. Může však trpět negativní kompresí.

Gencompress

Dalším ze slovníkových algoritmů je Gencompress [22]. Od algoritmů založených na tomto principu, jenž jsme si zmínili dříve, se ale liší použitím vyhledávání přibližných opakování, zatímco Biocompress a Cfact jsou založeny na detekci exaktních opakování. Díky tomu dosahuje větší kompresní účinnosti, protože opakující se motivy jsou často mutované, ať již pouze v jednom nebo i více nukleotidech. Na začátku je vybrána sekvence, která je porovnávána se zbytkem sekvence. Toto porovnání si lze představit jako globální zarovnání takových úseků. Přitom je sledováno, zda jsou znaky shodné nebo nikoliv, případně jestli došlo k deleci či inzerci znaku. Shodné znaky tak mohou být kódovány pouze jedním bitem. Díky tomu dosahuje dobrých kompresních výsledků.

Využití výše zmíněných kompresních algoritmů pro genomická data je navrženo pro zpracování znakových řetězců. Primárně tedy pro kompresi signálů použitelné není. Nabízí se ovšem možnost úpravy stávajících algoritmů pro použití na signálová data. Toto vyhodnocení si však uvedeme až po objasnění charakteristik genomických signálů. Doplníme i obecný příklad algoritmu pro kompresi signálů, který teď nemá smysl uvádět, neboť signálových reprezentací existuje velké množství a různé algoritmy nemusí být využitelné pro všechny typy signálů té samé sekvence DNA.

3. Konverze dat

V této kapitole se dostaneme k jádru problému zpracování genomických dat pomocí číslicových metod. Tím je konverze genomických sekvencí na genomické signály. Numerických map existuje větší množství a poskytují různé signály. Můžeme je dělit podle různých kritérii.

Degenerativní vs. nedegenerativní

Toto dělení bylo zmíněno již v předchozí kapitole. Při použití degenerativní konverze dat, již signál není možné znovu převést na sekvenci. V tomto případě bude komprese ztrátová (vůči původní znakové sekvenci). Nedegenerativní konverzi pak můžeme využít i pro kompresi bez ztráty informace.

Dimensionalita signálu

Představě číslicového signálu nejvíce odpovídá reprezentace v jedné rovině, tedy ve 2D prostoru. Mnoho numerických metod je založeno právě na dvourozměrných vektorech, které ve výsledku takovéto signály poskytují. Výjimkou nejsou ani signály vyjádřené v prostoru, tedy 3D signály. Možností je i 1D numerická reprezentace. Alternativní formou pak může být převod sekvence do barevného kódu.

Ukázky jednotlivých signálů budou předvedeny na vybraných testovacích sekvencích získaných z databáze GenBank. Jako zástupce mitochodriálních genomů byl vybrán mitochodriální genom lidský s přístupovým kódem NC_012920 (*Homo sapiens* mitochondrion, complete genome.) o délce 16 569 bp. Jako zástupce genomů prokaryotních pak genom *E. coli* kmene K-12 s přístupovým kódem NC_000913 (*Escherichia coli* str. K-12 substr. MG1655, complete genome.) o délce 4 639 675 bp. Pro porovnání je navíc možné najít jednotlivé signály pro lidský 21. chromosom v příloze A, jedná se o sekvenci s přístupovým kódem NC_000021 (*Homo sapiens* chromosome 21, GRCh37.p10 Primary Assembly.) o délce 48 129 895 bp

3.1 H křivka

I když je signálové pojetí genomu věc především 21. století spojená s rozvojem stolních počítačů, první numerické mapy se začaly objevovat s nástupem Sangerovy sekvenace. Takovým příkladem může být H křivka [27]. Jedná se 3D signálové zobrazení DNA, které vzniklo již v roce 1983 kvůli stále narůstající délce známých sekvencí, které již nebyly celé dobře vizualizovatelné v podobě sekvence znaků. Na tehdejší dobu se ovšem jako o dlouhých sekvencích bavíme o posloupnostech čítajících okolo 1000 znaků. Tedy v porovnání s dnešními mnoho miliónů nukleotidů čítajícími sekvencemi se jedná o velice krátké úseky.

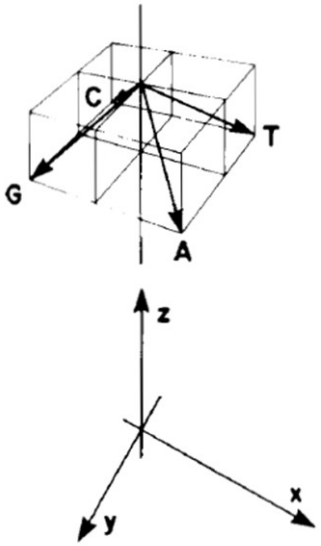
Jednotlivé nukleotidy jsou reprezentovány vektory, které dostaneme složením tří jednotkových vektorů $\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$ mířících ve směrech os x,y, respektive z.

$$\vec{a} = \vec{\imath} + \vec{\jmath} - \vec{k}$$

$$\vec{t} = \vec{\imath} - \vec{\jmath} - \vec{k}$$

$$\vec{c} = -\vec{\imath} - \vec{\jmath} - \vec{k}$$

$$\vec{g} = -\vec{\imath} + \vec{\jmath} - \vec{k} \qquad (3.1)$$



Obr. 3.1: Bázové vektory H křivky [27]

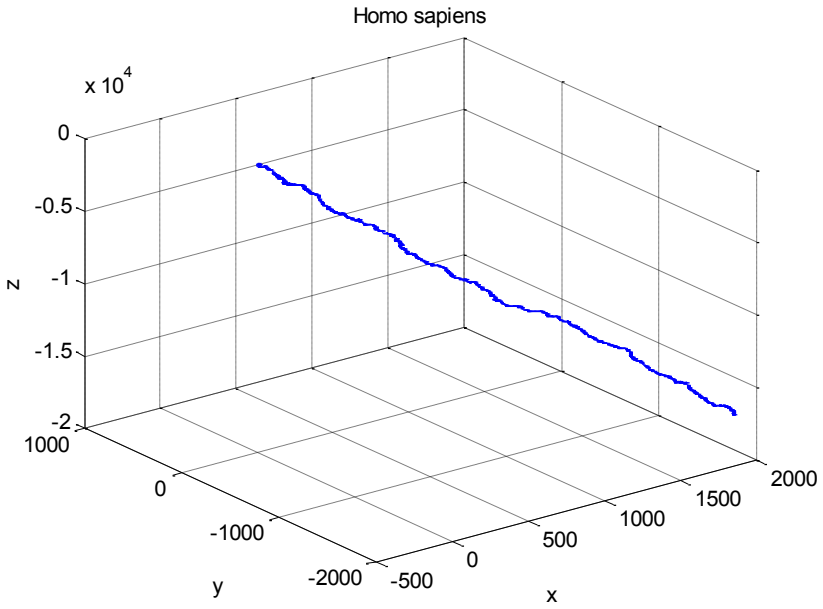
Sumací těchto bázových vektorů podél celé sekvence dostaneme křivku H. Protože $\vec{k}$ je vždy záporné, směr výsledného signálu je v záporném směru osy z. Absolutní hodnota poslední z souřadnice tak udává délku celé sekvence. Zbylé dvě souřadnice posledního prvku nesou také informaci. Hodnota y udává poměr purinů a pyrimidinů, přičemž kladná hodnota znamená nadbytek purinů. Hodnota v ose x pak vyjadřuje poměr slabé a silné vazby, kladná hodnota znamená převahu A/T.

Tyto poznatky lze využít také při pozorování celého signálu. Přední pohled, v rovině x-z, může pomocí v odhadu kódujících a nekódujících oblastí. Píky v tomto pohledu znamenají změnu oblastí bohatých na slabovazebné a silnovazebné nukleotidy. Píky doprava znamenají přechod od GC bohatých k AT bohatým oblastem, píky doleva pak ukazují opačnou situaci. Stejným způsobem v rovině y-z rozlišíme regiony bohaté na puriny nebo pyrimidiny.

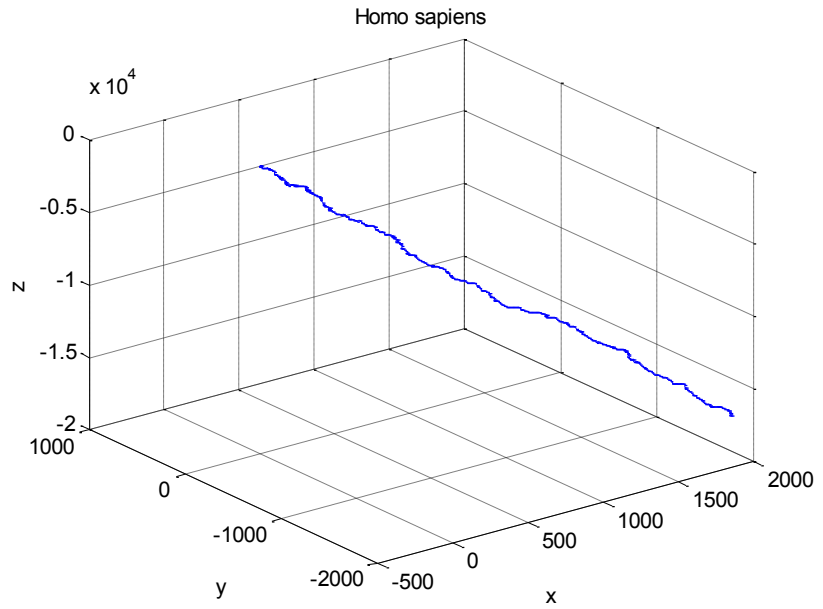
Z předpisu vektorů je jasné, že H křivka je nedegenerativním signálem a může být znovu převedena na znakovou sekvenci. Účel vzniku H křivky byl primárně v dobré vizualizaci dlouhých sekvencí. Ta ale nemusí být optimální v případě delších sekvencí, které chceme zobrazit v jednom okně. Detaily vedou ke tlustší, špatně rozlišitelné křivce. Proto

vznikla ještě vyhlazená forma H křivky. Ta využívá plovoucího okna o předem definované délce, ze kterého podle poměrného zastoupení nukleotidů počítá náhradní vektor. Tato forma signálu ovšem již nesplňuje bezezbytku podmínky pro zpětné mapování na sekvenci znaků, je tak formou mapy degenerativní, ovšem pouze v některých ohledech jak zjistíme níže.

Na následující sérii obrázků můžeme porovnat rozdíly mezi H křivkami. Skripty pro výpočet křivek jsou dostupné v příloze, stejně jako originální H křivka pro 21. lidský chromosom.

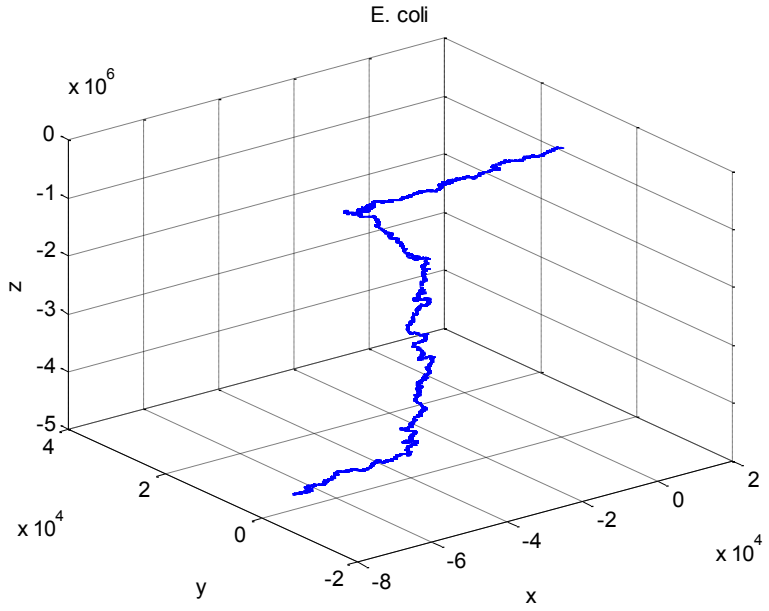


Obr. 3.2: Originální H křivka, *Homo sapiens* mitochondrie



Obr. 3.3: Vyhlazená H křivka, *Homo sapiens* mitochondrie

Pro vyhlazování bylo použito plovoucí okno o délce 19 bp. Originální křivka na Obr. 3.2 působí jako silnější, jelikož se v ní projevují jednotlivé zákmity způsobené změnou nukleotidů mezi jednotlivými sousedícími pozicemi. Vyhlazená křivka na Obr. 3.3 tyto zákmity redukuje, a proto ve velkém měřítku působí užším dojmem. H křivka pro lidskou sekvenci působí oproti sekvenci pro *E. coli* na Obr. 3.4 téměř lineárním dojmem. Jedná se však pouze klam způsobený natočením 3D prostoru grafu.



Obr. 3.4: Originální H křivka, *E. coli*

Rozdíl mezi standardní a vyhlazenou H křivkou je také v náročnosti obou algoritmů. Klasický přístup je algoritmem lineárním $O(n)$, vyhlazování však zavádí v každém cyklu soubor vnořených příkazů, což tento přístup dělá algoritmem v asymptotě kvadratickým $O(n^2)$. Vyhlazení celého prokaryotického chromosomu je proto v praxi neproveditelné. Sekvence je mnohanásobně delší než mitochondriální genom a její výpočet je tak velmi problematický. Výpočetní náročnost vyhlazování je projevem zastaralosti této metody, která byla původně určena pro velmi krátké sekvence.

Při převodu originální H křivky dostaneme podle předpokladu zcela stejnou sekvenci nukleotidů jako před konverzí. Navíc ke zpětnému mapování postačují pouze vzorky v osách x a y, díky ekvidistantnímu vzorkování v ose z. Informace o délce sekvence uložená v poslední hodnotě vzorku v ose z je nadbytečná, neboť může být nahrazena informací délky vektoru souřadnice v ose x, případně ose y.

Zpětná konverze vyhlazené křivky může být provedena také bez nadbytečné osy z. Trasování v osách x a y už je ovšem složitější, neboť jednotlivé vzorky již nenabývají definovaných celočíselných hodnot. Jedná se o různé neceločíselné prvky dané různým sklonem křivky v jednotlivých částech signálu, ten je navíc ještě ovlivňován délkou okna.

Navíc plovoucí okno o délce *w* (liché číslo) způsobí ztrátu *w-1* nukleotidů. Neztracené vzorky signálu lze ovšem zpětně převést porovnáním vždy dvou po sobě jdoucích prvků vektorů nesoucích informaci o x a y souřadnici prvku signálu. Rozdíly kladné lze chápat jako hodnotu 1 a rozdíly záporné jako hodnotu -1. Díky tomuto lze vzorky konvertovat zpět na nukleotidovou sekvenci bez ztráty informace. Nebýt tedy ztráty nukleotidů dané plovoucím oknem, i tato metoda by byla metodou nedegenerativní, což ukazuje velkou robustnost tohoto signálu a prostor pro případné podvzorkování, které by také nemuselo znamenat ztrátu informace.

3.2 Z křivka

Z křivka je podobná předcházející metodě, ovšem numerická mapa byla pozměněna, co se týče výpočtů ve směru osy z. Výstupem je tedy také 3D signál. Reprezentace délky sekvence pomocí hodnoty v ose z, jako je tomu u H křivky, není potřeba, neboť křivka je složena z *N* uzlů, jejichž počet je roven délce symbolické sekvence. Metoda byla vyvinuta již roku 1994 a od této doby je stále vylepšována o nové možnosti analýzy, například Fourierovou transformací [29].

Výpočet Z křivky je dán tzv. Z transformací DNA sekvence. Ta je postupně počítána po jednotlivých uzlech P_0 až P_N, kde N je délka sekvence. Souřadnice uzlů v jednotlivých osách vychází z kumulativního výskytu jednotlivých nukleotidů.

$$\begin{aligned} x_n &= (A_n + G_n) - (C_n + T_n) \\ y_n &= (A_n + C_n) - (G_n + T_n) \\ z_n &= (A_n + T_n) - (G_n + C_n) \\ x_n, y_n, z_n &\in \langle -N, +N \rangle, n = 0, 1, 2, \dots, N \end{aligned} \tag{3.2}$$

Nultý uzel P_0 odpovídá nultému nukleotidu, tedy všechny jeho souřadnice jsou rovny nule. Z křivka tak vždy začíná v počátku souřadné soustavy. Další významnou vlastností je, že suma výskytu jednotlivých typů nukleotidů pro *n*-tý znak je rovna právě *n*, tedy

$$A_n + C_n + G_n + T_n = n. \tag{3.3}$$

Díky tomu může být signál zpětně převeden na sekvenci pomocí inverzní Z transformace, dané vztahem

$$\begin{pmatrix} A_n \\ C_n \\ G_n \\ T_n \end{pmatrix} = \frac{n}{4} \times \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4} \times \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}. \tag{3.4}$$

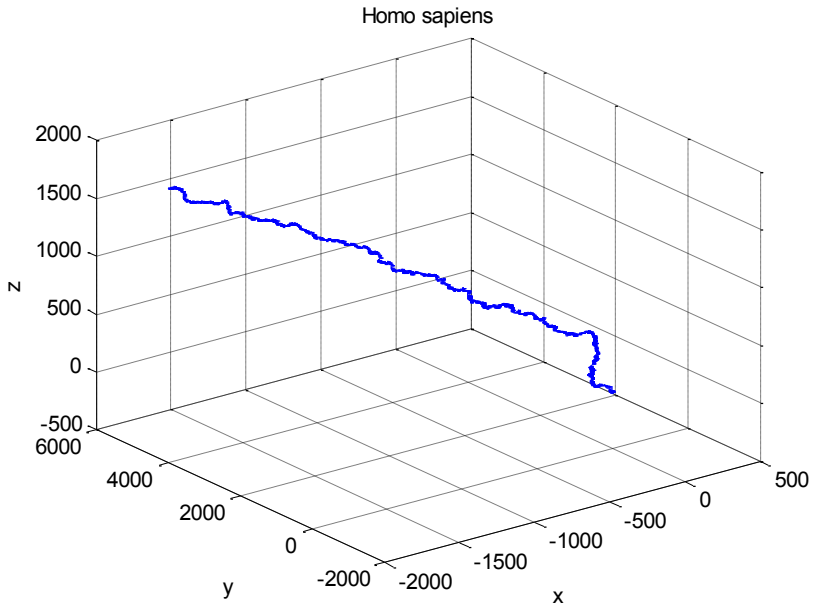
Ze souboru rovnic (3.2) je patrná analogie s předchozí numerickou mapou. Osa x ukazuje výskyt purinů/pyrimidinů. Tuto funkci u H křivky plnila osa y. Osa y u Z křivky

ukazuje poměr amino/keto skupiny, což je informace, kterou předchozí signál postrádal. Osa z pak patří vyjádření slabé/silné vazby, jež v předchozí mapě byla ukryta v ose x. Tento podobný charakter obou metod znamená, že i Z křivka nemusí být pro lidského operátora v základním rozlišení dobře čitelná a je potřeba ji vyhladit. Pro tuto operaci jsou zavedeny v jednotlivých osách B-splajnové funkce.

$$x_n = \frac{1}{6}x_{n-1} + \frac{2}{3}x_n + \frac{1}{6}x_{n+1}$$
$$y_n = \frac{1}{6}y_{n-1} + \frac{2}{3}y_n + \frac{1}{6}y_{n+1}$$
$$z_n = \frac{1}{6}z_{n-1} + \frac{2}{3}z_n + \frac{1}{6}z_{n+1} \tag{3.5}$$

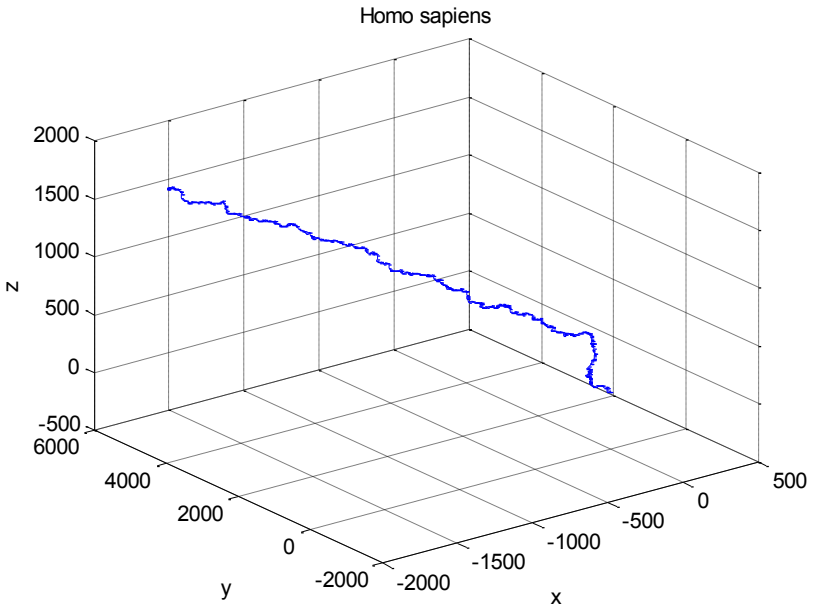
Díky počátku signálu v nultém prvku nedochází ke zkracování sekvence na začátku. Aplikací vyhlazování posledního uzlu pouze na základě jeho pozice a pozice uzlu předcházejícího se můžeme vyvarovat i zkracování sekvence od konce. Celý proces vyhlazování můžeme několikrát opakovat. Pro zobrazování celých lidských chromosomů je typické vyhlazování v řádu tisíců opakování.

Srovnání křivek poskytuje následující série obrázků, přitom signál pro lidskou jadernou DNA lze nalézt v příloze.



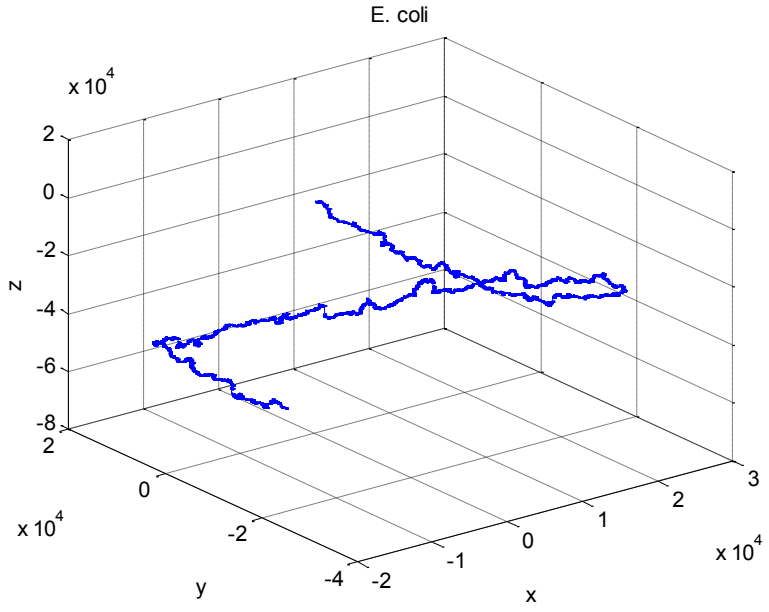
Obr. 3.5: Originální Z křivka, *Homo sapiens* mitochondrie

Z křivky nabývají podstatně odlišných prostorových souřadnic než korespondující H křivky. Nicméně i zde působí nevyhlazené křivky silnějším dojmem jak je patrné na Obr. 3.5 a Obr. 3.7. Vyhlazování zde spočívá ve snižování skoků mezi sousedícími vzorky signálu bez techniky plovoucího okna. Síla vyhlazování je tak dána počtem opakování vyhlazování celé sekvence. Na Obr. 3.6 a Obr. 3.8 jsou původní sekvence vyhlazeny 50×.

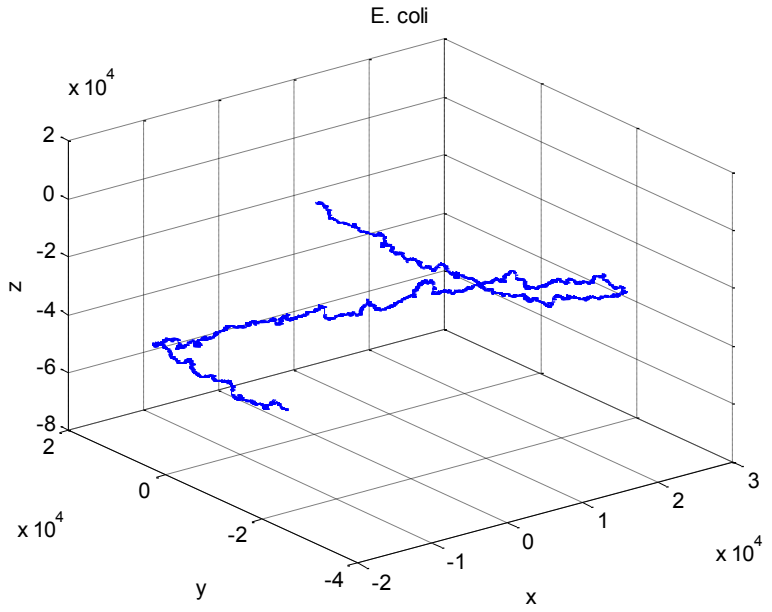


Obr. 3.6: Vyhlazená Z křivka, *Homo sapiens* mitochondrie

Inverzní Z transformace vyhlazeného signálu je ovšem problematická. Dochází k nenávratné ztrátě detailů. Při vyhlazení 50× je po zpětné konverzi výsledná sekvence pouze z 52% shodná se sekvencí původní.



Obr. 3.7: Originální Z křivka, *E. coli*



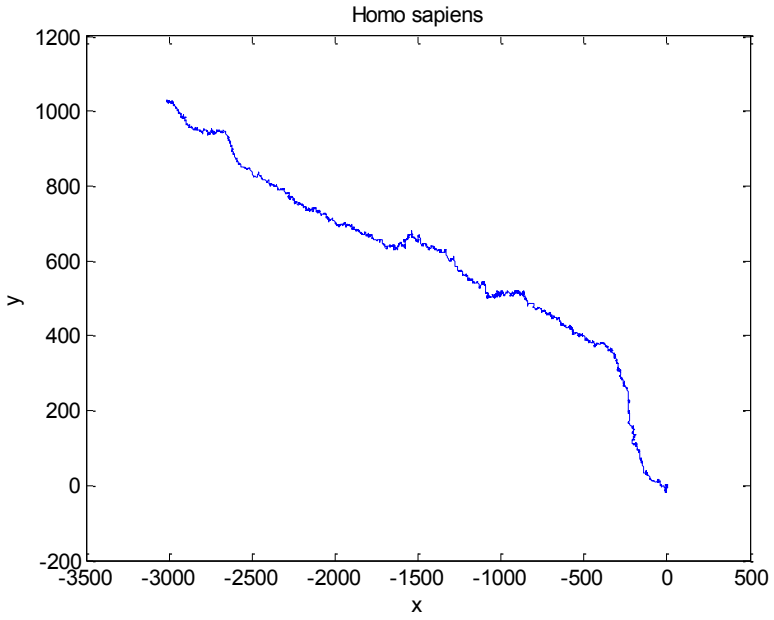
Obr. 3.8: Vyhlazená Z křivka, *E. coli*

Téměř polovina informace je tedy po vyhlazení ztracena. Vyhlazování má za následek vznik nesmyslných homopolymerních úseků. I při desetinásobném vyhlazení dochází ke ztrátě jedné čtvrtiny informace. Signál tak není robustní vůči změnám a je nevhodný k podvzorkování. Na druhou stranu umožňuje Z křivka mnohem hlubší nástroje pro analýzu než starší H křivka. Lze s výhodou počítat GC obsah bez nutnosti využití plovoucího okna po signálu. V kombinaci s Fourierovou transformací je pak tato reprezentace vhodná k označení kódujících částí DNA sekvence.

Co se týče náročnosti konverze, jsou obě verze, jak originální, tak filtrovaná, v asymptotě lineárně náročnými algoritmy.

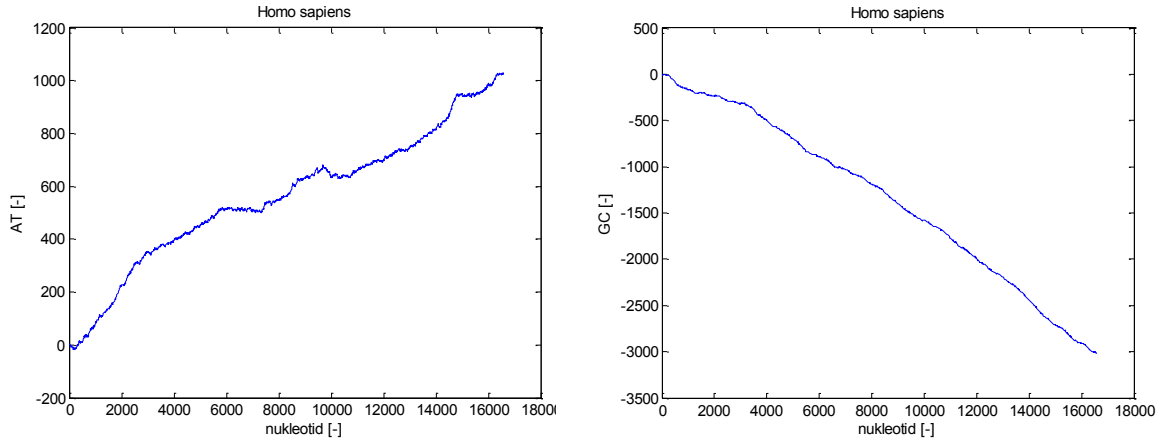
3.3 DNA walk

Pod pojmem DNA walk nalezneme soubor 2D reprezentací nukleotidových sekvencí. Jedná se také o velmi rozpracovanou metodu, neboť je intenzivněji vyvíjena od roku 1996 [30]. Samotné základy těchto vizualizačních metod však byly položeny již roku 1985 [31]. Numerická mapa DNA walk není přesně číselně dána. Je založena na vizualizaci AT podílu v jedné ose a GC podílu v ose druhé. Často se pak setkáváme s tím, že osa x reprezentuje vývoj GC, tedy výskyt cytosinu znamená pohyb po ose vlevo, detekce guaninu pak vpravo. Obdobně v ose y dochází k pohybu nahoru pro adenin a pohybu dolů pro thymin. Nejčastěji je pak používán jednotkový krok. Jelikož jde vývoj signálu do všech čtyř základních směrů, dochází k překryvu signálu sebou samým. Taková signálová reprezentace tak úplně neodpovídá konvenci, na kterou je lidské oko zvyklé při pozorování signálů. I tak je ale metoda vhodná na vizuální porovnávání různých sekvencí. Vizualizace testovacích sekvencí pomocí DNA walk, dle nejčastěji používaných pohybů, jsou na následujících obrázcích.



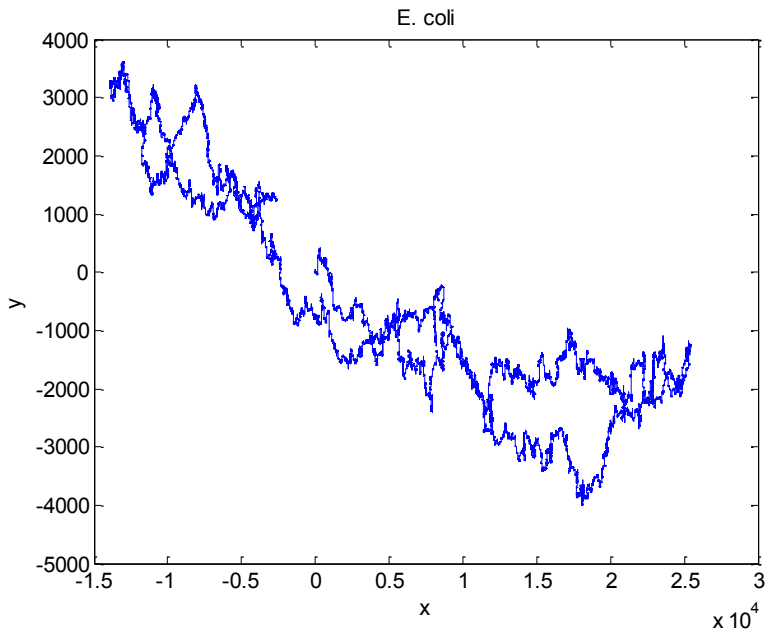
Obr. 3.9: DNA walk, *Homo sapiens* mitochondrie

V signálu lidského mitochondriálního genomu na Obr. 3.9 nejsou výrazné překryvy, křivka signálu je poměrně úzká. Další možností je vizualizace vývoje GC nebo AT podél sekvence zvlášť, tedy vizualizace os x a y nezávisle na sobě [32]. Výrazné píky v signálu vizualizujícím GC složku mohou pomoci v odhalování replikačních míst DNA. Rozklad signálu mitochondriálního genomu je na Obr. 3.10.



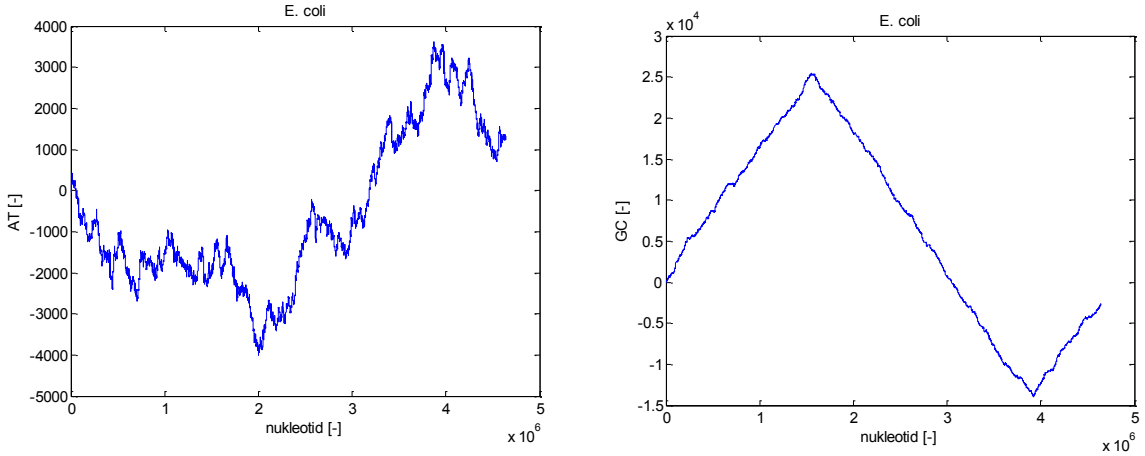
Obr. 3.10: *Homo sapiens* mitochondrie AT signál (vlevo), GC signál (vpravo)

Signály jednotlivých os poskytují o sekvenci zajímavé statistické informace. Z jejich tvarů je zřetelné, že sekvence mitochondriálního genomu obsahuje více adeninů než thyminů, neboť končí pro AT hodnotou 1000. Tedy obsahuje o 1000 adeninů více než thyminů. A obsahuje výrazně více guaninů než cytosinů, zhruba o 3000. Signály 21. chromosomu je opět možné najít v příloze.



Obr. 3.11: DNA walk, *E. coli*

Tvar bakteriálního genomu na Obr. 3.11 se od předchozího velmi liší. Zde jsou již patrné výrazné překryvy signálu a jeho křivka je silnější. Další informace tak znovu přinese až rozpis do jednotlivých os na Obr. 3.12.



Obr. 3.12: *E. coli* AT signál (vlevo), GC signál (vpravo)

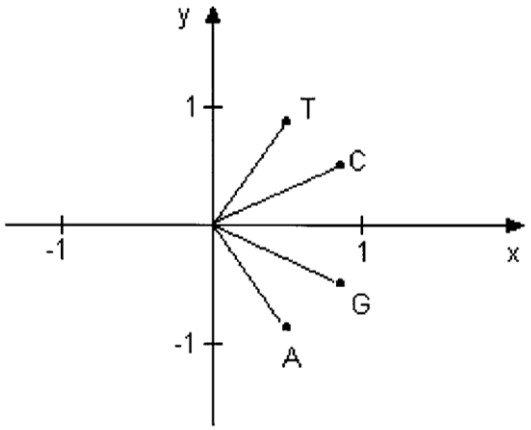
Z rozepsaných signálů je vidět, že sekvence *E. coli* obsahuje asi o 1000 víc adeninů než thyminů, přitom rozložení těchto nukleotidů podél sekvence je celkem vyrovnané, neboť velikost signálu je malá. GC signál nabývá mnohem větších hodnot a končí v nule, tedy počet cytosinů a guaninů je sice shodný, ale jejich rozložení podél sekvence nikoliv. První čtvrtina sekvence je velmi bohatá na cytosin, pak následuje dlouhý úsek s převažujícími guaniny a poslední část sekvence opět obsahuje více cytosinů než guaninů. Tato vlastnost je velice důležitá a může být s výhodou použita pro kompresi, jak si později ukážeme.

Vizualizovaný signál DNA walk je degenerativním vyjádřením sekvence. Kvůli překrytím nelze od oka přesně určit sekvenci nukleotidů. Počítačové zpracování však pracuje po jednotlivých uzlech, tedy s posloupností jednotlivých pohybů signálu. Díky tomu je celý signál zpětně mapovatelný na znakovou sekvenci bez ztráty informace. Celková náročnost algoritmu je definována lineární závislostí a to i při rozepsání signálu do obou os zvlášť.

Výrazné vylepšení ve vizualizaci ve 2D prostoru se povedlo skupince autorů [33] v roce 2003. Tato metoda nemá přesné pojmenování, svým charakterem se však dá zařadit mezi techniky DNA walk. Jedná se o genomický signál vyjádřený pouze ve dvou kvadrantech kartézské soustavy souřadnic. Díky tomu má charakter obecně vnímaného signálu, který se vyvíjí s rostoucí hodnotou v ose x. Celý signál je tvořen skládáním vektorů reprezentujících jednotlivé nukleotidy na sebe. Pro každý nukleotid je tedy definován vektor začínající v počátku souřadnic s koncovým bodem

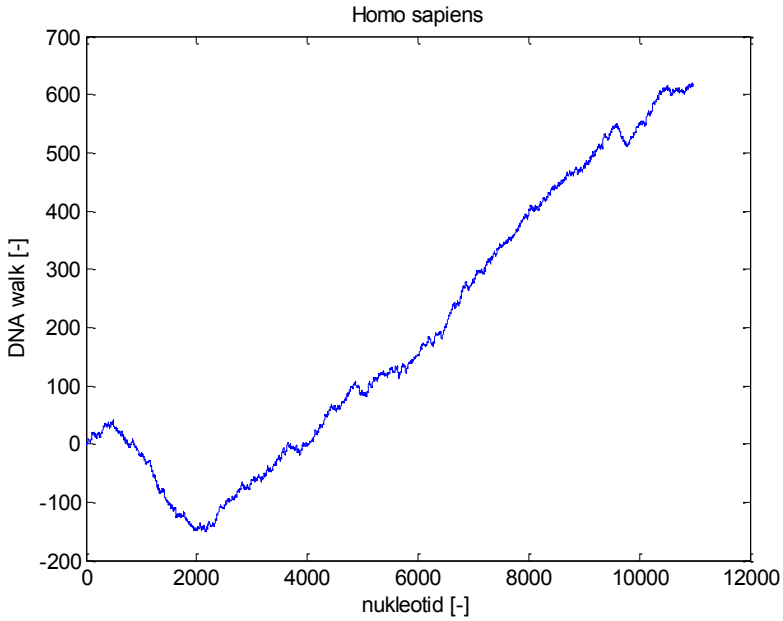
$$\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) \rightarrow A, \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right) \rightarrow C, \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \rightarrow G, \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) \rightarrow T.$$

Z toho vyplývá, že je využit pouze první a čtvrtý kvadrant souřadné soustavy. Přitom 1. kvadrant je využit pro pyrimidiny a 4. pro puriny, jak je ukázáno na Obr. 3.13.

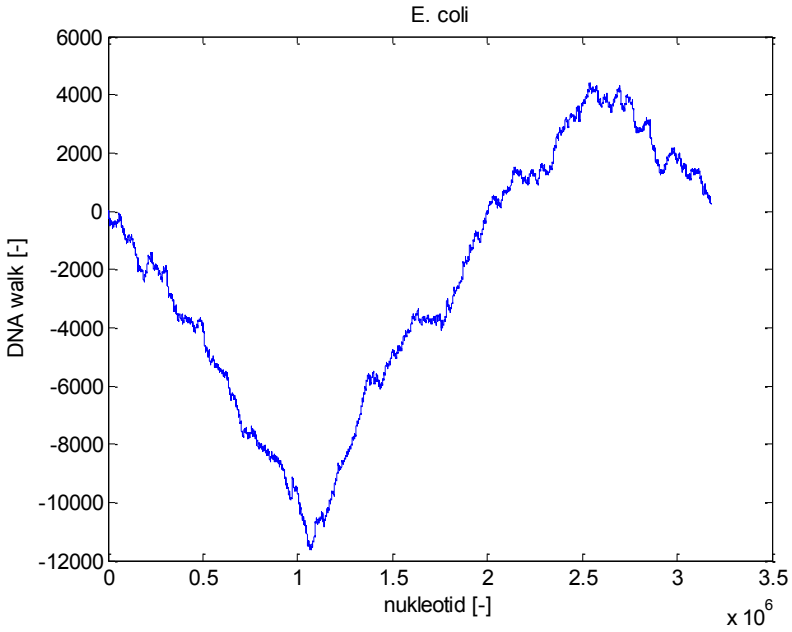


Obr. 3.13: Vizualizace vektorů DNA walk v 1. a 4. kvadrantu [33]

Nedegenerativnost signálu je dána stejným vztahem jako u Z transformace a vychází tak z rovnice (3.3). Zpětnou x a y projekcí tak dostaneme zcela stejnou nukleotidovou sekvenci jako před konverzí. Mitochondriální signál na Obr. 3.14 se žádnému předcházejícímu signálu nepodobá, hlavně kvůli nevyrovnanosti počtů jednotlivých nukleotidů. U prokaryotického chromosomu na Obr. 3.15 se ale projevuje vyrovnané rozložení adeninu s thyminem. Tvar signálu je tak dán především distribucí zbývajících nukleotidů a je tak podobný GC signálu s prohozeným znaménkem u jednotlivých vzorků signálu.



Obr. 3.14: DNA walk *Homo sapiens* mitochondrie, signál v 1. a 4. kvadrantu



Obr. 3.15: DNA walk *E. coli*, signál v 1. a 4. kvadrantu

3.4 Sekvence binárních indikátorů

Sekvence binárních indikátorů je nejpřirozenější formou DNA signálu co se formy vyjádření týče. Vzhledem k tomu, že máme 4 různé nukleotidy, je takový signál reprezentován 4 vektory binárních hodnot, tedy zpravidla nulami a jedničkami. Každý z těchto vektorů je stejně dlouhý jako původní nukleotidová sekvence a mapuje výskyt právě jednoho nukleotidu. Tedy sekvence pro adenin má jedničky na místech, kde se v původní sekvenci adenin

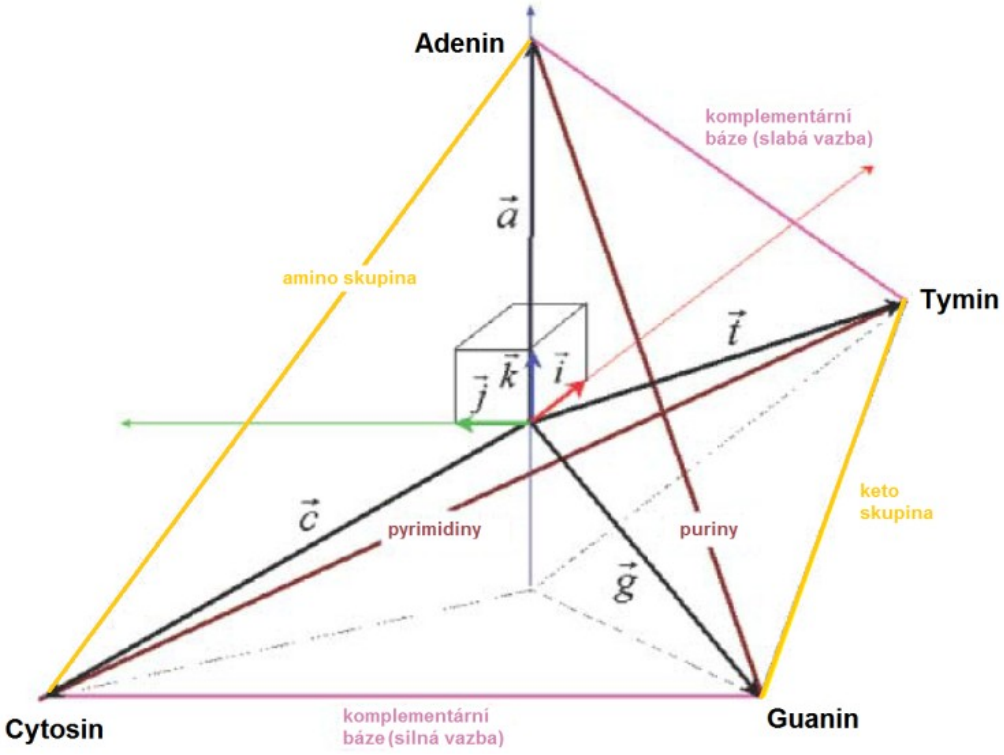
vyskytuje. Na ostatních pozicích obsahuje nuly. Obdobně jsou vytvořeny sekvence pro ostatní nukleotidy [34].

$$u_X(n) = \begin{cases} 1, & s(n) = X \\ 0, & s(n) \neq X \end{cases}, \quad X \epsilon \{A, C, G, T\} \tag{3.6}$$

Metoda je nedegenerativní s lineární náročností. Jejím největším úskalím je právě forma vyjádření. Jedná se totiž o matematicky korektní vyjádření, které je ovšem 4D a tak není možné takový signál vizualizovat. Je sice možné jednotlivé vektory vykreslit odděleně, ale takový signál nejeví žádnou strukturu. Hlavní snahou je tak korektní redukce dimenzionality takového signálu.

3.5 Tetrahedron

K redukci 4D do 3D můžeme a výhodou využít geometrický objekt s vhodnými vlastnostmi, takovým je například čtyřstěn na Obr. 3.16. Jedná se o objekt se 4 vrcholy, což je počet nukleotidů. Jejich položením do vrcholů získáme nukleotidový čtyřstěn. Jednotlivé hrany čtyřstěnu pak navíc nesou informaci o chemické podobnosti nukleotidů, co jsme si definovali v kapitole 1.4 Genomická data.



Obr. 3.16: Nukleotidový tetrahedron [18]

Důležité je natočení nukleotidu v prostoru, které ovlivňuje vyjádření jednotlivých nukleotidů ve třech osách. Položením středu pravoúhlé soustavy souřadnic do těžiště

čtyřstěnu získáme pro každý nukleotid jednotkový vektor. Každý nukleotid pak můžeme matematicky popsat, jak ukazuje soustava následujících rovnic.

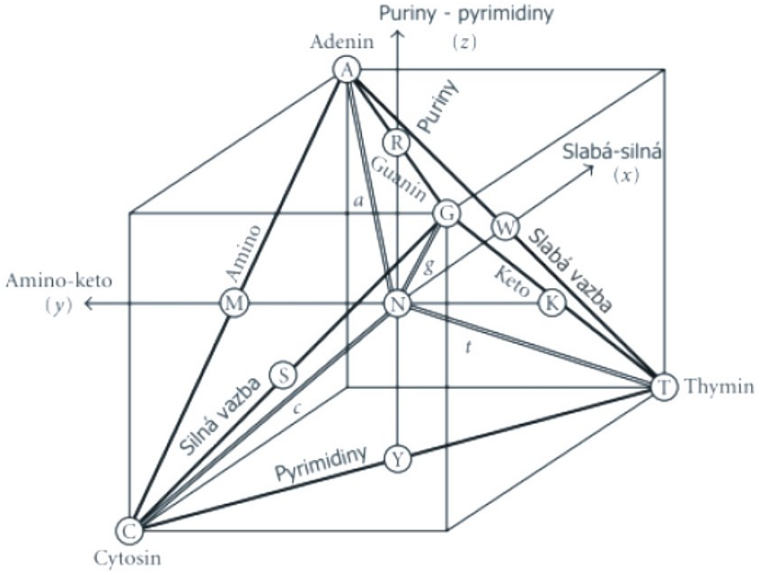
$$\vec{a} = \vec{k}$$

$$\vec{c} = -\frac{2\sqrt{2}}{3}\vec{\imath} + \frac{\sqrt{6}}{3}\vec{\jmath} - \frac{1}{3}\vec{k}$$

$$\vec{g} = -\frac{2\sqrt{2}}{3}\vec{\imath} - \frac{\sqrt{6}}{3}\vec{\jmath} - \frac{1}{3}\vec{k}$$

$$\vec{t} = \frac{2\sqrt{2}}{3}\vec{\imath} - \frac{1}{3}\vec{k} \tag{3.7}$$

Protilehlé strany čtyřstěnu vyjadřují základní chemické vlastnosti, tedy slabou či silnou vazbu, puriny nebo pyrimidiny a amino nebo keto skupinu. Orientace těchto hran v 3D prostoru je ale poměrně nevýhodná. Nukleotidový čtyřstěn ale můžeme s výhodou natáčet v prostoru podle potřeby. Rotace je zpravidla prováděna tak, aby byl čtyřstěn vepsán do jednotkové krychle. Tedy všechny 4 vrcholy tetrahedronu jsou zároveň 4 z 8 vrcholů krychle. Do takové krychle viz Obr. 3.17 pak kromě znaků pro jednotlivé nukleotidy můžeme umístit i ostatní znaky, jak jsme si je definovali v Tab. 1.1.



Obr. 3.17: Tetrahedron v jednotkové krychli [18]

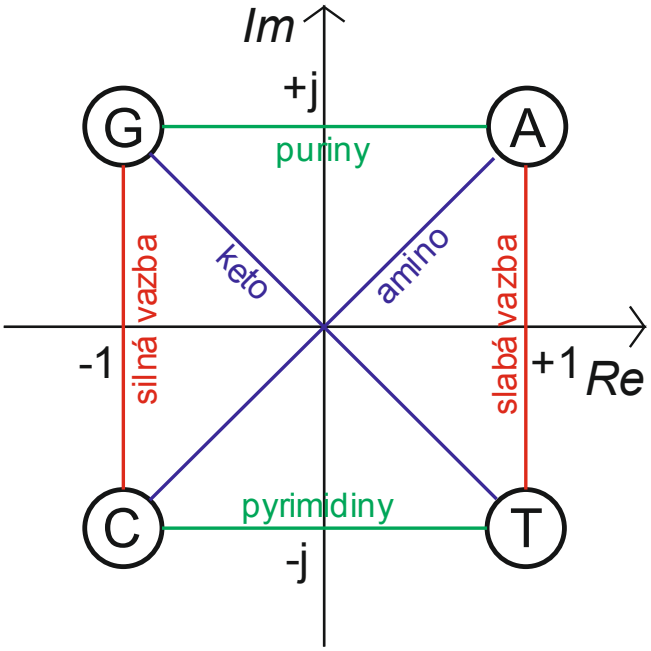
Využití jednotkové krychle je žádoucí i za cenu ztráty normalizační podmínky vektorů v euklidovském prostoru. Různou rotací nukleotidového čtyřstěnu tak v podstatě dostáváme numerické mapy zmíněné výše (H křivka, Z křivka). Jeho projekcí do 2D prostoru pak můžeme získat další numerické mapy, založené na vyjádření pomocí komplexních čísel.

Je také základem pro vizualizaci sekvence pomocí barev s využitím barevného prostoru RGB. Náznak je vidět již z obrázku výše označením jednotkových vektorů barvami.

3.6 Fázové signály

Fázové signály vychází z vizualizace fáze jednotlivých nukleotidů v sekvenci. Fázi mají komplexní čísla, jedná se tedy o numerické mapy založené na převodu znaků na komplexní čísla. Fázovou reprezentaci nukleotidů dostaneme právě výše zmiňovanou projekcí nukleotidového čtyřstěnu do jedné z ploch. Konverze z 3D do 2D prostoru znamená ztrátu části informace. Je přitom pouze na nás, kterou z os vypustíme. Často se využívá vypuštění osy s informací o amino a keto skupině. Dvě osy 2D prostoru pak nesou informaci o chemickém složení (puriny/pyrimidiny) a o komplementaritě bází (silná/slabá vazba). Stále se však pohybujeme v kartézské soustavě v oblasti reálných čísel. Osy prostoru však můžeme zaměnit za reálnou a imaginární osu komplexní roviny, čímž dostaneme výslednou komplexní numerickou mapu.

Obr. 3.18 ukazuje numerickou mapu, která je nadále použita pro vykreslení signálů. Takové signály totiž tvoří charakteristické tvary pro různé sekvence [35], jak si také později ukážeme. V této konvenci imaginární část čísla nese informaci o chemickém složení a reálná o komplementaritě bází. Informace o volných radikálech není ztracena, avšak nepodílí se na tvaru křivky.



Obr. 3.18: Komplexní numerická mapa

Nukleotidy jsou reprezentovány komplexními čísly

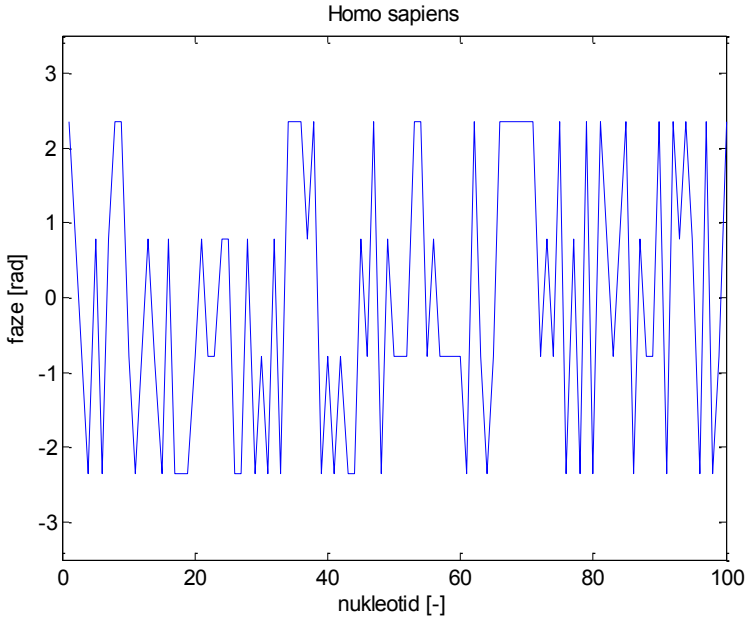
$$a = \ \ 1 + j,$$
$$c = -1 - j,$$
$$g = -1 + j,$$
$$t = \ \ 1 - j.$$

Podle definice fáze komplexních čísel nabývá hodnot $(-\pi, \pi\rangle$ rad. Pomocí goniometrických funkcí snadno dopočteme fázi našich 4 čísel.

$$\{\varphi_a, \varphi_c, \varphi_g, \varphi_t, \} = \left\{\frac{\pi}{4}, -\frac{3\pi}{4}, \frac{3\pi}{4}, -\frac{\pi}{4}\right\}$$

Fáze sekvence

První z možností jak fázi využít pro signálovou reprezentaci je vykreslení fáze po sobě jdoucích nukleotidů. Osa x tak reprezentuje pozici nukleotidu v sekvenci, hodnota v ose y je pak fází daného nukleotidu. V takovém signálu je obor hodnot dán intervalem $\langle -\frac{3\pi}{4}, \frac{3\pi}{4}\rangle$ rad. Pro dlouhé sekvence tak vypadá jako rovná čára. Na Obr. 3.19 si tak ukážeme signál o délce pouze 100 bází.



Obr. 3.19: Fáze sekvence, *Homo sapiens* mitochondrie

Výhodou tohoto signálu je jednoduché zpracování spočívající pouze v přiřazení fáze jednotlivým nukleotidům. Náročností se tak jedná o lineární algoritmus. Rychle a jednoduše může být zpětně konvertován v libovolném místě na znak bez nutnosti znalosti

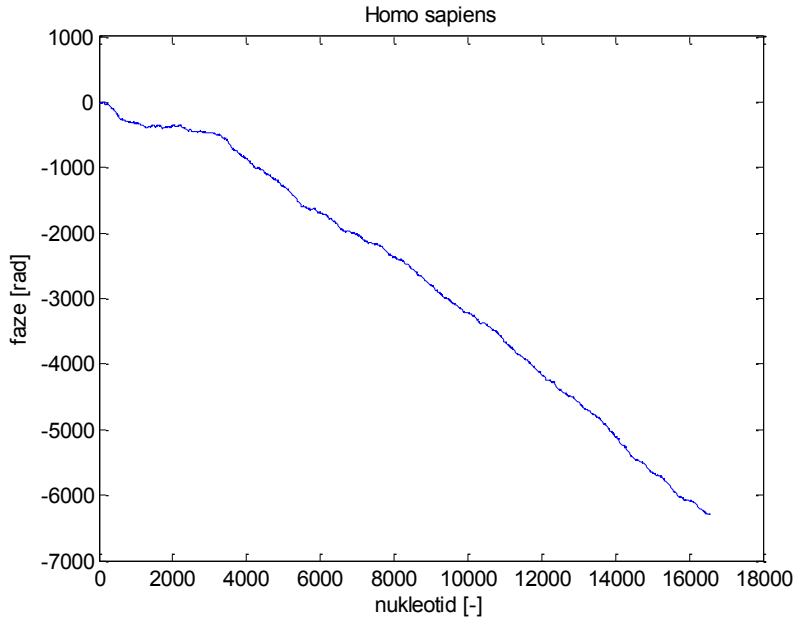
předcházejících nukleotidů. Nevýhodou je, že takový signál v podstatě nepřináší žádnou výhodu a svými vlastnostmi se velmi podobá sekvenci zapsané pomocí znaků.

Kumulovaná fáze

Kumulovaná fáze je suma fází od prvního po aktuální nukleotid. Dá se tedy počítat postupným přičítáním fáze aktuálního nukleotidu k hodnotě v předchozím vzorku signálu. Při znalosti četnosti výskytu jednotlivých nukleotidů n_A, n_C,n_G, n_T do aktuální pozice n lze získat kumulovanou fázi i ze vzorce

$$\theta_C = \frac{\pi}{4}[3(n_G - n_C) + (n_A - n_T)] \tag{3.8}$$

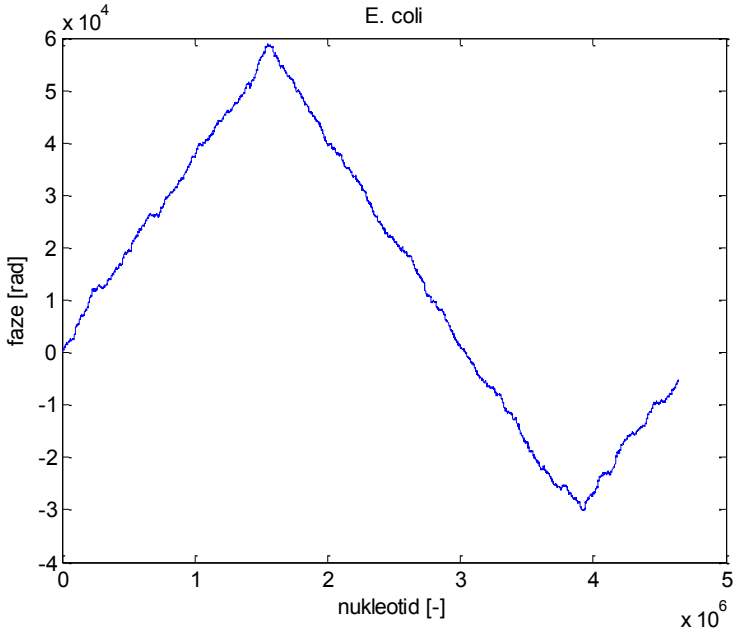
Signály kumulované fáze se u různých typů sekvencí podstatně liší, přitom ale vždy jeví výrazně typický tvar v měřítku délky celé sekvence. Jak již vyplývá ze vzorce (3.8), tento signál vizualizuje aktuální podíl nukleotidů v sekvenci. Tento signál tak končí zase zhruba v nulové fázi podle 2. Chargaffova zákona [36], který říká, že podíl purinů a pyrimidinů v rámci jednoho vlákna DNA má tendenci se vyvažovat. Tento fakt, však neplatí pro mitochondriální genom, jak můžeme pozorovat na Obr. 3.20. Ze signálu je jasné, že takový genom obsahuje více pyrimidinů než purinů, ty mají tendenci se vyvažovat pouze v počáteční části sekvence. Zbytek sekvence má záporný, téměř lineární spád, což je předpoklad pro dobrou kompresi. Celkově je tak tento signál tvarově velmi zajímavý.



Obr. 3.20: Kumulovaná fáze, *Homo sapiens* mitochondrie

Sekvence prokaryotického chromosomu již 2. Chargaffův zákon splňuje, stejně tak i lidská jaderná DNA (viz přílohy). Podíl purinů a pyrimidinů je celkově vyrovnaný, ovšem lokálně nabývá zajímavých hodnot, jak je vidět na signálu pro *E. coli* na Obr. 3.21. Ten

ukazuje, že v sekvenci jsou místa s výrazně převládajícími puriny a místa s výrazně převládajícími pyrimidiny. Zároveň si připomeňme, že signál velmi podobného tvaru, jsme si již ukázali v kapitole zabývající se vizualizací pomocí DNA walk, konkrétně se jednalo o Obr. 3.12 vpravo. Tento signál ovšem nese informaci o výskytu GC podílu. Z toho můžeme vyvodit závěr, že charakteristický podíl purinů/pyrimidinů vedoucí k typickému tvaru signálu je tvořen pouze podílem GC, přičemž podíl AT se uplatí pouze nevýrazně. A protože celkově je podíl nulový, můžeme tvar signálu zachovat i při vypuštění všech adeninů a thyminů ze sekvence, tedy při 50% ztrátové kompresi.



Obr. 3.21: Kumulovaná fáze, *E. coli*

Signály kumulované fáze se od signálů fáze velmi liší a mají mnohem lepší předpoklady pro kompresi a interdruhové porovnávání. Přitom si ponechávají v asymptotě lineární složitost a jejich výpočet je tak velmi rychlý.

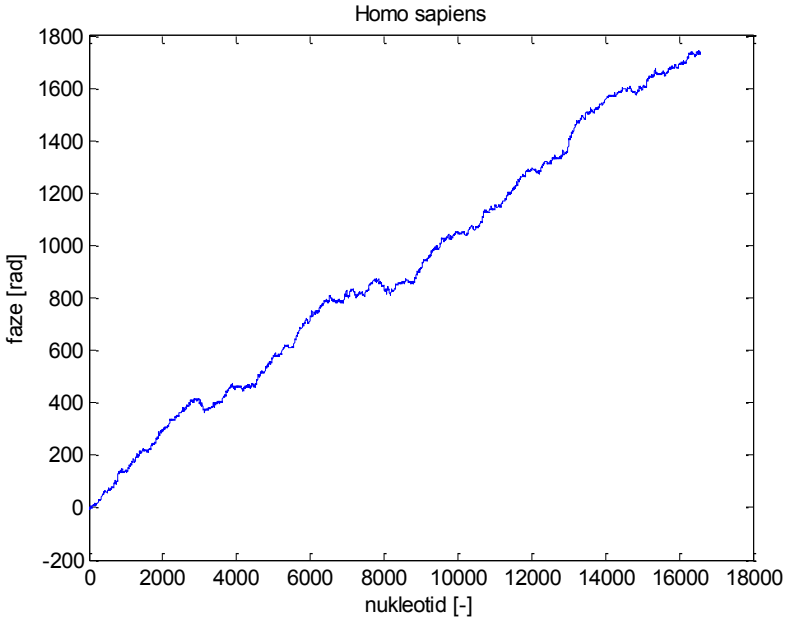
Rozbalená fáze

Rozbalená fáze je korigovanou fází eliminující fázové skoky. Signál je získán na základě fázových posunů mezi jednotlivými nukleotidy. Rozlišujeme tak transice pozitivní A→G, G→C, C→T, T→A znamenající zvětšení fáze o $\pi/2$ a transice negativní A→T, T→C, C→G, G→A znamenající zmenšení fáze o $\pi/2$. A posuny fázově neutrální. Ty jsou navíc dvou druhů. U transic 1. druhu (zachování báze) A↔A, C↔C, G↔G, T↔T je rozdíl fází 0. U transic 2. druhu je roven $\pm\pi$. Přitom musí být pevně rozlišeny transice C→A, T→G jako $+\pi$ a transice A→C, G→T jako $-\pi$.

Signál rozbalené fáze je lineárně rostoucí pro všechny typy sekvencí, přičemž jeho spád je typický pro jednotlivé organismy a chromozómy. Zpětná mapovatelnost na sekvenci znaků není nemožná, ale je problematičtější než u předcházejících dvou fázových signálů.

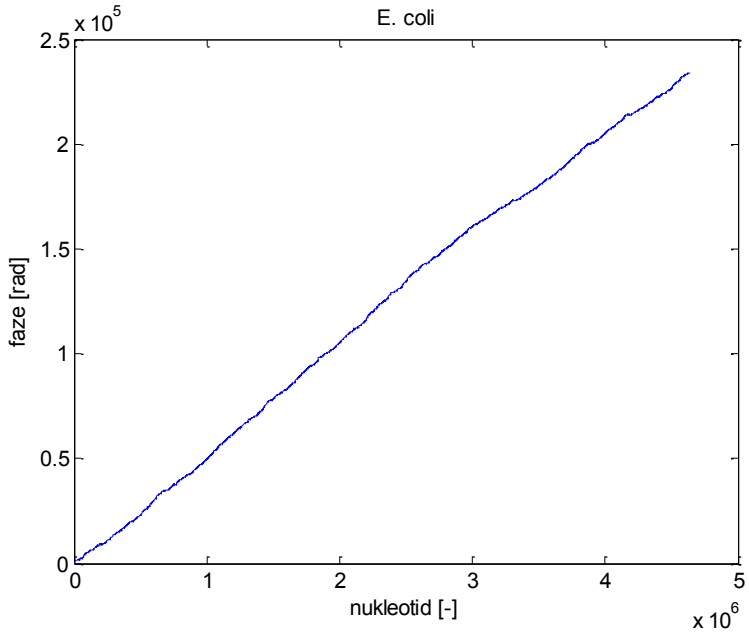
Vychází z předpokladu nulového nultého vzorku, pak může být signál korektně mapován na sekvenci znaků, je ovšem potřeba konvertovat od prvního vzorku. Tato vlastnost je podmíněna odlišným charakterem tohoto signálu. Předcházející fázové signály totiž ukazují informaci přímo o nukleotidech. Rozbalená fáze je postavena na vizualizaci informace o párech nukleotidů [37].

Na Obr. 3.22 můžeme pozorovat signál rozbalené fáze pro naši testovací lidskou mitochondriální sekvenci. Lineární růst dává předpoklad extrémní možnosti komprese až na jedinou hodnotu, kterou by byl spád rozbalené fáze. Při takové kompresi by ale došlo k výrazné ztrátě informace, protože signál není zcela lineární.



Obr. 3.22: Rozbalená fáze, *Homo sapiens* mitochondrie

Ani signál *E. coli* není zcela lineární, byť tak na první pohled může působit. Musíme si ovšem uvědomit, že jeho délka je mnohanásobně větší a detaily nelinearit tak ve velkém měřítku nejsou výrazné. Využitelnost maximální komprese na jedinou hodnotu v podobě spádu můžeme vyloučit, vzhledem k obrovskému počtu druhů prokaryotních organismů, která by vyžadovala jednoznačně nedosažitelnou přesnost určení spádu.



Obr. 3.23: Rozbalená fáze, *E. coli*

Fázové signály, i když stojí na stejném základu, se velmi liší. Prostá fáze je založena na informaci o nukleotidech a nezobrazuje žádnou statistickou informaci, kromě samotné délky signálu. Oproti tomu jsou signály kumulované a rozbalené fáze statistickými obrazy nukleotidů. Rozbalená fáze obrazem o párech nukleotidů a kumulovaná přímo statistickým obrazem nukleotidů. Proto je ze všech tří fázových signálů nejzajímavější.

3.7 Reprezentace barevným kódováním

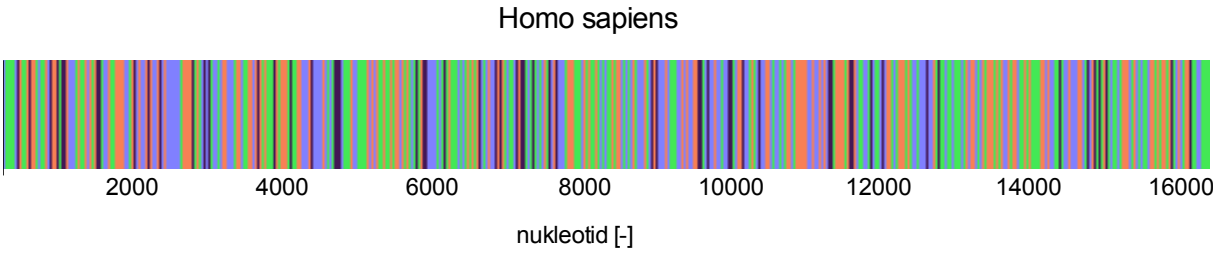
Barevné kódování dostaneme převodem 4D vyjádření do 3D prostoru tvořeného z os RGB. Tedy jedná se o další formu převodu sekvencí binárních indikátorů do prostoru, který je pro člověka lépe pochopitelný [38]. Svou úrovní redukce dimenzionality pro počítačové zpracování se staví na úroveň nukleotidového čtyřstěnu, pro vizualizaci je však redukován na jednodimenzionální, kdy má každá pozice v sekvenci právě jednu výslednou barvu.

$$x_r(n) = \frac{\sqrt{2}}{3}(2u_T(n) - u_C(n) - u_G(n))$$

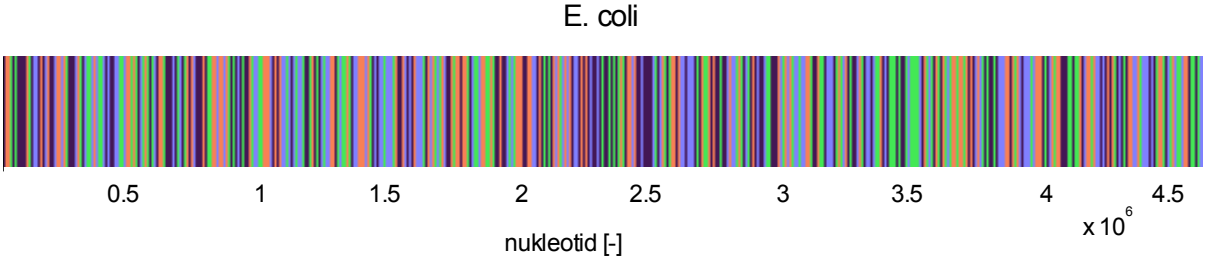
$$x_g(n) = \frac{\sqrt{6}}{3}(u_C(n) - u_G(n))$$

$$x_b(n) = \frac{1}{3}(3u_A(n) - u_T(n) - u_C(n) - u_G(n)) \qquad (3.9)$$

Barevně kódované testovací sekvence jsou vyobrazeny na Obr. 3.24 a Obr. 3.25.



Obr. 3.24: RGB reprezentace, *Homo sapiens* mitochondrie



Obr. 3.25: RGB reprezentace, *E. coli*

Byť je tato reprezentace jednoduše zpracovatelná ve formě obrázku a převod je rychlý, zatížený pouze lineární složitostí, je důležité si uvědomit, že každý nukleotid je reprezentován vektorem 3 hodnot. Paměťové nároky jsou tedy výrazně vyšší než například u fázových signálů.

4. Komprese genomických signálů

Ve čtvrté kapitole této práce se dostaneme k jádru problematiky. Na základě informací z kapitoly předchozí zvolíme signál nejvhodnější ke kompresi takový, který zároveň dokáže rozlišit rozdíly genomických sekvencí mezi jednotlivými druhy organismů. Zároveň si objasníme základní teorii použitých technik při zpracování testovaných signálů.

4.1 Kumulovaná fáze

Z rozboru jednotlivých numerických map v předchozí kapitole je patrné, že se signály prokaryot a mitochondrií sestrojených na základě stejné numerické mapy často tvarově liší. Zpravidla zcela odlišné jsou pak signály eukaryotních jaderných sekvencí, které je možné nalézt v příloze. Pokud ale vyjdeme z předpokladů o podobnosti prokaryotního kruhového chromosomu a kruhové mitochondriální DNA probraných v první kapitole, můžeme vyslovit myšlenku, že tyto signály mají podobné nebo stejné vlastnosti. Rozdíl ve tvaru signálů je pak způsoben rozdílem v délkách, kdy se vlastnosti mitochondriálních neprojeví tolik jako vlastnosti prokaryotních, jejichž délka nabývá o 3 řády vyšších hodnot. Při výběru vhodného signálu se tedy zaměříme na vizualizace delších prokaryotních sekvencí a budeme předpokládat, že podobné vlastnosti pak budou mít i kratší mitochondriální.

Vzhledem k tomu, že v následujícím textu již plně opustíme znakové sekvence a začneme se zabývat analýzou signálů, je potřeba změnit některé používané termíny. V sekvenci znaků mluvíme o posloupnosti nukleotidů. Signál je ovšem tvořen posloupností hodnot, neboli vzorků. Je nutné si uvědomit, že pojmy vzorek a nukleotid jsou ekvivalentní pouze do doby, než je signál komprimován a jeho počet vzorků se tak sníží.

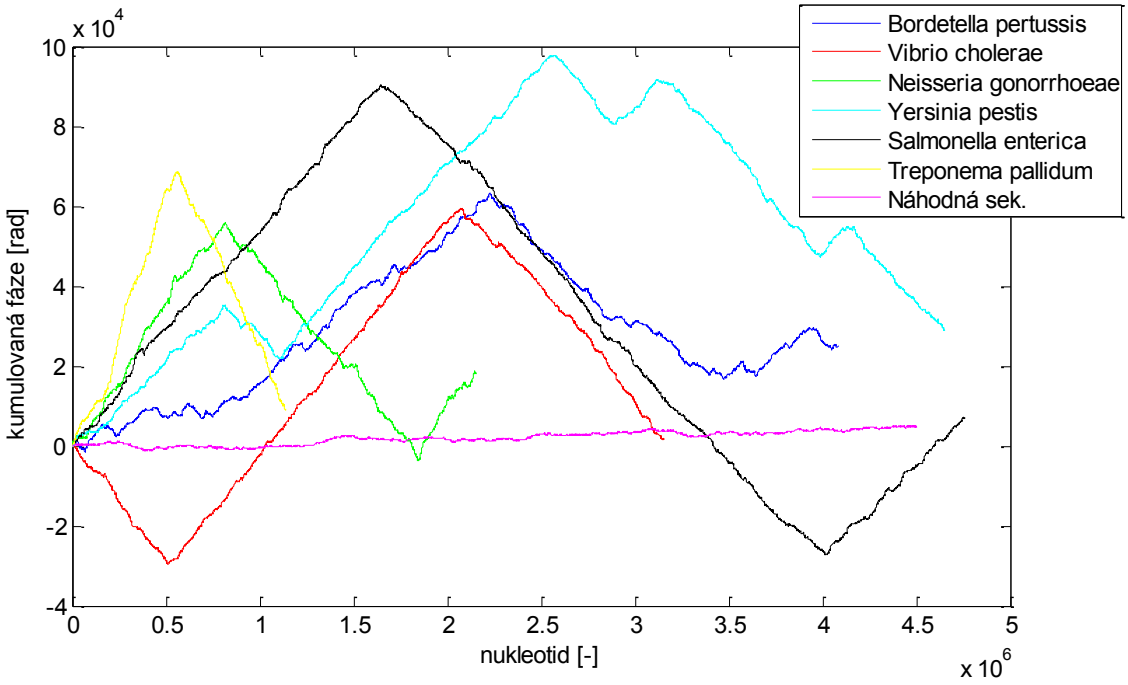
Výběr signálu

Z výběru vyloučíme numerické mapy definované ve 3D prostoru. Takové signály sice mohou dobře rozlišovat jednotlivé organismy, nicméně jejich komprese a zpracování je náročné právě kvůli třem parametrům pro každý nukleotid. Do takových signálů, kromě H a Z křivek, patří i reprezentace v RGB prostoru. Ve 2D prostoru pak můžeme stejným způsobem vyloučit signál DNA walk, neboť popisuje každý nukleotid dvěma parametry. Zbylé signály totiž kódují ve vodorovné ose pořadí nukleotidů v sekvenci a každý nukleotid je tak popsán pouze jedním parametrem. Vhodný signál navíc musí být něčím charakteristický v měřítku celé sekvence. Proto můžeme rovnou vyloučit signál prosté fáze a rozbalené fáze, které jsou ve velkém měřítku lineární a tedy málo charakteristické pro různé organismy. Na výběr zůstává signál DNA walk v 1. a 4. kvadrantu a signál kumulované fáze.

Jako vhodnější se jeví signál kumulované fáze a to hned z několika důvodů. Prvním důvodem je do značné míry subjektivní názor, že signál kumulované fáze působí uhlazenějším dojmem. Protože genomické signály vznikly především kvůli vizualizaci

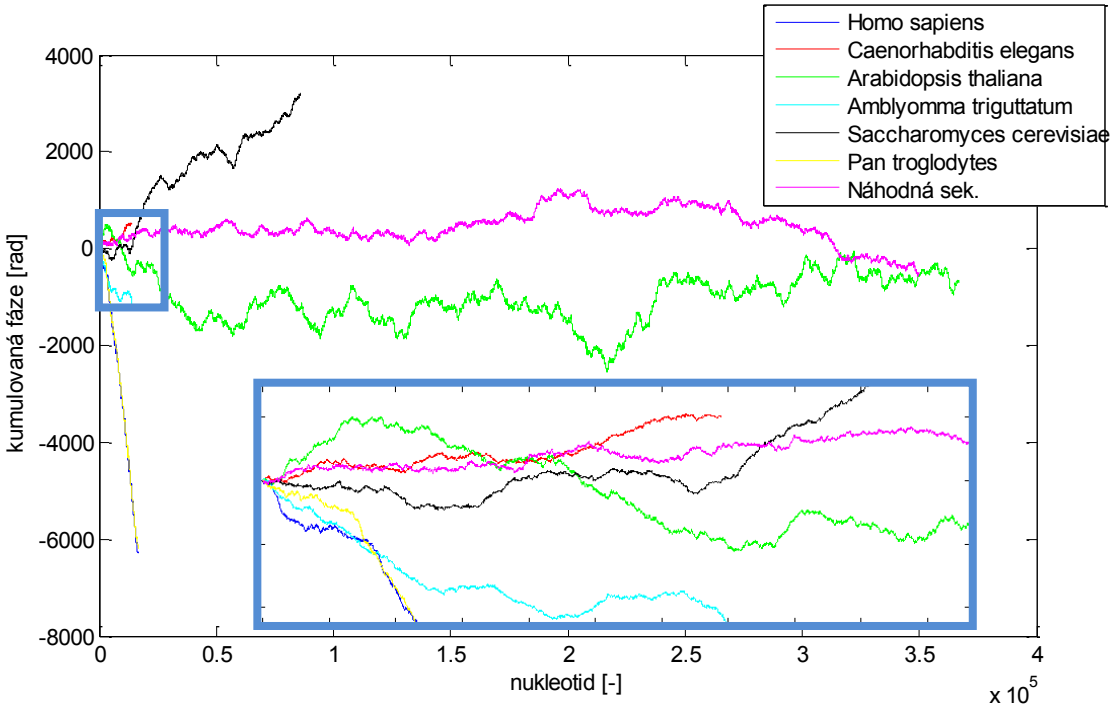
pro lidské oko, můžeme tento důvod považovat do jisté míry za objektivní. Signál kumulované fáze má také mnohem lépe definované numerické hodnoty jednotlivých nukleotidů. Ty jsou založeny na chemické podobnosti jednotlivých nukleotidů. Vzhledem k tomu, že při matematickém zpracování a porovnávání signálů bude velmi záležet na číselných rozdílech mezi příslušejícími si vzorky, je tato vlastnost nesmírně důležitá. DNA walk v 1. a 4. kvadrantu takto silně obhájené numerické hodnoty nemá. Třetím důvodem je velmi dobrý předpoklad kumulované fáze k úspěšné kompresi. Ten lze odvodit z podobnosti tvaru kumulované fáze a GC profilu sekvence, který jsme měli možnost porovnat na Obr. 3.21 a Obr. 3.12. Z těchto grafů vyplývá, že tvar kumulované fáze je tvořen především rozložením cytosinu a guaninu podél sekvence, přičemž rozložení adeninu a thyminu tvoří v signálu stejnosměrnou nebo nízkofrekvenční složku. Vzhledem k tomu, že podíl AT ku GC je zhruba 50:50 máme předpoklad k dobré minimálně 50% kompresi.

Rozdílnost kumulovaných fází pro různé prokaryotní organismy si můžeme ukázat na Obr. 4.1. Ten ukazuje signály vybraných bakterií. Mezi signály jsou na první pohled vidět velké rozdíly, tedy takové signály jsou bezpochyby vhodné ke klasifikaci jednotlivých organismů. Obrázek poskytuje důležité informace. Především potvrzuje platnost 2. Chargaffova zákona. A přestože prokaryota zahrnují pouze 2 domény organismů [39] a jejich fylogenetika na vyšší úrovni je dlouho a jasně daná, délky genomů různých organismů se velmi liší. Na obrázku je také signál náhodné sekvence, která vznikla uměle opakovaným náhodným výběrem z množiny znaků {A ,C ,G ,T}. Oproti reálným signálům však takový signál ve velkém měřítku netvoří žádnou výraznější strukturu. To dokazuje, že tvar signálů kumulované fáze není náhodný.



Obr. 4.1: Kumulované fáze různých prokaryot

Klasifikaci eukaryot jsme si zjednodušili použitím mitochondriální DNA, jejíž sekvence je pro každý druh typická a odstraňuje problém s jadernou DNA, která je u eukaryot členěná do různého počtu chromosomů. I tak je klasifikace eukaryot stále velmi složitá, jelikož se jedná o fylogeneticky mnohem širší skupinu velmi odlišných organismů. Rozdíly mezi 3 základními říšemi hub, rostlin a živočichů může poskytnout Obr. 4.2, který ukazuje, že mitochondriální genom rostlin (*Aradopsis thaliana*) je delší než genom hub (*Saccharomyces cerevisiae*) a genom živočichů (*Homo sapiens*).

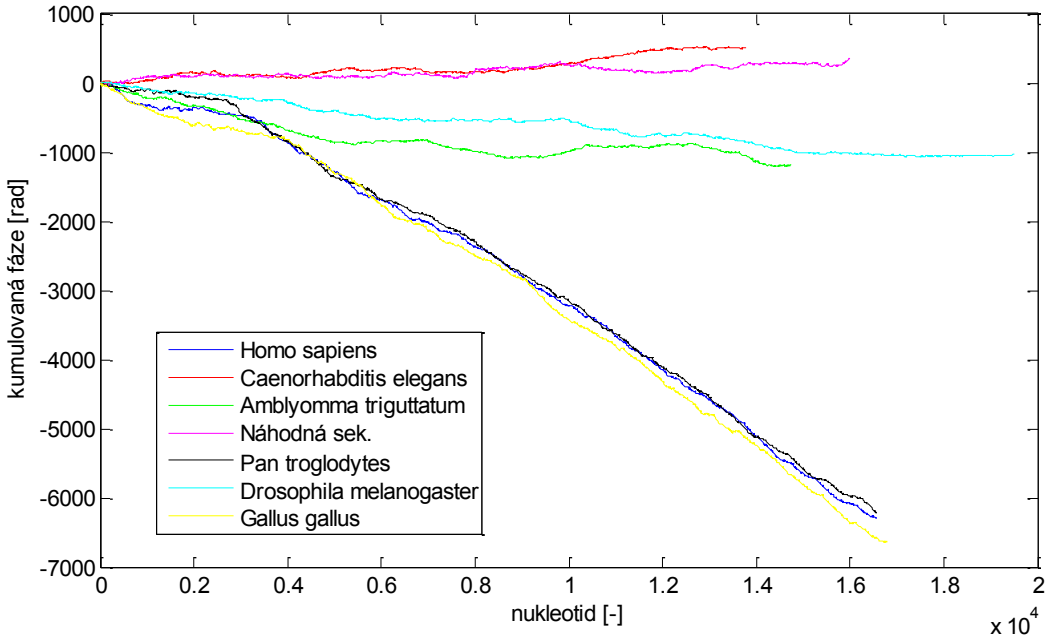


Obr. 4.2: Kumulované fáze různých eukaryot

2. Chargaffův zákon plní pouze sekvence DNA z mitochondrií rostlin. Zbylé 2 říše mají příliš krátké mitochondriální genomy. Samozřejmě klasifikace huseníčku s člověkem v jedné analýze je otázkou nanejvýš nepravděpodobnou. Délkou je signál člověka mnohem bližší kvasince patřící do říše hub. To také odpovídá novějšímu třídění eukaryot na 6 infratříd, kdy oba tyto organismy patří do skupiny *Opisthokonta* [40]. Také zde se všechny reálné signály již od pohledu výrazně liší od signálu náhodné sekvence.

Pro jednodušší taxonomii se zaměříme na klasifikaci živočišné říše, která by měla být díky nejkratší délce sekvencí nejproblematičtější. Pokud přesto bude klasifikace dostatečně robustním nástrojem, bude fungovat i na delší sekvence. V použitém měřítku na Obr. 4.2 se jeví signál kumulované fáze pro klasifikaci živočichů nevhodně, neboť se signály překrývají. Z bližšího pohledu na Obr. 4.3 je vidět, že signály obratlovců jsou si velmi podobné. Možnosti komprese to omezuje, neboť u takových signálů bude požadavek na zachování mnohem větší míry původní informace. Další problém pak odhaluje signál háďátka (*Caenorhabditis elegans*), jehož sekvence se velmi podobá sekvenci náhodné. V molekulární

biologii je háďátko často používaný modelový organismus, protože se jedná o nejjednoduššího vícebuněčného živočicha. Klasifikace vývojově nižších eukaryotních živočichů na základě kumulované fáze tak může být po kompresi problematická.



Obr. 4.3: Kumulované fáze vybraných živočichů

4.2 Analýza signálu

Základem správného zpracování signálu je jeho dostatečně podrobná analýza, na základě které zvolíme ideální metodu dalšího zpracování. Pro účely analýzy se tak vraťme zpět k našim testovacím organismům *E. coli* a *H. sapiens.*

Transformace signálu

Signál kumulované fáze jsme vybrali kvůli jeho výše zmíněným vlastnostem. Těchto vlastností nabývá signál v originální oblasti. Také jsme díky signálům GC a AT profilu odhadli jakých vlastností by mohl signál nabývat v oblasti frekvenční. Pro potvrzení těchto vlastností a získání dalších vlastností je potřeba signál transformovat. Obecně je cílem transformace signálu z originální oblasti do určité obrazové oblasti získání jiného pohledu na signál, snaha o získání dalších informací, které nejsou v originální oblasti rozpoznatelné.

Lineární spojitou transformaci signálu *f(t)* můžeme vyjádřit pomocí integrálu [28]

$$\langle f(t), \psi(t) \rangle = \int_{-\infty}^{+\infty} f(t)\psi^*(t)dt \tag{4.1}$$

kde $\psi(t)$ patří mezi tzv. bázové funkce určující typ transformace.

Taková transformace je definována pro spojité signály s parametrem *t*, tedy spojitým časem. Genomické signály ovšem spojité nejsou. Navíc nejsou definované časovým vývojem, ale vývojem podél délky znakové sekvence. Již dlouholetým standardem je naštěstí číslicové zpracování signálů, které pracuje se signály vzorkovanými. Takové signály nemají spojitý čas, nástroje pro jejich zpracování jsou tak použitelné i pro genomické signály. Obecnou spojitou integrální transformaci, tak můžeme nahradit diskrétní transformací pomocí řady

$$\langle f(n), \psi(n) \rangle = \sum_{-\infty}^{+\infty} f(n)\psi(n) \tag{4.2}$$

Spektrum

Signály chápeme jako aditivní směs (obecně nekonečného počtu) harmonických složek. Převod signálu na zobrazení jeho harmonických složek je zprostředkován Fourierovou transformací a označujeme jej jako spektrum signálu. Bázová funkce spojité Fourierovy transformace (FT)

$$\psi(t) = e^{-j\omega t} \tag{4.3}$$

závisí na parametru ω, jenž se mění s kmitočtem a na čase *t*, který podmiňuje, že je takové spektrum neperiodické.

Bázová funkce pro Fourierovu transformaci diskrétního signálu (DTFT, discrete-time Fourier transform) [41] je také transformací frekvence ω

$$\psi(n) = e^{-j\omega nT} \tag{4.4}$$

která se ale opakuje s periodou $2\pi/T$ označovanou jako vzorkovací frekvence f_{vz}. Spektrum diskrétního signálu, a tedy i genomického signálu, je proto vždy periodické. Fourierovu transformaci s diskrétním časem, pak definujeme jako

$$DTFT\{f(n)\} = F(\omega) = \sum_{-\infty}^{+\infty} f(n)\, e^{-j\omega nT} \tag{4.5}$$

Vzorkovací frekvence je podmíněná charakteristikami signálu a je volena tak, aby byl splněn Nyquistův teorém

$$f_{vz} \geq 2f_{max}, \tag{4.6}$$

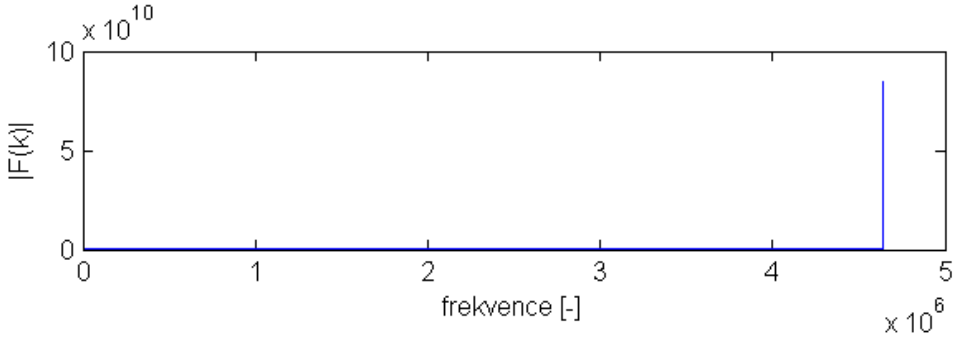
kde f_{max} značí nejvyšší frekvenci v signálu. V opačném případě dojde k překrytí period spektra signálu vedoucí k aliasingu a znehodnocení signálu. Zde se genomické signály liší od ostatních biologických signálů. Například při akvizici signálu EKG víme, jaká jsou zajímavá frekvenční pásma a vzorkovací frekvenci podle toho zvolíme. Všechny signály tak mají stejnou vzorkovací frekvenci. Genomický signál získáme konverzí z genomické sekvence, přičemž vzorkovací frekvenci nemáme dopředu možnost ovlivnit žádným

antialiasingovým filtrem a je tak vždy rovna délce signálu, neboli počtu nukleotidů sekvence. Tato úvaha je při výpočtu spektra genomického signálu velmi důležitá a hned si ukážeme proč.

Genomické signály jsou neperiodické. Přitom spektrum neperiodických signálů je spojité, jak vyplývá z definice DTFT. Pokud si ale uvědomíme, že signál kumulované fáze sekvence DNA není definovaný na intervalu $(+\infty, -\infty)$, ale pouze na intervalu $\langle 1, N \rangle$, kde N je počet nukleotidů sekvence DNA, pak můžeme tento interval brát jako jednu periodu signálu, který by se jinak periodicky opakoval na intervalu $(+\infty, -\infty)$. Díky tomu můžeme kmitočtovou osu rozdělit na N stejných dílů $\Omega = 2\pi/NT$ a spektrum vypočítat pomocí diskrétní Fourierovy transformace (DFT), která posloupnosti N vzorků v originální oblasti přiřazuje stejně dlouhou posloupnost ve frekvenční oblasti

$$DFT\{f(n)\} = F(k) = \sum_{n=1}^{N} f(n)\, e^{-jk\Omega nT} \tag{4.7}$$

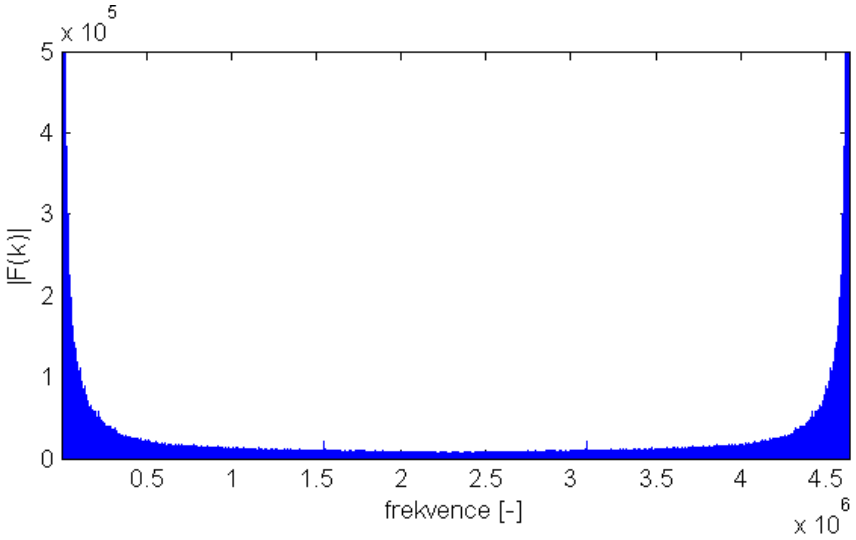
Nyní jsme již schopni vypočítat spektra našich dvou testovacích signálů. Spektrum prokaryotického chromosomu E. coli je ukázáno na Obr. 4.4. Délka této sekvence je 4 639 675 bp, proto f_{vz}=4 639 675. Tato vzorkovací frekvence je několikanásobně vyšší než vzorkovací frekvence jiných biologických signálů, jako EKG, EEG apod. a je bezrozměrná, protože signál v originální oblasti není definován v čase, ale podél sekvence. Takové spektrum tedy vyjadřuje změny nukleotidů podél sekvence.



Obr. 4.4: Spektrum kumulované fáze *E. coli*

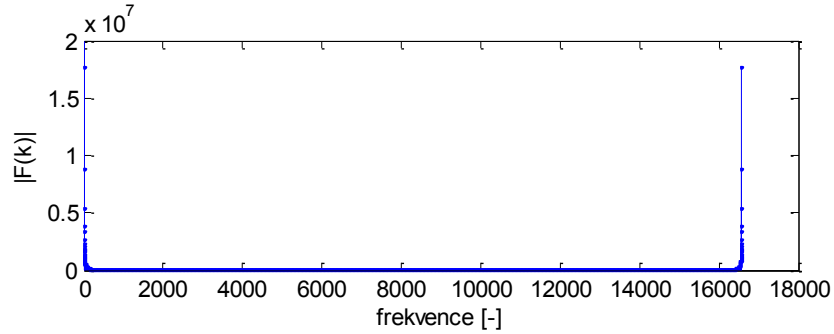
V zobrazeném měřítku vypadá spektrum jako by bylo tvořeno pouze dvěma spektrálními čarami na 0 a f_{vz} (kvůli symetrii spektra). To by samozřejmě svědčilo o tom, že signál může být vyjádřen pouze jedním vzorkem. Toto zdání je ale způsobeno nevhodným měřítkem, kdy velikost těchto spektrálních čar je o několik řádů vyšší než těch zbylých. Zajímavostí a důkazem, že ve spektru je toho mnohem víc může být samotný obrázek, který oproti ostatním grafům působí rozmazanějším dojmem. Tento graf musel být pro zobrazení v textu převeden z vektorového na rastrový a tak komprimován, neboť jeho původní velikost

byla 380 MB a nebyl tak zpracovatelný. Detailnější pohled tak může nabídnout přiblížené spektrum na Obr. 4.5, kde je vertikální osa oříznuta o nejvyšší hodnoty. Velikost nejvýraznějších spektrálních čar je až $9 \cdot 10^{10}$ (viz Obr. 4.4) a překračuje tak zobrazený limit.



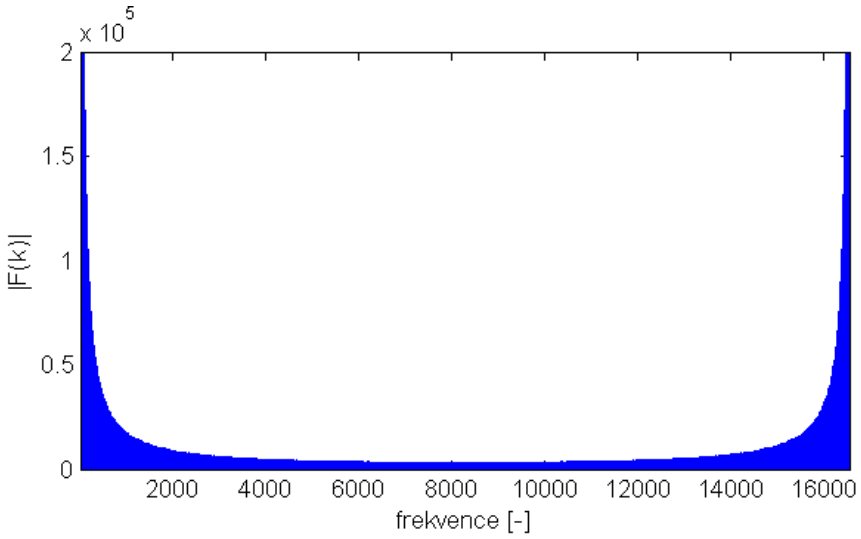
Obr. 4.5: Přiblížené spektrum kumulované fáze *E. coli*

Při pohledu na přiblížené spektrum je jasné, že užitečných spektrální čar je mnohem větší množství, nicméně stále můžeme větší množství spektra odebrat bez toho aniž by signál ve velkém měřítku výrazně změnil tvar.



Obr. 4.6: Spektrum kumulované fáze *H. sapiens* mitochondrie

U signálu z lidské mitochondriální DNA na Obr. 4.6 je sice již v původním měřítku vidět více spektrálních čar, i zde jsou však krajní spektrální čáry řádově mnohem větší než zbylé a pro lepší pohled je tak potřeba spektrum přiblížit, jak je tomu na Obr. 4.7. I tento obrázek tak ořezává nejvýraznější spektrální čáry o velikostech až $1{,}8 \cdot 10^{10}$.



Obr. 4.7: Přiblížené spektrum kumulované fáze *H. sapiens* mitochondrie

Samozřejmě vzorkovací frekvence mitochondriálního signálu je mnohem nižší, neboť délka sekvence je pouze 16 569 bp. I tak je f_{vz}= 16 569 pořád mnohem větší než vzorkovací frekvence pro jiné biologické signály. Bez ohledu na měřítko, mají obě spektra na Obr. 4.6 a Obr. 4.7 stejný tvar, a tedy stejné předpoklady ke kompresi.

4.3 Komprese číslicových signálů

Ve druhé kapitole jsme probrali teorii o kompresi dat a zmínili i několik algoritmů pro kompresi genomických znakových sekvencí. Tyto algoritmy byly bezeztrátové, a to statistické nebo se slovníkovým kódováním. Obdoba těchto algoritmů by byla použitelná i na signál kumulované fáze, neboť se jedná o bezeztrátovou numerickou mapu a signálový zápis je tak ekvivalentní tomu sekvenčnímu. Samozřejmě použitelnost statistických metod a metod se slovníkovým kódováním je universální a mohou být použity na signály obecně. Potom ale dostaneme zase data se změněným kódováním, které není použitelné pro další zpracování. Takto komprimovaný signál by byl pro účely klasifikace tedy nepoužitelný. Nepoužitelnost slovníkového a statistického kódování vychází z toho, že nechceme data přenášet ani ukládat, ale porovnávat mezi sebou. Sice můžeme chtít uložit banku vybraných signálů, se kterou budou analyzované signály při analýze porovnávány, potom si ale musíme upřesnit fakt, proč by takto uložená komprimovaná data nebyla použitelná.

Expanze dat způsobená konverzí na signál

Data by bylo potřeba před každou analýzou znovu dekomprimovat do původního stavu, protože slovník nebo statistika každého genomického signálu by byla jiná, jak plyne z velmi rozdílných tvarů signálů. A porovnávání celých signálů nás staví na úroveň zpracování celých sekvencí znaků, které je výpočetně náročné kvůli zarovnávání různě dlouhých řetězců. Zarovnání je potřeba respektovat i pro různě dlouhé signály. K tomu přidávají genomické signály další nevhodnou vlastnost a to svoji velikost co se objemu dat týče.

Převod sekvence na signál kumulované fáze znamená obrovskou expanzi v zápisu našich dat. Ta je způsobená rozdílným dynamickým rozsahem. V sekvenci se vyskytují pouze 4 znaky, tedy počet kvantizačních hladin je roven 4 a můžeme použít standardně zpracovatelné 8 bitové kódování. Dynamický rozsah kumulované fáze je jiný pro každý signál a počet kvantizačních hladin několikanásobně převyšuje 4 hladiny u korespondující sekvence. To má za následek nutnost použití vícebitového kódování. Potom můžeme začít narážet na problém s nároky na velikost operační paměti pro zpracování takových dat.

Ztrátová komprese

Charakteristický tvar signálu ve velkém měřítku na druhou stranu ukazuje redundanci jeho zápisu pomocí stejného počtu vzorků jako je počet znaků v sekvenci, která žádnou charakteristiku ve velkém měřítku nemá. Tato nadbytečnost může být redukována při ztrátě informace na úrovni rozlišení jednotlivých nukleotidů.

Tento předpoklad se však vymyká klasickému pojmu komprese, jelikož se bavíme o snaze redukovat počet vzorků ve vodorovné ose, zatímco komprese v pravém slova smyslu znamená redukci počtu kvantizačních hladin v ose svislé. Z pohledu signálového zpracování se tak bavíme o podvzorkování signálu. Signál kumulované fáze sekvence se od konvenčních biologických signálů v mnohém liší, na což jsme narazili již při spektrální analýze. Také zpracování genomických signálů se bude výrazně lišit od zpracování jiných signálů. Chápání slova komprese jako podvzorkování signálu proto nebude neoprávněné. Záleží na našem pohledu na dimensionalitu signálu. K tomu se chováme jako k 1D signálu vyvíjejícímu se podél sekvence namísto v čase. Také pro analýzu jsme zatím využili algoritmy ve tvaru pro zpracování 1D signálů. Numerická mapa kumulované fáze je ale definovaná ve 2D prostoru, kde jednu dimenzi dostáváme náhradou roviny reálné za rovinu komplexní. Pokud po podvzorkování nezachováme měřítko vodorovné osy, což je i vzhledem k její bezrozměrnosti dost problematické, ale mezi sousedícími vzorky ponecháme bezrozměrnou vzdálenost jedné jednotky, můžeme zcela opodstatněně takové podvzorkování označit za kompresi. Z takového pohledu vnímáme náš signál jako 2D na nějž je aplikováno 1D zpracování. Takový přístup je velmi netypický, ale pro naše účely vhodný.

Současné algoritmy pro ztrátovou kompresi

Přestože v odstavci výše jsme vyloučili možnost použití kompresních algoritmů pro signály, můžeme se jimi alespoň inspirovat. Dobrých výsledků při kompresi EKG, jiného biologického signálu, se dosahuje například při využití algoritmů SPIHT nebo EZW [11]. Tyto algoritmy jsou oba založené na využití vlnkové transformace, pomocí které je signál rozložen na různá frekvenční pásma. Ta jsou potom zvlášť kódována pomocí menšího počtu bitů. Zpracování EKG signálu je ale od zpracování našeho signálu dost odlišné. Jedná se totiž o zpracování mnoha po sobě jdoucích opakujících se úseků. Komprese pak může být prováděna například online vždy po několika periodách přímo při akvizici signálu [42].

Při porovnávání signálů pak teoreticky můžeme porovnávat dekomprimované úseky postupně. Nakolik je takové ušetření paměti výhodné oproti zvýšení časové výpočetní náročnosti ponechme stranou. Genomický signál periodický není, respektive ho pro signálové zpracování chápeme jako jednu periodu. Taková perioda je řádově mnohem delší než periody konvenčních signálů což nás omezuje v použití již rozvinutých metod, které by pak nikdy nevedlo k úspoře výpočetní časové či paměťové náročnosti při klasifikaci.

Zde můžeme navázat na problém podvzorkování. Pokud podvzorkujeme EKG signál s faktorem 2, bude mít jeho jedna perioda stále stejnou délku co se časové osy týče. Tato osa však bude mít o polovinu menší počet vzorků. Jedna sekunda signálu tak bude mít například oproti původním 512 hodnotám pouze 256 hodnot. Nicméně časová délka signálu zůstane nezměněna. Při podvzorkování signálu kumulované fáze budeme mít místo signálu o délce 512 vzorků pouze signál o délce 256 vzorků. Přitom jeho zarovnání na původní vodorovnou osu pro nás není podstatné, pokud všechny porovnávané signály podvzorkujeme stejným faktorem, tedy jejich proporcionální délky zůstanou stejné. Absolutní délka genomického signálu pro nás tedy není důležitá.

4.4 Vlnková transformace

Využití v současných algoritmech nalezneme při využití vlnkové transformace. Stejně jako Fourierova transformace, je i vlnková transformace lineární spojitou transformací. Liší se pouze v použití jiné bázové funkce, tzv. vlnky

$$\psi_{\lambda,\vartheta}(t) = \frac{1}{\sqrt{\lambda}}\psi\left(\frac{t-\vartheta}{\lambda}\right) \tag{4.8}$$

kde parametr λ podmiňuje dilataci (stlačení nebo roztažení) mateřské vlnky a parametr ϑ její časové posunutí.

Dyadická vlnková transformace

Speciální případ vlnkové transformace dostaneme při využití parametrů, které jsou mocninami dvou, tedy $\lambda=2^m$ a $\vartheta=2^m kT$, pro $m>0$. Tu označujeme jako dyadickou vlnkovou transformaci (dyadickou DWT)

$$y(m,k) = \frac{1}{\sqrt{2^m}} \int_{-\infty}^{+\infty} x(t)\, \psi^*(2^{-m}t - kT)dt \tag{4.9}$$

kde m, k reprezentují kmitočtové a časové měřítko. Přitom se ve vzorci uplatňuje ještě konstanta T. Ta závisí na šířce pásma mateřské vlnky a určuje hustotu vzorkování koeficientů na časové ose pro jednotlivá frekvenční pásma daná indexem m.

Stále se pohybujeme v oblasti transformací spojitých signálů, i když diskrétní vlnkovou transformací. Než si uvedeme vlnkovou transformaci pro diskrétní signály je

vhodné nejprve prozkoumat některé vlastnosti ve spojité oblasti, které pro nás budou důležité i v oblasti diskrétní.

Označme spektrum, tedy Fourierovu transformaci, mateřské vlnky jako *Ψ(ω)*, potom

$$FT\left[\frac{1}{\sqrt{2^m}}\psi\left(\frac{t-2^m kT}{2^m}\right)\right]=\frac{1}{\sqrt{2^m}}\int_{-\infty}^{+\infty}\psi\left(\frac{t-2^m kT}{2^m}\right)e^{-j\omega t}dt=\sqrt{2^m}\Psi(2^m\omega)e^{-j\omega 2^m kT} \quad (4.10)$$

časová expanze *m*-té vlnky na 2^m násobnou délku znamená kompresi spektra této vlnky na $1/2^m$ násobek výchozí šířky a jeho posunem k nižším frekvencím se středním kmitočtem na $1/2^m$ násobku původního. Díky normalizaci vzroste modul spektra na $2^{m/2}$ násobek. To je důležitá vlastnost, kterou později s výhodou použijeme. DWT se tedy vyznačuje oktávovou podobou spekter soustavy vlnek.

Pokud použijeme vztah mezi korelací a konvolucí, můžeme korelace signálu *x(t)* s vlnkami realizovat jako jeho konvoluci s funkcemi časově reverzními

$$y(m,k)=\int_{-\infty}^{+\infty}x(t)\,h_m(2^m kT-t)dt=\int_{-\infty}^{+\infty}h_m(t)\,x(2^m kT-t)dt \quad (4.11)$$

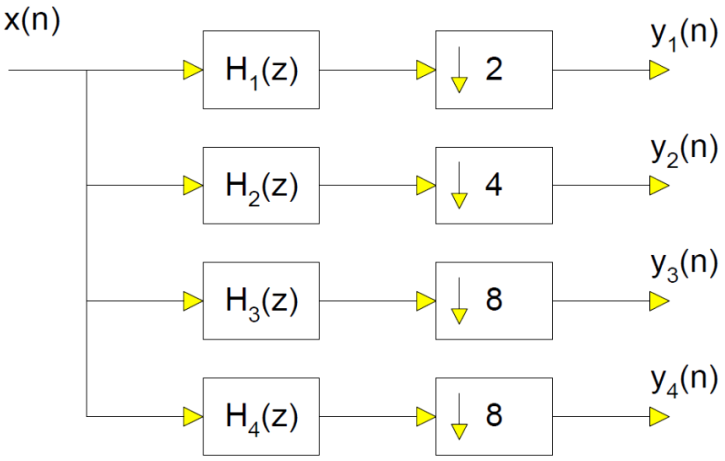
a takovou dyadickou DTW lze realizovat rozkladem signálu bankou lineárních spojitých oktávových filtrů s impulsními charakteristikami $h_m(t)$.

Dyadickou vlnkovou transformaci s diskrétním časem (dyadickou DTWT) můžeme definovat analogicky s využitím diskrétní konvoluce

$$y_m(n)=\sum_{i=-\infty}^{+\infty}x(i)\,h_m(2^m n-i)=\sum_{i=-\infty}^{+\infty}h_m(i)\,x(2^m n-i) \quad (4.12)$$

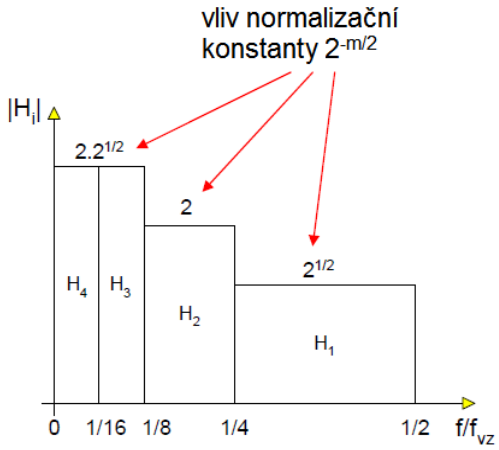
tedy rozkladem signálu bankou diskrétních oktávových filtrů s impulsními charakteristikami $h_m(n)$. Potom vzorkovací frekvence signálu $y_m(n)$ na výstupu *m*-tého filtru je 2^m krát nižší než vzorkovací frekvence f_{vz} vstupního signálu *x(n)*. [43]

Ukázku realizace třístupňové dyadické DTWT s využitím banky FIR filtrů můžeme vidět na Obr. 4.8. Počet pásem je vždy o jedna větší než stupeň rozkladu, tedy v našem příkladě je signál rozložen na 4 pásma. Výstupní signály jsou posloupnostmi koeficientů dyadické DTWT.



Obr. 4.8: Dyadická DTWT pro tři stupně rozkladu [43]

Modulové frekvenční charakteristiky příslušných ideálních filtrů jsou na Obr. 4.9. Důležitý je vliv normalizační konstanty, který se uplatňuje i při zpracování diskrétních signálů.



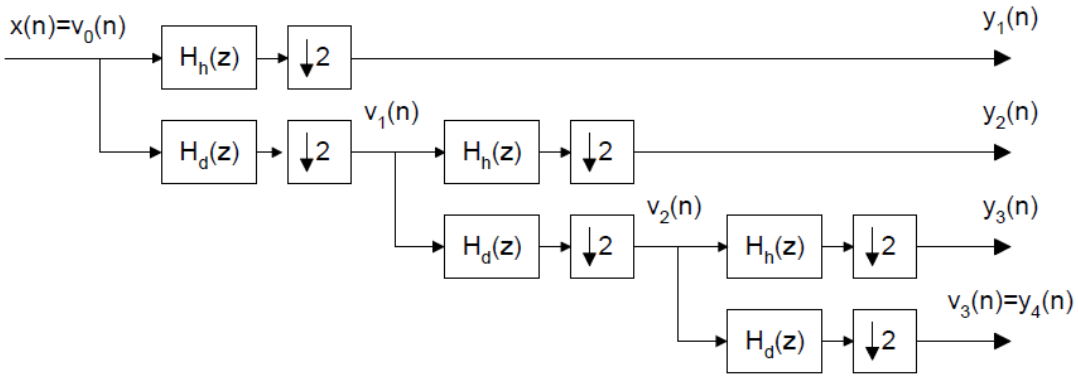
Obr. 4.9: Dyadická DTWT, modulové charakteristiky filtrů [43]

4.5 Komprese kumulované fáze

Využitím DTWT na naše testovací signály si ukážeme možnosti rozkladu u prokaryotních chromosomových signálů a u mitochondriálních signálů živočichů zvlášť. Ze spektrální analýzy na Obr. 4.5 a Obr. 4.7 vyplývá, že nejvýznamnější frekvenční pásmo, podmiňující tvar signálu se nachází v oblasti nejnižších frekvencí signálů. Pro analýzu tak budeme signál postupně rozkládat dvojicí zrcadlových půlpásmových filtrů a zkoumat vlastnosti výstupu filtru dolní propusti.

Takovým rozkladem je i celá vlnková transformace často počítána. Výpočet se potom označuje jako tzv. rychlá DTWT, protože používáme dvojici stále stejných zrcadlových filtrů

$H_h(z)$ a $H_d(z)$. Schéma popsaného rozkladu ukazuje . V i-tém kroku tedy analyzujeme signál $v_i(n)$.

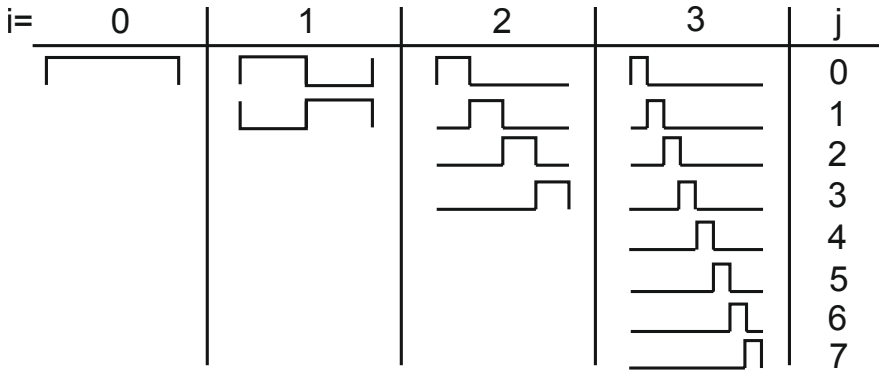


Obr. 4.10: Rychlá DTWT, [43]

Zvolená vlnka

Volba vlnky a počet stupňů rozkladu jsou naše jediné vstupní parametry. Právě tato jednoduchost se hodí při využití automatické komprese v rámci výpočtů pro klasifikaci organismů. Přitom nás zajímá pouze rozdělení spektra na dvě poloviny, jako logické potom využití nejjednodušší, tzv. Haarovy vlnky. Jedná se o nejstarší a nejjednodušší vlnku popsanou již v roce 1909 [44]. Nachází ovšem uplatnění i v dnešní době, neboť díky její jednoduchosti je výpočet velmi rychlý. Haarova vlnka je popsána i v pozdější době jako speciální případ vlnky Daubechiesové [45]. Využití složitějších vlnek nemá smysl, neboť jejich výpočet je vždy náročnější a nepřináší žádnou výhodu. Naopak použití vlnek, které mění tvar signálu v jednotlivých pásmech je přímo nežádoucí, neboť jak jsme prokázali, signály různých organismů se již dostatečně liší.

Haarova vlnka je obdélníková vlnka, jejíž tvar pro tři stupně rozkladu ukazuje Obr. 4.11. Pro i-tý stupeň rozkladu má vlnka $j=2^i$ tvarů, neboli rozklad probíhá na j pásem.



Obr. 4.11: Haarova vlnka

Posunutí vlnky a její dilatace, která není v našem případě definovaná v čase, ale zase podél sekvence, závisí tedy na stupních rozkladu a přizpůsobuje se délce původního signálu. To je velmi vhodné vzhledem k tomu, že každý signál je jinak dlouhý. Samozřejmě výpočet

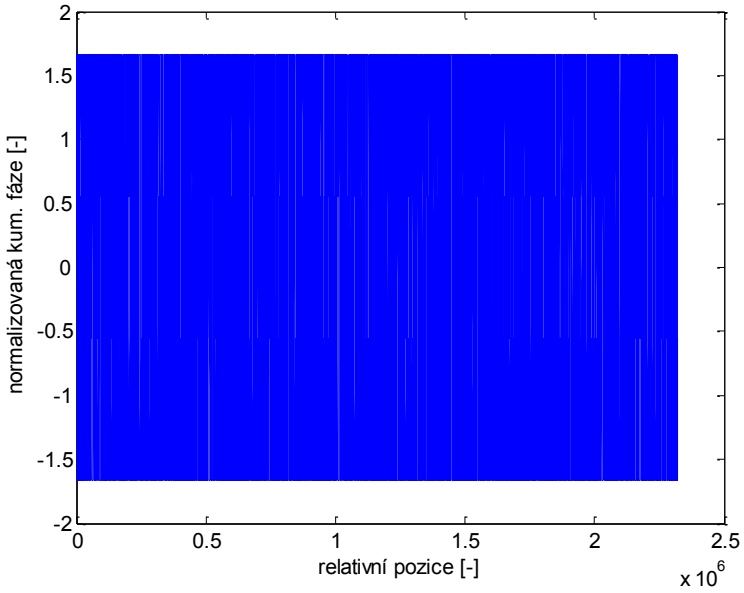
DTWT probíhá pomocí dvojice filtrů. Haarově vlnce odpovídá dvojice filtrů s impulzními charakteristikami:

$$h_h(n) = \{-0{,}7071; 0{,}7071\}$$

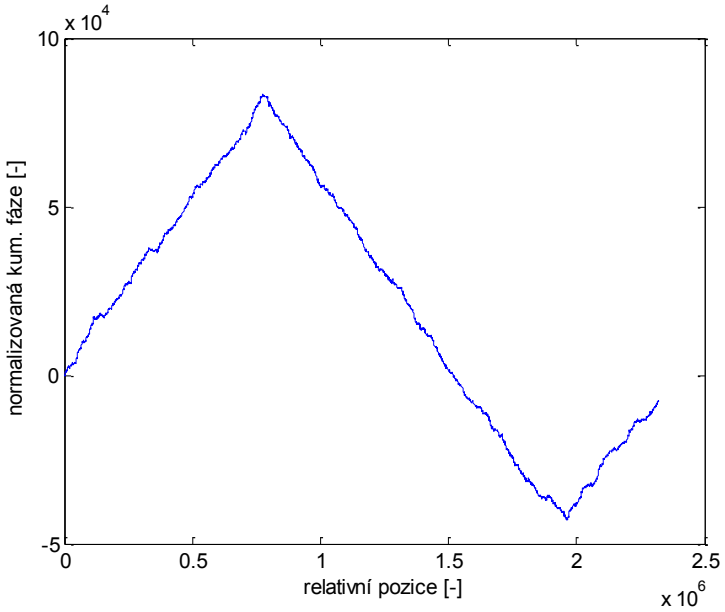
$$h_d(n) = \{0{,}7071; 0{,}7071\}$$

Prokaryotní signály

Analýzu si ukážeme na našem testovacím signálu organismu *E. coli*. Obě pásma prvního stupně rozkladu jsou na Obr. 4.12 a Obr. 4.13. Obě pásma jsou podvzorkována, proto jsou signály o polovinu kratší.



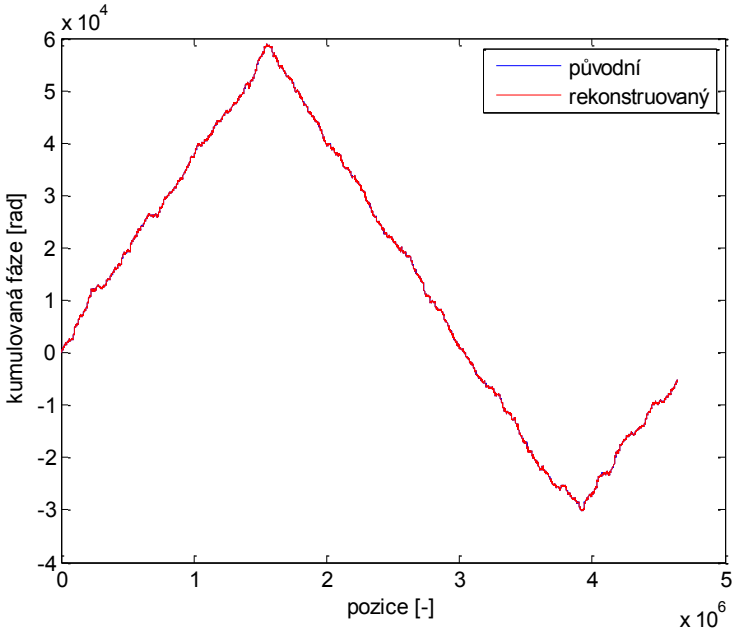
Obr. 4.12: *E. coli*, DTWT 1. stupeň, horní pásmo



Obr. 4.13: *E. coli*, DTWT 1. stupeň, dolní pásmo

Horní pásmo obsahuje periodický vysokofrekvenční šum s velmi malou amplitudou oproti původnímu signálu. Takové pásmo tedy můžeme z další analýzy vypustit. Dolní pásmo se tvarově velmi podobá původnímu signálu, proto ho můžeme označit za první stupeň komprese původního signálu. Velikost signálu dolního pásma je stejná jako signálu původního, protože při prvním stupni rozkladu se ještě neuplatní vliv normalizační konstanty. Všimněme si také, že zde hovoříme o velikosti a ne amplitudě signálu. Tyto signály totiž nejsou periodické, je proto vhodnější použít pojem velikost. Tento pojem také více odpovídá našemu pseudo1D/2D pohledu na signál. Pokud bychom se na signál dívali čistě z 1D pohledu, který jsme použili pro Fourierovu transformaci, potom jej chápeme jako jednu periodu a pojem amplituda můžeme oprávněně použít.

Hodnocení ztráty informace můžeme získat při zpětné rekonstrukci signálu. Při této rekonstrukci ovšem použijeme pouze dolní pásmo a horní nahradíme nulovým signálem potřebné délky. Srovnání původního a rekonstruovaného signálu ukazuje Obr. 4.14. Přitom se oba signály překrývají a je mezi nimi pouze malý rozdíl.



Obr. 4.14: *E. coli*, DTWT 1. stupeň, rekonstruovaný signál

Rozdíl signálů můžeme zhodnotit pomocí jejich procentuální odchylky [46]

$$PRD = \sqrt{\frac{\sum_{i=1}^{n}(x_0(i) - x_r(i))^2}{\sum_{i=1}^{n}(x_0(i) - \overline{x_o})^2}} \cdot 100\% \qquad (4.13)$$

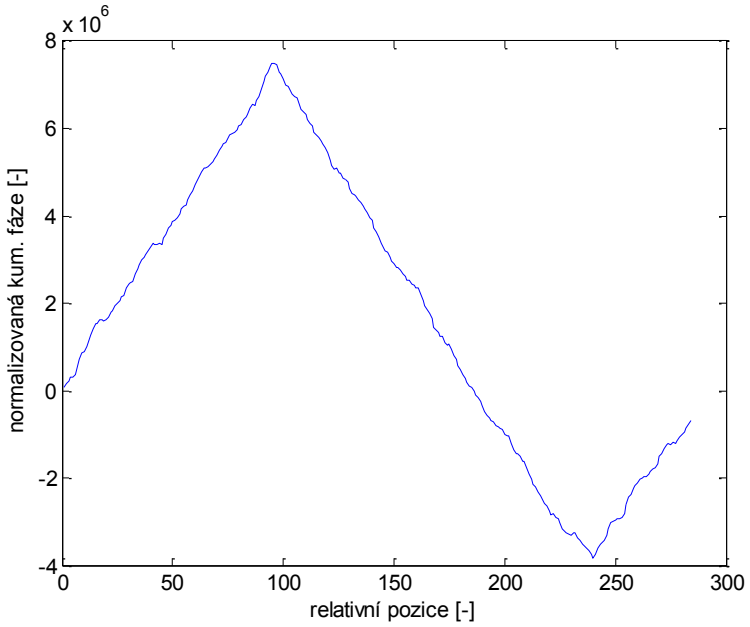
kde x_0 je originální a x_r rekonstruovaný signál o délce n vzorků.

Pro náš signál dosahuje *PRD=0,0034%*. Což je sice velmi malá odchylka, ale rekonstruovaný signál již nelze zpětně převést na znakovou sekvenci, protože kvantizační hladiny rekonstruovaného signálu už nejsou pouze násobky *π/4*. Tedy bylo by to možné,

ale pouze za cenu dalšího zaokrouhlování. Pro naše potřeby však signál zpětně konvertovat nechceme a spokojíme se tak pouze s hodnocením *PRD* v signálové reprezentaci.

Zároveň je pro nás obtížné definovat popisné parametry komprese (viz druhá kapitola), jako je průměrná délka slova, nebo kompresní poměr. Protože jsme se vzdali pohledu na původní sekvenci, kde slovo znamenalo právě jeden znak, nebudeme hodnotit průměrnou délku slova. Cílem naší komprese je především ušetření časové a ne paměťové náročnosti. Proto kompresní poměr (CR - compression ratio) definujeme jako snížení počtu vzorků, což je v našem případě prozatím $CR=1:2^1=0,5$.

Otázkou zůstává, kolikrát můžeme rozklad pomocí DTWT opakovat, abychom signál dostatečně komprimovali, ale zároveň zachovali význačné informace. Oblast genomických signálů je velmi málo rozvinutá oblast a klasifikace genomických signálů není dosud popsána. Vzhledem k tomu, že signálové reprezentace vznikly kvůli dobré vizualizaci genomů, požadavkem může být právě zachování detailů ve velkém měřítku a procentuální odchylky menší než 1%. Pro signály prokaryotních kruhových to odpovídá až 14 stupňovému rozkladu, kde výsledné 15. pásmo zastupuje původní signál. Signál 15. pásma je na Obr. 4.15.

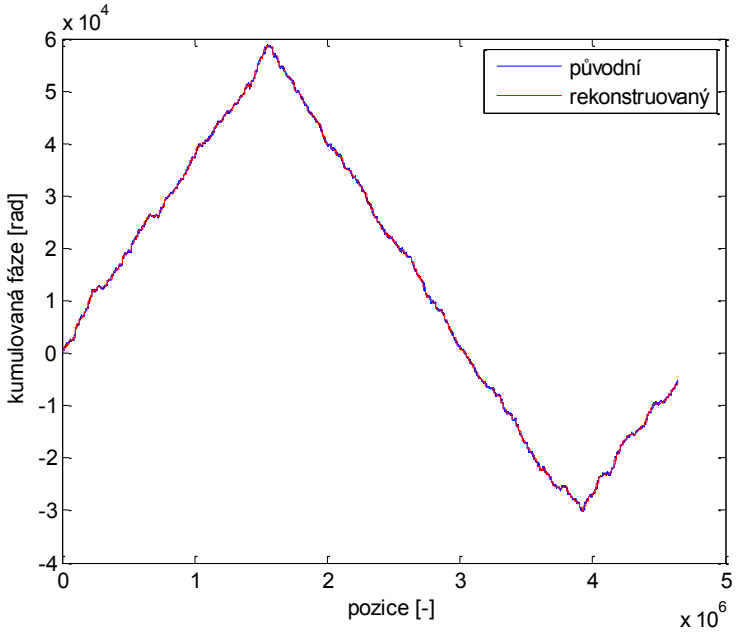


Obr. 4.15: *E. coli*, DTWT 15. pásmo

Takto komprimovaný signál tvarově stále zachovává požadované detaily a je dostatečně specifický pro klasifikaci organismu z nějž byl signál získán, jak si ukážeme na příkladu s reálnými sekvencemi na konci práce. Zde se uplatňuje výhoda vlnkové transformace v podobě normalizační konstanty. Signál v 15. pásmu je tvarově stejný jako původní signál. Obsahuje ale 2^{14} krát méně vzorků, jejichž kumulovaná fáze je normalizována. Tedy nabývá řádově vyšších hodnot. Tato vlastnost redukuje změnu

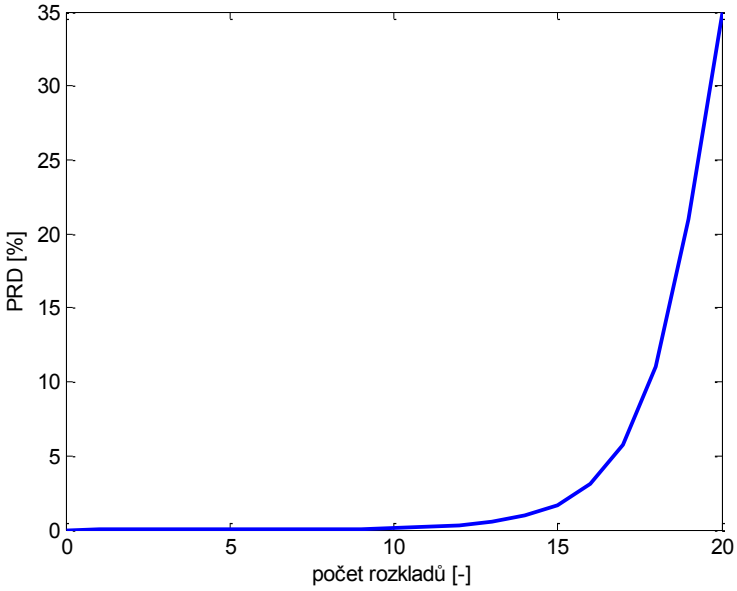
proporcionální vzdálenosti dvou signálů, kdy porovnáváme sice menší počet vzorků, nicméně rozdíly mezi nimi jsou mnohem větší. Toho s výhodou využijeme právě pro klasifikaci.

Signál 15. pásma s doplněnými nulovými pásmy můžeme znovu rekonstruovat na původní signál na Obr. 4.16.



Obr. 4.16: *E. coli*, DTWT 15. pásmo, rekonstruovaný signál

Procentuální odchylka zůstala menší než 1%, konkrétně *PRD=0,9568%*. Pro více stupňů rozkladu začíná procentuální odchylka rekonstruovaného signálu prudce stoupat a její vývoj ukazuje Obr. 4.17. Ztráta informace pro vyšší než 15. pásmo už je tak výrazná, že by znesnadňovala klasifikaci takto komprimovaných signálů.

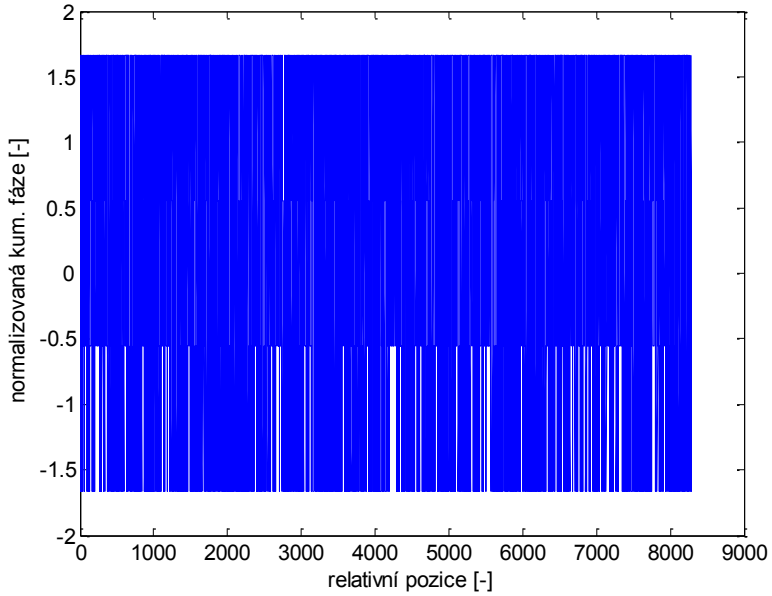


Obr. 4.17: *E. coli*, vývoj PRD

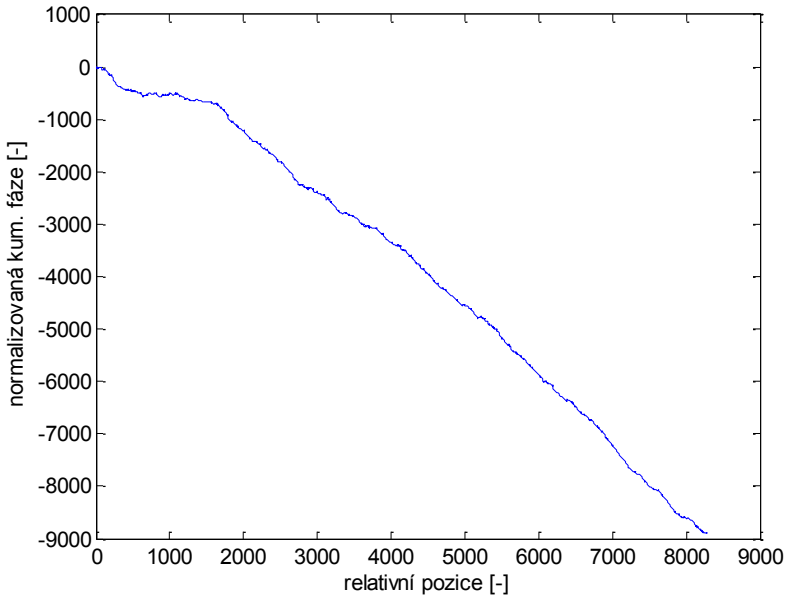
Pro prokaryota jsme dosáhli použitelného kompresního poměru $CR=1:2^{14}$. Tato komprese tak mnohem předčí odhady původní alespoň 50% komprese a v klasifikaci bude znamenat významné ušetření výpočetního času.

Mitochondriální signály

Stejný postup nyní můžeme provést pro signál lidské mitochondriální DNA.



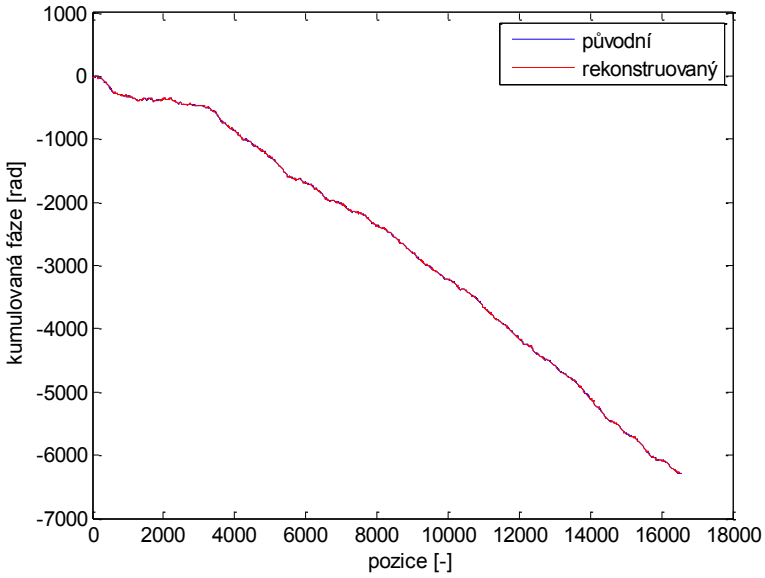
Obr. 4.18: *H. sapiens* mitochondrie, DTWT 1. stupeň, horní pásmo



Obr. 4.19: *H. sapiens* mitochondrie, DTWT 1. stupeň, dolní pásmo

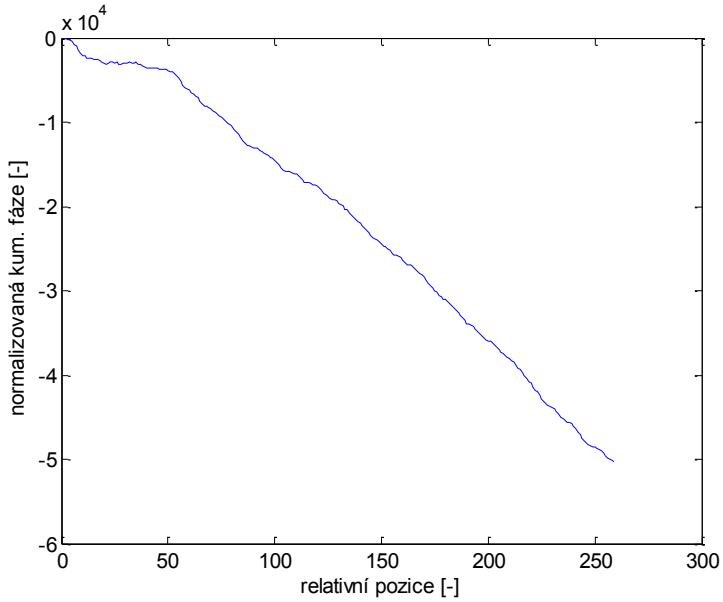
Jednostupňovým rozkladem na Obr. 4.18 a Obr. 4.19 odstraníme rušivou složku se stejnou amplitudou jako u signálů prokaryot. Obě pásma jsou zase podvzorkována

faktorem 2 a jejich délka je tak o polovinu menší. Pokud rušivé pásmo nahradíme nulovým signálem a provedeme inverzní vlnkovou transformaci dosáhneme *PRD=0,0441%*, což je o řád vyšší odchylka než u jednostupňového rozkladu předchozího signálu. Potvrzuje se tedy fakt o problematičtější kompresi pro mitochondriální sekvence, které na rozdíl od prokaryot obsahují mnohem větší podíl kódující ku nekódující sekvenci. Rekonstruovaný a původní signál ukazuje Obr. 4.20.



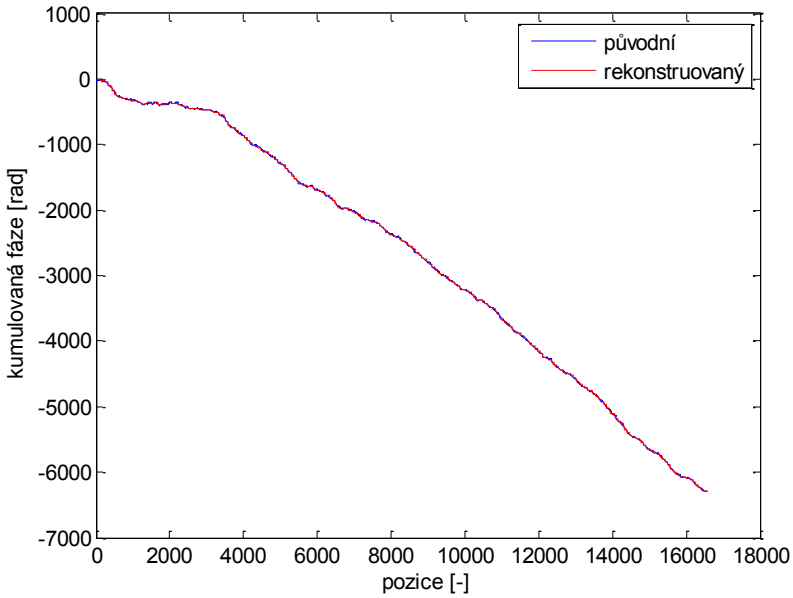
Obr. 4.20: *H. sapiens* mitochondrie, DTWT 1. stupeň, rekonstruovaný signál

Hloubka rozkladu signálu co se počtu stupňů týče je také limitována délkou signálu. Protože signál v *i*-tém pásmu má 2^i krát méně vzorků než signál původní, nesmí být tato hodnota vetší než délka signálu.



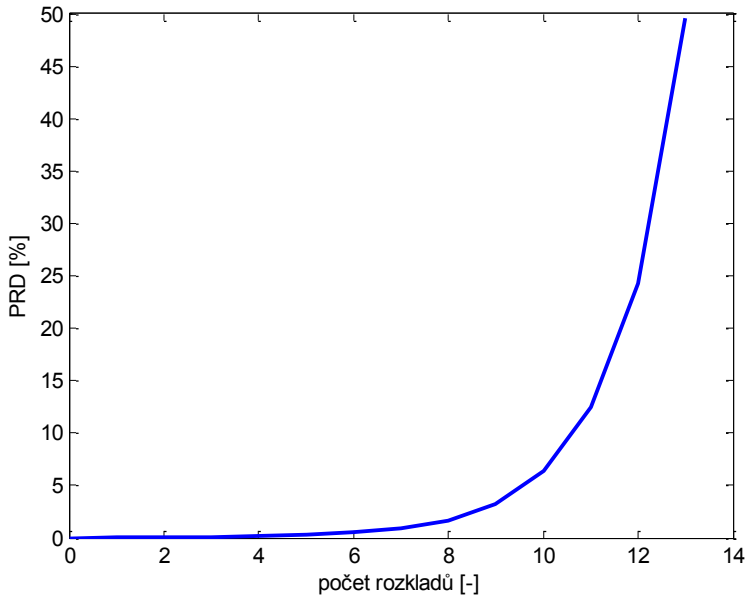
Obr. 4.21: *H. sapiens* mitochondrie, DTWT 7. pásmo

U mitochondriálních signálů se tedy spokojíme se šestistupňovým rozkladem a využitím výsledného 7. pásma, viz Obr. 4.21. Komprimovaný signál člověka teď dosahuje stejné délky jako komprimovaný signál E. coli, i když byl předtím několikanásobně kratší. Vlivem normalizační konstanty se opět značně mění velikost signálu. Rekonstrukcí 7. pásma (pro které je $CR=1:2^6$) na Obr. 4.22 dosáhneme *PRD=0,5115%.*



Obr. 4.22: *H. sapiens* mitochondrie, DTWT 7. stupeň, rekonstruovaný signál

Vývoj procentuální odchylky až do 13. stupně rozkladu ukazuje.



Obr. 4.23: *H. sapiens* mitochondrie, vývoj PRD

5. Klasifikace genomických signálů

Kumulovaným fázím komprimovaným naším způsobem zůstaly proporčně zachované vlastnosti a očekáváme tak, že míra původní genetické informace, jež nesly sekvence znaků, zůstala i přes ztrátovost komprese zachována na dostatečné úrovni pro klasifikaci. Postup klasifikace si probereme v následující kapitole.

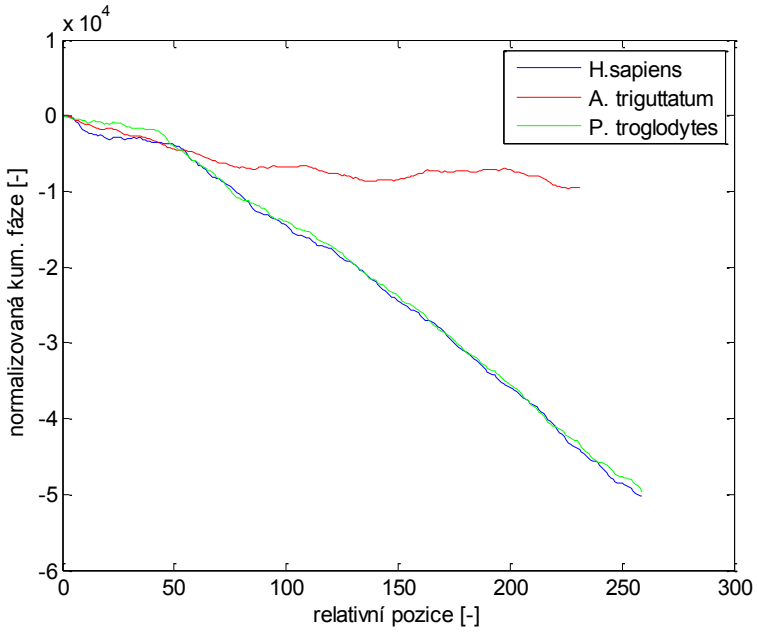
5.1 Zarovnání signálů

Standardním postupem v bioinformatice je před jakoukoli analýzou více sekvencí tyto sekvence nejdříve zarovnat [47]. Je to velmi důležité kvůli povaze DNA, která jak jsme si zmínili v první kapitole obsahuje pohyblivé elementy. Pokud by nedošlo k zarovnání, porovnávali bychom části sekvence, které sice v DNA porovnávaných organismů leží na stejném místě, ve skutečnosti ale spolu mohou mít pramálo společného. Tento fakt je potřeba akceptovat i v signálové reprezentaci, přičemž je velký rozdíl mezi různými druhy sekvencí. Při porovnávání více organismů, je časté využití vícenásobného zarovnání, kdy jsou všechny klasifikované organismy zarovnány najednou. Tento fakt vychází hlavně z toho, že algoritmy pro znakové zpracování jsou uzpůsobeny na analýzu pouze krátkých úseků sekvence, například určitých genů. Přitom délka těchto úseků je ve všech organismech přibližně stejná.

Mitochondriální signály

Mitochondriální sekvence jsou oproti jaderným mnohem více konzervované. Vliv pohyblivých elementů se u nich uplatňuje pouze výjimečně. Navíc mitochondriální sekvence příbuzných organismů mají velmi podobné délky sekvencí, jejichž rozdíly se ještě minimalizují při kompresi. Proto zarovnávání mitochondriálních sekvencí můžeme provést pouze porovnáním dvojice signálů od prvního po poslední vzorek kratšího signálu a k tomuto porovnání přičíst penalizaci v podobě maximálního rozdílu mezi signály násobeného rozdílem počtu vzorků obou signálů. Pro výpočet matice vzdáleností pak využijeme vždy pouze příslušné dvojice porovnávaných signálů. Vícenásobné zarovnání by vyžadovalo složitější způsob hodnocení penalizací při rozdílné délce sekvencí, což není žádoucí.

Ukažme si příklad na naší testovací sekvenci člověka (*Homo sapiens,* NC_012920, délka sekvence: 16 569 bp), kterou budeme porovnávat se sekvencemi šimpanze (*Pan troglodytes*, NC_001643, délka sekvence: 16 554 bp) a klíštěte (*Amblyomma triguttatum*, NC_005963, délka sekvence: 14 740 bp). Soubor komprimovaných signálů ukazuje Obr. 5.1.



Obr. 5.1: Komprimované signály eukaryot

Komprimované signály člověka a šimpanze mají stejnou délku, neboť jejích rozdíl v délce DNA sekvencí o 5 nukleotidů se kompresí zcela smazal. Signál klíštěte je stále kratší. Jeho délka je 231 vzorků oproti 259 vzorkům signálů člověka a šimpanze. Při porovnávání signálu klíštěte s ostatními tak musíme zbývající 2 signály vždy ořezat o koncových 28 vzorků a výpočet doplnit o penalizační konstantu. Přitom při porovnávání člověka a šimpanze využijeme oba signály celé a penalizace tak není potřeba.

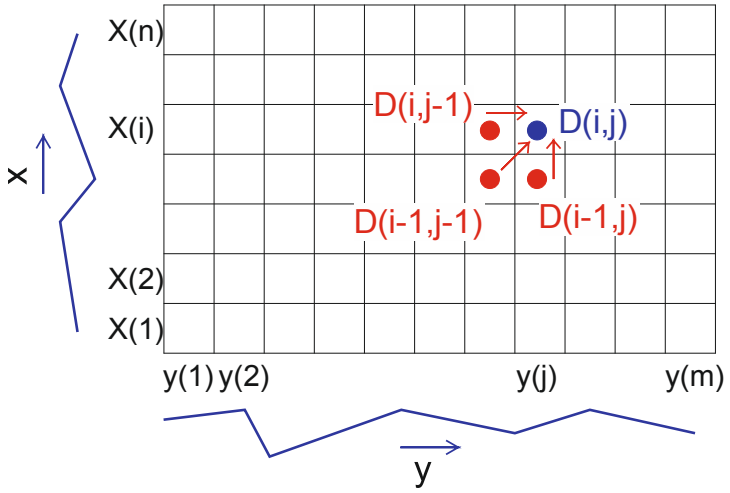
Prokaryotní signály

U prokaryot je situace zcela odlišná. Genomy jednotlivých organismů se velmi liší co se týče délky i tvaru signálů. Rozdíly v délce jsou často tak velké, že nejsou redukovány ani kompresí. Také se zde velmi uplatňuje pohyb jednotlivých částí genomů, který se udál při fylogenetickém vývoji. Pro maximalizaci využité genetické informace tak budeme zase využívat zarovnání pouze dvojic signálů namísto vícenásobného zarovnání. Zarovnání těchto dvojic však nemůžeme provádět odstraněním koncových vzorků [48], ale potřebujeme nástroj ekvivalentní k lokálnímu nebo globálnímu zarovnání Smith-Watermanovým (SW) [49] nebo Needleman-Wunschovým (NW) [50] algoritmem pro znakové sekvence. Potřebné vlastnosti má algoritmus dynamického borcení časové osy (DTW- dynamic time warping), který dokáže vhodně rozpoznávat v signálech různé vzory [51].

DTW je optimalizační algoritmus také označovaný jako dynamické programování. Jeho cílem je zarovnání dvou signálů takovým způsobem, aby byla minimalizována vzdálenost mezi těmito dvěma signály. Základem algoritmu je výpočet tabulky vzdáleností dvou signálů. Již z názvu je jasné, že algoritmus byl vytvořen pro analýzu 1D signálů

s vývojem v čase. Pokud se na genomické signály podíváme opět z pohledu 1D vývoje podél sekvence nebude problém algoritmus aplikovat i na naše data.

Tabulka akumulované vzdálenosti mezi signály se zpravidla počítá z levého spodního rohu, protože signál do prvního sloupce zapisujeme od koncového vzorku. Jinak se ale neliší od tabulky akumulovaných vzdáleností počítané NW a SW algoritmy, kde sekvenci zapisujeme do sloupce od začátku a matici tak počítáme z levého horního rohu. Schéma výpočtu pro signály *x(n)* a *y(m)* ukazuje Obr. 5.2.



Obr. 5.2: Schéma výpočtu DTW

Protože se algoritmus snaží o minimalizaci vzdálenosti dvojice signálů, je výpočet kumulované vzdálenosti podmíněn vztahem

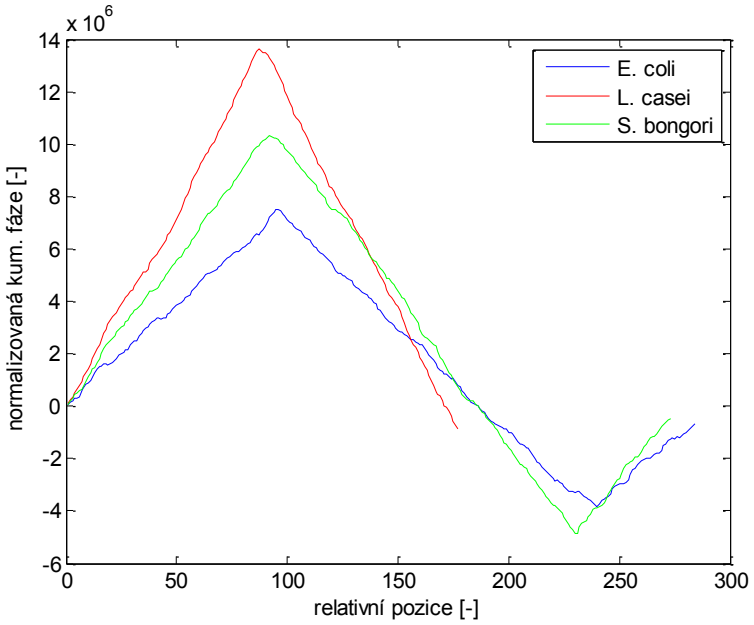
$$D(i,j) = min \begin{cases} D(i,j-1) + d(i,j) \\ D(i-1,j-1) + d(i,j) \\ D(i-1,j) + d(i,j) \end{cases} \tag{5.1}$$

kde *d(i,j)* vyjadřuje lokální vzdálenost *i*-tého vzorku signálu *x(n)* a *j*-tého vzorku signálu *y(m)* danou vztahem

$$d(i,j) = |x(i) - y(j)| \tag{5.2}$$

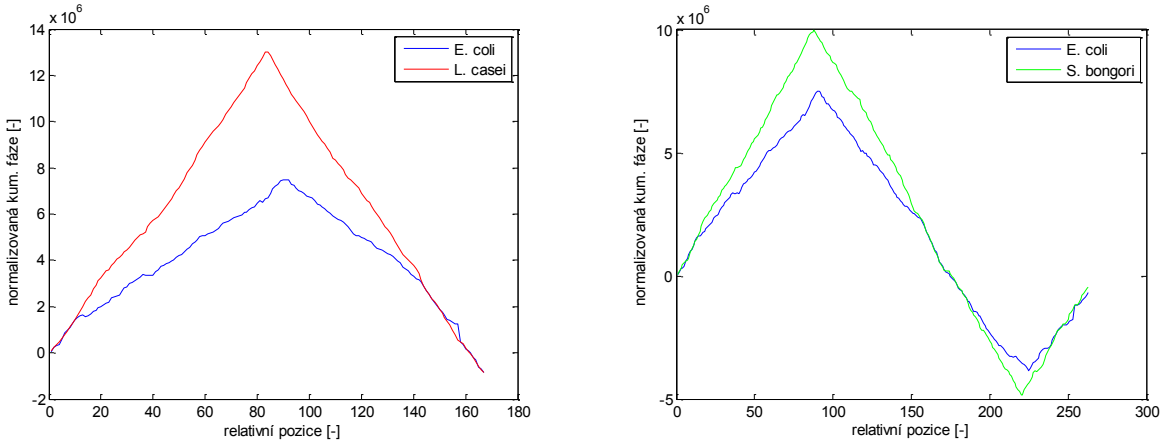
Zarovnání signálů je provedeno zpětným trasováním z tabulky akumulovaných vzdáleností.

Reálný příklad zarovnání si ukážeme pro náš testovací signál *Escherichia coli* (NC_000913, délka sekvence: 4 639 675 bp) se signálem organismů *Lactobacillus casei* (NC_008526, délka sekvence: 2 895 264 bp) a *Salmonella bongori* (NC_015761, délka sekvence: 4 460 105 bp). Uvedené komprimované signály jsou uvedeny na Obr. 5.3.



Obr. 5.3: Komprimované signály prokaryot

Signál *L. casei* je výrazně kratší, vícenásobné zarovnání by znamenalo příliš velkou ztrátu informace při porovnávání zbylých dvou signálů, způsobenou zahrnutím zbytečně velkého počtu delecí do těchto signálů. Navíc by došlo k velkému nárůstu výpočetní náročnosti. Zarovnání *E. coli* ve dvojici se zbývajícími dvěma signály pomocí DTW ukazuje Obr. 5.4. Přitom je patrné, že je vždy využita maximální možná informace. Při zarovnání *E. coli* a *L. casei* došlo k výraznému zkrácení prvního signálu, ale zůstaly vhodně zarovnány rostoucí a klesající části signálů. Takové zarovnání se podobá lokálním zarovnání SW algoritmem. U druhé dvojice je signál *E. coli* zkrácen pouze nepatrně, protože oba signály mají přibližně stejnou délku. Zarovnání se tak podobá spíše NW algoritmu pro globální zarovnání. Rostoucí a klesající části signálu jsou taktéž optimálně zarovnány.



Obr. 5.4: Zarovnané signály pomocí DTW

5.2 Výpočet vzdálenosti a shluková analýza

Výpočet vzdálenosti

Zarovnáním komprimovaných signálů jsme splnili požadavek, abychom při měření vzdálenosti mezi signály porovnávali příslušející si genetické informace obou porovnávaných organismů. Při měření vzdálenosti využijeme výhodu, kterou nám poskytla vlnková transformace a tou je redukce ztráty proporcionálních vlastností signálu při kompresi. Dvojice porovnávaných signálů má vždy stejnou délku, ať se jedná o prokaryota nebo eukaryota. Proto je výpočet pro obě skupiny organismů stejný. Vzdáleností 2 signálů tak budeme chápat jejich Euklidovskou vzdálenost, jejíž výpočet je dán vztahem

$$d = \sqrt{\sum_{i=1}^{n} [x(i) - y(i)]^2} \tag{5.3}$$

kde *x(n)* a *y(n)* je dvojice zarovnaných signálů o stejné délce *n* vzorků.

Shluková analýza

Vzdálenosti dvojic signálů můžeme snadno zapsat do trojúhelníkové matice vzdáleností všech porovnávaných organismů. Protože vzdálenosti mohou nabývat velmi rozdílných hodnot je vhodné matici před samotnou hlukovou analýzou normalizovat na rozsah hodnot $\langle 0,1 \rangle$ podělením každého prvku matice největším prvkem.

Na matici následně provedeme shlukovou analýzu pro vytvoření skupin stejných signálů, ze kterých vypočítáme konsenzuální signály reprezentující daný shluk. Takový signál pak bude danou skupinu reprezentovat v bance signálů vytvořené za účelem rychlé klasifikace dalších sekvencí. Vhodné vlastnosti pro ideální rozdělení do shluků v našem případě poskytuje metoda nejvzdálenějšího souseda (CLINK, complete linkage) podmíněná vztahem

$$D(X,Y) = max_{x \in X, y \in Y}\, d(x,y) \tag{5.4}$$

kde vzdálenost shluků X a Y je podmíněna největší vzdáleností prvků x a y náležících do těchto shluků.

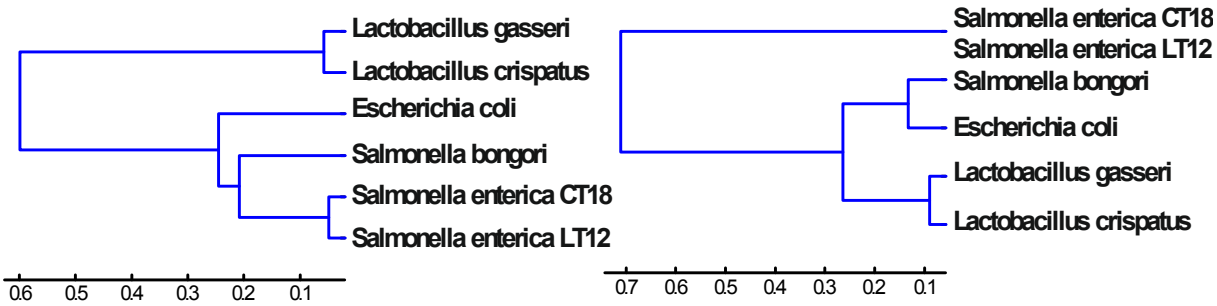
5.3 Srovnání metody

Ještě před závěrečným vyhodnocením klasifikace podle výše popsaného postupu si ukážeme výhodu tohoto postupu oproti zpracování znakových sekvencí na reálném příkladu několika prokaryotních organismů viz Tab. 5.1. Přitom budeme srovnávat stromový graf sestrojený z celogenomových komprimovaných signálů a fylogram sestrojený metodou spojování sousedů (NJ, Neighbor-joining) postavený na 16S rRNA datech.

Tab. 5.1: Testovací organismy

Lactobacillus gasseri
Lactobacillus crispatus
Escherichia coli
Salmonella bongori
Salmonella enterica CT 18
Salmonella enterica LT 12

16S rRNA jsou krátké úseky chromosomální prokaryotické DNA o délce okolo 1500 bp. Obsahují jak konzervované tak i variabilní úseky a vhodně tak dokážou reprezentovat celý organismus. Přesto jejich klasifikace občas selhává jak ukazuje Obr. 5.5.



Obr. 5.5: Dendrogramy klasifikovaných signálů (vlevo) a znakových sekvencí (vpravo)

Při znakovém zpracování nelze rozeznat dva různé kmeny *Salmonella enterica*, protože jejich 16S sekvence je totožná. Zároveň dochází k nesprávné klasifikaci druhu *Salmonella bongori*, který je klasifikován jako příbuznější s *Echerichia coli*. Při zpracování informace z celého chromozómu je klasifikace správná. Navíc délka větví obou stromů je velmi podobná. Přitom strom sestrojený pomocí NJ se snaží prezentovat fylogenetický vývoj právě pomocí délky větví. Toho je ovšem dosaženo při využití matematických modelů, které opět přinášejí vyšší výpočetní náročnost. Zdá se, že signálová reprezentace nese tuto informaci přímo v sobě díky nukleotidovým změnám podél sekvence, které se v signálu kumulují.

5.4 Algoritmus řešení

S pomocí teoretických poznatků z předchozích kapitol byly vytvořeny dvě základní funkce v programovém prostředí MATLAB sloužící ke zpracování dat podle popsaného postupu. Na těchto funkcích bude později zhodnocena funkčnost celého algoritmu.

První funkce *clustersignal.m* slouží k vytvoření banky signálů, na základě které mohou být klasifikovány další organismy. Funkce má 4 vstupní a 2 výstupní parametry uvedené i s popisem v Tab. 5.2 a Tab. 5.3. Celý zdrojový kód funkce je pak dostupný v příloze.

Tab. 5.2: Vstupní parametry funkce clustersignal

parametr	vysvětlení parametru
seq	soubor vstupních sekvencí uložených v proměnné typu cell
nclusters	požadovaný počet shluků
ncompress	stupeň komprese ve smyslu počtu stupňů rozkladu při DWT
type	textová proměnná, 'p' pro výpočet prokaryot, 'e' pro výpočet eukaryot

Tab. 5.3: Výstupní parametry funkce clustersignal

parametr	vysvětlení parametru
bank	banka komprimovaných konsenzuálních signálů, proměnná typu cell
T	vektor hodnot s informací, do které skupiny byl příslušející signál zařazen

Druhou funkcí je *seqclass.m* se 4 vstupními a 1 výstupním parametrem uvedenými v Tab. 5.4 a Tab. 5.5. Tato funkce slouží ke klasifikaci dalších sekvencí na základě zvolené banky signálů. Je důležité zvolit stejný stupeň komprese jaký mají signály v bance. V opačném případě bude vést klasifikace k nesmyslným výsledkům. Celý zdrojový kód je také dostupný v příloze.

Tab. 5.4: Vstupní parametry funkce seqclass

parametr	vysvětlení parametru
seq	vstupní sekvence v proměnné typu string
bank	banka signálů pro klasifikaci
ncompress	stupeň komprese ve smyslu počtu stupňů rozkladu při DWT
type	textová proměnná, 'p' pro výpočet prokaryot, 'e' pro výpočet eukaryot

Tab. 5.5: Výstupní parametry funkce seqclass

parametr	vysvětlení parametru
clust	číslo klástru, do kterého byla sekvence přiřazena

5.5 Banka signálů eukaryot

Banka signálů

Na souboru organismů uvedených v Tab. 5.6 si ukážeme vytvoření banky signálů s různými parametry. Z vypočtených statistik pak budeme moci odhalit vliv komprese a přesnost samotné klasifikace.

Všechny organismy pochází z živočišné říše, na kterou jsme se zaměřili z dříve zmíněných důvodů. Dle systematické biologie můžeme vybrané organismy rozdělit na 3 různé kmeny, 4 různé podkmeny a 5 různých tříd. Díky tomuto rozdělení můžeme statisticky porovnat, zda shlukování signálů odpovídá i biologické hierarchii.

Tab. 5.6: Seznam eukaryot pro banku signálů

Říše	Kmen	Podkmen	Třída	Druh
Animalia	Cnidaria	Cnidaria	Anthozoa	Metridium senile
				Chrysopathes formosa
				Paraminabea aldersladei
	Arthropoda	Chelicerata	Arachnida	Amblyomma triguttatum
				Ixodes persulcatus
				Ixodes uriae
				Rhipicephalus sanguineus
		Hexapoda	Insecta	Dermatobia hominis
				Haematobia irritans
				Hypoderma lineatum
				Philaneus spumarius
	Chordata	Vertebrata	Aves	Anas formosa
				Carduelis sinica
				Meleagris gallopavo
				Pica pica
				Sturnus tristis
			Mammalia	Camelus bactianus
				Equus caballus
				Gorilla gorilla
				Mus musculus
				Rattus norvegicus

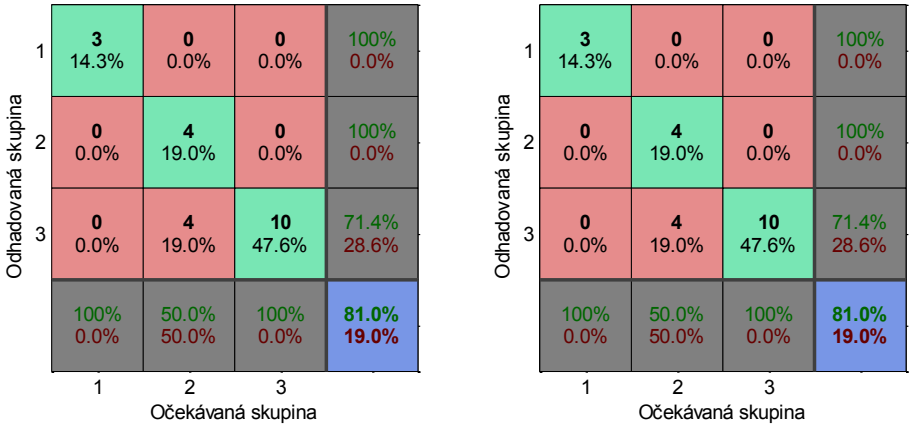
Identifikace kmenů

Při shlukování do 3 skupin budeme očekávat, že tyto skupiny odpovídají skupinám na úrovni kmene dle hierarchie dělení. Kmeny si pro tento účel označíme čísly dle Tab. 5.7.

Tab. 5.7: Značení kmenů eukaryot

kmen	číslo
Cnidaria	1
Arthropoda	2
Chordata	3

Správnost klasifikace při vytváření konsenzuálních signálů můžeme zhodnotit pomocí matic zmatení na Obr. 5.6. Přitom porovnáme rozdíly při shlukování nekomprimovaných signálů (vlevo) a signálů komprimovaných 6 stupňovým rozkladem pomocí DTW (vpravo).



Obr. 5.6: Matice zmatení, 3 skupiny, bez komprese (vlevo), s kompresí (vpravo)

Z matic je patrné, že komprese na úroveň 7. pásma DWT nezanáší do analýzy žádnou chybu, protože obě matice jsou totožné. V obou případech byla každá ze 3 sekvencí ze skupiny 1 (*Cnidaria*) zařazena správně. Dále všechny sekvence ze skupiny 3 (*Chordata*) byly zařazeny správně do této skupiny. Navíc ovšem byly do této skupiny zařazeny i 4 sekvence ze skupiny 2 (*Arthropoda*).

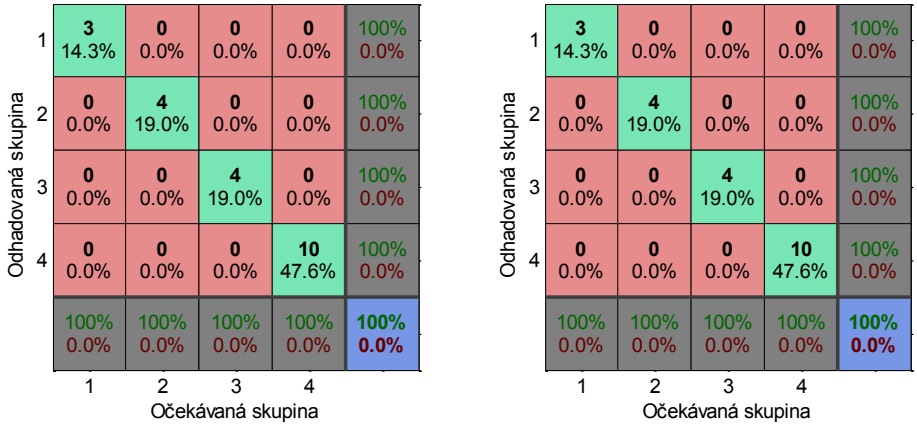
Identifikace podkmenů

Při shlukování do 4 skupin budeme očekávat, že tyto skupiny odpovídají skupinám na úrovni podkmene dle hierarchie dělení. Podkmeny si pro tento účel označíme čísly dle Tab. 5.8.

Tab. 5.8: Značení podkmenů eukaryot

podkmen	číslo
Cnidaria	1
Chelicerata	2
Hexapoda	3
Vertebrata	4

Stejně jako v předchozím případě porovnáme správnost klasifikace pomocí matic zmatení na Obr. 5.7 pro nekomprimované (vlevo) i komprimované signály (vpravo).



Obr. 5.7: Matice zmatení, 4 skupiny, bez komprese (vlevo), s kompresí (vpravo)

Zařazení do první skupiny opět respektuje biologické členění organismů. Navíc respektují rozdělení i všechny ostatní skupiny. To znamená, že klasifikace nerespektující biologické členění v předchozím testu byla způsobena podobností jednoho podkmene ze skupiny *Arthropoda*, který má sekvence podobnější sekvencím podkmene *Vertebrata* skupiny *Chordata*. Obě matice jsou zase totožné, čili komprese do analýzy nezavádí žádnou chybu.

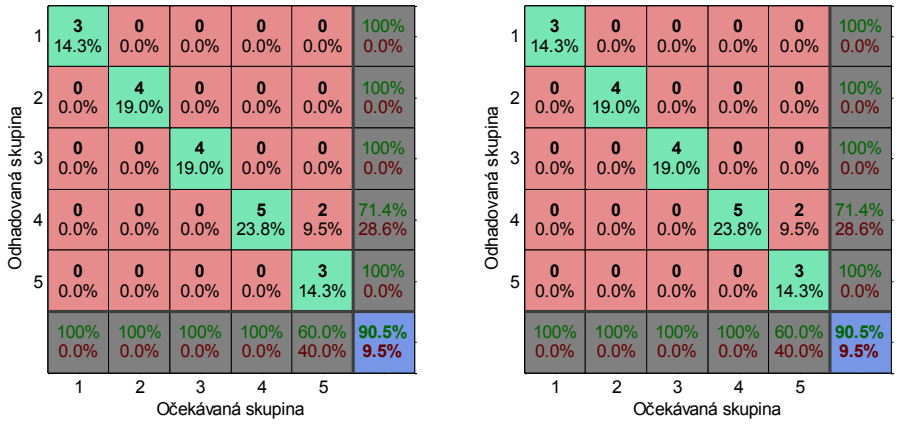
Identifikace tříd

Při shlukování do 5 skupin budeme očekávat, že tyto skupiny odpovídají skupinám na úrovni třídy dle hierarchie dělení. Třídy si pro tento účel označíme čísly dle Tab. 5.9.

Tab. 5.9: Značení tříd eukaryot

třída	**číslo**
Anthozoa	1
Arachnida	2
Insecta	3
Aves	4
Mammalia	5

Opět porovnáme správnost klasifikace pomocí matic zmatení na Obr. 5.7 pro nekomprimované (vlevo) i komprimované signály (vpravo).



Obr. 5.8: Matice zmatení, 5 skupin, bez komprese (vlevo), s kompresí (vpravo)

Klasifikace opět proběhla až na výjimky podle biologické hierarchie. Ovšem 2 savci (*Mammalia*) byli zařazeni do shluku s ptáky (*Aves*). Ani při tomto shlukování nezavádí komprese žádnou chybu.

Diskuse výsledků

Na základě dosažených výsledků můžeme vyvodit hned několik závěrů. I když jsme se z celé skupiny eukaryotních organismů zaměřili pouze na živočišnou říši, stále se jedná o velmi početnou a rozmanitou skupinu organismů, co se jejich systematického členění týče. Navíc toto členění nerespektuje podobnost mitochondriálních sekvencí jednotlivých organismů. Tedy přinejmenším na základě naší použité numerické mapy a algoritmu měření vzdálenosti mezi jednotlivými sekvencemi. Sekvence organismů podkmene *Hexapoda* jsou podle naší analýzy více podobné sekvencím organismů podkmene *Vertebrata* než sekvencím podkmene *Chelicerata*. Přitom *Hexapoda* a *Chelicerata* patří do stejného kmene, zatímco *Vertebrata* do jiného. Snažit se tedy klasifikovat celou živočišnou říši na základě banky signálů je nemožné. Důvodem je hlavně nejasné členění celé této skupiny, které se dost často mění podle aktuálních trendů. Právě signálová reprezentace by se mohla stát takovým trendem, protože dokáže hodnotit jednotlivé skupiny objektivně.

Při podrobnějším shlukování do více skupin je patrné, že organismy systematicky spadající do určité skupiny na nižší úrovni klasifikačního systému mají pořád tendenci vytvářet shluky. Zcela jistě proto můžeme vytvořit banku signálů pro klasifikaci na nižší úrovni systematického dělení živočichů a tato klasifikace bude dobře odpovídat dělení podle systematické biologie.

Posledním a velmi sympatickým závěrem plynoucím z našich výsledků je velká redundance sekvenčního zápisu. Pomocí signálové reprezentace dosahujeme stejných výsledků i při použití 2^6 krát zkrácených signálových reprezentací, jako při použití nezkrácených. Vzhledem k tomu, že klasifikační algoritmus při vytváření matice vzdáleností dosahuje podle velké O notace kvadratické závislosti $O(n^2)$, vede naše komprese k 2^{12} krát

menšímu počtu výpočetních operací, což je velmi výrazné ušetření výpočetního času. Zároveň jsme tím potvrdili, že nastavení míry komprese pomocí procentuální odchylky bylo správné a neplatí pouze pro lidský mitochondriální genom, ale pro mitochondriální genomy zástupců živočišné říše obecně.

5.6 Klasifikace eukaryot bankou signálů

Klasifikaci živočichů dle vytvořených bank signálů provedeme abychom potvrdili, že na základě konsenzuálních signálů jsme schopni zařadit i organismy, které se nepodílely na vytvoření těchto signálů. Proto tuto klasifikaci vyzkoušíme na nových sekvencích získaných z databáze GenBank. Jejich seznam je v Tab. 5.10.

Tab. 5.10: Seznam eukaryot pro klasifikaci

Říše	Kmen	Podkmen	Třída	Druh
Animalia	Cnidaria	Cnidaria	Anthozoa	Nematostella sp. JVK-2006
	Arthropoda	Chelicerata	Arachnida	Damon diadema
				Phytoseiulus persimilis
		Hexapoda	Insecta	Philaenus spumarius
				Diadegma semiclausum
	Chordata	Vertebrata	Aves	Lanius tephronotus
				Larus brunnicephalus
			Mammalia	Pan troglodytes
				Homo sapiens

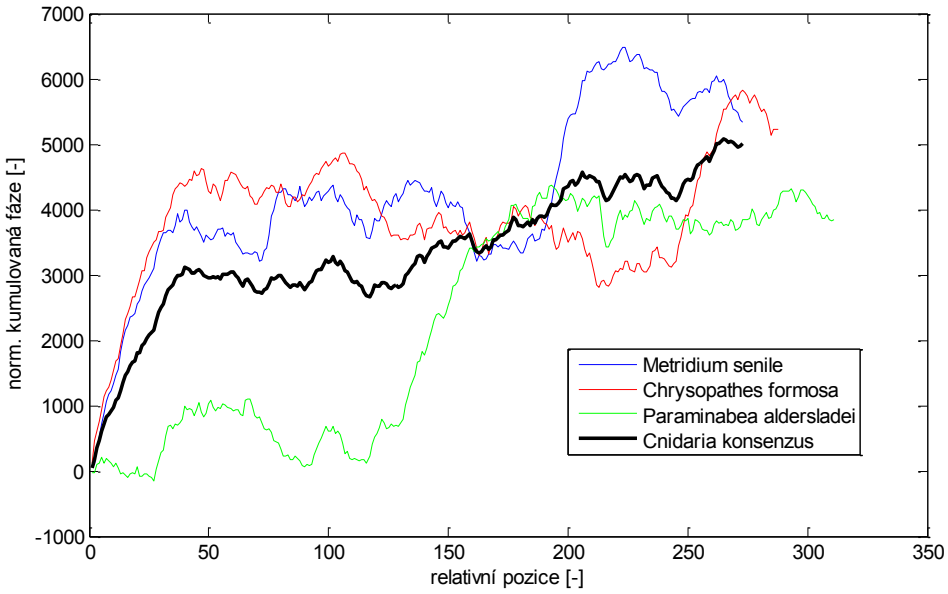
Výsledky shrnuje Tab. 5.11. Protože jsme prokázali že komprese nemá na analýzu vliv, budeme klasifikovat pouze komprimované signály.

Tab. 5.11: Výsledky klasifikace bankou signálů, eukaryota

organismus	3 skupiny		4 skupiny		5 skupin	
	očekávaná	odhad	očekávaná	odhad	očekávaná	odhad
Nematostella sp. JVK-2006	1	1	1	1	1	1
Damon diadema	2	2	2	2	2	2
Phytoseiulus persimilis	2	1	2	2	2	1
Philaenus spumarius	3	3	3	3	3	3
Diadegma semiclausum	3	1	3	3	3	3
Lanius tephronotus	3	3	4	4	4	4
Larus brunnicephalus	3	3	4	4	4	4
Pan troglodytes	3	3	4	4	5	4
Homo sapiens	3	3	4	4	5	4

Nesprávně klasifikované signály jsou v tabulce vyznačeny červeně. Při klasifikaci na kmeny vznikly 2 chyby, klasifikace na podkmeny byla bezproblémová a klasifikace na třídy přinesla další chyby.

Příklad komprimovaných signálů pro skupinu organismů *Cnidaria* a jejich konsenzuální signál, který organismy zastupuje v bance ukazuje Obr. 5.9. Signál *P. aldersladei* je rozdílný od ostatních, vytvořený konsenzuální signál se snaží reprezentovat přibližný signál skupiny organismů.



Obr. 5.9: Konsenzuální signál pro *Cnidaria*

Diskuse výsledků

Při klasifikaci na kmeny se znovu projevuje fakt, že při klasifikaci příliš mnoha odlišných taxonomických skupin je klasifikace na základě zadaných parametrů nefunkční. Chyba klasifikace je zde způsobena pravděpodobně tím, že signály použité při vytváření konsenzu jsou si vzájemně příliš vzdálené.

Pro identifikaci podkmenů pak klasifikace funguje bezchybně. I při vytváření banky byla klasifikace s těmito parametry bezchybná. Konsenzuální signály pro tuto banku tak byly sestrojeny ze správných signálů. Zavádějící banka signálů byla vytvořena pro rozdělení na třídy, kdy se do banky pro skupinu 4 dostaly i dva signály, které měly být klasifikovány do skupiny 5. Tato chyba se projevila špatnou klasifikací dvou posledních signálů.

Pro správnou klasifikaci je velmi důležité, aby vytvořená banka signálů věrně reprezentovala požadované skupiny. Přitom nemusí být vytvořena na základě velkého počtu signálů, ale na signálech, které dobře reprezentují rozdělení do požadovaných skupin.

V našem případě pro rozdělení do 3 a 5 skupin tomu tak nebylo a chyba se projevila při klasifikaci na základě těchto špatných bank signálů.

5.7 Banka signálů prokaryot

Podobnou analýzu jako u eukaryotních organismů nyní provedeme i pro organismy prokaryotní. Jejich klasifikace je složitější úkon, především kvůli velmi rozdílným délkám jednotlivých genomů.

Jejich taxonomie se od eukaryot liší. Budeme tak zkoumat rozdíly na úrovni říše, kmenů a tříd mezi organismy uvedenými v Tab. 5.12. Dynamické borcení časové osy je výpočetně velmi náročný algoritmus a soubor našich testovacích organismů je tak velmi omezený. Ani tak není možné zpracovat nekomprimované signály a budeme proto porovnávat komprimované signály se stupněm rozkladu 7, což je nejmenší stupeň rozkladu, při kterém lze data porovnat na běžném počítači. A se stupněm rozkladu 14, který jsme si definovali pomocí procentuální odchylky komprimovaného signálu od původního. Výpovědní hodnota statistiky je bohužel kvůli omezenému souboru malá, nicméně i tak může ukázat potenciál metody.

Tab. 5.12: Seznam prokaryot pro banku signálů

Říše	**Kmen**	**Třída**	**Druh**
Archaea	Euryarchaeota	Thermococci	Thermococcus gammatolerans EJ3
		Thermococci	Thermococcus sp. 4557
		Thermococci	Pyrococcus furiosus COM1
Bacteria	Proteobacteria	Gammaproteobacteria	Bibersteinia trehalosi
			Mannheimia haemolytica
			Proteus mirabilis HI4320
		Betaproteobacteria	Bordetella pertussis Tohama I
			Acidovorax ebreus TPSY
			Thauera sp. MZ1T
	Fusobacteria	Fusobacteriales	Ilyobacter polytropus DSM 2926
			Leptotrichia buccalis C-1013-b
			Fusobacterium nucleatum ATCC 25586

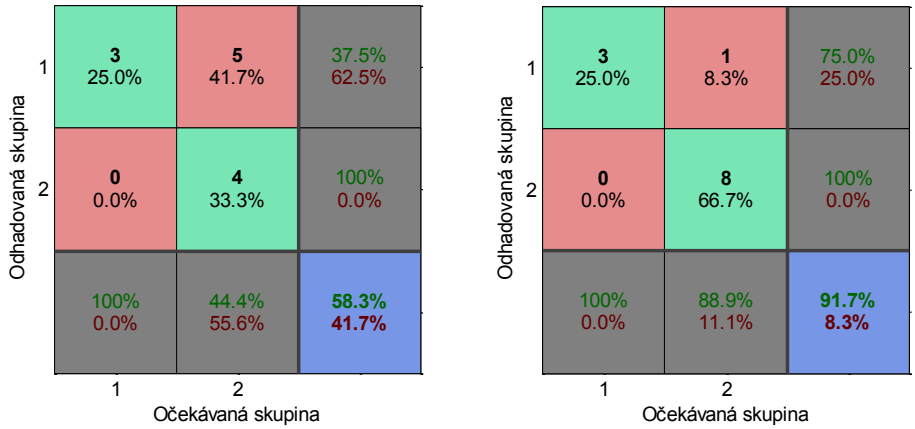
Identifikace říší

Při rozdělení na 2 skupiny bychom podle taxonomického členění očekávali rozdělení do shluků podle říše. Proto si říše označíme dle Tab. 5.13.

Tab. 5.13: Značení říší prokaryot

říše	číslo
Archea	1
Bacteria	2

A výsledky zhodnotíme pomocí matic zmatení na Obr. 5.10.



Obr. 5.10: Matice zmatení, 2 skupiny, komprese 7 (vlevo), komprese 14 (vpravo)

Výsledek je na první pohled velmi překvapivý, protože klasifikace více komprimovaných signálů proběhla lépe. Při 7 stupňovém rozkladu bylo klasifikováno správně 58,3% procenta organismů, při 14 stupňovém 91,7%. Tato analýza ukazuje, že zcela chybně byla v obou případech klasifikována jedna sekvence. Další chyba byla způsobena přesunem celé třídy do jiné skupiny (*výstupní parametr T funkce clustersignal*). Ta se ovšem projevila pouze u méně komprimovaných signálů. Tento fakt je pravděpodobně způsoben dynamickým borcením osy, pro které jsou tak málo komprimované signály příliš dlouhé a algoritmus tak nepracuje optimálně.

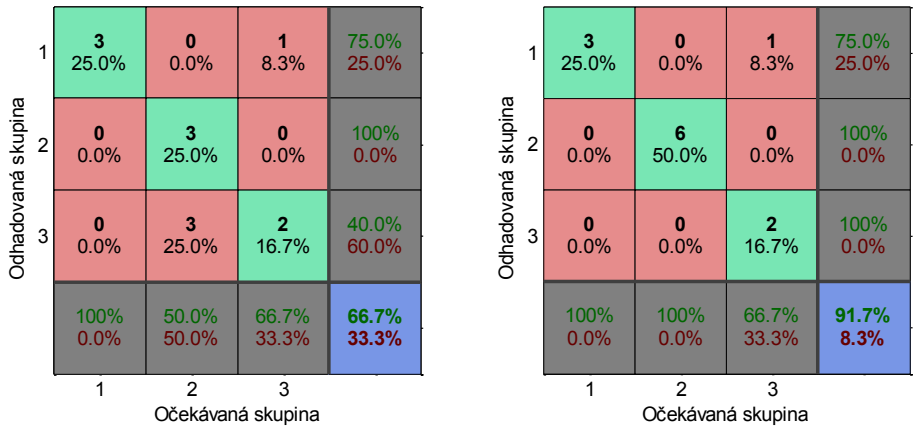
Identifikace kmenů

Rozdělení na 3 skupiny pro nás znamená rozdělení na taxonomické kmeny. Jejich značení je dostupné v Tab. 5.14.

Tab. 5.14: Značení kmenů prokaryot

kmen	číslo
Euryarchaeota	1
Proteobacteria	2
Fusobacteria	3

Rozdělení je zhodnoceno pomocí matic na Obr. 5.11.



Obr. 5.11: Matice zmatení, 3 skupiny, komprese 7 (vlevo), komprese 14 (vpravo)

Protože u nižší komprese došlo ke špatnému zařazení třídy, je klasifikace chybná i na úrovni kmenů. U vyššího stupně komprese je klasifikace dobrá. Obsahuje pouze jednu špatně zařazenou sekvenci.

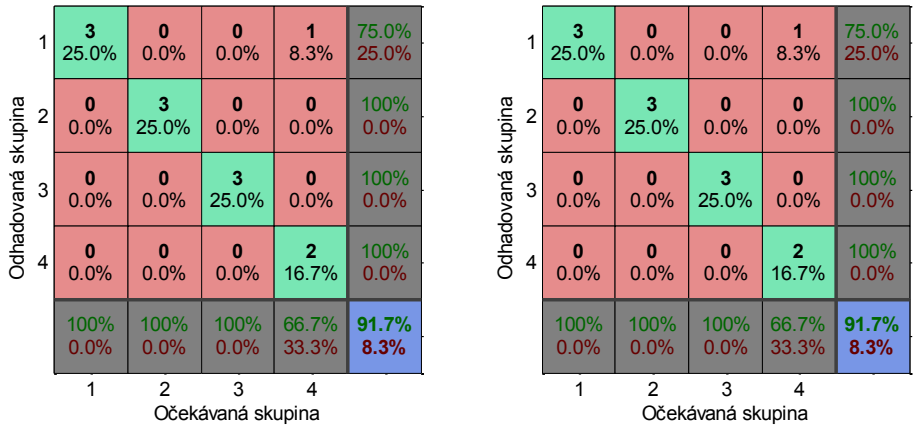
Identifikace tříd

Nakonec provedeme rozdělení na úrovni tříd. Budeme tak požadovat rozdělení do 4 skupin. Ty označíme čísly podle Tab. 5.15.

Tab. 5.15: Značení tříd prokaryot

třída	**číslo**
Thermococci	1
Gammaproteobacteria	2
Betaproteobacteria	3
Fusobacteriales	4

Výsledek analýzy zapsaný v maticích zmatení ukazuje Obr. 5.12.



Obr. 5.12: Matice zmatení, 4 skupiny, komprese 7 (vlevo), komprese 14 (vpravo)

Na úrovni tříd je pro oba stupně komprese klasifikace správná až na jednu problémovou sekvenci.

Diskuse výsledků

Z našich výsledků vyplývá několik myšlenek. Klasifikace prokaryot je mnohem více postavena na molekulárních znacích, na rozdíl od klasifikace eukaryot, která je stále ovlivněna původním hodnocením pomocí znaků morfologických. Ani tak není taxonomie jednoznačná. Některé nižší taxonomické skupiny nejsou dokonce ani v dnešní době klasifikovány do skupin vyšších. Přitom se ale ukázalo, že komprese dat, pro kterou máme optimalizované borcení časové osy, může respektovat taxonomické větvení. Tento výsledek je velmi zajímavý, protože organismy z jedné skupiny mohou mít velmi rozdílné délky genomů, i v řádu miliónů párů bází. To je případ i naší analýzy, a přesto byly sekvence zařazeny do stejných shluků.

Komprese genomických signálů tak může pomoci zjednodušit výpočty vzdáleností mezi jednotlivými organismy, která následně vede ke klasifikaci a která opravdu reprezentuje taxonomické dělení. Při ztrátě malého množství informace tak může vést k lépe interpretovatelným výsledkům než klasifikace dat původních. Přitom je velmi důležité porovnávat příslušející si informace z genomů a algoritmus zarovnání je tak nedílnou součástí celého klasifikačního procesu.

Proces zarovnání odstraňuje problémy s velkým množstvím pohyblivých fragmentů v rámci prokaryotní DNA, protože pouze jedna sekvence byla klasifikována zcela chybně. Pohyblivé fragmenty jsou největším problémem pro sekvenační techniky, proto tato špatně klasifikovaná sekvence může mít také špatnou assembly. Na základě našeho malého souboru testovacích organismů ale toto nemůžeme nijak potvrdit.

Protože zarovnání a klasifikace je v asymptotě kvadratický algoritmus $O(n^2)$, kompresí na úroveň 15. pásma jsme dosáhli ušetření 2^{28} výpočetních kroků. Díky tomu je možné zpracovat celé genomy, což by v původní formě nebylo vůbec možné.

5.8 Klasifikace prokaryot bankou signálů

Klasifikaci bankou si ukážeme opět na jiných organismech než byli použity při vytváření banky signálů. A provedeme ji pouze pro komprimované signály do 15. pásma, kde byla klasifikace do bank méně problematická. Soupis testovaných organismů obsahuje Tab. 5.16.

Tab. 5.16: Seznam prokaryot pro klasifikaci

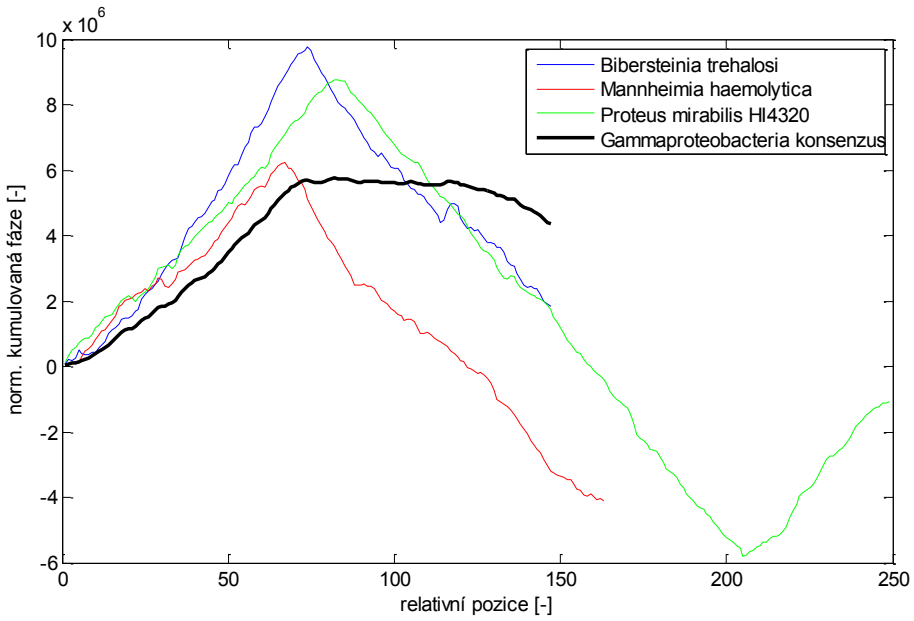
Říše	**Kmen**	**Třída**	**Druh**
Archaea	Euryarchaeota	Thermococci	Thermococcus onnurineus NA1
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea ananatis LMG 20103
		Betaproteobacteria	Bordetella bronchiseptica RB50
	Fusobacteria	Fusobacteriales	Ilyobacter polytropus DSM 2926

Výsledky klasifikace shrnuje Tab. 5.17. Přitom špatně klasifikované případy jsou vyznačeny červeně.

Tab. 5.17: Výsledky klasifikace bankou signálů, prokaryota

organismus	**2 skupiny**		**3 skupiny**		**4 skupiny**	
	očekávaná	**odhad**	**očekávaná**	**odhad**	**očekávaná**	**odhad**
Thermococcus onnurineus NA1	1	2	1	1	1	1
Pantoea ananatis LMG 20103	2	2	2	2	2	2
Bordetella bronchiseptica RB50	2	2	2	2	3	3
Ilyobacter polytropus DSM 2926	2	2	3	1	4	3

Ukázku signálu skupiny *Gammaproteobacteria* ukazuje Obr. 5.13. Ukazuje také jejich konsenzuální signál. Ten se jeví jako hodně deformovaný. Je nutné si uvědomit, že takový signál se snaží o to, aby byl dost specifický pro svou skupinu signálů a dostatečně ji odlišoval od skupin ostatních. Podle výše ukázané klasifikace se mu to spolehlivě daří, i když tvarově neodpovídá žádnému ze signálů, ze kterého byl vytvořen.



Obr. 5.13: Konsenzuální signál pro *Gammaproteobacteria*

Diskuse výsledků

V klasifikaci na základě vytvořený bank se projevila jedna zvláštnost. Testovací organismus říše *Archea* byl správně klasifikován při rozdělení na více skupin. Při rozlišení říší byl ale zařazen nesprávně. To ukazuje, že pro správnou klasifikaci je nutné kromě podobných signálů v cílové skupině, aby ostatní skupiny nebyly vytvořeny z příliš vzdálených signálů. V našem případě signál pro říši *Bacteria* byl vytvořen z příliš odlišných signálů a výsledný pseudosignál je tak náhodou více podobný organismu z druhé říše.

Chybná klasifikace *Fusobacteria* je pravděpodobně způsobena vytvořením konsenzuálního signálů pouze ze dvou signálů. Organismus je tak správně klasifikován pouze na úrovni říše. Je proto také potřeba vytvářet zástupné signály z dostatečného počtu organismů.

Závěr

Cílem této práce bylo zhodnotit využití komprimovaných genomických signálů při klasifikaci a identifikaci organismů. Dílčími úkoly tak byly výběr vhodné signálové reprezentace, která dokáže reprezentovat interdruhové rozdíly a definice vhodné techniky pro kompresi takových signálů, aby mohli být zpracovány informace celých genomů.

V prvních dvou kapitolách jsem se zabýval teoretickými poznatky z genomiky a teorie informace. Definování základních pojmů a osvětlení zvláštností, kterými se různé genomické sekvence vyznačují je důležité pro výběr vhodných sekvencí pro klasifikaci. Základy teorie informace jsou nezbytné pro pochopení, jakým způsobem mohou být data komprimována a proč není možné využít současné algoritmy.

Třetí kapitola obsahuje rešerši numerických map, což jsou postupy konverze znakové sekvence na numerický signál. Zpracoval jsem podrobný přehled technik, které se výrazně liší zejména tvarem signálů, které poskytují. Rozdíly mezi jednotlivými signály jsem demonstroval a vyhodnotil na testovacích sekvencích zvlášť pro prokaryota a eukaryota.

Ve čtvrté kapitole jsem vybral kumulovanou fázi jako ideální signál pro rozlišení jednotlivých organismů. Pomocí frekvenční analýzy těchto signálů jsem odhalil možnosti jejich komprese. Jako ideální techniku jsem nakonec vybral vlnkovou transformaci, jejíž výhodou je rychlý výpočet a normalizace modulu spektra komprimovaného signálu. Míra použitelné komprese se liší pro prokaryota a eukaryota. U obou skupin organismů jsem však aplikoval stejné pravidlo maximálně 1% odchylky komprimovaného signálu od původního. Následně jsem pro porovnávání rozdílně dlouhých signálů prokaryot aplikoval algoritmus dynamického borcení osy pro jejich zarovnání. Toto zarovnání je nezbytné pro správnou klasifikaci. Celá kapitola tedy popisuje nový přístup, jakým mohou být genomická data zpracována.

Poslední část práce je věnována ukázce navrženého algoritmu na reálných datech získaných z veřejné databáze GenBank. Přitom jsou zvlášť vyhodnoceny obě skupiny požadovaných organismů. Zhodnocení obsahuje ověření, zda je míra komprese oprávněná a ztráta informace nevede k nesprávné klasifikaci. Výsledky dokazují, že přiměřená ztráta informace je v tomto případě výhodná, zjednodušuje klasifikaci a poskytuje výsledky lépe odpovídající taxonomickému dělení organismů, jak prokazuje srovnání s metodou klasifikace na základě 16S rRNA. Práce tak ukazuje zcela nové možnosti pro porovnávání celogenomové informace, které nemá v současné době ekvivalent v podobě znakového zpracování z důvodu přílišné výpočetní náročnosti.

Reference

[1] WATSON J.D. AND CRICK F.H.C.. A Structure for Deoxyribose Nucleic Acid. Nature, 1953, roč. 171, čís. 4356, s. 737–738.

[2] SANGER, F. Nobel lecture: Determination of nucleotide sequences in DNA, 1980, Nobelprize.org,

[3] CVRČKOVÁ, F. Úvod do praktické bioinformatiky. Praha : ACADEMIA, 2006, ISBN 80-200-1360-1

[4] THE HUMAN GENOME MANAGEMENT INFORMATION SYSTEM (HGMIS). *About the Human Genome Project* [online]. 18.7.2011. [cit. 2012-11-11]. Dostupné z: http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

[5] DE MAGALHÃES, J. P., FINCH, C. E., & JANSSENS, G. . Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing research reviews*, 2010, *9*(3), 315–23. doi:10.1016/j.arr.2009.10.006

[6] ABEEL, T., SAEYS, Y., BONNET, E., ROUZÉ, P., & VAN DE PEER, Y. (2008). Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Research*, *18*(2), 310–323.

[7] PROVAZNÍK, I.; KUBICOVÁ, V.; ŠKUTKOVÁ, H.; NEDVĚD, J.; TKACZ, E.; BABULA, P.; KIZEK, R. Detection of Short Exons in DNA Sequences Using Complex Wavelet Transform of Structural Features. In *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics 2012*. 2012. s. 1-4. ISBN: 978-1-4673-0491- 7.

[8] JECK, W. R., REINHARDT, J. A, BALTRUS, D. A, HICKENBOTHAM, M. T., MAGRINI, V., MARDIS, E. R., DANGL, J. L., et al. (2007). Extending assembly of short DNA sequences to handle error. *Bioinformatics (Oxford, England)*, *23*(21), 2942–4.

[9] WILLING, E.-M., HOFFMANN, M., KLEIN, J. D., WEIGEL, D., & DREYER, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics (Oxford, England)*, *27*(16), 2187–93.

[10] BRITTEN, RJ; DAVIDSON, EH. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly review of biology* 46**,** 1971, (2): 111–38. doi: 10.1086/406830

[11] SALOMON, D. Data Compression: The Complete Reference. London: Springer Science+Business Media, LLC, 2007, ISBN 1-84628-602-6

[12] Prokaryota. [online]. [cit. 2013-05-06]. Dostupné z: http://cs.wikipedia.org/wiki/Prokaryota

[13] WANG, L., ZHANG, K., & ZHANG, L. (2001). Perfect phylogenetic networks with recombination. *Journal of computational biology a journal of computational molecular cell biology*, *8*(1), 69–78.

[14] ZHOU, J., WU, L., DENG, Y., ZHI, X., JIANG, Y.-H., TU, Q., XIE, J., et al. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal*, *5*(8), 1303–1313.

[15] Živočišná buňka. [online]. [cit. 2013-05-06]. Dostupné z: http://biolog.blog.cz/0706/bunka-rostlinna-a-zivocisna

[16] WUITSCHICK JD, KARRER KM . Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila. J. Eukaryot. Microbiol., 1999, 46 (3): 239–47. doi:10.1111/j.1550-7408.1999.tb05120.x. PMID 103779

[17] VANIN EF . Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet., 1985, 19: 253–72. doi:10.1146/annurev.ge.19.120185.001345. PMID 3909943.

[18] CRISTEA, P. D., Conversion of nucleotides sequences into genomic signals. *Journal of Cellular and Molecular Medicine*, 2002, 6(2), 279–303

[19] ŠEDA, O., LIŠKA, F., ŠEDOVÁ, L., Aktuální genetika: Multimediální učebnice lékařské biologie, genetiky a genomiky. [online]. [cit. 2012-11-18]. Dostupné z: http://biol.lf1.cuni.cz/ucebnice/index.htm

[20] SHANNON, C. E., WEAVER, W., The Mathematical Theory of Communication. Univ of Illinois Press, 1949. ISBN 0-252-72548-4

[21] JONES, N. C., PEVZNER, P. A., An Introduction to Bioinformatics Algorithms. MIT Press, 2004, ISBN 0-262-10106-8

[22] CHEN, X., KWONG, S., LI, M., A. compression algorithm for DNA sequences and its applications in genome comparison,Genome Informatics ,10:52 –61,1999.

[23] P, RAJA RAJESWARI, ALLAM APPARAO a KUMAR V.K. Genbit compress tool (GBC): A java-based tool to compress DNA sequences and compute compression ratio (bits/base) of genomes. *International Journal of Computer Science and Information Technology*. 2010-06-07, vol. 2, no. 3, pp. 181-191. ISSN 09753826. DOI: 10.5121/ijcsit.2010.2313.

[24] GRUMBACH, S., TAHI, F. Compression of DNA sequences, Proc. IEEE Symp. on Data Compression, 340-350, 1993.

[25] GRUMBACH, S., TAHI, F. A new challenge for compression algorithms: genetic sequences, J. Information Processing and Management, 30(6):875-866, 1994.

[26] RIVALS, E., DELAHAYE, J.-P., DAUCHET, M., DELGRANGE, O., A Guaranteed Compression Scheme for Repetitive DNA Sequences, LIFL Lille I University, technical report IT-285, 1995.

[27] HAMORI, E., RUSKIN, J., H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *The Journal of biological chemistry*, 1983, 258(2), 1318–1327.

[28] VUT BRNO, FEKT, UBMI. *Přednášky z předmětu FACS*: *Teorie informace*. Brno, 2011.

[29] ZHANG, C.-T., R. ZHANG a H.-Y. OU. The Z curve database: a graphic representation of genome sequences. *Bioinformatics* [online]. 2003-03-22, roč. 19, č. 5, s. 593-599 [cit. 2012-12-03]. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btg041.

[30] LOBRY, J.R. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 1996, 660-665

[31] GATES,M.A. Simpler DNA sequence representations. Nature, 1985, 316, 219.

[32] LOBRY, J.R. Genomic landscapes. Microbiology Today, 26, 1999, 164-165.

[33] YAU, S. S.-T., WANG, J., NIKNEJAD, A., LU, C., JIN, N., HO, Y.-K. DNA sequence representation without degeneracy. Nucleic Acids Research, 31(12), 2003, 3078–3080.

[34] VOSS R. F., "Evolution of Long-range Fractal Correlations and 1/f noise in DNA base sequences," Physical Review Leters, vol. 68, 1992, pp. 3805-3808.

[35] SEDLÁŘ, K.; ŠKUTKOVÁ, H. Digital Signal Processing Technics for DNA Compression. In *Conference book:Conference on Biotechnology and Biomedicine.* Brno: Masaryk University, 2013. s. 76-77. ISBN: 978-80-210-6200- 9.

[36] SORIMACHI, K. (2009). A Proposed Solution to the Historic Puzzle of Chargaff's Second Parity Rule. *The Open Genomics Journal*, *2*(1), 12–14. doi:10.2174/1875693X00902010012

[37] CRISTEA, P. D. (2003). Large scale features in DNA genomic signals. *Signal Processing*, *83*, 871–888. doi:10.1016/S0165-1684(02)00477-2

[38] ANASTASSIOU, D. Genomic signal processing. Signal Processing Magazine, IEEE, 2001, Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=939833

[39] H. F. Copeland. *The Classification of Lower Organisms*. Palo Alto : Pacific Books, 1956.

[40] VUT BRNO, FEKT, UBMI. *Přednášky z předmětu FMOL*: *Úvod do molekulární biologie*. Brno, 2011.

[41] JAN, J. Číslicová filtrace, analýza a restaurace signálů, VUTIUM, Brno 2002

[42] HRUBEŠ, J.; VÍTEK, M.; KOZUMPLÍK, J. Possibilities of Wavelet Decomposition for SPIHT Compression of ECG Signals. *Analysis of Biomedical Signals and Images,* 2008, roč. 19, č. 1, s. 1-4. ISSN: 1211- 412X.

[43] VUT BRNO, FEKT, UBMI. *Přednášky z předmětu FACS*: *Vlnková transformace*. Brno, 2011.

[44] HAAR, A. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*. 1910, roč. 69, čís. 3, s. 331–371. ISSN 0025-5831

[45] DAUBECHIES, I.: Ten lectures on wavelets, CBMS-NSF conference series in applied mathematics. SIAM Ed, 1992

[46] HRUBEŠ, J.; VÍTEK, M.; KOZUMPLÍK, J. Multipoint Validation of Decompressed ECG Signal. In *Proceedings of Biosignal 2010: Analysis of Biomedical Signals and Images.* Brno: UBMI FEKT VUT v Brně, 2010. s. 1-5. ISBN: 978-80-214-4106- 4.

[47] DURBIN, R., EDDY, S., KROGH, A., AND MITCHISON, G. (1998). Biological Sequence Analysis (Cambridge University Press).

[48] SEDLÁŘ, K.; ŠKUTKOVÁ, H. Classification of prokaryotic organisms based on compressed whole genome signals. In *Sborník z konference: Student EEICT 2013.* Brno: VUT Brno, 2013. s. 185-187. ISBN: 978-80-214-4694- 6.

[49] SMITH, T.F. a M.S. WATERMAN. Identification of common molecular subsequences. *Journal of Molecular Biology*. roč. 147, č. 1, s. 195-197. ISSN 00222836. DOI: 10.1016/0022-2836(81)90087-5

[50] NEEDLEMAN, Saul B. a Christian D. WUNSCH. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. roč. 48, č. 3, s. 443-453. ISSN 00222836. DOI: 10.1016/0022-2836(70)90057-4.

[51] BERNDT, D. & CLIFFORD, J. Using dynamic time warping to find patterns in time series. *New York* 398, 359-370 (1994).

Seznam zkratek

cDNA	komplementární deoxyribonukleová kyselina
CLINK	shluková analýza metodou nejvzdálenějšího souseda
CR	kompresní poměr
DFT	diskrétní Fourierova transformace
DNA	deoxyribonukleová kyselina
DTFT	Fourierova transformace diskrétního signálu
DTW	dynamické borcení časové osy
DTWT	vlnková transformace diskrétního signálu
DWT	diskrétní vlnková transformace
FT	Fourierova transformace
mRNA	mediátorová ribonukleová kyselina
mtDNA	mitochondriální deoxyribonukleová kyselina
NJ	neighbor-joining
PRD	procentuální odchylka signálů
RLE	run-length encoding
RNA	ribonukleová kyselina
WT	vlnková transformace

Seznam příloh

A. Signálové reprezentace 21. lidského chromozómu

B. Zdrojové kódy

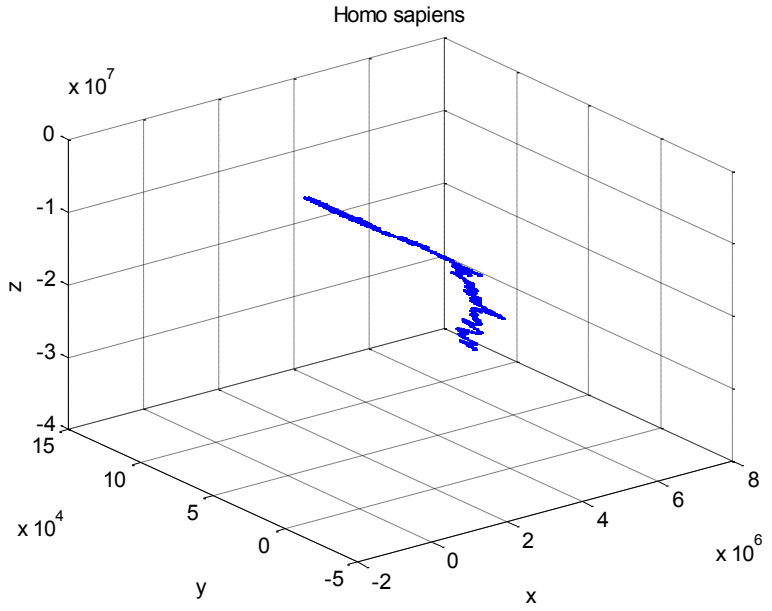
C. Přiložené CD

C 1. elektronická verze diplomové práce

C 2. skripty numerických map

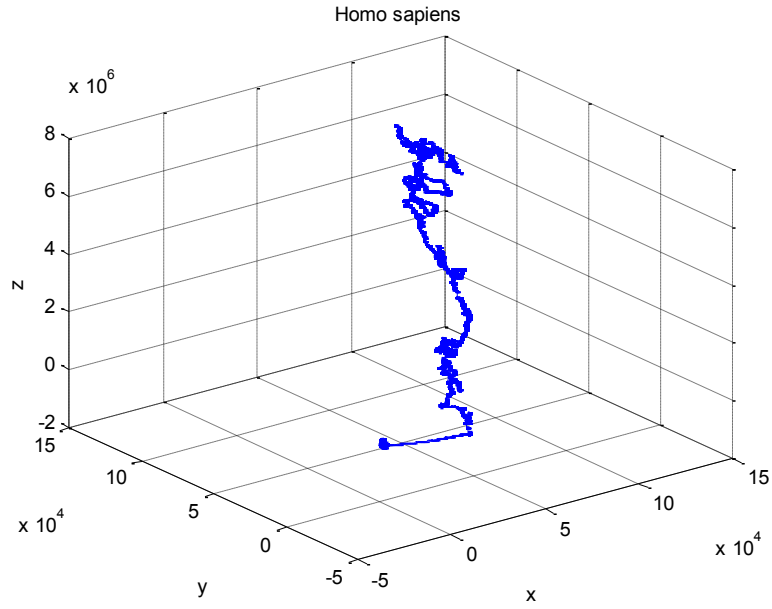
C 3. hlavní funkce a vnořené funkce

A 1. H křivka



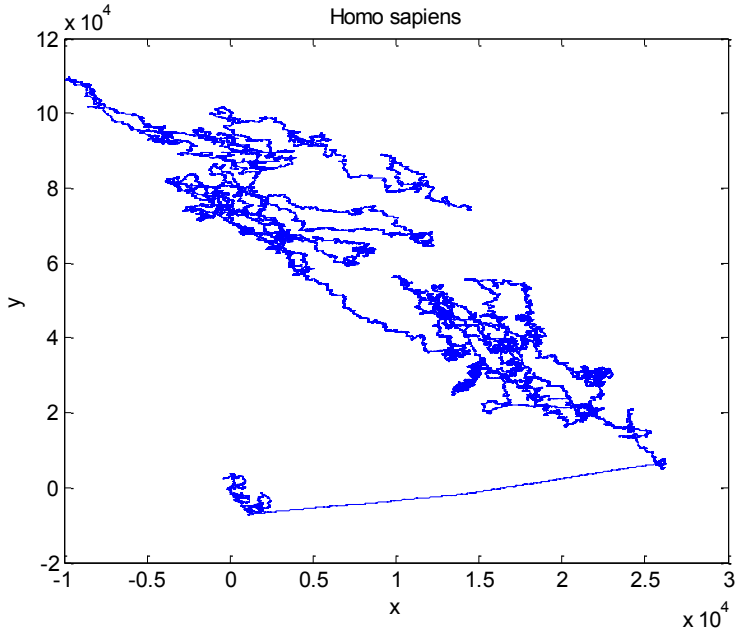
Obr I: H křivka 21. chromozómu

A 2. Z křivka



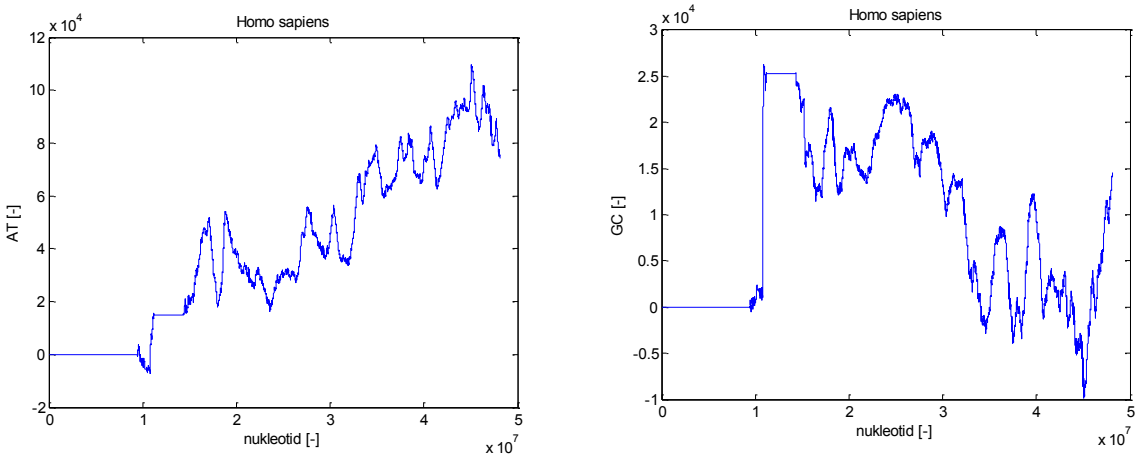
Obr II: Z křivka 21. chromozómu

A 3. DNA walk



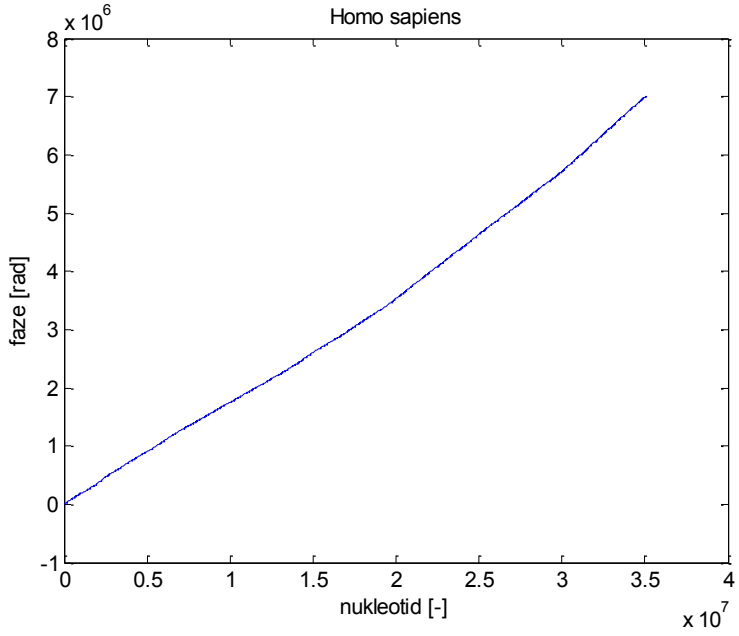
Obr III: DNA walk 21. chromozómu

A 4. DNA walk AT a GC profil



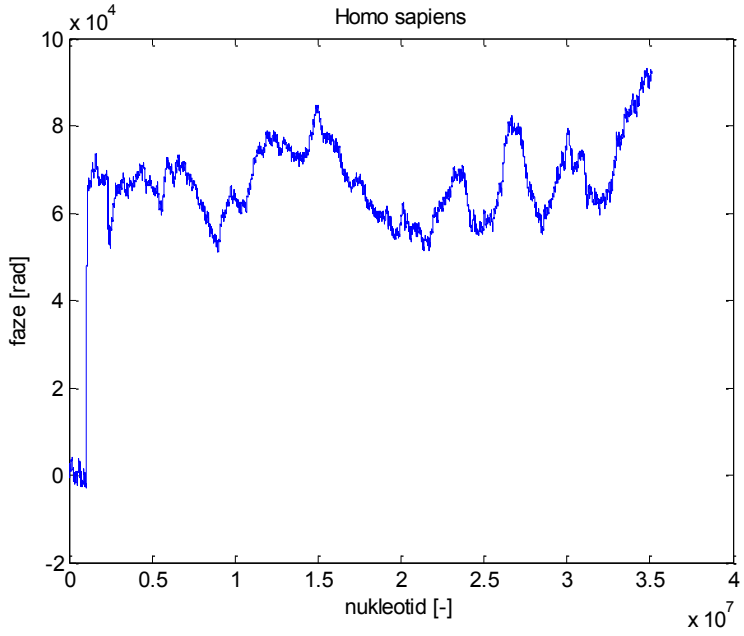
Obr IV: DNA walk 21. chromozómu, AT profil (vlevo), GC profil (vpravo)

A 5. Rozbalená fáze



Obr V: Rozbalená fáze 21. chromozómu

A 6. Kumulovaná fáze



Obr VI: Kumulovaná fáze 21. chromozómu

B 1. clustersignal

```
function [bank,T]=clustersignal(seq,nclusters,ncompress, type)

% vstupni parametry:
% seq...       (cell) soubor vstupních sekvencí
% nclusters... (num) počet požadovaných skupin
% ncompress... (num) stupneň rozkladu DWT pro kompresi
% type...      (str) 'p' pro prokaryota, 'e' pro eukaryota

%výstupní parametry:
% bank...      (cell) banka komprimovaných konsenzuálních signálů pro
                klasifikaci
% T...         (num) číslo skupiny kam byl signál zařazen pro vytvoření
                konsenzu

n=length(seq); % počet sekvencí

for i=1:n
    origsig{i}=cumphase(seq{i});          %převod sekvence na kumulovanou
end                                        fázi

for i=1:n
    sig{i}=comprim(origsig{i},ncompress); %komprese signálu
end

dist=[];

switch type
    case 'e'
        for i=1:n-1
            for j=i+1:n
                currentdist=eukaryotadist(sig{i},sig{j});   %výpočet
                dist=[dist currentdist];                     vzdálenosti
            end
        end
    case 'p'
        for i=1:n-1
            for j=i+1:n
                currentdist=prokaryotadist(sig{i},sig{j});  %výpočet
                dist=[dist currentdist];                     vzdálenosti
            end
        end
end

maxim=max(dist);
norm_dist=dist./maxim;                %normalizace distanční matice

Z = linkage(norm_dist,'complete');    %shlukování
T = cluster(Z,'maxclust',nclusters);  %výběr shluků

for i=1:nclusters
    bank{i}=consense(sig,T,i);        %banka signálů
end
```

B 2. seqclass

```
function [clust]=seqclass(seq, bank, ncompress, type)

% vstupni parametry:
% seq...       (str) vstupních sekvence
% bank...      (cell) banka signálů
% ncompress... (num) stupneň rozkladu DWT pro kompresi
% type...      (str) 'p' pro prokaryota, 'e' pro eukaryota

%výstupní parametry:
% clust...     (num) číslo skupiny kam byl signál zařazen

origsig=cumphase(seq);          %převod sekvence na kumulovanou fázi
sig=comprim(origsig,ncompress); %komprese
n=length(bank);
dist=[];

switch type
    case 'e'
        for i=1:n
            currentdist=eukaryotadist(sig,bank{i}); %vzdálenost od clusteru
            dist=[dist currentdist];
        end
    case 'p'
        for i=1:n
            currentdist=prokaryotadist(sig,bank{i}); %vzdálenost od
            dist=[dist currentdist];                  clusteru
        end
end


[~,clust]=min(dist);    %výběr nejbližšího clusteru
```