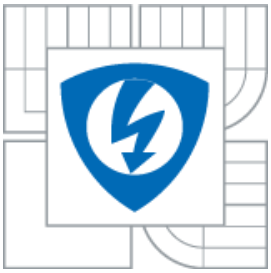
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

POROVNÁNÍ METOD PRO KONSTRUKCI BAREVNÝCH DNA SPEKTROGRAMŮ

COMPARISON OF METHODS FOR RGB SPECTROGRAM CONSTRUCTION OF DNA

DIPLOMOVÁ PRÁCE
MASTER´S THESIS

AUTOR PRÁCE
AUTHOR
Bc. TEREZA REICHLOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR
Ing. VLADIMÍRA KUBICOVÁ

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

Studentka: Bc. Tereza Reichlová ***ID:*** 115112
Ročník: 2 ***Akademický rok:*** 2012/2013

NÁZEV TÉMATU:

Porovnání metod pro konstrukci barevných DNA spektrogramů

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši o numerických reprezentacích DNA sekvencí, o možnostech konstrukce barevných DNA spektrogramů a o vzorech detekovatelných ze spektrogramů. Diskutujte také posloupnost kroků vedoucích k vytvoření barevného spektrogramu. 2) Pojednejte o výhodách a nevýhodách jednotlivých metod pro konstrukci DNA spektrogramů. Vyberte numerické reprezentace vhodné pro vytvoření barevných spektrogramů. 3) Navrhněte programové řešení metod konstrukce spektrogramů v programovém prostředí Matlab. 4) Porovnejte spektrogramy zkonstruované různými metodami a zhodnoťte účinnost a využitelnost řešení.

DOPORUČENÁ LITERATURA:

[1] DIMITROVA, Nevenka, CHEUNG, Yee H. and ZHANG, Michael. Analysis and Visualization of DNA Spectrograms: Open Possibilities for the Genome Research. In: Proceedings of the 14th annual ACM international conference on Multimedia. 2006, pp. 1017-1024. ISBN 1-59593-447-2.
[2] ZHOU, Honxia at al. Detection of tandem repeats in DNA sequences based on parametric spectral estimation. IEEE Transactions on Information Technology in Biomedicine. 2009, vol. 13, pp. 747-55. ISSN 1089-7771.
[3] ANNASTASSIOU, Dimitrij. Frequency-domain analysis of biomolecular sequences. Bioinformatics. 2010, vol. 16, pp. 1073-1081. ISSN 1367-4803.

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

Vedoucí práce: Ing. Vladimíra Kubicová
Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.

předseda oborové rady

Abstrakt

Tato práce pojednává o možnostech konstrukce barevných DNA spektrogramů a o vzorech, které z nich detekujeme. Spektrogramy jako nástroje spektrální analýzy nám umožňují současný pohled na lokální frekvence napříč celou nukleotidovou sekvencí. Jsou vhodné pro identifikaci genů či jejich regionů, určování globálních vlastností celých chromozomů, ale také dávají možnost objevit nové dosud neznámé regiony s potenciálním významem. Za účelem takovéto analýzy DNA lze použít techniky číslicového zpracování signálů. Jejich použití však musí předcházet metody konvertování DNA sekvence do numerické reprezentace. Výběr správné numerické reprezentace ovlivní, jak dobře budou dané biologické vlastnosti reflektovány v numerickém zápisu potřebném pro další použití v analýze zpracování signálů.

Abstract

This thesis discusses about possibilities of construction colour DNA spectrograms and about patterns which can be detected there. Spectrograms as tools of spectral analysis give us a simultaneous view of the local frequency throughout the nucleotide sequence. They are suitable for gene identification or gene regions identification, determination of global character about whole chromosomes and also give us a chance for the discovery of yet unknown regions of potential significance. For purpose of this kind of DNA analysis is possible to use digital signal processing methods. We can apply them on only after conversion of DNA sequence to numerical representation. Selection of correct numerical representation affects how well will be reflected biological features in numerical record which we need for another use in digital signal analysis.

Klíčová slova

DNA sekvence, numerická reprezentace, spektrogram, Fourierova transformace, autoregresní model

Key words

DNA sequence, numerical representation, spectrogram, Fourier transform, autoregressive model

REICHLOVÁ, T. *Porovnání metod pro konstrukci barevných DNA spektrogramů: diplomová práce.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 95 s. Vedoucí diplomové práce Ing. Vladimíra Kubicová.

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma Porovnání metod pro konstrukci barevných DNA spektrogramů jsem vypracovala samostatně pod vedením vedoucí diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

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

V Brně dne 24. května 2013 .....

podpis autorky

Poděkování

Děkuji Ing. Vladimíře Kubicové za velmi kvalitní vedení práce, za čas a ochotu i za poskytnutí důležitých informací a materiálů nutných k její realizaci.

V Brně dne 24. května 2013 .....

podpis autorky

Obsah

1. Úvod

Potřeba technologií zpracovávajících biologické informace se stává více a více naléhavou. Existující výzkumy v bioinformatice se zabývají mnohými typy analýz DNA sekvencí. Hlavními disciplínami jsou zarovnávání sekvencí, vyhledávání specifických genů, kompletování genomu, predikce struktury genů, analýza genové exprese, proteinové interakce a modelování biologické evoluce.

Analýza využívající číslicové zpracování signálů zahrnuje stále se rozvíjející techniky vhodné ke kvalitnímu zpracování množství komplikovaných informací v genomu. Spektrální analýza nabízí nové postupy pro hledání specifických míst v sekvencích DNA, která mohou korespondovat s určitou biologickou funkcí. Nástrojem spektrální analýzy je spektrogram. Jde o dvojrozměrný obraz, v němž jedna souřadnice odpovídá frekvenci a druhá pozici dané báze v sekvenci DNA. Spektrogramy nám umožňují současný pohled na lokální frekvence napříč celou nukleotidovou sekvencí. Jsou vhodné pro identifikaci genů či jejich regionů, určování globálních vlastností celých chromozomů, ale také dávají možnost objevit nové dosud neznámé regiony s potenciálním významem. Hlavní výhodou využití spektrogramů je určitě možnost vizualizace celého chromozomu a eventuálně všech chromozomových nebo genomových vzorů.

Frekvenční analýze ale překáží vyjádření DNA sekvence jako řetězce bází značených prvními písmeny A, C, T, G. Proto jí předchází metody konvertování do numerické reprezentace, kdy přiřadíme každému symbolu jeho odpovídající numerickou hodnotu. Výběr správné numerické reprezentace DNA sekvence ovlivní, jak dobře budou dané biologické vlastnosti reflektovány v numerickém zápisu potřebném pro další použití v analýze zpracování signálů. Hlavní myšlenkou je tedy uvažovat výskyty nukleotidových bází v sekvenci jako individuální číslicové signály a potom je transformovat do frekvenční oblasti.

Tato práce se věnuje optimalizaci dvou metod pro konstrukci barevných DNA spektrogramů a jejich porovnání. Jedná se o metody, které využívají k vykreslení frekvenčních spekter Fourierovu transformaci a autoregresní model.

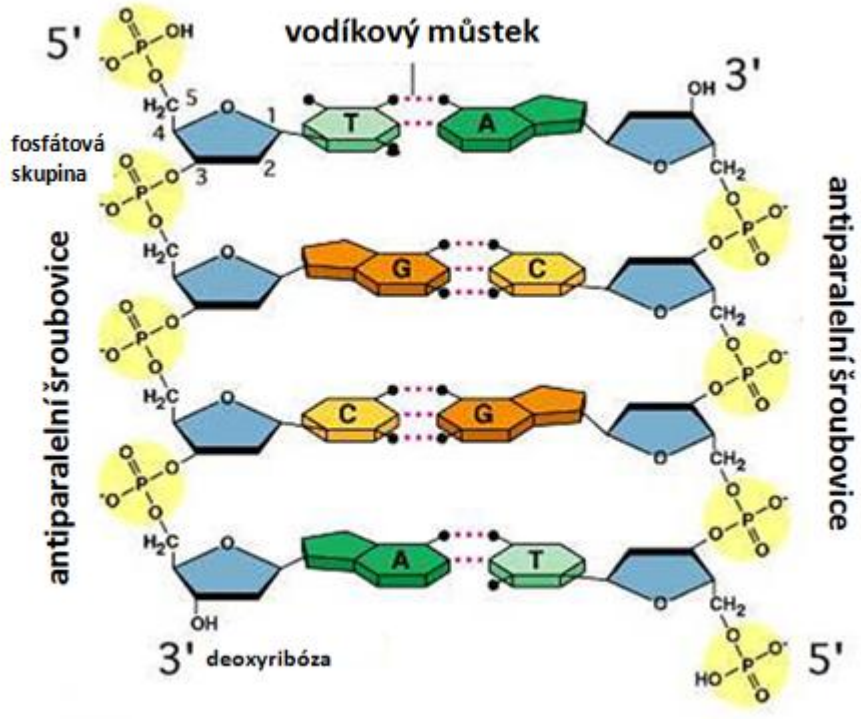
Teoretická část obsahuje literární rešerši o numerických reprezentacích DNA sekvencí, o možnostech konstrukce barevných DNA spektrogramů a o vzorech, které z nich můžeme detekovat.

V programové části se pojednává o výhodách a nevýhodách obou metod pro konstrukci DNA spektrogramů. Řešení metod je navrženo v programovém prostředí Matlab verze R2010a. O využitelnosti obou metod se rozhoduje pomocí vizuální detekce vzorů nacházejících se v daných testovaných sekvencích, které jsou ve spektrogramech vyobrazeny s barevnými odlišnostmi.

2. Deoxyribonukleová kyselina (DNA)

Deoxyribonukleová kyselina, běžně označovaná zkratkou DNA, řídí a udržuje při životě celý organismus vydáváním pokynů buňce pro vytváření základních molekul bílkovin. Je tzv. nositelkou genetické informace všech organismů, látkou pro život nezbytnou. Dlouhé molekuly této nukleové kyseliny jsou uloženy v chromozomech uvnitř buněčného jádra.

DNA je biologická makromolekula – polymer v podobě řetězce nukleotidů. Nukleotidy se skládají z cukru deoxyribózy (monosacharid odvozený z ribózy), fosfátové skupiny (sůl kyseliny fosforečné po odtržení kyselých vodíků) a jedné ze čtyř dusíkatých bází. Právě báze mají informační funkci. Jedná se o adenin (A), guanin (G), cytosin (C) a thymin (T). První dvě patří mezi puriny, zbylé mezi tzv. pyrimidiny. DNA tvoří dvě navzájem spletené protisměrně orientované šroubovice. Mezi protilehlými bázemi obou vláken se vytvářejí vodíkové můstky – tři mezi guaninem a cytosinem a dva mezi adeninem a thyminem (viz *Obr.1*). [7]



Obr.1 Struktura DNA šroubovice [7]

DNA je středem zájmu mnoha vědců z mnoha biologických oborů a to jistě proto, že je zásadním nástrojem pro diagnostiku onemocnění, pro určování otcovství, v kriminalistice při vyšetřování zločinů či v zemědělství v genovém inženýrství. Byly vytvořeny různé promyšlené techniky pro studium DNA – pro její separaci, izolaci, syntézu či klonování. Jednou z biologických věd věnujících se DNA je také

bioinformatika. Zahrnuje metody jako ukládání, vyhledávání a analýzu biologických dat využitím algoritmů, databází, webových informačních systémů či technik umělé inteligence. [7]

Při analýze DNA sekvencí si můžeme DNA představit jako sled za sebou jdoucích bází reprezentovaných jejich prvními písmeny, tedy vstupní data pro analýzu vypadají např. takto: ACCTTTGCATTACAGGTACCTGGGGGTGTGTCTAATAA... Sekvence DNA je tvořena miliardami takovýchto bází. Genová výbava člověka obsahuje přibližně $3{,}2 \times 10^9$ vazebných párů. Kdyby se jejich začátečními písmeny měla popsat jejich struktura, vznikla by kniha s více než 500 000 stranami. Přehledné zobrazení takového množství hodnot je nemožné, i když použijeme pro zobrazení pouze užitečnou – kódující část DNA (3%). [6] Pro analýzu je nutné data upravit do přijatelnější podoby vhodné pro zobrazení. Řešením jsou grafické a numerické reprezentace, které nám poskytují snadný pohled na celou sekvenci, umožňují klasifikaci a porovnání různých živočišných druhů.

3. Numerické reprezentace DNA sekvencí

Numerická reprezentace je vhodný nástroj pro předzpracování genomických dat pro následnou analýzu, tedy pro metody analýz používaných původně ve zpracování elektrických signálů jako je Fourierova transformace, filtrace a další. Numerickou reprezentací rozumíme převod symbolického zápisu bází na numerickou formu podle dopředu stanovené numerické mapy. Numerická mapa (numerické mapování) je konvence, podle které přiřadíme každému symbolu jeho odpovídající numerickou hodnotu. Výběr správné numerické reprezentace DNA sekvence ovlivní, jak dobře budou dané biologické vlastnosti reflektovány v numerickém zápisu potřebném pro další použití v analýze zpracování signálů. Ideální numerická mapa by měla nést stejné množství informace jako sekvence v symbolickém zápisu, neměla by zavádět další informace nad rámec symbolického zápisu, umožňuje rychlé zpracování a je čitelná i pro lidské oko. [8]

Metody numerické reprezentace klasifikujeme do dvou hlavních skupin – metody s pevně stanoveným mapováním a metody založené na chemicko-fyzikálních vlastnostech DNA molekul.

3.1. Metody s pevně stanoveným mapováním

Patří sem binární reprezentace, nukleotidový čtyřstěn, 2D reprezentace reálnými čísly, 2D reprezentace komplexními čísly, 1D reprezentace. Nukleotidy jsou transformovány do série numerických sekvencí určených náhodnými čísly na rozdíl od druhého typu chemicko-fyzikálního mapování, kde přiřazujeme čísla podle různých vlastností DNA molekul. [4]

3.1.1. 4D binární reprezentace (Vossova reprezentace)

Tento typ reprezentace se vyskytuje nejčastěji. Používá se především před zpracováním sekvencí Fourierovou transformací. Metoda vytváří čtyři indikační vektory $u_A(n)$, $u_C(n)$, $u_G(n)$ a $u_T(n)$, které ukazují přítomnost nebo nepřítomnost dané báze na pozici n:

$$u_X(n) = 1 \; pro \; s(n) = X, \tag{1}$$

kde $s(n)$ pro $n = 0,1,...,N-1$ je symbolická sekvence o délce N. Jestliže na pozici n v sekvenci nalezneme znak X, přiřadíme symbolu číslo 1. Je-li nalezen jiný znak, přiřadíme mu hodnotu 0. Takto se projde celá sekvence postupně prohledáváním míst se znaky A, dále se vše opakuje pro znak C, pro znak G a nakonec pro znak T. Např. máme-li sekvenci CCATGTCAAG, jednotlivé indikační vektory jsou: u_A = 0010000110, u_C = 1100001000, u_G = 0000100001 a u_T = 0001010000. [4],[8]

3.1.2. 3D numerická reprezentace redukcí 4D binární reprezentace

4D reprezentaci lze zredukovat bez ztráty informace na 3D reprezentaci. Každé bázi přiřadíme jednotkový 3D vektor směřující ze středu do jednoho ze čtyř vrcholů pravidelného čtyřstěnu (viz *Obr.2*) takto:

$$A = (a_R, a_G, a_B) = (0,0,1)$$

$$C = (c_R, c_G, c_B) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$$

$$G = (g_R, g_G, g_B) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$$

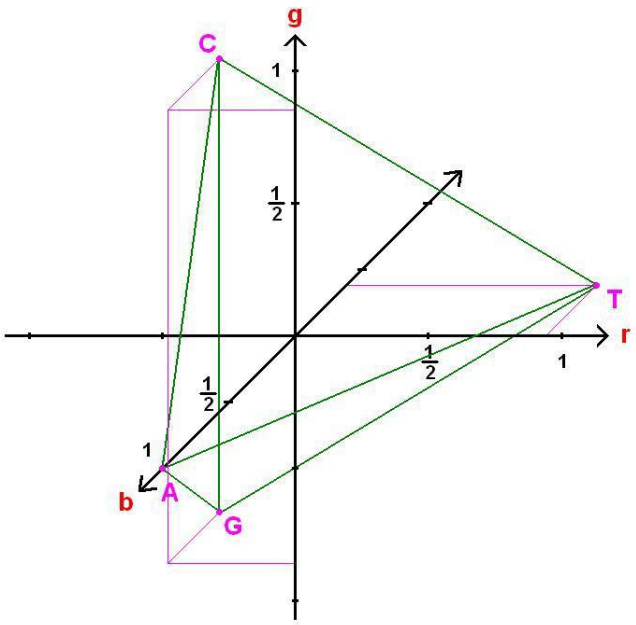
$$T = (t_R, t_G, t_B) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right) \tag{2}$$

4D binární reprezentaci převedeme na 3D beze ztráty informace tak, že DNA je potom reprezentována třemi numerickými sekvencemi x_R, x_G, x_B pro barevné složky RGB pozičně odpovídajícími lineárnímu zápisu symbolické sekvence: [9]

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

$$x_G(n) = \frac{\sqrt{6}}{3}[u_C(n) - u_G(n)]$$

$$x_B(n) = \frac{1}{3}[3u_A(n) - u_T(n) - u_C(n) - u_G(n)] \tag{3}$$

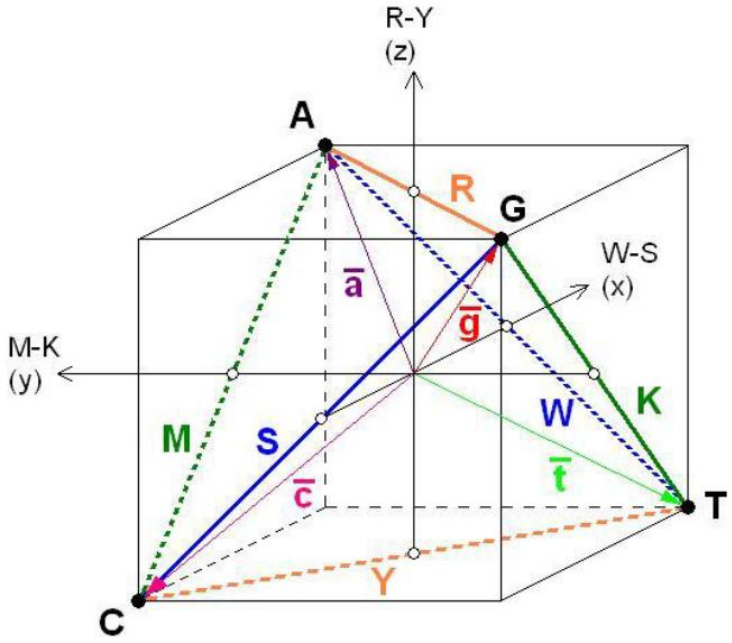


Obr.2 3D numerická reprezentace [8]

3.1.3. Nukleotidový čtyřstěn

Jde o grafickou prezentaci numerické formy tvořené čtyřmi vektory symetricky rozloženými ve 3D prostoru. Vektory jsou orientovány k vrcholům čtyřstěnu. Každá z šesti hran potom odpovídá jedné ze tříd zahrnující pár nukleotidů se stejnou chemickou vlastností, což je znázorněno na *Obr.3*. Tato metoda tedy zachovává důležité chemické vlastnosti bází, jako jsou:

- molekulární struktura – báze A a G patří mezi puriny (R), báze C a T patří mezi pyrimidiny (Y)
- síla vazby – mezi bázemi A a T se tvoří dva vodíkové můstky – jde o vazbu slabou (W) a mezi bázemi C a G se tvoří tři vodíkové můstky – jde tedy o vazbu silnou (S)
- obsah radikálů – báze A a C obsahují amino skupinu NH_3 (M), báze T a G obsahují keto skupinu C=O (K)



Obr.3 Nukleotidový čtyřstěn [8]

Výsledná reprezentace je třírozměrná a osy souřadného systému popisujeme rovnicemi:

$$x = W - S,\ y = M - K,\ z = R - Y. \qquad (4)$$

Báze jsou přiřazeny k vrcholům čtyřstěnu tak, aby jejich vzdálenost od středu byla rovna jedné, a jsou reprezentovány vektory ve tvaru:

$$\vec{a} = \vec{\imath} + \vec{\jmath} + \vec{k},\ \vec{c} = -\vec{\imath} + \vec{\jmath} - \vec{k},\ \vec{g} = -\vec{\imath} - \vec{\jmath} + \vec{k},\ \vec{t} = \vec{\imath} - \vec{\jmath} - \vec{k}. \qquad (5)$$

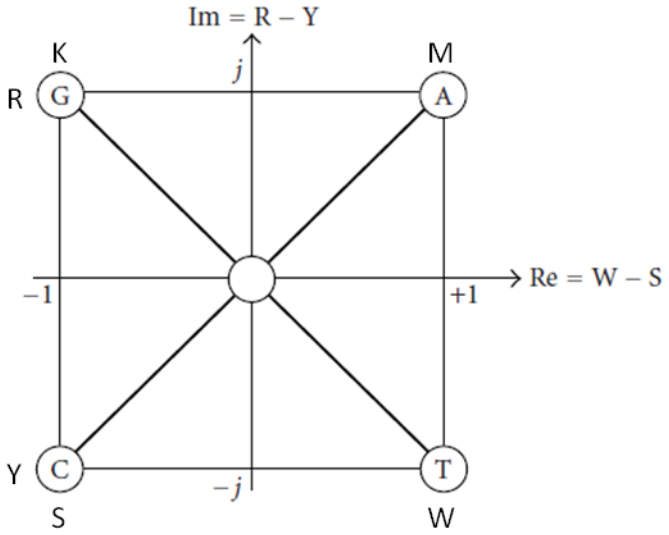
Střed čtyřstěnu je potom stanoven jako bod [0,0,0] a souřadnice nukleotidů jsou:

$$A = [1,1,1];\ C = [-1,1,-1];\ G = [-1,-1,1];\ T = [1,-1,-1]. \qquad (6)$$

[10]

3.1.4. 2D reprezentace komplexními čísly

Pro porovnávání sekvencí jsou výhodnější 2D úpravy. Předchozí 3D reprezentaci tedy zredukujeme do 2D reprezentace projekcí nukleotidového čtyřstěnu do vybrané 2D plochy. Výběr této plochy je podmíněn sledovanými parametry. Například chceme-li sledovat odlišnosti typu S-W (silná-slabá vazba) a Y-R (pyrimidiny-puriny), tak provedeme projekci do x-z plochy.



Obr.4 Redukce nukleotidového čtyřstěnu do 2D reprezentace [10]

Popis nukleotidů je pak v komplexní rovině:

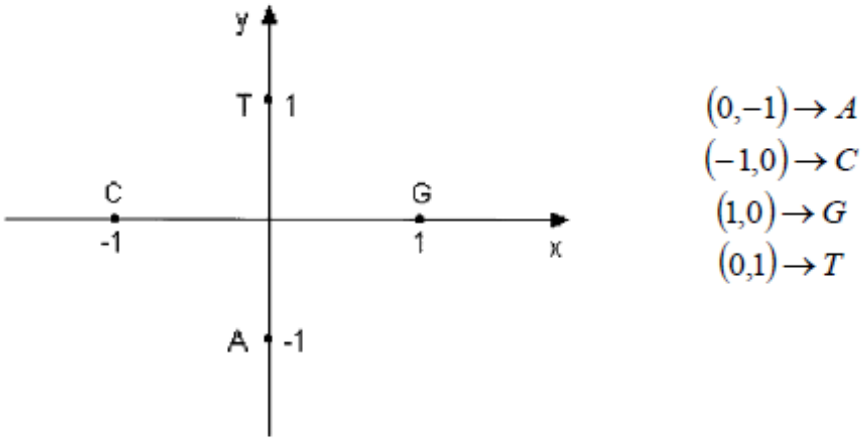
$$A = 1 + j, C = -1 - j, G = -1 + j, T = 1 - j. \qquad (7)$$

Pro nukleotid adenin tedy platí, že patří mezi puriny (v imaginární části má kladné znaménko), vytváří slabou vazbu (v reálné části má kladné znaménko) a obsahuje amino skupinu. Podobně z obrázku *Obr.4* vyčteme vlastnosti zbylých nukleotidů. [10]

Mapování nukleotidů komplexními čísly se využívá také pro fázovou analýzu DNA sekvencí.

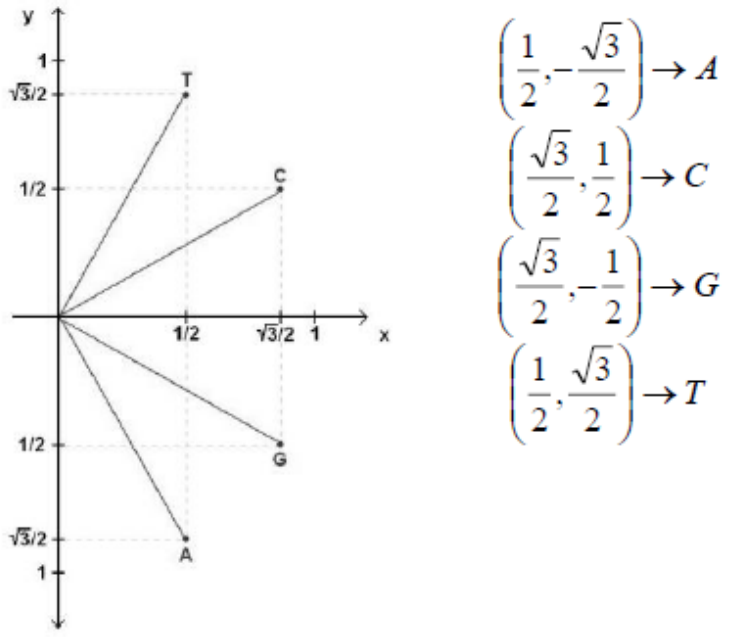
3.1.5. 2D reprezentace reálnými čísly

Tato metoda je pro porovnání sekvencí, tak jako reprezentace komplexními čísly, výhodnější než 3D reprezentace, ale za cenu částečné ztráty informací o chemických vlastnostech nukleotidů. Jak vidíme na *Obr.5*, osu x reprezentuje silná vazba, osu y slabá vazba, kladná znaménka na obou osách znamenají obsah keto skupiny, záporná naopak určují obsah amino skupiny, v prvním kvadrantu se nacházejí pyrimidiny a ve čtvrtém kvadrantu jsou puriny.



Obr.5 2D reprezentace reálnými čísly [8]

Reprezentaci ve všech kvadrantech můžeme zredukovat na reprezentaci v prvním a čtvrtém kvadrantu (viz *Obr.6*). Tato je vhodnější než předchozí, protože nezpůsobuje degeneraci sekvence, tzn., že nemůže dojít k tomu, aby měly různé sekvence stejnou grafickou reprezentaci (např. AGTC, AGTCA a AGTCAG). [8],[11]



Obr.6 2D reprezentace reálnými čísly v prvním a čtvrtém kvadrantu [8]

3.1.6. 1D reprezentace

Reprezentaci sekvencí nukleotidů můžeme dále redukovat až do 1D reprezentace. Jde o nejjednodušší reprezentaci, kdy používáme mapování DNA bází reálnými čísly:

$$\{A, C, G, T\} \rightarrow \{0, 1, 2, 3\}. \quad (8)$$

Máme celkem 4! = 24 možností, jak přiřadit číslice 0, 1, 2 a 3 bázím A, C, G, T. Jako nejvhodnější přiřazení se v literatuře uvádí: T = 0, C = 1, A = 2, G = 3. [10]

3.2. Metody založené na chemicko-fyzikálních vlastnostech DNA molekul

Při těchto metodách jsou pro mapování použity chemické a fyzikální vlastnosti DNA sekvencí. Jde o mapování metodami EIIP, atomovým číslem, numerickým párem, DNA-Walk modelem a reprezentací digitálním Z-signálem. [4]

3.2.1. Metoda EIIP

EIIP je zkratka pro metodu Electron-Ion Interaction Potential. Metoda využívá k reprezentaci distribuci energií volných elektronů podél DNA sekvence. Jednotlivé nukleotidy jsou zastupovány danými fyzikálními hodnotami energií:

$$A = 0.1260, C = 0.1340, G = 0.0806, T = 0.1335. \quad (9)$$

Tato metoda se používá pro detekci kódujících a nekódujících míst v sekvenci DNA. Uvádí se, že poskytuje lepší rozlišení těchto míst než často používaná 4D binární reprezentace. [15] Máme-li např. sekvenci AATGCATCA, výsledný vektor bude vypadat takto:

$$X = [0.1260, 0.1260, 0.1335, 0.0806, 0.1340, 0.1260, 0.1335, 0.1340, 0.1260].$$

3.2.2. Metoda atomového čísla

DNA sekvence je reprezentována vektorem atomových čísel nukleotidů, kdy pro jednotlivé nukleotidy byly stanoveny hodnoty:

$$A = 70, C = 58, G = 78, T = 66. \quad (10)$$

Jde o metodu používající se pro konkrétní techniky. Např. byla aplikována ve výzkumu radiačních rezistentních-opravných genů. Každé bázi studované sekvence daného

genu byla přiřazena hodnota podle vztahu 10. Výsledná numerická reprezentace byla základem pro statistickou analýzu. [4],[21]

3.2.3. Metoda numerického páru

Nukleotidy jsou mapovány ve smyslu komplementarity bází, kdy A a T je přiřazena hodnota +1 a C a G přiřazujeme hodnotu -1. Tato metoda také uvažuje známý fakt, že introny jsou bohaté na báze A a T, zatímco exony obsahují častěji báze C a G. [4] DNA sekvence potom může být reprezentována dvěma způsoby:

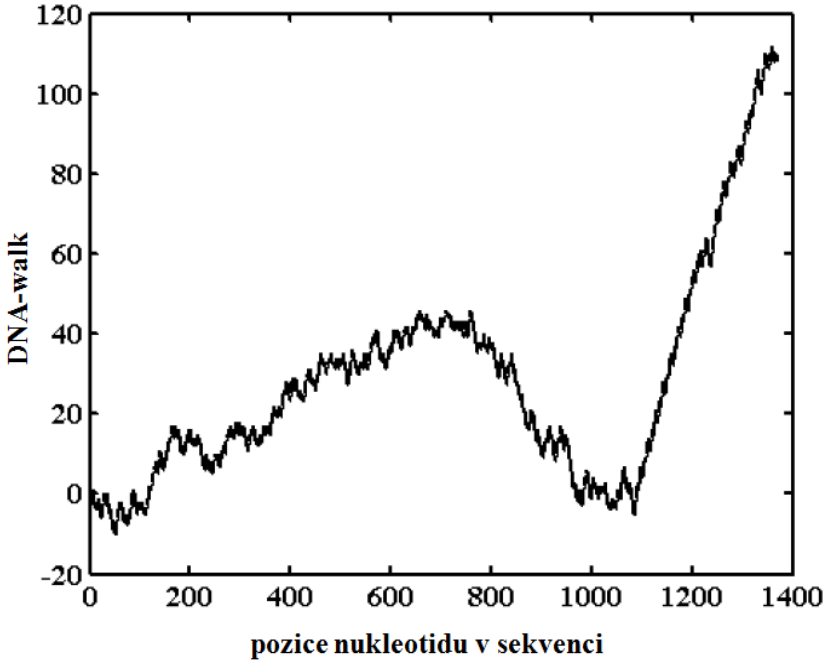
Příklad pro sekvenci CGAT: $X_1 = [-1, -1, 1, 1]$

$X_2 = [-1, -1, 0, 0]$ a $[0, 0, 1, 1]$.

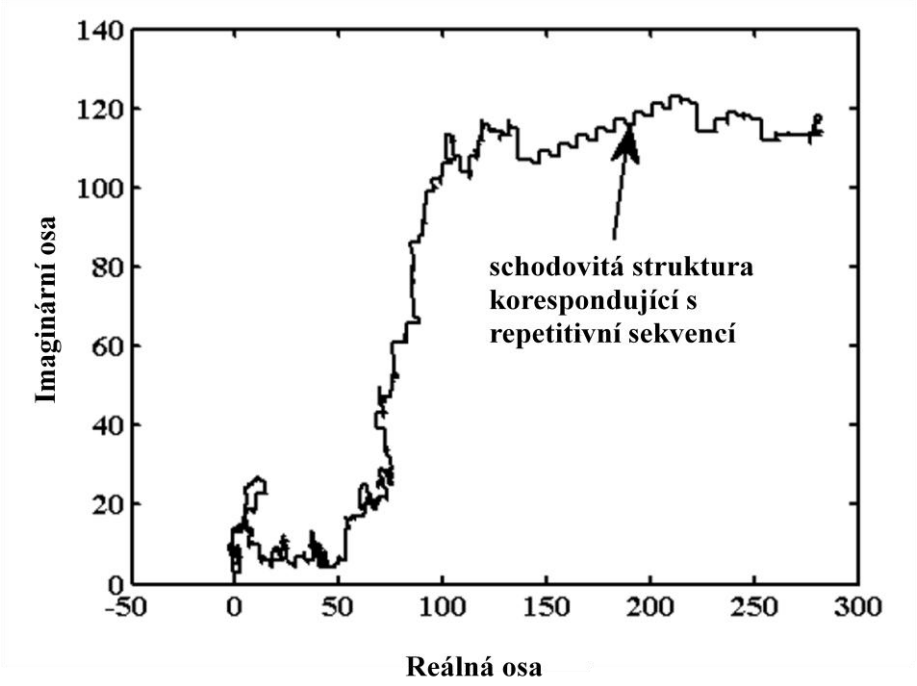
3.2.4. Metoda DNA-Walk

Jde o grafickou metodu, kdy vykreslujeme kroky podél osy x čtením sekvence nukleotid po nukleotidu. Pro pyrimidiny (C a T) platí krok směrem nahoru (+1), pro puriny (A a G) směrem dolů (-1). Ve výsledném vektoru jsou za sebou zapsány hodnoty pro dané báze tak, že se průběžně přičítá nebo odčítá hodnota pro danou pozici k předchozí hodnotě. Takto je popsána metoda pro jednodimenzionální kroky (viz *Obr.7*). Existuje také metoda vícedimenzionálních kroků. Na *Obr.8* vidíme metodu dvojdimenzionálních kroků. [4],[12]

Např. pro sekvenci CGAT je výsledný vektor: X = [1, 0, -1, 0].



Obr.7 Metoda DNA-walk pro jednodimenzionální kroky [18]



Obr.8 Metoda DNA-walk pro dvojdimenzionální kroky [18]

3.2.5. Reprezentace digitálním Z-signálem

Digitální Z-signál rozkládá DNA sekvenci do tří digitálních signálů tzv. Z-křivek. Je to vlastně třídimenzionální křivka určená třemi vektory (Z-křivkami). Tyto tři vektory Δx_n, Δy_n, Δz_n nabývají pouze hodnot +1 a -1. DNA sekvence je nejprve převedena do binární reprezentace (Vossovy reprezentace) do čtyř vektorů $x_A(n)$, $x_C(n)$, $x_G(n)$ a $x_T(n)$, potom jsou vypočítány Z-křivky pomocí převodu:

$$\begin{bmatrix}\Delta x_n \\ \Delta y_n \\ \Delta z_n\end{bmatrix} = 2 \times \begin{bmatrix}1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}x_A(n) \\ x_C(n) \\ x_G(n) \\ x_T(n)\end{bmatrix} - \begin{bmatrix}1 \\ 1 \\ 1\end{bmatrix}. \qquad (11)$$

Komponenty Δx_n, Δy_n, Δz_n zobrazují distribuci purinů versus pyrimidinů, amino versus keto skupin, slabých versus silných bází v celé sekvenci. [5],[13]

3.3. Výhody a nevýhody jednotlivých metod

Numerické reprezentace DNA sekvencí s následným použitím metod číslicového zpracování signálů nabízejí možnost odhalit skryté periodicity, nukleotidové distribuce a jiné znaky sekvencí, které nemohou být odkryty jinými technikami. Každá ze zmíněných reprezentací má své výhody i nevýhody.

Nejprve k metodám s pevně stanoveným mapováním. Vossova binární reprezentace je nejvhodnější pro použití ve spektrální analýze DNA sekvencí, stejně tak může být dobře

použita reprezentace nukleotidovým čtyřstěnem. Např. po vykreslení výkonnostního spektra můžeme detekovat kódující a nekódující části sekvence. Naproti tomu 1D reprezentace reálnými čísly může zavádět matematické vlastnosti, které neexistují v sekvencích, a proto je u nich použití technik číslicového zpracování signálů limitováno. Reprezentace komplexními čísly zahrnuje v sobě vlastnosti komplementarity bází, což je výhodou této metody.

Obecně o metodách založených na chemicko-fyzikálních vlastnostech DNA molekul můžeme říci, že obsahují méně redundantních informací než druhá skupina metod. Metoda EIIP stejně jako metoda atomového čísla mohou být vyjádřeny vektorem podobným vektoru Vossovy binární reprezentace. Ale na rozdíl od binární reprezentace může mít metoda EIIP lepší rozlišovací schopnost pro vyhledávání genů v genomech a může snížit výpočetní nároky až o 75%. [15] Metoda atomového čísla je nedávno vzniklá technika, která potřebuje hlubší prozkoumání pro zjištění jejího potenciálu. Stejně tak jako metoda numerického páru, která má ale jistou známou výhodu. Tou je snížení výpočetní náročnosti při signálovém zpracování díky redukci do dvou hodnot pro čtyři báze. Metoda DNA-walk je určitě vhodná pro vizualizaci korelací a nukleotidových záměn ve velkém rozsahu, ale lze ji použít jen pro krátké sekvence o stovkách párů bází. [4] Stoupající části křivky svědčí o přítomnosti pyrimidinů a klesající části zase o větší koncentraci purinů. U metody dvojdimenzionálních kroků můžeme v obrázcích identifikovat repetitivní sekvence, což se projevuje jako schodovité struktury. A poslední metoda za použití Z-křivek ve snadno postřehnutelné formě zobrazuje ve svém zápisu informace o chemických vlastnostech bází. V 3D obrázku Z-křivky můžeme také odhadnout místa předpokládané replikace. Tato reprezentace se nepoužívá pro spektrální analýzu. Některé literatury uvádí, že tato metoda je výhodnější než spektrální analýza – je graficky přehlednější než výkonnostní spektrum či samotný barevný spektrogram. [18]

4. Spektrální analýza

Nástrojem spektrální analýzy je spektrogram. Jde o dvojrozměrný obraz, v němž jedna souřadnice odpovídá frekvenci a druhá času, mluvíme-li o časově-frekvenční analýze. V případě analýzy DNA sekvencí časová osa odpovídá pozici dané báze v sekvenci. Barva nebo úroveň jasu odpovídá amplitudě odpovídajících koeficientů spekter. Signál (sekvence DNA) je rozdělen na úseky o délce N vzorků z celkového počtu M vzorků odpovídajícímu délce celé sekvence. Jednotlivé úseky (=délka okna) se mohou překrývat. U každého úseku stanovíme jeho spektrum. Získaná spektra zobrazíme ve sledu bází. Pozorovací interval je určen kompromisem mezi požadavkem na dostatečnou rozlišovací schopnost ve frekvenční oblasti (kdy frekvence je nepřímo úměrná délce okna) a snahou dobře rozlišit pozice bází v sekvenci DNA (kdy minimální rozlišitelný rozdíl je

úměrný délce okna). Jinak řečeno se vzrůstající délkou okna se zvyšuje frekvenční rozlišení spektrogramu a klesá rozlišení pozic bází v sekvenci. [16]

4.1. Analýza DNA pomocí spektrogramů

Spektrální analýza je stále se rozvíjející metoda pro systematické hledání specifických míst v sekvencích DNA, která mohou korespondovat s určitou biologickou funkcí. Spektrogramy nám umožňují simultánní pohled na lokální frekvence napříč celou nukleotidovou sekvencí. Jsou vhodné pro identifikaci genů či jejich regionů, určování globálních vlastností celých chromozomů, ale také dávají možnost objevit nové dosud neznámé regiony s potenciálním významem. Existuje několik výhod využití spektrogramů. Tou hlavní je určitě možnost vizualizace celého chromozomu. Např. lidský chromozom 1 je dlouhý 150 miliónů párů bází. Sledovat takto dlouhou sekvenci bází adeninu, guaninu, cytosinu a thyminu v lineárním zápise není moc praktické a neodhalí nám mnoho z vlastní struktury ani jiné skryté informace. Naproti tomu v barevném spektrogramu můžeme vyčíst pohodlně vlastnosti celého chromozomu. Např. opakující se vzory a oblasti s neobvyklými nukleotidovými kompozicemi jsou ve vykresleném frekvenčním spektru DNA sekvence viditelné jako místa s ostrými kontrasty oproti pozadí v daném spektrogramu. Světlé čáry můžeme identifikovat jako místa, kde se objevují opakující se vzory. Obrázky potom mohou být generovány s různým rozlišením a proměnnou velikostí okna. [1],[14]

Základní myšlenkou je uvažovat výskyty jednotlivých bází v sekvenci DNA jako čtyři individuální číslicové signály, které získáme vybranou numerickou reprezentací a potom je transformovat do frekvenční oblasti. [1]

4.2. Postup pro vytvoření barevného DNA spektrogramu

První příklad algoritmu pro vytvoření barevného DNA spektrogramu se skládá z následujících kroků:

1) konvertování DNA sekvence do numerické reprezentace
2) výpočet frekvenčního spektra (výkonnostního spektra)
3) mapování DFT hodnot do RGB prostoru
4) normalizace hodnot pixelů

Existují další algoritmy s jiným pořadím těchto kroků, např. 1,3,2,4 nebo 1,2,4,3. Tyto budou popsány dále v kap. 4.2.5.

4.2.1. Konvertování DNA sekvence do numerické reprezentace

Nejčastěji používanou numerickou reprezentací pro vytváření spektrogramů je 4D binární reprezentace definující pro každou bázi vektory $u_A(n)$, $u_C(n)$, $u_G(n)$ a $u_T(n)$. Vektory nabývají hodnot 1 pro pozice, kde se nachází daný nukleotid a hodnot 0 pro pozice, kde nukleotid chybí. (viz kap. 3.1.1)

Pro vytvoření výkonnostního spektra je ekvivalentní metodou k 4D reprezentaci 3D numerická reprezentace, která popisuje DNA třemi vektory $x_R(n)$, $x_G(n)$, $x_B(n)$. (viz kap. 3.1.2) Tato metoda vede přímo k získání barevného DNA spektrogramu, z něhož vyčteme informace o lokálních frekvencích bází. Ty byly vypočteny pomocí superpozice tří barevných vektorů.

Hlavní aplikací pro komplexní reprezentaci je detekce exonů, tedy kódujících míst v sekvenci DNA, a predikce genů. Ve výkonnostním spektru je možno lépe rozeznat exony na rozdíl od spektra vytvořeného z 4D reprezentace. Je to proto, že komplexní reprezentace v sobě zahrnuje vlastnosti bází, které jsou představovány matematickým zápisem. Číselné hodnoty báze C a G jsou si podobnější než báze A a T, což je důležité z hlediska toho, že exony obsahují větší počet C a G. (viz kap. 3.1.4)

1D reprezentace (viz kap. 3.1.6) není vhodná pro spektrální analýzu pomocí AR modelu. Je to proto, že vektor hodnot autoregresního členu a vektor hodnot aktuálního členu mají stejnou lineární závislost a stejné parametry.

Vzhledem k tomu, že reprezentace nukleotidovým čtyřstěnem zachovává důležité chemické vlastnosti bází, je usuzováno, že tato metoda může zlepšit detekci vzorů v DNA ve spektrogramech. (viz kap. 3.1.3)

Reprezentace vytvořená metodou EIIP a metodou atomového čísla má podobná výsledná výkonnostní spektra. Srovnáme-li tyto metody s 4D binární reprezentací, mají lepší rozlišovací schopnost a nižší výpočetní nároky, jak je zmíněno v kap. 3.3.

Co se týká reprezentace metodou numerického páru, je ekvivalentní ke komplexní reprezentaci. [18]

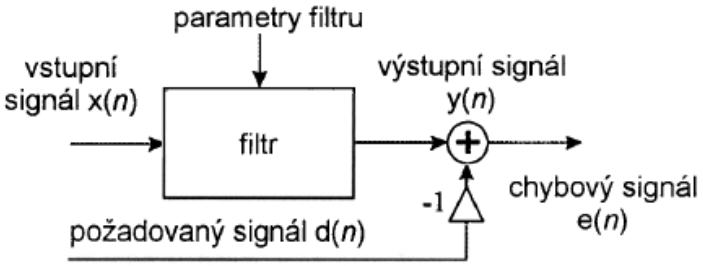
Ostatní metody jsou zhodnoceny v kap. 3.3.

4.2.2. Výpočet frekvenčního spektra

K výpočtu frekvenčních spekter signálů jsou obecně k dispozici parametrické metody a neparametrické metody. Druhé jmenované jsou založeny na pásmových filtrech pro zaznamenaný signál. Patří k nim i Fourierova transformace. Ta slouží pro převod signálů z časové oblasti do oblasti frekvenční. Signál může být buď ve spojitém či diskrétním čase.

Aplikace parametrických metod bývá v literatuře často označována jako Wienerova filtrace. Cílem je nalezení parametrického modelu systému popsaného ve tvaru racionální

lomené přenosové funkce. Tomu odpovídá zapojení na *Obr.9*. Hledáme tedy koeficienty racionální lomené přenosové funkce *H(f)* filtru, který je buzen procesem *x(n)* tak, aby na výstupu filtru byl signál *y(n)*, který se blíží požadovanému procesu *d(n)*. Nalezení parametrů posuzujeme podle minima chybové funkce *e(n)*. Metody jsou tedy založeny na výpočtu autoregresního modelu časové řady vzorků, které jsou považovány za výstup lineárního dynamického systému se vstupním signálem typu bílého šumu. Spektrum výstupního signálu je pak dáno frekvenční přenosovou funkcí této soustavy a faktorem jejího zesílení, který je dán rozptylem vstupního bílého šumu. [34]



Obr.9 Wienerův filtr [34]

Diskrétní Fourierova transformace (DFT)

Základem pro výpočet frekvenčních spekter DNA spektrogramu je diskrétní Fourierova transformace (DFT). Bývá často nazývána krátkodobou diskrétní Fourierovou transformací (STFT). Jako STFT se označuje Fourierova transformace aplikovaná na analyzovanou funkci postupně po krátkých úsecích, které vybírá pomocí reálného symetrického okna. Vlastní výpočet se často v praxi realizuje rychlými algoritmy FFT (rychlá Fourierova transformace). Jedná se o efektivní výpočet DFT a označení FFT nemá z hlediska jejích vlastností (kromě rychlosti) opodstatnění.

DFT je definovaným algoritmem nad jistými vektory čísel. Transformace je funkcí frekvence, kdy koeficienty diskrétního spektra přísluší jistým frekvencím a vzorky signálu jistým časovým okamžikům (v případě analýzy DNA pomocí spektrogramu jde o jisté pozice bází jdoucích v sekvenci za sebou). [16]

Frekvenční spektrum každé báze získáme dosazením vektoru numerické reprezentace do vzorce pro výpočet DFT:

(12)

$$U_X[k] = \sum_{n=0}^{N-1} u_X\,[n] e^{-j\frac{2\pi}{N}k\,n},$$

kde $k = 0,1,...,N/2$ a $X = A,C,T,G$.

Symbol *N* označuje délku okna, tedy část sekvence, pro kterou se počítá DFT; *k* je koeficient spektra z intervalu 0 až N/2, který určuje pořadí vzorků v kmitočtové oblasti; *j* je

imaginární číslo; *n* značí pořadí jednotlivých oken jdoucích v čase za sebou; $u_X(n)$ je *n*-tá hodnota v indikační sekvenci vymezená oknem *N*. [1],[16]

Výsledné výkonnostní spektrum pro všechny báze se vypočítá podle vzorce:

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2. \quad (13)$$

[3],[25]

Autoregresní (AR) model

Jde o čistě rekurzivní model. Je vyjádřen parametrickou rovnicí frekvenčního spektra signálu. Koeficienty této rovnice lze vypočítat několika různými algoritmy. V AR modelu je hodnota signálu v čase *t* vyjádřena lineární kombinací předchozích hodnot (dopředná predikce), kombinací následujících hodnot (zpětná predikce) nebo kombinací obojího (dopředně-zpětná predikce). Jako dopředná predikce vstupního signálu *x(n)* je definován vztah:

$$x(n) = \tilde{x}(n) + e(n) \quad (14)$$

$$\tilde{x}(n) = -\sum_{k=1}^{p} a_k x(n-k),$$

kde *p* je řád AR modelu, a_k jsou parametry modelu, *e(n)* reprezentuje náhodnou chybu a $\tilde{x}(n)$ je odhadovaný signál. Výkonovou spektrální hustotu neboli výkonnostní spektrum (PSD) AR modelu na výstupu soustavy určíme jako součin výkonnostního spektra na vstupu a druhé mocniny absolutní hodnoty frekvenční přenosové funkce:

$$P_{AR}(\omega) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} a_k exp(-j\omega k)\right|^2}, \quad (15)$$

kde $P_{AR}(\omega)$ je výkonnostní spektrum při úhlové frekvenci ω, σ^2 je celkový výkon signálu bílého šumu a součin $(-j\omega k)$ je frekvenční přenosová funkce. [2],[26]

Prvním krokem výpočtu je odhad parametrů a_k AR modelu. Existují čtyři hlavní algoritmy pro tento odhad – Yule-Walkerova metoda, Burgova metoda, kovarianční metoda (známá také jako metoda nejmenších čtverců) a modifikovaná kovarianční metoda. Burgova metoda je výpočetně nejefektivnější. Algoritmus funguje tak, že ze vztahů 14 je potřeba vypočítat náhodné chyby *e(n)* podle přesných rovnic a ve výsledku minimalizovat součet dopředných a zpětných chyb. Hlavním principem je odhad rozdílu mezi signálem a ideálním modelem signálu, který je formulován rovnicemi 14. [2],[27]

Dalším krokem je výpočet řádu AR modelu. Řád *p* není nijak globálně znám, musíme jej vypočítat v souladu s charakteristikami dat. Příliš nízká hodnota řádu má za

následek přehnaně vyhlazený spektrální odhad, zatímco příliš vysoká hodnota řádu způsobuje falešné píky ve spektru. Existuje mnoho kritérií, podle kterých se vybírá hodnota řádu. Patří k nim finální predikce chyby (FPE), Akaikeho informační kritérium (AIC), korigované Akaikeho informační kritérium (AICc), minimální popis délky (MDL) – což je to samé jako Bayesovo informační kritérium (BIC) a posledním je Hannan-Quinnovo kritérium (HQC).
Jsou definovány vztahy:

$$FPE(p) = V_p \frac{N+p}{N-p}, \tag{16}$$

$$AIC(p) = log(V_p) + \frac{2p}{N}, \tag{17}$$

$$AICc(p) = log(V_p) + \frac{2(p+1)}{N-p-2}, \tag{18}$$

$$BIC(p) = MDL(p) = log(V_p) + \frac{p\ log\,(N)}{N}, \tag{19}$$

$$HQC(p) = log(V_p) + \frac{2p\ log\,[log\,(N)]}{N}, \tag{20}$$

kde V_p je ztrátová funkce pro modelový *p*-tý řád – normalizovaný součet druhých mocnin náhodných chyb *e(n)*, tedy rozdíl mezi odhadovaným $\tilde{x}(n)$ a vstupním signálem *x(n)*. *N* je počet vzorků (délka sekvence). Když kriteriální funkce dosáhne minima, korespondující hodnota *p* je považována jako nejlepší řád. Vždy je ale potřeba si získané hodnoty ověřit experimentálně.

Je známo, že vypočtená spektra pomocí Fourierovy transformace (FT) jsou nepřesná, mohou obsahovat falešné píky a mají slabé rozlišení. AR model nemá tento problém, což bylo dokázáno mnohým měřením, jak uvádí literatura. [2],[26]

Mezi další metody číslicového zpracování signálů používané v analýze DNA spektrogramů patří AMFD funkce (Average Magnitude Difference Function) a algoritmus TDP (Time Domain Periodogram). Tyto metody produkují ekvivalentní výsledky jako FT nebo AR model, ale dosáhnou jich jen za použití oken s velmi malou délkou. Jsou výhodnější pro predikci takových exonů v sekvenci DNA, které jsou krátké a rozmístěny v těsné vzdálenosti. [20]

4.2.3. Mapování DFT hodnot do RGB prostoru

Spektra čtyř binárních vektorů $U_A[k]$, $U_C[k]$, $U_G[k]$ a $U_T[k]$ hodnot DFT zredukujeme do tří sekvencí v RGB prostoru pomocí rovnic:

$$X_r[k] = a_r|U_A[k]| + t_r|U_T[k]| + c_r|U_C[k]| + g_r|U_G[k]|$$

$$X_g[k] = a_g|U_A[k]| + t_g|U_T[k]| + c_g|U_C[k]| + g_g|U_G[k]|$$

$$X_b[k] = a_b|U_A[k]| + t_b|U_T[k]| + c_b|U_C[k]| + g_b|U_G[k]|, \tag{21}$$

kde (a_r, a_g, a_b), (t_r, t_g, t_b), (c_r, c_g, c_b) a (g_r, g_g, g_b) jsou barevné mapovací vektory pro báze A, T, C, G. Výsledná barva pixelu $X_{r,g,b}[k]$ je dána superpozicí mapovacích vektorů barev váhovaných frekvenční složkou báze na dané pozici. Pro barevné mapování je obecně doporučeno, aby barevné mapovací vektory byly zvoleny jako vrcholy pravidelného čtyřstěnu (viz *Obr.2*). [1],[14]

Hodnoty jsou:

$$\begin{aligned} a_r &= 0 & t_r &= 0.911 & c_r &= 0.244 & g_r &= -0.817 \\ a_g &= 0 & t_g &= -0.244 & c_g &= 0.911 & g_g &= -0.471 \\ a_b &= 1 & t_b &= -0.333 & c_b &= -0.333 & g_b &= -0.471 \end{aligned} \quad (22)$$

Používají se i jiné hodnoty. [3] Např.:

$$\begin{aligned} a_r &= 0 & t_r &= 0.943 & c_r &= -0.471 & g_r &= -0.471 \\ a_g &= 0 & t_g &= 0 & c_g &= 0.816 & g_g &= -0.816 \\ a_b &= 1 & t_b &= -0.333 & c_b &= -0.333 & g_b &= -0.333 \end{aligned} \quad (23)$$

Nebo nejčastěji:

$$\begin{aligned} a_r &= 0 & t_r &= 1 & c_r &= 0 & g_r &= 0.333 \\ a_g &= 0 & t_g &= 0 & c_g &= 1 & g_g &= 0.333 \\ a_b &= 1 & t_b &= 0 & c_b &= 0 & g_b &= 0.333 \end{aligned} \quad (24)$$

Výběr vhodných barevných mapovacích vektorů závisí na tom, co chceme ze spektrogramů identifikovat. Většinou se jejich volba řeší prakticky při tvorbě a testování nových algoritmů.

4.2.4. Normalizace hodnot pixelů

Před vykreslením spektrogramu musíme RGB hodnotu každého pixelu $X_{r,g,b}$ normalizovat na rozsah 0-1. Můžeme použít klasickou normalizaci pro každou barevnou složku zvlášť pomocí rovnice:

$$X_{r,g,b}(i,j) = \frac{X_{r,g,b}(i,j) - min\ [X_{r,g,b}]}{max\ [X_{r,g,b}] - min\ [X_{r,g,b}]}. \quad (25)$$

4.2.5. Záměna pořadí kroků

Algoritmy pro vytvoření barevných DNA spektrogramů mohou být i s jiným pořadím kroků, např. 1,3,2,4 nebo 1,2,4,3. Zaměníme-li základní pořadí kroků na 1,3,2,4, změní se počet vektorů vstupujících do kroku výpočtu spektra (2) ze čtyř na tři. Je to proto, že po prvním kroku (konvertování DNA sekvence do numerické reprezentace) následuje

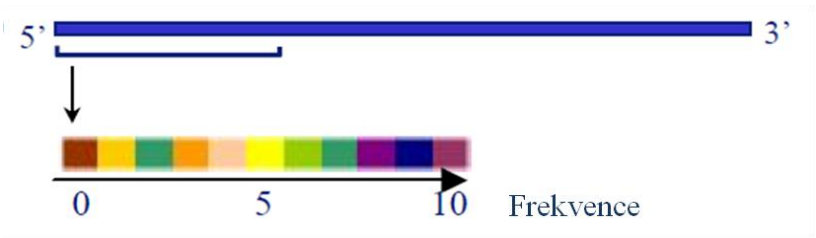
jako druhý krok redukce 4D reprezentace do 3D, tedy mapování hodnot do RGB prostoru (3). Posledním krokem je potom normalizace hodnot pixelů (4).

Zaměníme-li pořadí na 1,2,4,3, vstupují do kroku výpočtu spektra (2), stejně jako v základní verzi pořadí kroků, čtyři vektory. Změna je ale v poslední části, kdy do kroku mapování hodnot do RGB prostoru (3) vstupují čtyři vektory hodnot normalizovaných na rozsah 0-1. Z posledního kroku (3) budou tedy k dispozici hodnoty mimo rozsah 0-1 a to bude mít vliv na vykreslení RGB spektrogramu. V literatuře se nejvíce používá algoritmus s pořadím kroků 1,2,3,4. [1],[14],[32]

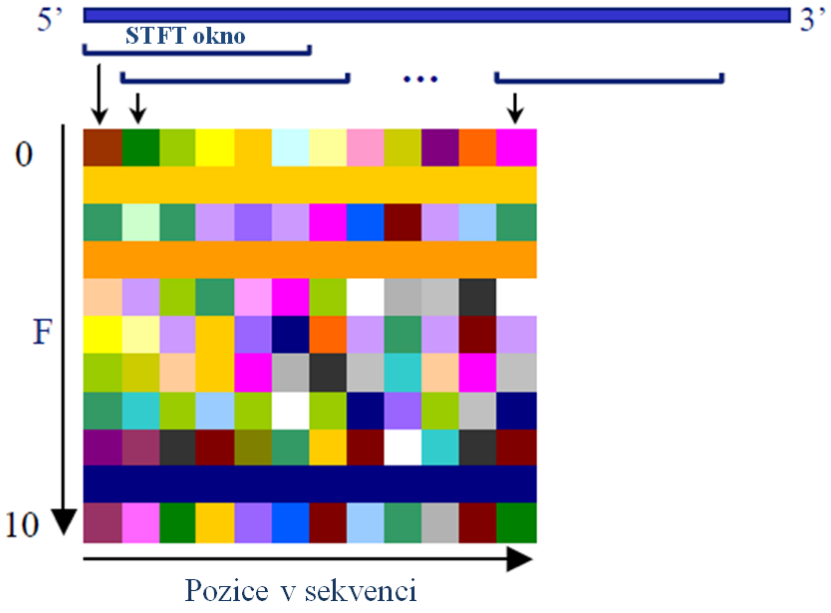
4.2.6. Finální vytvoření spektrogramu

K samotnému vytvoření spektrogramu potřebujeme okno s danou délkou, které se pohybuje po sekvenci dat. Spektrum DNA segmentu je ukázáno na *Obr.10*. Je tvořeno pásem za sebou jdoucích barevných pixelů. Pro dlouhé DNA sekvence opakujeme kroky 1) – 4) algoritmu pro vykreslení spektrogramu se zvolenou velikostí okna podél celé sekvence. Okna se mohou překrývat. Horizontální pásy (segmenty DNA) jsou skládány vertikálně vedle sebe, přičemž každý pás je definován frekvenčním spektrem daného segmentu DNA (viz *Obr.11*).

Vzhled spektrogramu je ovlivněn výběrem velikosti okna, zvolením délky překrývajících se sekvencí mezi sousedními okny a použitými hodnotami barevných mapovacích vektorů. Velikost okna ovlivňuje frekvenční rozlišení a velikost pixelů reprezentujících dané báze. Hledáme-li např. v sekvenci DNA opakující se vzory, je lepší použít okno několikrát větší než je délka repetitivního vzoru, ale menší než je velikost regionu, ve kterém se vzor nachází. Dalším sledovaným parametrem je velikost překrytí dvou sousedních oken. Se zvyšující se hodnotou, se zlepšuje rozlišení jednotlivých pozic bází v sekvenci. [1]



Obr.10 Barevné spektrum DNA segmentu [1]



Obr.11 Spektrogram DNA sekvence [1]

4.3. Vzory detekovatelné ze spektrogramů

Spektrogramy umožňují vizualizaci a detekci biologicky významných oblastí DNA přehledně v rámci celého genomu nebo podrobněji pro určitý úsek sekvence DNA.

DNA eukaryot obsahuje značný podíl nekódujících sekvencí. Nejsou většinou transkribovány a jejich funkce není ve většině případů známá. Mohou být unikátní nebo se v genomu nachází ve více identických či podobných kopiích. Sekvence DNA s vysokým množstvím takovýchto kopií se nazývají repetitivní sekvence. Pokud jsou kopie sekvenčního motivu uspořádány za sebou, hovoříme o tandemových repeticích. Sekvence, které se neseskupují a jsou lokalizované na mnoha místech genomu nazýváme rozptýlenými repeticemi. [17]

Funkce repetitivních sekvencí v DNA stále není úplně objasněna, ale v posledních letech bylo zjištěno, že počet tandemových repeticí může souviset se vznikem různých onemocnění a hrát důležitou roli v genové regulaci. Některé tandemové repetice byly používány jako důležité genetické markery pro studie mapování genů, analýzu genových vazeb a testování totožnosti. Proto má jistě analýza repeticí v DNA sekvenci značný význam. [2]

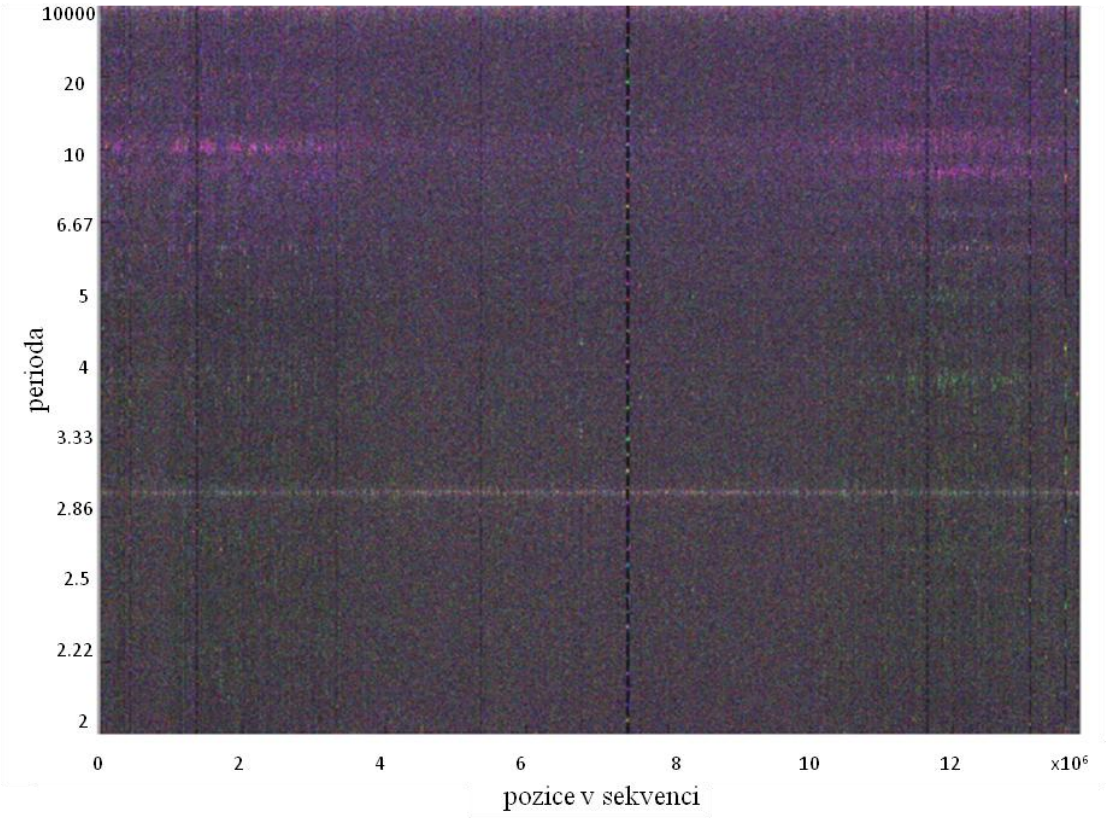
4.3.1. Tandemové repetice

U tandemových repeticí sledujeme čtyři vlastnosti – délku vzoru, strukturu vzoru, počet kopií a pozice vzorů. Uvažujeme-li délku vzoru, repetice můžeme rozdělit do tří typů – satelity, minisatelity a mikrosatelity.

Satelitní DNA je hojná v oblasti centromer a konstitutivního heterochromatinu. Vytváří bloky s délkou v rozmezí 100 kbp až 1 Mbp (párů bází). Jednotlivé vzory mají délku větší než 100 bp. Z mnoha satelitů nacházených v oblasti centromer, tvoří rodina alfa satelitu (s primární jednotkou dlouhou 171 bp) pravděpodobně funkční jádro centromery, je tedy důležitá během buněčného dělení. Funkce ostatních satelitů je neznámá, jsou považovány obvykle za odpadní (junk) DNA.

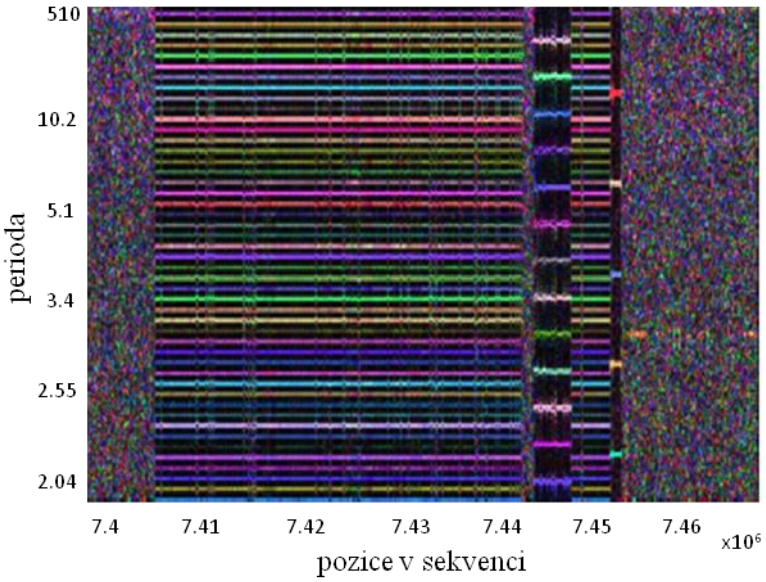
Minisatelity jsou kratší tandemové repetice vytvářející bloky v rozsahu od 1 kbp do 20 kbp složené ze vzorů o délce 9-80 bp. Vyskytují se v subtelomerických oblastech chromozomů. Jsou obvykle mnohotvárné co do počtu opakování jednotky repetice a mohou být použity jako genetické markery (VNTR). Někdy se uvažuje o tom, že by některé minisatelity mohly mít regulační funkce, jako např. VNTR v promotoru inzulinového genu, kde byla různá délka VNTR asociována s různými typy diabetu.

Mikrosatelity jsou tvořeny 1-6 bp, které se opakují v délce až 150 bp. Jsou nejčastější formou repetitivních sekvencí a vyskytují se na různých místech všech chromosomů. Nejčastější jsou dinukleotidové repetice, ze kterých převažuje typ CA (TG na komplementárním vlákně). Z mononukleotidových repetic se vyskytují nejčastěji A a T, naopak G a C jsou vzácné. Mikrosatelity jsou v genomu velice časté, vysoce polymorfní a jsou často používány jako genetické markery. [2],[17]



Obr.12 DNA spektrogram chromozomu III C. elegans, okno 10000, posun okna 0 [14]

Příklad spektrogramu, ve kterém můžeme rozeznat oblast minisatelitu, vidíme na *Obr.12*. Uprostřed spektrogramu chromozomu III modelového organismu Caenorhabditis elegans se nachází vertikální linie na pozici 7,4 Mbp identifikující minisatelit o délce zhruba 50 kbp. Osa y reprezentuje periodu a osa x pozice bází. Takto jsme si zobrazili celý chromozom. Část chromozomu III C. elegans s detailním záběrem na minisatelit je potom ukázána na *Obr.13*.



Obr.13 DNA spektrogram znázorňující minisatelit v chromozomu III C. elegans - detail z Obr.12, okno 510, posun okna 400 [14]

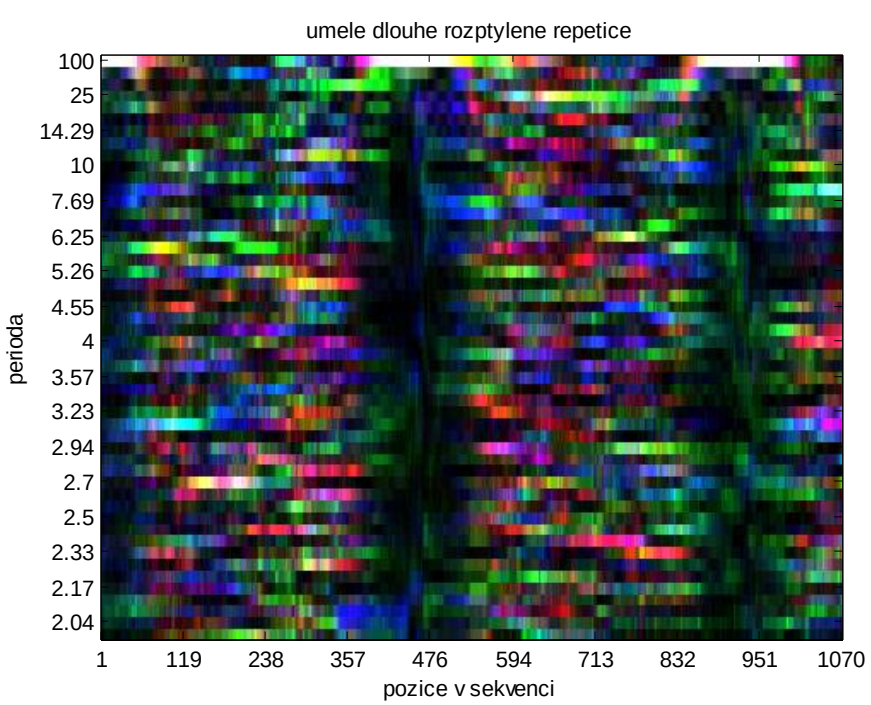
4.3.2. Rozptýlené repetice

Tyto sekvence se neseskupují, jsou lokalizované na mnoha místech genomu. Většina rozptýlených repetic vzniká procesem transpozice, což můžeme označit jako skákání segmentu DNA na jiné místo genomu. Rozlišujeme dva typy rozptýlených repetic – DNA transpozony a retrotranspozony.

DNA transpozony jsou v lidském genomu považovány za inaktivní v důsledku akumulace mutací v průběhu fylogeneze obratlovců. Můžeme ale vyrobit aktivní transpozon odvozený z lidských fosilních elementů, např. transpozon "Sleeping Beauty", který se může stát základem další generace genové terapie. Jádrem DNA transpozonu je sekvence kódující enzym transpozázu. Tento enzym se váže k oběma koncům repetitivního elementu. Komplex transpozon-transpozáza se váže na specifický sekvenční motiv jinde v genomu. Transpozáza štěpí hostitelskou DNA a váže transpozon na nové místo. Takto se transpozon pohybuje mechanismem vyjmout-vložit.

Pro **retrotranspozony** je typické, že mechanismus jejich šíření je velice podobný životnímu cyklu retrovirů. DNA retrotranspozonů se přepíše do RNA (pomocí RNA polymeráz II či III), potom tato RNA kopie podléhá reverzní transkripci do DNA, která je

vložena do genomu na nové místo. V lidské DNA tvoří kolem 45% celkové sekvence a mají tedy obrovský význam pro pochopení nekódující DNA vůbec. Na rozdíl od DNA transpozonů jsou v lidském genomu stále aktivní. Proces retrotranspozice je ale náchylný k různorodým chybám. Proto jsou nově vzniklé kopie většinou inaktivovány delecemi nebo bodovými mutacemi. Rozlišujeme LTR retrotranspozony, LINE retrotranspozony (dlouhé rozptýlené jaderné elementy) a SINE retrotranspozony (krátké rozptýlené jaderné elementy). Tyto elementy představují velkou část DNA označovanou někdy jako junk (odpadní) DNA – úseky, které nejsou přepisovány v proteiny. [17] Příklad umělého LINE retrotranspozonu je na *Obr.14*. Jedná se o repetici dlouhou 71 bp, vyskytující se na pozicích 1-71, 427-497, 853-923 bp, vloženou do náhodné sekvence o délce 1070 bp.

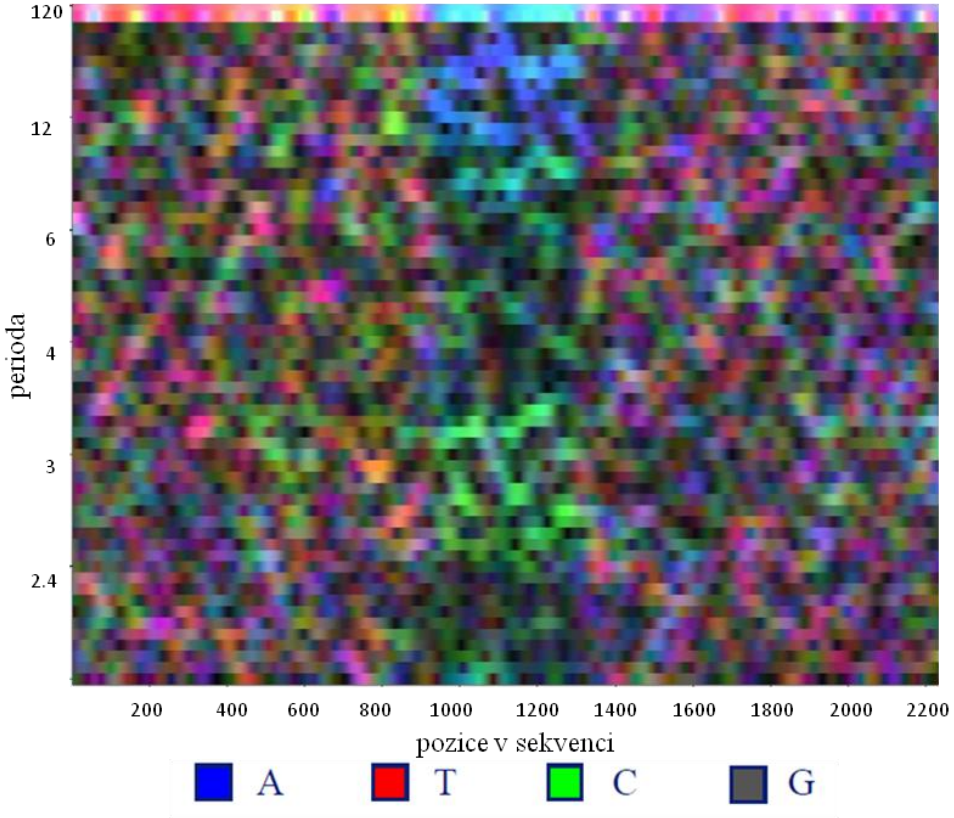


Obr.14 DNA spektrogram s umělými rozptýlenými repeticemi, okno 100, posun okna 1

4.3.3. CpG ostrůvky

Jedná se o regiony v DNA bohaté na CpG dinukleotidy. CpG vznikají metylací DNA, při které dochází k napojení metylových skupin na cytosin, po kterém následuje guanin. Písmeno p označuje fosfátovou vazbu mezi oběma nukleotidy. CpG ostrůvky se nachází v blízkosti téměř 70 % promotorů lidských genů a umožňují vazbu transkripčních faktorů. Většina CpG míst je v lidském genomu metylována, kromě CpG míst označovaných jako CpG ostrůvky, které jsou nemetylovány. Změna nemetylovaných CpG ostrůvků na metylované je doprovázena ztrátou genové exprese. Tyto změny také souvisí s nádorovým onemocněním. [19]

CpG ostrůvky jsou významné pro studium genové regulace. Hrají důležitou roli při buněčné diferenciaci a regulaci genové exprese u obratlovců. Podle jedné teorie jsou CpG ostrůvky označovány jako oblasti s nejméně 200 bp a obsahem C+G více než 50%. Jiná teorie říká, že jsou to regiony delší než 500 bp s obsahem C+G minimálně 55%. Poměr C+G (počet C x počet G / délka segmentu) je dán hodnotou 0,6-0,65. Příklad CpG ostrůvku nacházejícího se v segmentu lidského chromozomu 21 (9905604-9907958 bp) vidíme ve spektrogramu na *Obr.15*. V centrální části spektrogramu vidíme zelenou oblast, což je právě CpG ostrůvek obsahující báze C mapované zelenou barvou a báze G mapované šedou barvou. [1]



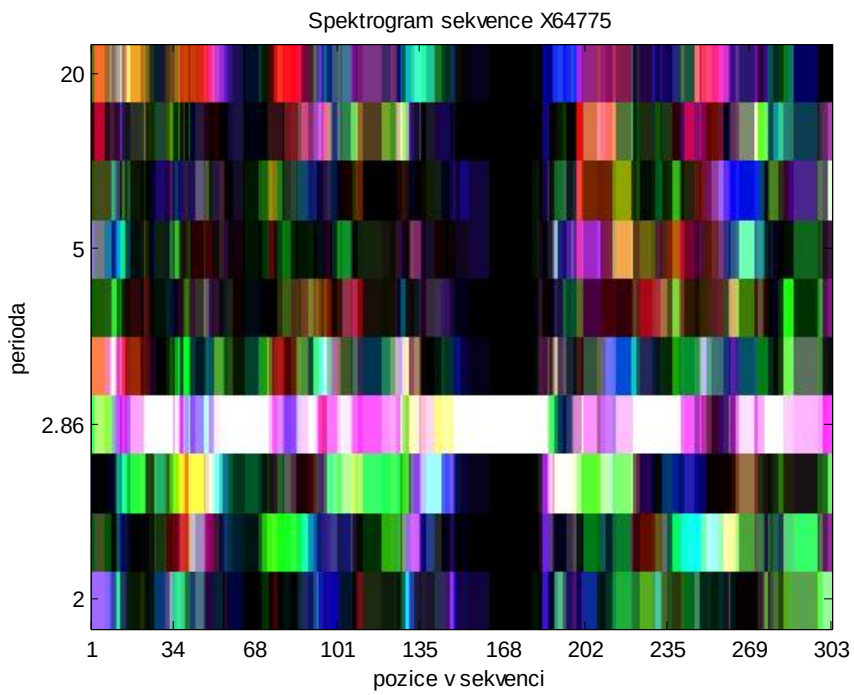
Obr.15 Spektrogram CpG ostrůvku v segmentu chr.21, okno 120, posun okna 1 [1]

4.4. Praktická ukázka DNA spektrogramu

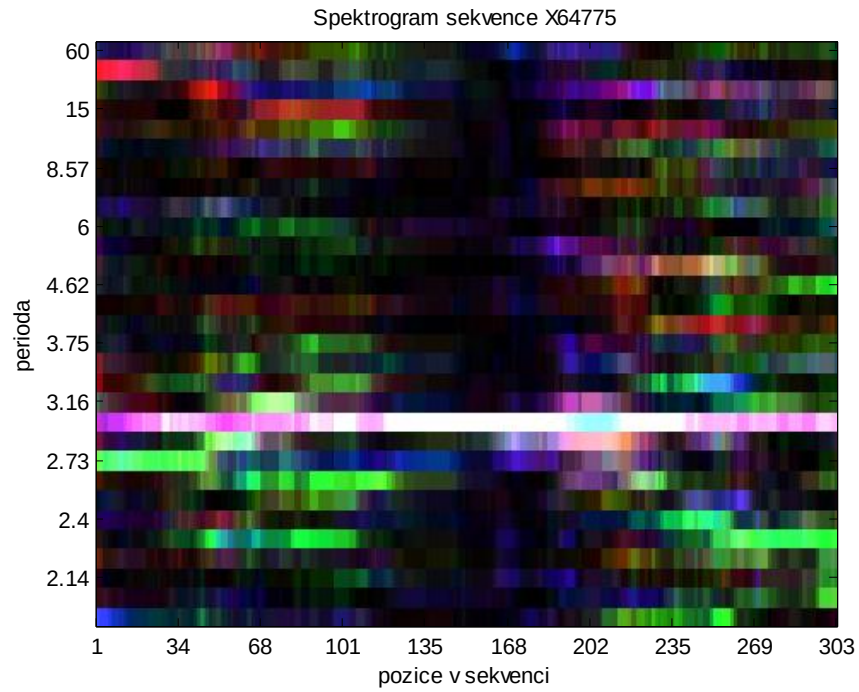
V programovém prostředí Matlab verze R2010a byla navržena první možnost řešení metody konstrukce spektrogramu. Algoritmus respektuje pořadí kroků 1) – 4) uvedených v kap. 4.2. Numerická reprezentace byla zvolena 4D binární (Vossova). Pro výpočet frekvenčního spektra byla použita Fourierova transformace. Pro mapování DFT hodnot do RGB prostoru byly využity rovnice 21 a hodnoty barevných mapovacích vektorů byly zvoleny ze vztahu 24. Výsledné hodnoty pixelů byly normalizovány s využitím rovnice 25.

Jako testovací sekvence byla vybrána Oryza sativa (rýže setá) z databáze NCBI ve *.*fasta* formátu. [22] Tato sekvence obsahuje tandemové repetice a dá se na ní dobře ověřit funkčnost algoritmu. Její označení je *X64775.1* a je dlouhá 303 bp.

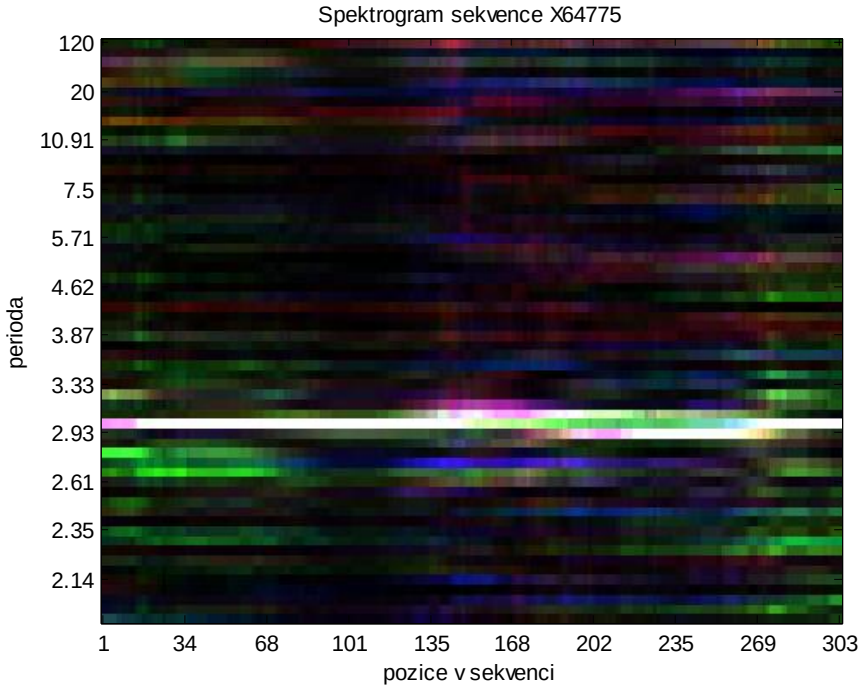
Před vykreslením spektrogramu bylo potřeba nastavit vhodně délku okna a posun oken, pro které se počítala FT. Délka okna by měla s délkou sekvence korespondovat tak, aby nebyla příliš malá nebo naopak velká. Malá délka okna způsobí, že program nenajde repetice delší než zvolená délka okna. Velká délka okna má za následek nepřehlednost spektrogramu. Hodnoty pro délku okna byly tedy pro porovnání nastaveny na 20 (viz *Obr.16*), 60 (viz *Obr.17*) a 120 (viz *Obr.18*) s posunem oken 1. Hodnota 1 pro posun oken je pro tuto krátkou testovací sekvenci nejvýhodnější pro dobré rozlišení obrázku. Čím větší hodnotu zvolíme, tím více se sníží počet spekter ve spektrogramu a výsledek je nepřesný (viz *Obr.19*). Avšak vyšší hodnoty je dobré použít u velmi dlouhých sekvencí, kdy se nám sníží doba výpočtu algoritmu.



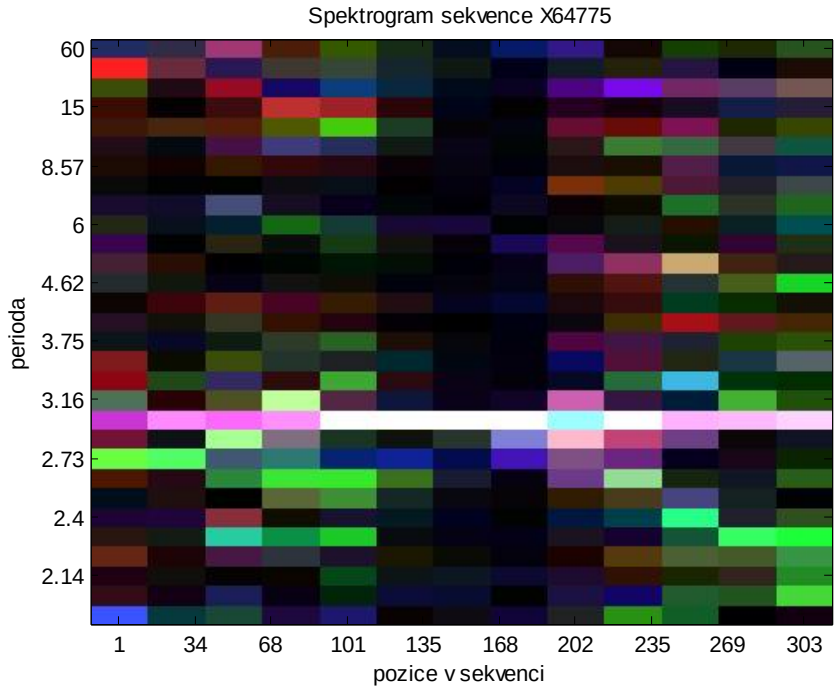
Obr.16 Spektrogram sekvence O. sativa s délkou okna 20



Obr.17 Spektrogram sekvence O. sativa s délkou okna 60



Obr.18 Spektrogram sekvence O. sativa s délkou okna 120



Obr.19 Spektrogram sekvence O. sativa s posunem oken 20

Nejvhodnější nastavení parametrů pro FT je délka okna 60 a posun oken 1, což vidíme z *Obr.17*. Na ose x jsou zobrazeny pozice nukleotidových bází v sekvenci, na ose y je znázorněna perioda opakované sekvence – podíl délky sekvence *N* a koeficientu spektra *k*. Význam barev je následující:

- zelená barva = cytosin
- červená barva = thymin
- modrá barva = adenin
- šedá barva = guanin

Pro lepší čitelnost výsledku je každá nukleotidová báze reprezentována jinou barvou a frekvenční spektrum všech čtyř bází je sloučeno do jednoho barevného spektrogramu. Výsledný barevný pixel potom má intenzitu odpovídající čtyřem bázím na dané frekvenci.

Úseky obsahující repetitivní části jsou znázorněny bílou barvou. Všechny tyto oblasti se vyskytují na svislé ose na pozici okolo hodnoty 3. Můžeme tedy říci, že všechny repetice v O. sativa mají délku opakované sekvence 3. Hlavní tandemová repetice s největší délkou opakování se vyskytuje na pozicích 142-186 bp v sekvenci a jde o trinukleotidovou GGC, což můžeme vidět v tab. 1 níže zvýrazněno modrou barvou. V repetici se může vyskytovat jedna mutace, např. místo GGC je na pozicích 142-144 bp GGA. Dále můžeme dvakrát najít mutaci GAC na pozicích 151-153 bp, 157-159 bp a mutaci GGT vidíme na pozici 181-183 bp.

Další velmi krátkou tandemovou trinukleotidovou repeticí v sekvenci O. sativa je CGG na pozicích 59-76 bp. Je zvýrazněna žlutou barvou a obsahuje jednu mutaci CGG na TGG.

Poslední nejkratší tandemovou repeticí, zvýrazněnou zeleně, je TAC na pozicích 49-57 bp, opět s jednou mutací TAC na GAC. Tabulka vznikla s pomocí veřejné databáze Tandem Repeats Database (TRDB), která slouží jako archiv informací o tandemových repeticích v genomech organismů a umožňuje díky širokému výběru metod jejich analýzu. [23],[2]

TABULKA1

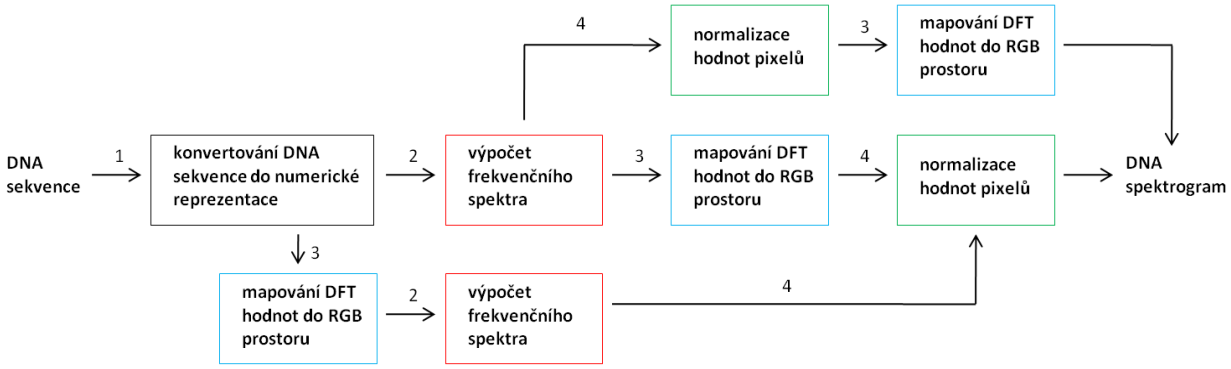
```
1   ATGGAG AGCGAC TGCCAG TTCTTG GTGGCG CCGCCG CAGCCG CACATG TACTAC GACACG
61  GCGGCG GCGGCG GTGGAC GAGGCG CAGTTC TTGCGG CAGATG GTGGCC GCGGCG GATCAC
121 CACGCG GCCGCC GCTGGG AGAGGA GGCGGC GACGGC GACGGC GGCGGC GGCGGC GGCGGC
181 GGTGGC GGGGAG AGGAAG CGGCGG TTCACG GAGGAG CAGGTG CGGTCG CTGGAG ACGACG
241 TTCCAC GCGCGG CGGGCC AAGCTG GAGCCG CGGGAG AAGGCG GACGTG GCGCGG GAGCTC
301 GGG
```

5. Programové řešení metod konstrukce spektrogramů

Všechny metody konstrukce spektrogramů byly řešeny v programovém prostředí Matlab verze R2010a, tak jako praktická ukázka DNA spektrogramu v kapitole předchozí. Metody se dají rozdělit na dvě hlavní podkapitoly podle typu operace použité na výpočet frekvenčního spektra. Tyto jsou dvě – Fourierova transformace a odhad spektra pomocí autoregresního modelu. Obě metody výpočtu spektra jsou realizovány jako samostatné funkce *DFT.m* a *p_burg.m* volané hlavní funkcí. Hlavní funkce je tvořena několika kroky:

1) konvertování DNA sekvence do numerické reprezentace
2) výpočet frekvenčního spektra (výkonnostního spektra)
3) mapování DFT hodnot do RGB prostoru
4) normalizace hodnot pixelů

Kroky 2,3,4 mohou za sebou následovat v různém pořadí, jak zobrazuje schéma na *Obr.20*.



Obr.20 Pořadí kroků pro vytvoření spektrogramu

Nejprve k popisu jednotlivých kroků zvlášť. Na začátku skriptu se načítá sekvence ve **.fasta* formátu do proměnné ***seq*** pomocí příkazu *fastaread*. První částí hlavní funkce a vždy prvním krokem pro všechny metody je převod DNA sekvence reprezentované písmeny A,C,T,G na numerickou reprezentaci jedniček a nul. Numerická reprezentace byla zvolena 4D binární (Vossova) pro všechny metody (kap. 3.1.1). Do matice ***num_repr*** o čtyřech řádcích a počtu sloupců podle délky testované sekvence se ukládají hodnoty 1 a 0. V prvním řádku se tedy přiřadí jedničky pozicím, kde se v sekvenci nachází adenin (A), na ostatních pozicích jsou nuly. Pro druhý řádek platí totéž, ale pro thymin (T). Pro třetí a čtvrtý řádek nápodobně pro cytosin (C) a guanin (G).

```
L = length(seq);
num_repr = zeros(4,L);
num_repr(1,:)= seq == 'A';
num_repr(2,:)= seq == 'T';
num_repr(3,:)= seq == 'C';
num_repr(4,:)= seq == 'G';
```

V případě algoritmu s pořadím kroků 1,2,3,4 je dalším krokem výpočet frekvenčního spektra. Před samotnými *for* cykly, které realizují tento výpočet, si definujeme hodnoty pro délku okna ***w*** a posun oken ***o***. Následující dva *for* cykly prochází numerickou reprezentaci – matici ***num_repr***, po sloupcích ***u*** a po řádcích ***v***.

```
for u = 1:o:L-w
    for v = 1:4
        x1 =(num_repr(v,u:u+w))-mean(num_repr(v,u:u+w)));
        x = (hann(length(x1)))'.*x1; % řádek jen pro FT!
        spektrum = DFT(x);
      % pro AR model:spektrum = p_burg(x,p)
        spektrum_celkem(v,:)= spektrum.^2;
    end
end
```

První *for* cyklus prochází numerickou reprezentaci po sloupcích ***u*** od první hodnoty s krokem ***o*** (posun oken) až do hodnoty ***L-w*** (délka sekvence mínus délka okna). Druhý vnořený *for* cyklus prochází reprezentaci po řádcích ***v***. Do proměnné ***x1*** se ukládají

hodnoty o délce ***u*** vždy z jednoho řádku matice ***num_repr***, tedy pro jeden typ nukleotidu, s odečtenou střední hodnotou proměnné ***num_repr***. Ta se odečítá proto, aby se v horní části spektrogramu nezobrazoval bílý pruh.

Vektor ***x1*** potom vynásobíme Hannovým oknem (`hann`), abychom dostali kvaziperiodický signál (abychom ho zkrátili na konečnou délku). Tento krok chybí ve skriptu, který používá AR model pro výpočet frekvenčního spektra.

Pro všechny hodnoty vektoru ***x*** se potom spektrum vypočítá voláním samostatné funkce *DFT.m* nebo *p_burg.m* (v případě odhadu spektra pomocí AR modelu), ve kterých jsou uloženy vztahy pro výpočet Fourierovy transformace a AR modelu. Jejich výstupy jsou ukládány do proměnné ***spektrum***. Jednotlivá spektra se umocní na druhou a uloží do matice ***spektrum_celkem*** na danou pozici. S touto maticí dále pracují kroky "mapování DFT hodnot do RGB prostoru" a "normalizace hodnot pixelů". Výpočet spektra oběma metodami bude podrobněji uveden v podkapitole 5.1 a 5.2.

Třetím krokem v případě algoritmu s pořadím kroků 1,2,3,4 je mapování DFT hodnot do RGB prostoru. Nejprve si definujeme hodnoty barevných mapovacích vektorů podle vztahů 22-24 z kap. 4.2.3. Hodnoty vektorů R, G, B získáme superpozicí mapovacích vektorů barev váhovaných frekvenční složkou báze na dané pozici, tedy odpovídajícím spektrem uloženým v matici ***spektrum_celkem***. Tento krok je v Matlabu popsán řádky:

```
R = ar.*(spektrum_celkem(1,:))+tr.*(spektrum_celkem(2,:))+
    +cr.*(spektrum_celkem(3,:))+gr.*(spektrum_celkem(4,:));
G = ag.*(spektrum_celkem(1,:))+tg.*(spektrum_celkem(2,:))+
    +cg.*(spektrum_celkem(3,:))+gg.*(spektrum_celkem(4,:));
B = ab.*(spektrum_celkem(1,:))+tb.*(spektrum_celkem(2,:))+
    +cb.*(spektrum_celkem(3,:))+gb.*(spektrum_celkem(4,:));
```

Vektory musíme transponovat, aby mohly být skládány vertikálně vedle sebe do výsledného spektrogramu.

Normalizace hodnot pixelů je realizována ve skriptu rovnicí 25 z kap. 4.2.4. Z vektorů hodnot z předchozího kroku jsou vypočteny hodnoty normalizované na rozsah 0-1:

```
MaxR = max(max(R)); MinR = min(min(R));
nR = R - MinR;
nnR = nR/(MaxR-MinR);

MaxG = max(max(G)); MinG = min(min(G));
nG = G - MinG;
nnG = nG/(MaxG-MinG);

MaxB = max(max(B)); MinB = min(min(B));
nB = B - MinB;
nnB = nB/(MaxB-MinB);
```

Ovšem v hlavním skriptu pro AR model bylo potřeba provést ještě jednu normalizaci, aby bylo možné lépe rozlišit oblasti zájmu od příliš tmavého pozadí spektrogramu. Tato normalizace je ve skriptu před předchozí zmíněnou, která hodnoty upravuje na rozsah 0-1. Napřed se provede vydělení každé hodnoty z vektoru R, G a B trojnásobkem normalizační konstanty ***str_hodn***, která slouží pro snížení rozptylu hodnot ve vektorech. Tato konstanta se vypočítá jako průměr středních hodnot z jednotlivých vektorů.

```
str_hodn = mean([mean(R) mean(G) mean(B)]);
R = R./(3*str_hodn);
G = G./(3*str_hodn);
B = B./(3*str_hodn);
```

Potom následuje vyhledání hodnot přesahujících maximální povolenou hodnotu 1 v každém vektoru R, G i B, což se uloží do proměnné ***m***. Další úprava barev je provedena třemi *for* cykly. V každém se vždy pro jeden vektor (jednu barvu z RGB) nastaví hodnota 1 jako maximum (zde v příkladu kódu je to řádek `G(m(i))= 1`) a ostatní dva vektory (zde `R` a `B`) se sníží o hodnoty větší než 1 uložené v dané proměnné, tedy vydělí danou proměnnou (zde je to `zelena`).

```
[m] = find(G>1);
pocet = length(m);
for i = 1:pocet
    zelena = G(m(i));
    G(m(i))= 1;
    R(m(i))= R(m(i))/zelena;
    B(m(i))= B(m(i))/zelena;
end
```

Po posledním kroku tedy může být vykreslen spektrogram. Před vykreslením příkazem *image* se tři normalizované vrstvy ***nnR***, ***nnG*** a ***nnB*** uloží do proměnné ***RGB***. Výsledná barva pixelů je dána superpozicí hodnot těchto tří vrstev. Přechod na další spektrum segmentu určené oknem ***w*** a posunem oken ***o*** se děje pomocí přičtení jedničky k pomocné proměnné ***pom*** definované ještě před samotným *for* cyklem na hodnotu 1.

```
RGB(:,pom,1) = nnR;
RGB(:,pom,2) = nnG;
RGB(:,pom,3) = nnB;
```

Na konci hlavního skriptu se ještě upravují hodnoty na obou osách. Svislou osu je třeba přepočítat z frekvence na periodu, aby se lépe odhadovaly délky repetic. Do pomocné proměnné ***periody*** se ukládají zaokrouhlené přepočtené hodnoty z frekvence na periodu. Příkazem *set* se přenastaví hodnoty na ose y tak, že v první části se funkcí *gca* získají současné hodnoty osy a v druhé části se přepíšou podle proměnné ***periody***.

```
periody = round((w./(1:w/2))*100)/100;
krok_osy_y = 10;
y_tick = [linspace(1,w/2,w/2)];
set(gca,'ytick',y_tick(1:krok_osy_y:end),
'yTicklabel',periody(1:krok_osy_y:end))
```

Vodorovnou osu je také potřeba nastavit, aby se na ní zobrazovaly hodnoty od 1 do délky sekvence ***L*** (bp) místo hodnot délky okna ***w***. Musíme určit velikost vykreslované matice ***RGB***. Konkrétně nás zajímá počet sloupců ***colms***, který použijeme pro zjištění současných souřadnic osy x příkazem *gca*. Osu potom přenastavíme na hodnoty proměnné ***osa_x***.

```
[rows colms layers] = size(RGB);
krok_osy_x = 10;
osa_x = fix(linspace(1,L,krok_osy_x));
set(gca,'xtick',[linspace(1,colms,krok_osy_x)],
'xTicklabel',osa_x)
```

5.1. Fourierova transformace

DFT.m je podfunkce volaná hlavními funkcemi označenými *hlavni_1234.m*, *hlavni_1243.m* a *hlavni_1324.m*. Skript obsahuje vztah 12 z kap. 4.2.2 realizující výpočet Fourierovy transformace. Nejprve si určíme parametr ***N*** označující délku okna, tedy část sekvence, pro kterou se počítá DFT. Tato délka okna je vlastně vektor ***x*** tvořený hodnotami sloupců ***u***, vzatý vždy z jednoho řádku ***v*** z celkových čtyř (pro daný nukleotid) z ***num_repr*** (viz řádek hlavního skriptu: `x =(num_repr(v,u:u+w));`). Dále si připravíme prostor pro vektor ***U***, do kterého se ukládají průběžně hodnoty po FT.

```
N = length(x);
U = zeros(1,floor(N/2));
```

Samotný vztah pro výpočet FT je uložen ve *for* cyklu:

```
for k = 1:N/2
    n = linspace(0,N-1,N);
    U(k) = sum(x.*exp(-2*1i*pi*k*n/N)); % FT
end
```

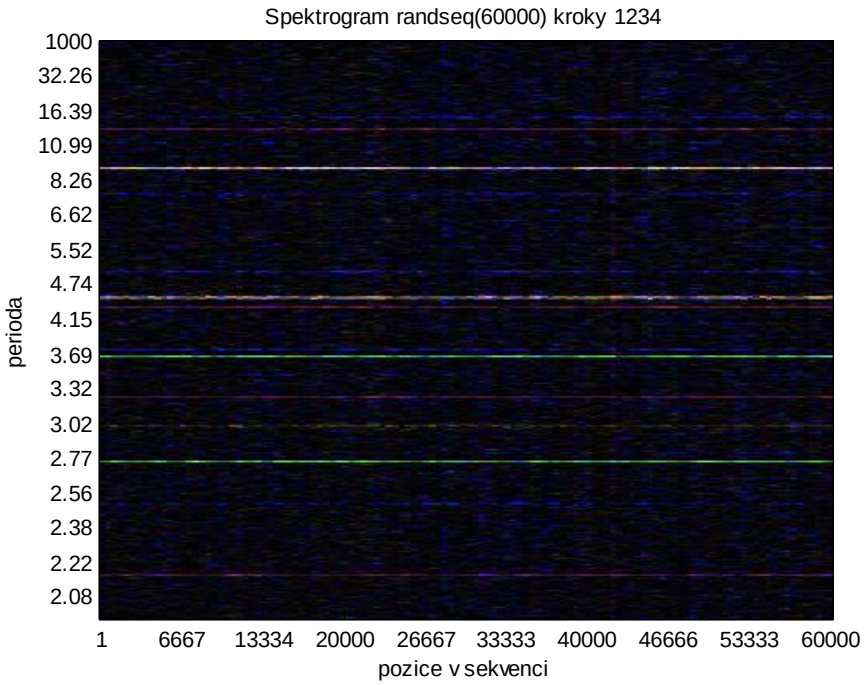
kde **k** je koeficient spektra z intervalu 0 až N/2, který určuje pořadí vzorků v kmitočtové oblasti a ***n*** značí pořadí jednotlivých oken jdoucích v čase za sebou (vektor N bodů v rozmezí a včetně hodnot 0 a N-1). Poslední řádek skriptu je výstupem pro hlavní funkci, kde se vytvoří výkonnostní spektrum součtem všech čtyř spekter pro každý nukleotid.

```
spektrum = abs(U);
```

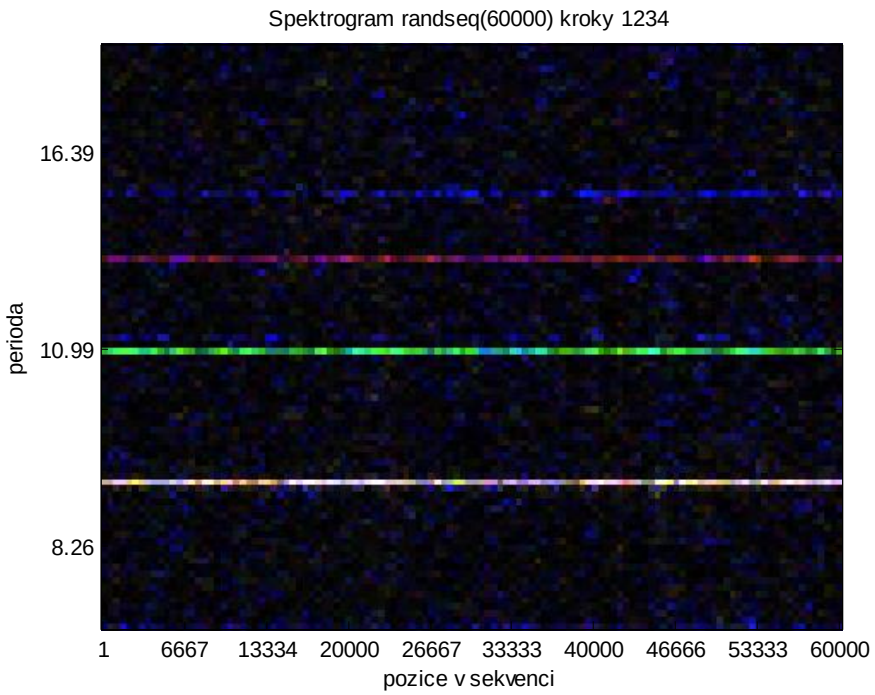
5.1.1. Záměna pořadí kroků pro vytvoření DNA spektrogramu

Prvním úkolem bylo srovnat spektrogramy vykreslené skripty s různě zaměněným pořadím kroků 1) – 4) podle schématu na *Obr.20*. Pro tento úkol byl krok výpočtu frekvenčního spektra vždy realizován stejnou metodou a to Fourierovou transformací. Hlavní skripty tedy volají funkci *DFT.m*. Délka okna byla nastavena na hodnotu 1000 a posun oken na hodnotu 500.

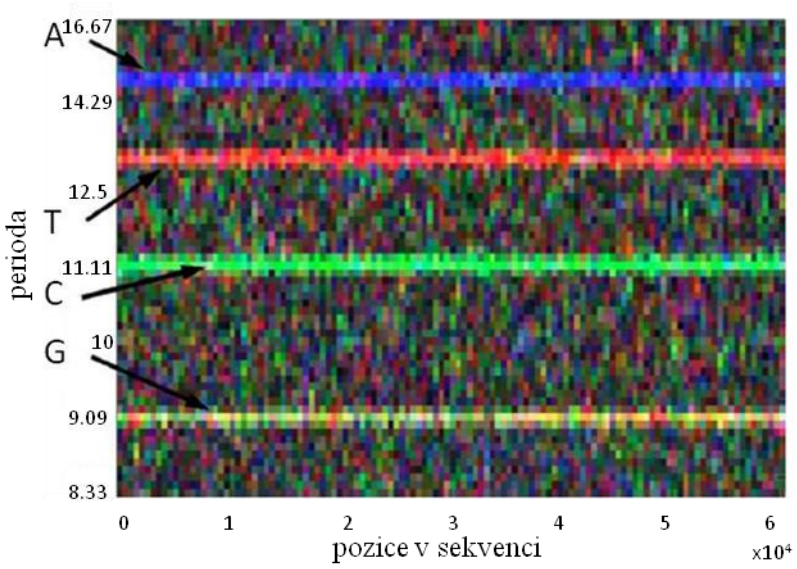
Nejprve aby bylo jisté, že základní skript *hlavni_1234.m* s klasickým pořadím kroků 1,2,3,4 funguje správně, byla vytvořena náhodná sekvence o délce 60 kbp s bázemi A, T, C a G opakujícími se s periodou 15, 13, 11 a 9, stejná jako použili autoři v lit. 14. Nukleotidy jsou reprezentovány barvami: A – modrá, T – červená, C – zelená a G – žlutá. Na *Obr.21* vidíme výsledný spektrogram. Přiblížíme-li si oblast kolem periody 9–15 na ose y (viz *Obr.22*), můžeme vyčíst, že nukleotid A znázorněný modrou barvou se opravdu opakuje s periodou 15, což pozorujeme jako vodorovný modrý pruh. Nukleotid T se opakuje v náhodné sekvenci s periodou 13, ve spektrogramu je to vykresleno jako červený pruh. Zadání je splněno i pro nukleotidy C a G. Pro porovnání je na *Obr.23* uveden spektrogram z lit. 14.



Obr.21 Spektrogram náhodné sekvence o délce 60 kbp s periodicky se opakujícím výskytem nukleotidů A(15), T(13), C(11), G(9)



Obr.22 Detail z Obr.21 v oblasti kolem periody 9-15

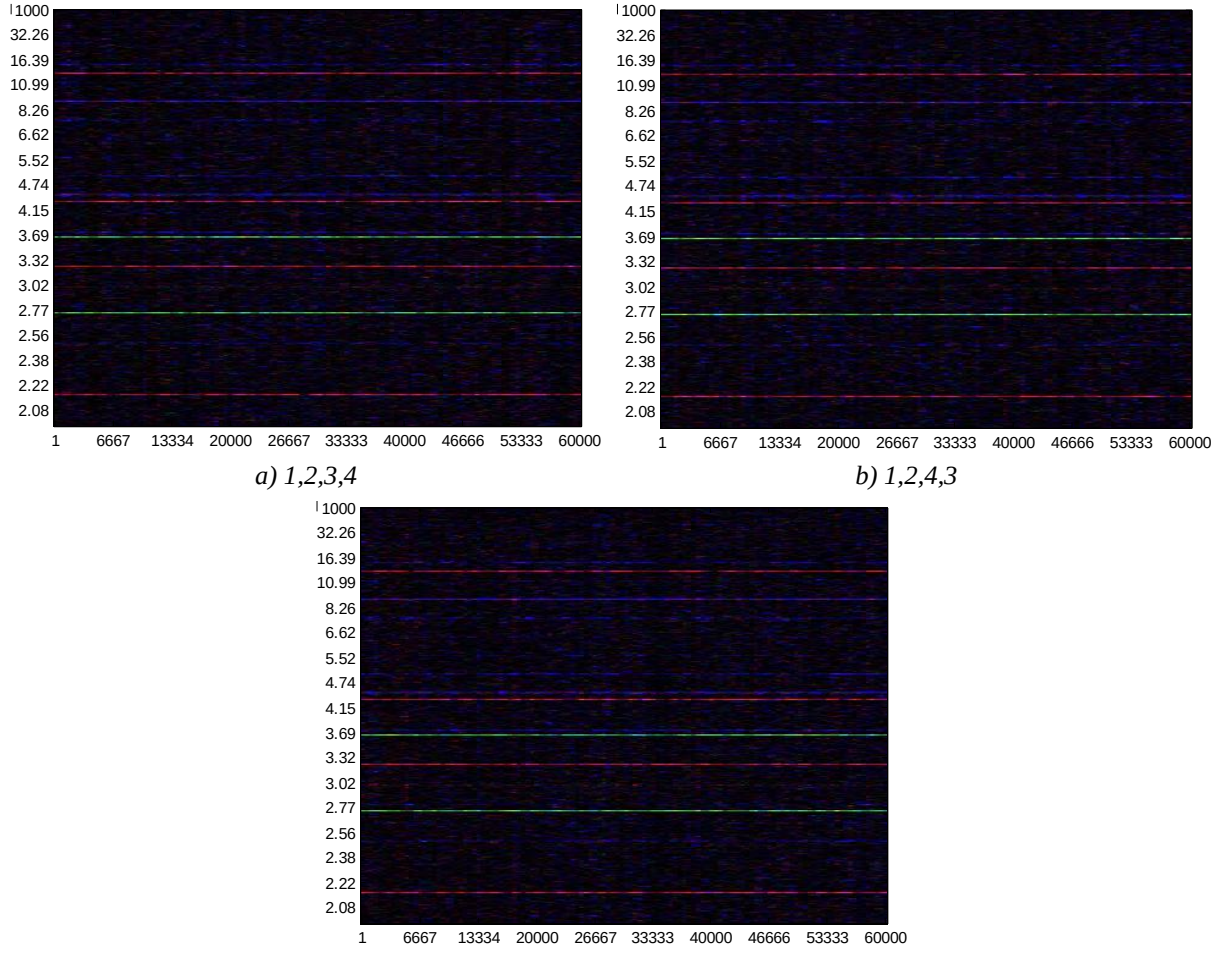


Obr.23 Spektrogram náhodné sekvence o délce 60 kbp s periodicky se opakujícím výskytem nukleotidů A(15), T(13), C(11), G(9) z lit.14

V Matlabu je náhodná sekvence vytvořena pomocí příkazu *randseq* a periodicky opakující se nukleotidy realizují řádky:

```
seq(15:15:end) = 'A'; % každý 15. nukleotid je A
seq(13:13:end) = 'T'; % každý 13. nukleotid je T
seq(9:9:end) = 'G';   % každý 9. nukleotid je G
seq(11:11:end) = 'C'; % každý 11. nukleotid je C
```

Potom, co bylo ověřeno, že program funguje správně, následovala hlavní část úkolu – prohození jednotlivých kroků pro vykreslení spektrogramu. Výsledkem jsou tři obrázky s pořadím kroků 1,2,3,4 (viz *Obr.24a*), 1,2,4,3 (viz *Obr.24b*) a 1,3,2,4 (viz *Obr.24c*) vykreslené pomocí skriptů *hlavni_1234.m, hlavni_1243.m* a *hlavni_1324.m*. Vidíme, že spektrogramy jsou identické. Hodnoty barevných mapovacích vektorů byly oproti předchozímu spektrogramu trochu odlišné: A – modrá, T – červená, C – zelená, G – černá. Je to proto, že skript 1,2,4,3 neobsahuje krok normalizace hodnot RGB jako poslední a musí tedy fungovat pouze s černou [0,0,0] barvou pro nukleotid G, aby byly hodnoty typu 'double' vstupující do příkazu *image* v rozsahu 0-1. *Obr.24* jasně potvrzuje, že záměna pořadí kroků nemá na výsledný spektrogram vliv, ovšem pomineme-li to, že nukleotid G splývá s černým pozadím spektrogramu. V literatuře je nejčastěji dodržováno pořadí kroků 1,2,3,4. Většinou je žádoucí rozlišit všechny čtyři nukleotidy, proto je vhodné použít skripty s pořadím kroků 1,2,3,4 nebo 1,3,2,4.



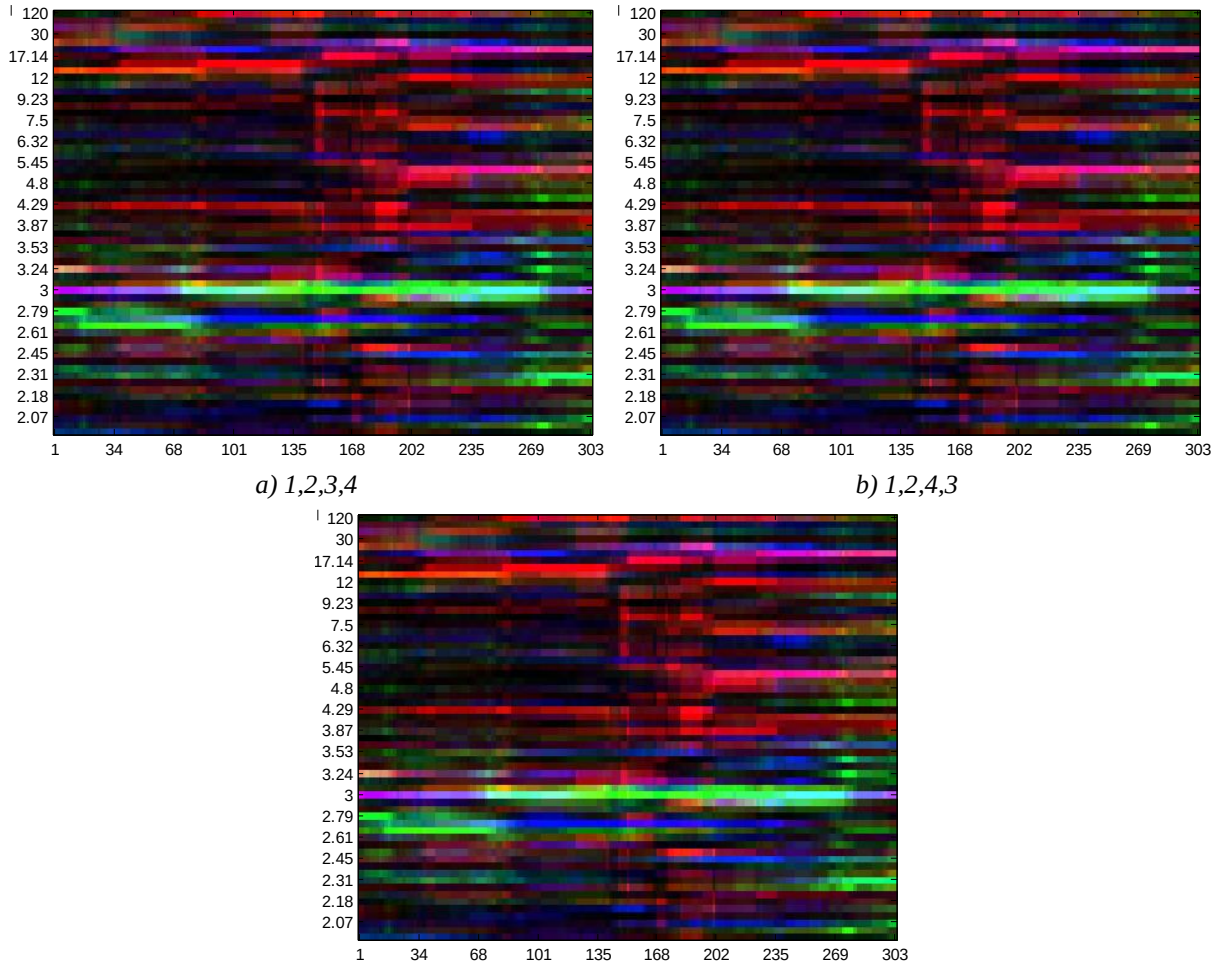
a) 1,2,3,4

b) 1,2,4,3

c) 1,3,2,4

Obr.24 Spektrogramy náhodné sekvence s různým pořadím kroků

Pro ověření, zda má vliv záměna pořadí kroků na výsledný spektrogram byla otestována i reálně existující sekvence O. sativa [22]. Z *Obr.25* je vidět, že všechny tři spektrogramy jsou identické. Délka okna byla nastavena na hodnotu 120 a posun oken na hodnotu 1. Barevné mapovací vektory byly stejné jako při testování předchozí náhodné sekvence.



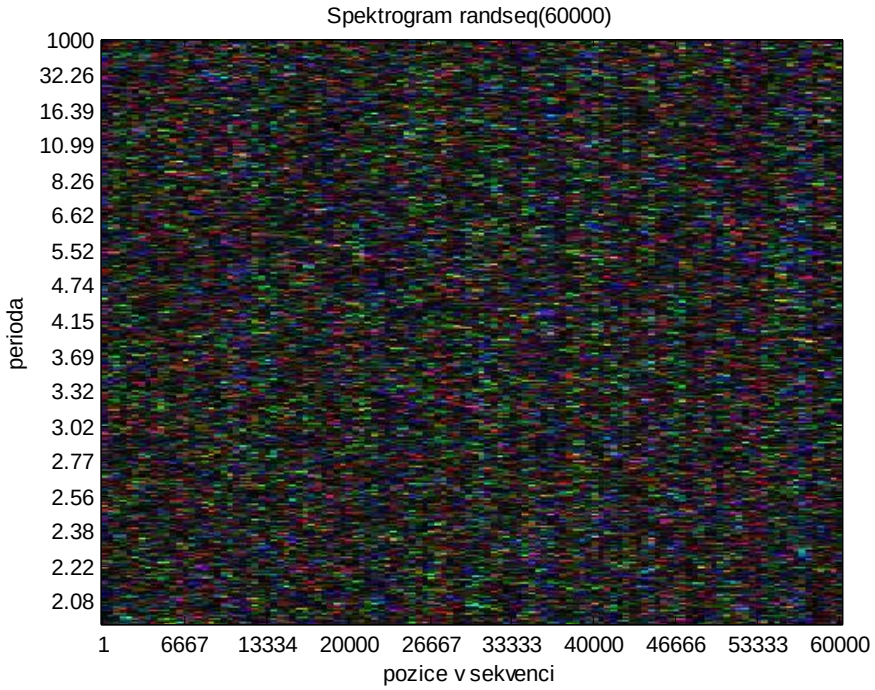
a) 1,2,3,4 *b) 1,2,4,3*

c) 1,3,2,4

Obr.25 Spektogramy O. sativa s různým pořadím kroků

5.1.2. Detekce vzorů ve spektrogramech pomocí FT

Hlavní program by měl sloužit pro detekci vzorů ve spektrogramech. Pro otestování, jak se tyto vzory v obrázcích projevují, byla vytvořena umělá náhodná sekvence o délce 60 kbp. Její spektrogram je možno vidět na *Obr.26*. Délka okna byla nastavena na hodnotu 1000 a posun oken na hodnotu 500, protože jde o dlouhou sekvenci, která se pomalu vykresluje. Sekvence se generuje v Matlabu příkazem *randseq*, kdy do závorky uvedeme počet bp. Tato úloha je zařazena ve skriptu *hlavni_1234.m*.



Obr.26 Spektrogram náhodné sekvence o délce 60 kbp

Tandemové repetice

V této sekvenci byly potom uměle vytvořeny vzory, které se ve spektrogramech mohou vyskytovat. Na *Obr.27a)* je vidět vložená umělá repetice CGG jako tmavý pruh na pozici okolo 10 kbp s výrazným zeleným místem okolo periody 3. Jde o tandemovou trinukleotidovou repetici opakující se v délce 150 bp = mikrosatelit. Protože repetice se opakuje s periodou 3 (tři nukleotidy CGG), nachází se zelený proužek na svislé ose okolo hodnoty 3. Tato barva nám potvrzuje, že jde o trinukleotid CGG, protože v barevných mapovacích vektorech je pro bázi C použita zelená a pro G šedá. Zelená barva patrně převáží šedou, proto se repetice vyskytuje jako zelený proužek. Můžeme ji detailněji pozorovat v *Obr.27b).*

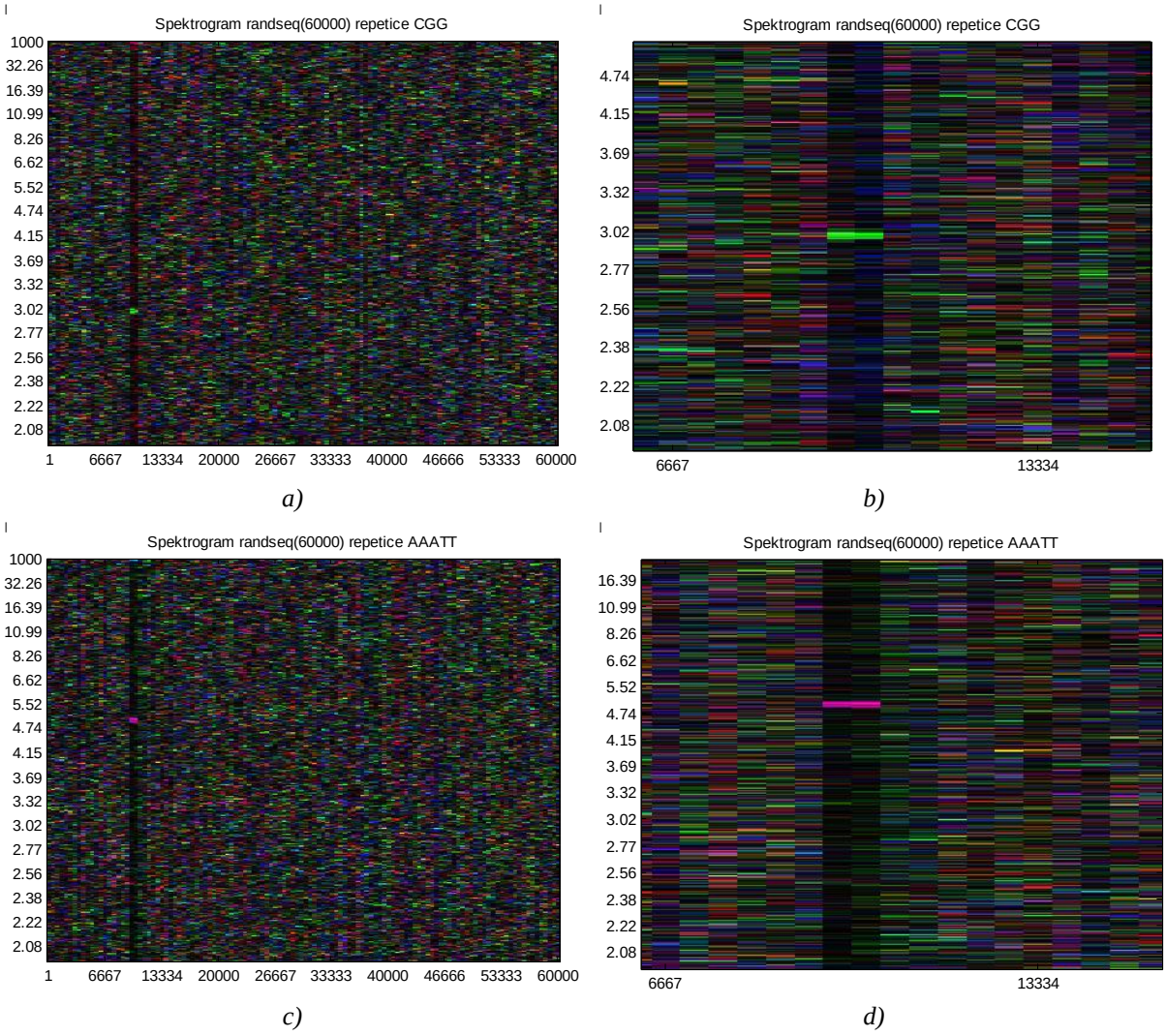
Vložení umělé repetice o délce tři nukleotidy opakující se v rozsahu 150 bp realizují ve skriptu *hlavni_1234.m* řádky:

```
seq(10000:3:10150) = 'C';
seq(10001:3:10151) = 'G';
seq(10002:3:10152) = 'G';
```

Pro porovnání byla vytvořena jiná repetice – umělý mikrosatelit AAATT. Ten můžeme pozorovat v *Obr.27c)* jako tmavý pruh opět okolo pozice 10 kbp na ose x s fialově zbarvenou částí okolo periody 5 na ose y, protože opakující se část má pět nukleotidů a barvy mapovacích vektorů pro A a T jsou modrá a červená, což nám dá dohromady v obrázku fialovou. Detail repetice vidíme v *Obr.27d).*

Repetice je také realizována ve skriptu *hlavni_1234.m* řádky:

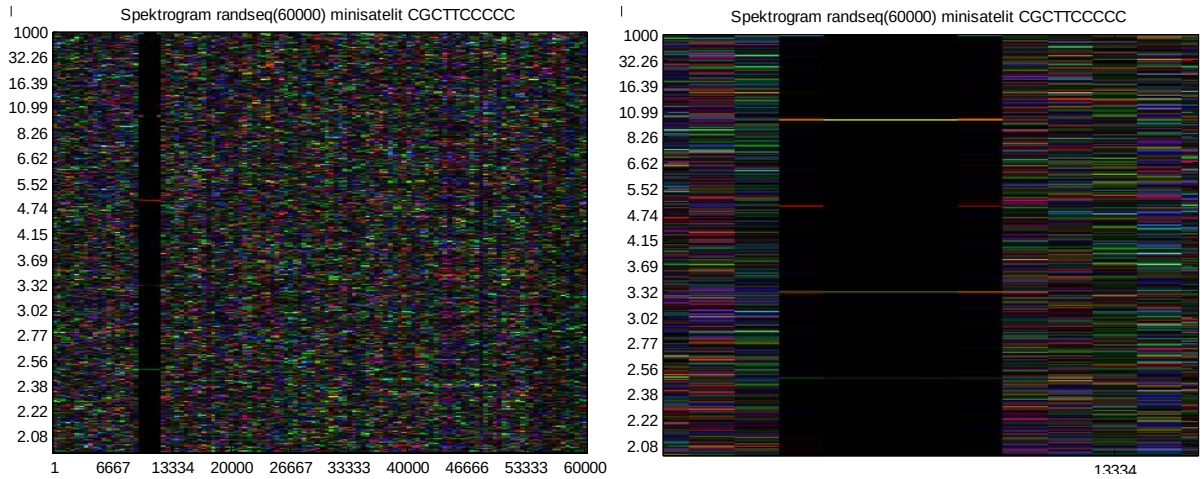
```
seq(10000:5:10150) = 'A';
seq(10001:5:10151) = 'A';
seq(10002:5:10152) = 'A';
seq(10003:5:10153) = 'T';
seq(10004:5:10154) = 'T';
```



Obr.27 Spektrogramy náhodné sekvence s umělými repeticemi a)CGG, c)AAATT s detailním výřezem okolo pozice 10 kbp pro b)CGG, d)AAATT

Na předchozích obrázcích jsou ukázány jen mikrosatelity, proto pro porovnání byly vytvořeny spektrogramy s minisatelity. Náhodná sekvence s minisatelity tvořenými repeticemi CGCTTCCCCC opakujícími se tandemově v délce 2 kbp je na *Obr.28*. Tato sekvence je označena jako *randomseq_minisatelit.fasta* a načítá se příkazem *fastaread* v hlavním skriptu *hlavni_1234.m*. Minisatelit se projeví jako tmavý svislý pruh, širší než předchozí mikrosatelity. Protože délka opakujícího se vzoru je větší (10) než u mikrosatelitů (3 nebo 5), obsahuje v sobě tento svislý pruh i více malých vodorovných proužků. Jejich počet by podle [14] měl odpovídat délce opakující se repetice. Tedy je-li

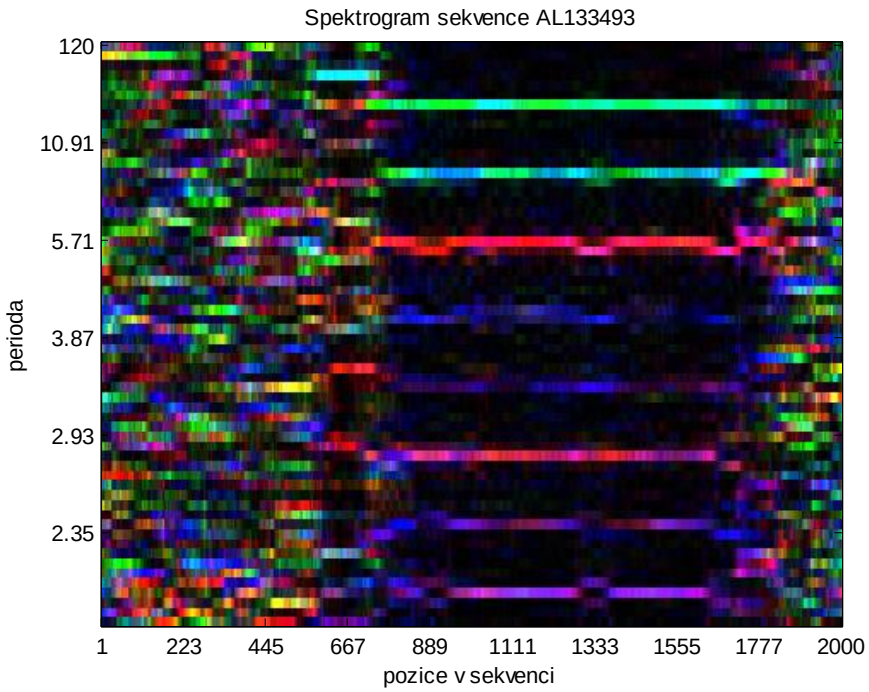
počet vodorovných proužků L sudý, vypočteme délku vzoru jako 2L. Je-li počet L lichý, potom délka vzoru bude 2L+1. Pro tuto sekvenci *randomseq_minisatelit.fasta* tedy můžeme odhadnout počet L = 5 z *Obr.28b*, potom délka vzoru by měla být 2L+1 = 11. Správná délka je deset nukleotidů, ale výsledek předchozího výpočtu může být ovlivněn nepřesným určením počtu vodorovných proužků. Jeden z nich totiž kopíruje horní okraj spektrogramu, tak není jisté, zda-li se má ještě počítat.



a) *b)*

Obr.28 a)Spektrogram náhodné sekvence s minisatelity CGCTTCCCCC, b)detail okolo pozice 10 kbp

Další *Obr.29* ukazuje tentokrát reálnou sekvenci *AL133493_human_chr21.fasta* s minisatelity [24], na rozdíl od předchozí uměle generované. Jedná se o segment (pozice 76001 až 78000 bp) ze sekvence lidského chromozomu 21 označené jako *AL133493*. Tato sekvence byla použita pro testování i v lit. 2. Parametry pro výpočet spektra byly: délka okna 120 a hodnota posunu oken 1. Ze spektrogramu můžeme vyčíst dva regiony mezi pozicemi 650-750 bp a 800-1800 bp, které obsahují minisatelity. První region obsahuje vzory o délce 32 bp, druhý má vzory s délkou 17 bp. Srovnáním s databází TRDB [23], která vypočetla regiony jako 623-737 bp a 802-1768 bp, vykresluje skript *hlavni_1234.m* spektrogram správně.

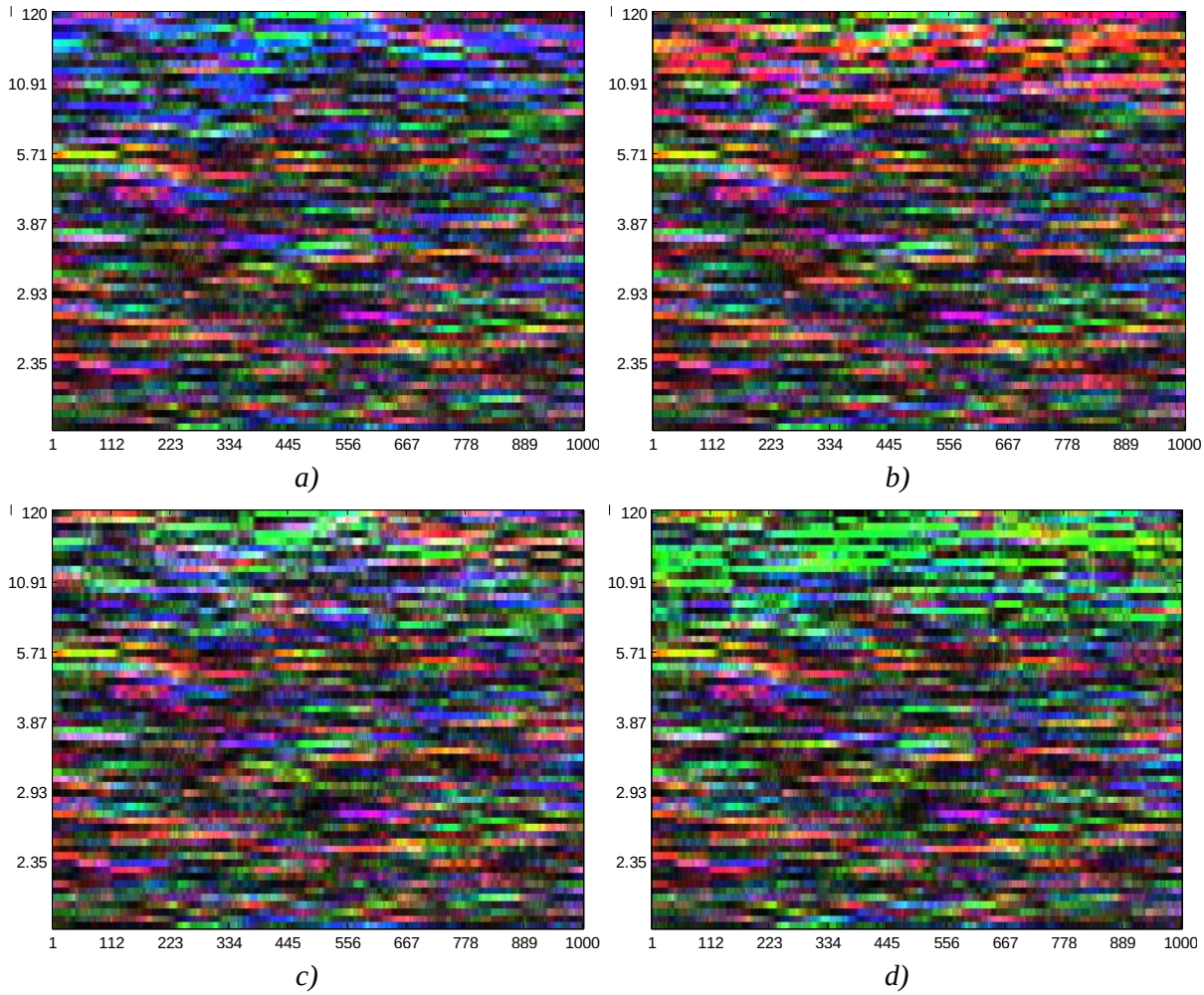


Obr.29 Spektrogram segmentu sekvence AL133493 lidského chromozomu 21 s minisatelity

Rozptýlené repetice

Doposud byly pro testování algoritmů použity sekvence obsahující tandemové repetice. Pro zjištění, jak se ve spektrogramech projevují repetice rozptýlené byly nejprve vytvořeny umělé sekvence. Tyto jsou dvě – *umele_rozptylene_repetice.fasta* obsahující krátké repetice o délce 5 bp vyskytující se v sekvenci 31krát a *umele_rozptylene_repetice_dlouhe.fasta* obsahující dlouhé repetice o délce 71 bp opakující se v sekvenci 3krát. Pro obě sekvence byly postupně vykreslovány čtyři spektrogramy, vždy se změnou nukleotidů v repeticích.

První testovanou byla sekvence *umele_rozptylene_repetice.fasta*, kde opakující se pětinukleotidová repetice byla *a)* AAAAA, *b)* TTTTT, *c)* GGGGG, *d)* CCCCC (viz *Obr.30*). Podle barevných mapovacích vektorů ze vztahu 24 je báze A znázorněna ve spektrogramu modře, repetice jsou tedy detekovatelné z *Obr.30a)* jako oblasti modře zbarvené nacházející se ve vrchní části spektrogramu nad periodou 5. Podobně to platí pro ostatní báze, kdy T je mapováno červeně (*Obr.30b*), G je mapováno šedou barvou (*Obr.30c*) a C, které je jako opakující se repetice CCCCC vidět nejlépe (*Obr.30d*), je mapováno zeleně. Parametry pro výpočet spektra byly nastaveny na hodnoty $\boldsymbol{w} = 120$, $\boldsymbol{o} = 0$.



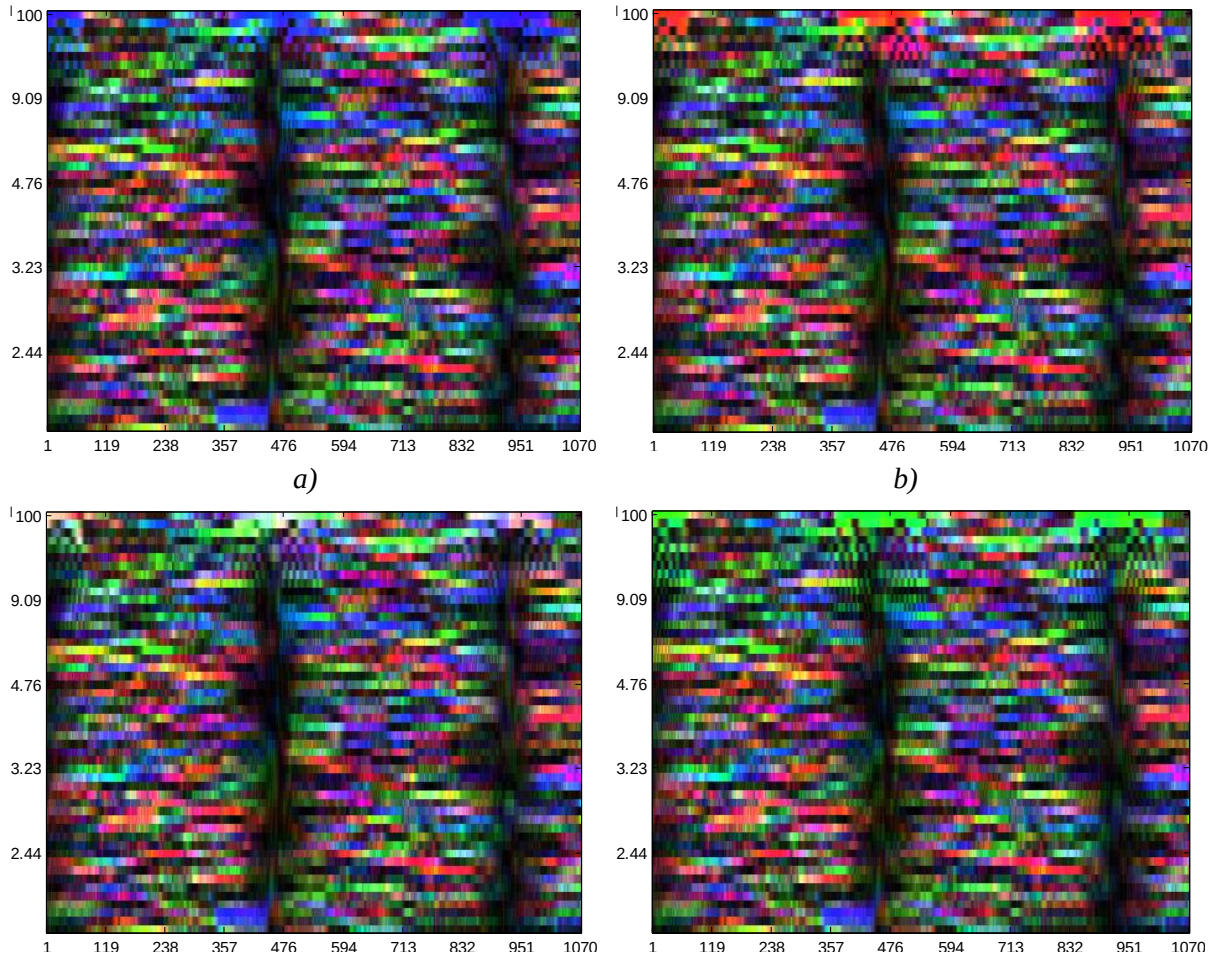
Obr.30 Spektrogramy s umělými krátkými rozptýlenými repeticemi

Druhou testovanou sekvencí byl soubor *umele_rozptylene_repetice_dlouhe.fasta* obsahující dlouhé rozptýlené repetice. Repetice byla tvořena stejnými nukleotidy, tak jako u předchozího příkladu, a opět bylo vyzkoušeno, jak se barevně projeví, obsahuje-li jeden daný nukleotid. Na *Obr.31* jsou čtyři spektrogramy s repeticí dlouhou 71 bp tvořenou nukleotidy *a)* A, *b)* T, *c)* G, *d)* C. Parametry pro výpočet spektra byly ***w*** = 100 a ***o*** = 0. Barvy pro jednotlivé nukleotidy jsou stejné jako v případě první sekvence.

Na rozdíl od předchozí krátké repetice, je v těchto spektrogramech možno detekovat pozice dlouhé repetice. Nachází se v sekvenci 3krát v místech 1-71 bp, 427-497 bp a 852-924 bp. Podle toho, jakým nukleotidem je tvořena sekvence opakujícího se vzoru, je barevně rozlišena horní část pruhů znázorňujících oblasti repetic.

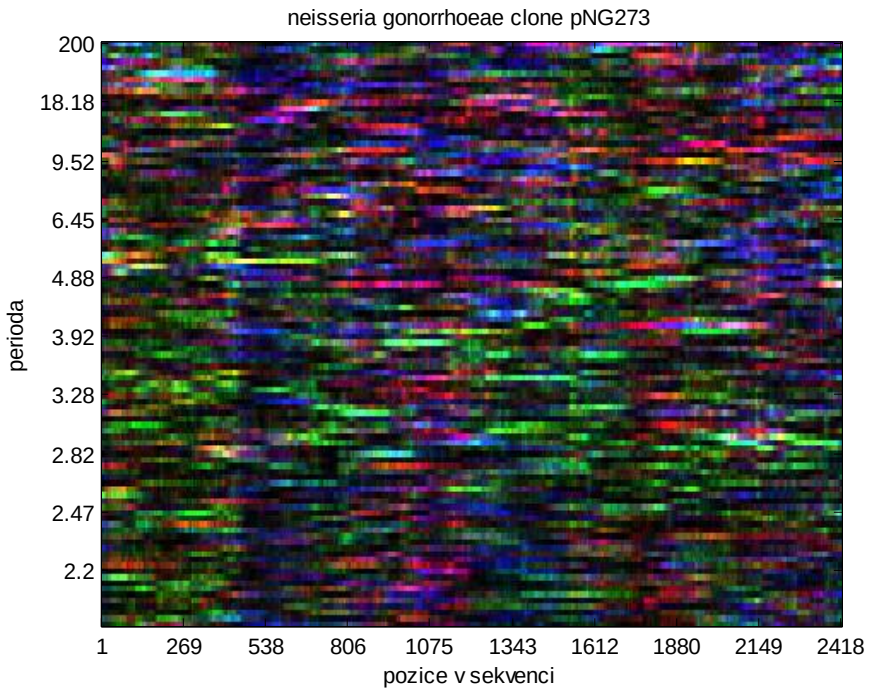
Spektrogramy byly výjimečně vykresleny pomocí skriptu s normalizací umístěnou mimo *for* cyklus, protože díky tomu se zobrazí odlišená oblast s repeticemi s barevně mapovanými bázemi od pozadí spektrogramu. Skript je označen *FT_normalizace_mimo_for.m* a bude dále použit na detekci CpG ostrůvků. Pro jiná měření

nemá jeho použití význam, protože vykresluje stejné obrázky jako hlavní skript *hlavni_1234.m*.

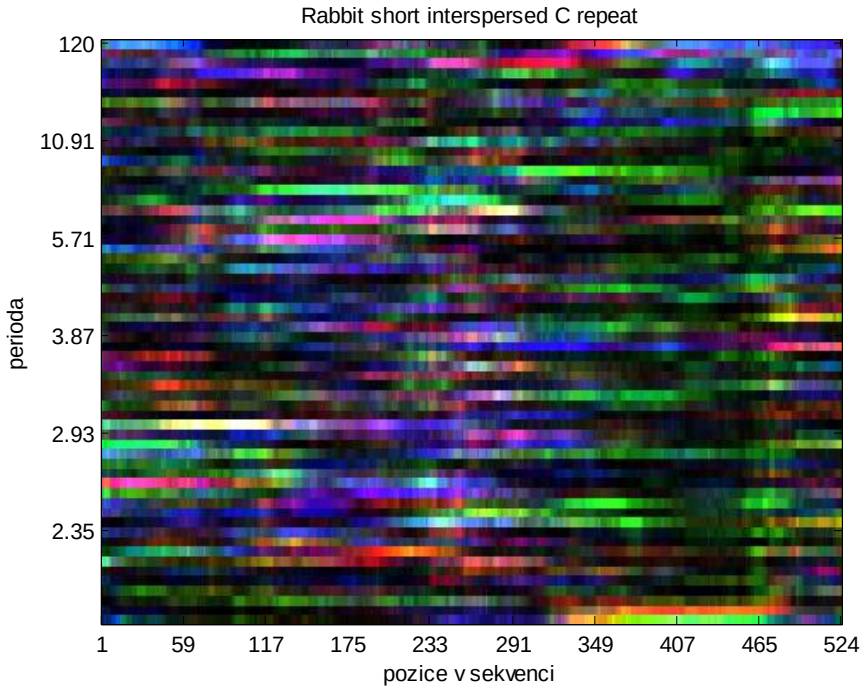


Obr.31 Spektrogramy s umělými dlouhými rozptýlenými repeticemi

Dále byly otestovány reálné sekvence s rozptýlenými repeticemi. Příklad je na *Obr.32*. Jedná se o sekvenci bakterie Neisseria gonorrhoeae označenou *M19675.1* z databáze NCBI. [33] Sekvence se načítá v hlavním skriptu *hlavni_1234.m*. Pro vykreslení bylo okno ***w*** nastaveno na hodnotu 200, posun oken ***o*** na hodnotu 1 a použity barevné mapovací vektory ze vztahu 24. Rozptýlené repetice v sekvenci by měly být dvou typů o délkách 152 bp a 25 bp. Delší z nich se nachází na pozicích 1668-1820 bp a 2257-2408 bp. Kratší je na pozicích 1795-1820 bp a 2383-2408 bp (podle informací o sekvenci na NCBI [33]). Ve výsledném spektrogramu ale bohužel nejsou tyto pozice nějak odlišné od pozadí. Je to pravděpodobně tím, že repetice se skládá z nukleotidů s rovnoměrně zastoupeným počtem od každého z nich a nemůže tedy být nijak barevně výraznější než ostatní části sekvence.



Obr.32 Spektrogram sekvence M19675.1 Neisseria gonorrhoeae obsahující rozptýlené repetice

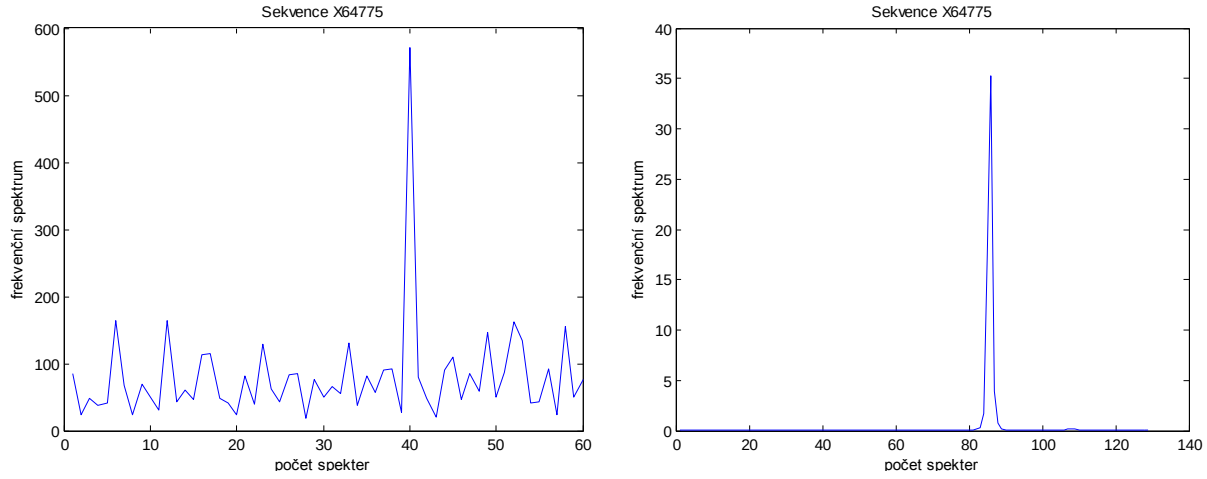


Obr.33 Spektrogram sekvence X02216.1 Rabbit short interspersed C repeat obsahující rozptýlené repetice

Druhou použitou reálnou sekvencí byla králičí DNA označená v databázi NCBI *X02216.1*. [33] Tato sekvence obsahuje rozptýlené repetice typu SINE o délce 13 bp na pozicích 64-77 bp a 432-445 bp. V repetici je s nejvyšším zastoupením nukleotid A, měla by být tedy v obrázku odlišena červenou barvou. Tato místa ve spektrogramu na *Obr.33* ale vyčíst nedokážeme. Nejspíše proto, že počet repetic je moc malý. Jediné, co je zde dobře vidět je zelená oblast mezi pozicemi cca 300-500 bp, která značí přítomnost nadměrného počtu nukleotidů C vůči ostatním nukleotidům. Pro vykreslení bylo okno ***w*** nastaveno na hodnotu 120, posun oken ***o*** na hodnotu 1 a použity barevné mapovací vektory stejné jako pro předchozí sekvenci.

Kódující sekvence

Identifikace genových regionů (exonů a intronů) probíhá podle literatury většinou z vykreslených frekvenčních (výkonnostních) spekter, která získáme sečtením spekter pro každý nukleotid umocněných na druhou podle rovnice 13. Výkonnostní spektrum odhaluje píky na frekvenci k = N/3, což koresponduje s periodou tří vzorků - tedy tří bází, které zastupují jeden kodón. Kódující části sekvence se určují podle polohy těchto píků ve spektru. Tato analýza se ale netýká této práce. Příklad frekvenčního spektra vidíme na *Obr.34*. Obrázek *a)* znázorňuje frekvenční spektrum sekvence O. sativa [22] získané pomocí FT, na obrázku *b)* je potom spektrum získané využitím AR modelu.



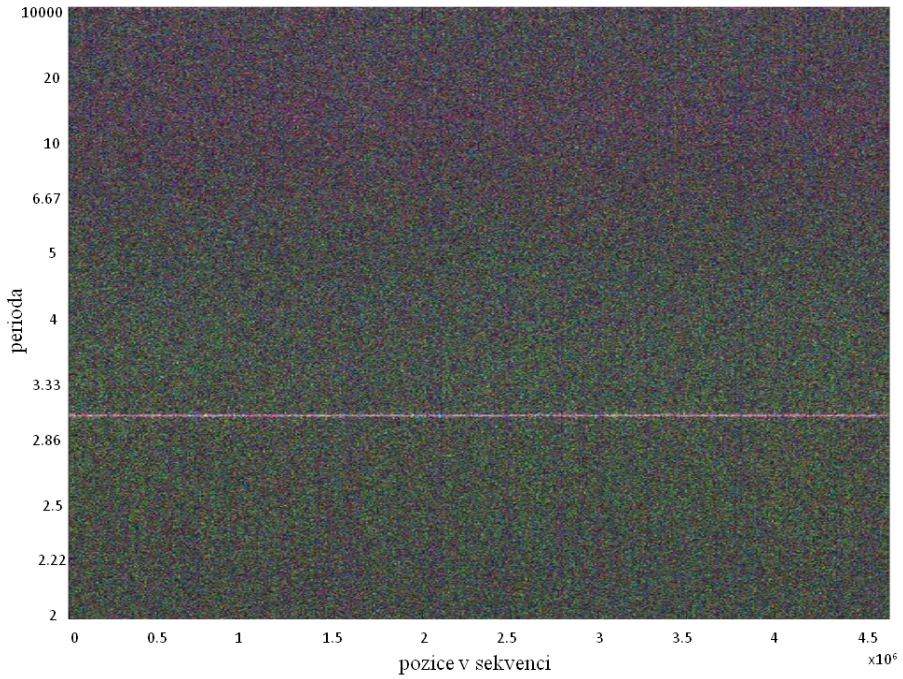
a) *b)*

Obr.34 Frekvenční spektrum O. sativa: a)FT, b)AR model

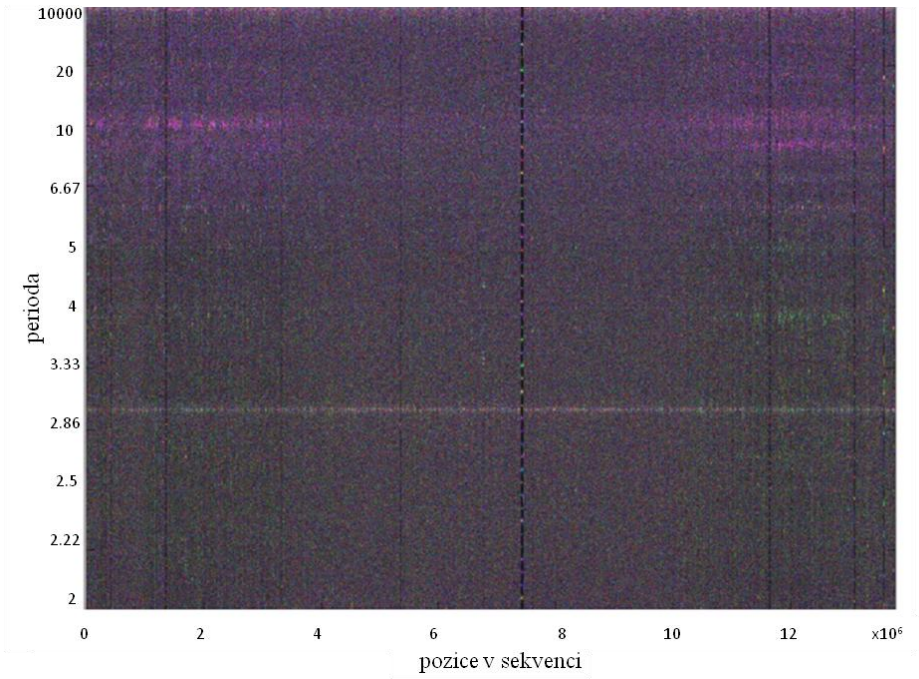
Z barevných DNA spektrogramů se identifikace kódujících regionů také provádí. Většinou se informace o kódujících místech v genech vyhledávají ze spektrogramů celých chromozomů. Doba výpočtu algoritmu pro tak dlouhé sekvence nukleotidů je však velmi dlouhá.

Kódující části v sekvenci jsou z barevných spektrogramů rozlišitelné podle přítomnosti světlé vodorovné linie v místě periody 3 na svislé ose. Čím více je tato linie

výrazná, svědčí to o tom, že více bází se podílí na kódování proteinů. Např. chromozom K12 E. coli má velmi výraznou linii (viz *Obr.35*) na rozdíl od chromozomu III C. elegans (háďátka obecného), ve kterém se jakožto v eukaryotickém organismu nachází více mezigenových nekódujících částí a proto je linie slabší (viz *Obr.36*). [14] Tyto celé chromozomy byly vybrány pro ukázku z literatury, protože výpočet spektrogramu by byl příliš náročný.

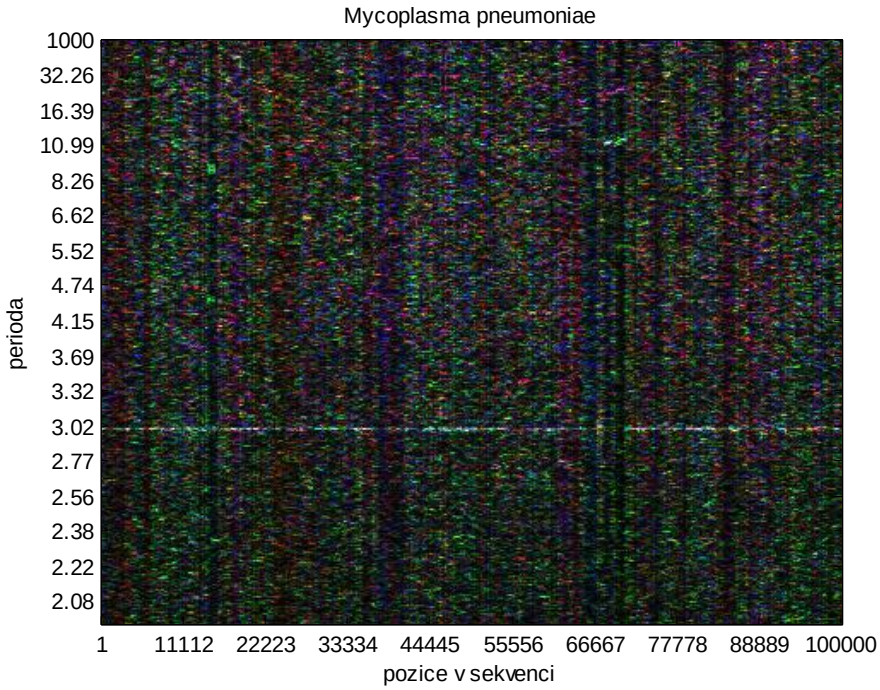


Obr.35 Spektrogram chromozomu K12 E. coli, okno 10000, posun oken 0 [14]



Obr.36 Spektrogram chromozomu III C. elegans, okno 10000, posun oken 0 [14]

Na dalším *Obr.37* už je vlastní spektrogram DNA sekvence bakterie Mycoplasma pneumoniae. [33] Jedná se o část jediného chromozomu této bakterie, tedy část jejího genomu vymezenou pozicemi 1-100000 bp. Pro vykreslení byly nastaveny hodnoty okna ***w*** = 1000, posun okna ***o*** = 10, barevné mapovací vektory ze vztahu 24. Soubor načítaný hlavním skriptem *hlavni_1234.m* se jmenuje *Mycoplasma pneumoniae.fasta*. Ve spektrogramu je možno vidět bílý vodorovný pruh v místě periody 3, tedy sekvence této bakterie obsahuje kódující regiony.

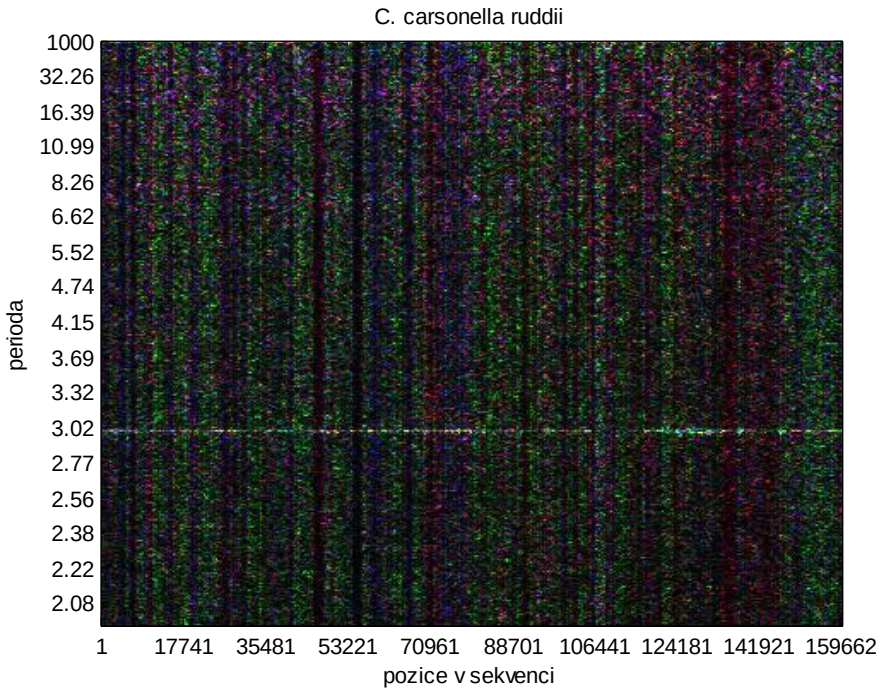


Obr.37 Spektrogram části genomu M. pneumoniae

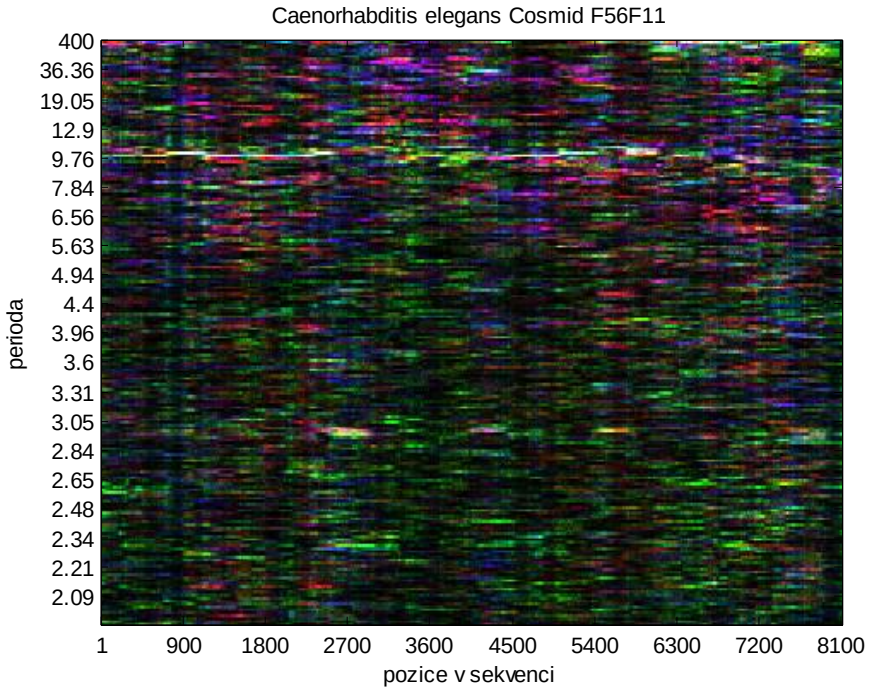
C. carsonella ruddii.fasta je soubor obsahující sekvenci proteobakterie Candidatus Carsonella ruddii s nejmenším genomem ze všech buněčných organismů. [33] *Obr.38* reprezentuje spektrogram celého genomu tohoto organismu s patrnou vodorovnou linií v místě periody rovné třem, což značí přítomnost kódujících částí sekvence. Jelikož se jedná o prokaryota, která obsahují velmi malé množství mezigenových nekódujících částí, je linie celkem výrazná. Totéž platí i pro předchozí sekvenci bakterie Mycoplasma pneumoniae. Spektrogram byl vykreslen s parametry ***w*** = 1000, ***o*** = 100, barevnými mapovacími vektory jako u M. pneumoniae.

Jiné vlastnosti můžeme vyčíst z *Obr.39*, který znázorňuje spektrogram sekvence genu F56F11 organismu C. elegans. [33] V tomto genu můžeme rozlišit oblasti s vyšším výskytem bází C a G, které se nacházejí pod bílou vodorovnou linií na periodě 10. Nad touto linií v oblasti s nižší frekvencí se objevuje spíše fialový odstín určující přítomnost bází A a T. Bílá linie v místě periody 10 může souviset s helikální strukturou DNA. Ta se

totiž opakuje právě po deseti bázích (podle lit. 14). Parametry pro vykreslení byly $\boldsymbol{w}$ = 400 a $\boldsymbol{o}$ = 10. Sekvence je uložena pod názvem *Caenorhabditis elegans Cosmid F56F11.fasta*.



Obr.38 Spektrogram genomu proteobakterie C. Carsonella ruddii



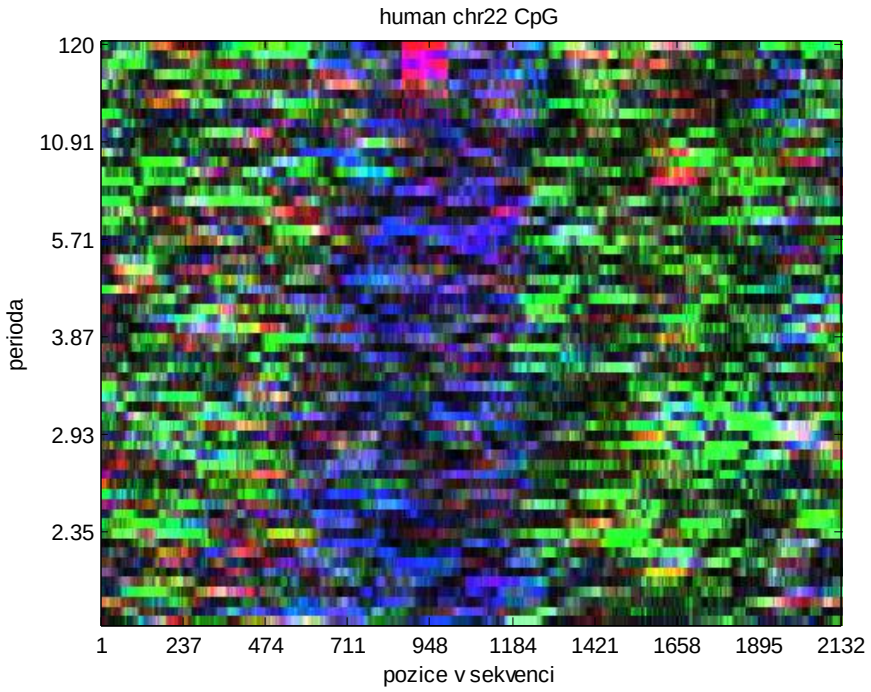
Obr.39 Spektrogram sekvence genu F56F11 C. elegans

CpG ostrůvky

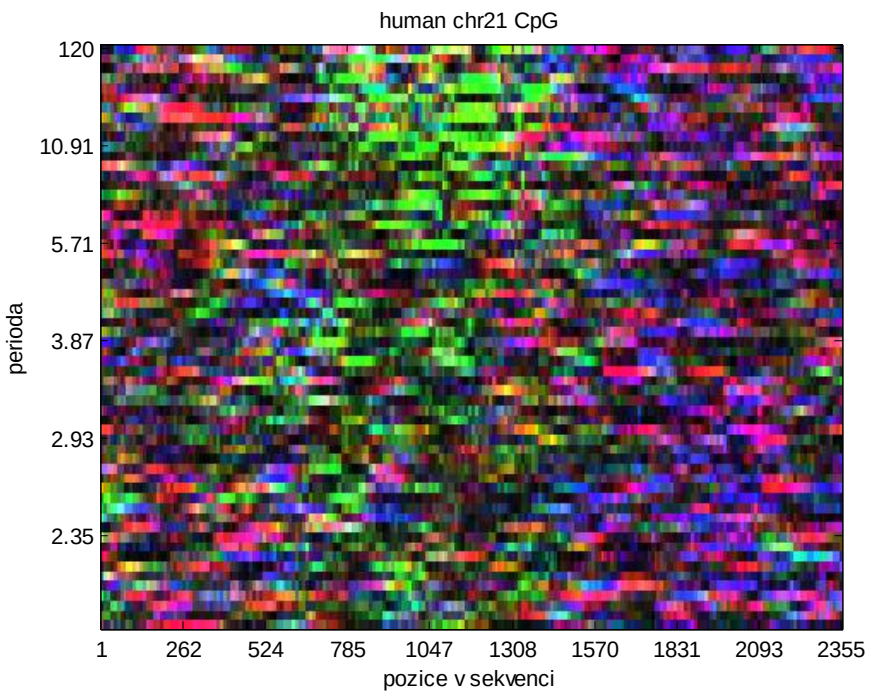
CpG ostrůvky jsou ve spektrogramech rozeznávány jako zeleno-šedé oblasti s délkou okolo 200 bp. Vyskytují se ve většině promotorů lidských genů. V segmentu sekvence lidského chromozomu 22 vymezeného pozicemi 2894684-2896815 bp jsou takovéto oblasti s vyšším obsahem dinukleotidů CG ve dvou širších pruzích v levé a v pravé části spektrogramu, což vidíme na *Obr.40*. [33] Sekvence je uložena pod názvem *NT_011519.9 human chr22.fasta*. Pro vykreslení bylo použito okno o velikosti 120, posun oken s nastavenou hodnotou 1 a barevné mapovací vektory ze vztahu 24.

Další testovanou sekvencí byla *Homo sapiens chromosome 21.fasta*. [33] Spektrogram s CpG ostrůvky ve střední části vidíme na *Obr.41*. Parametry pro výpočet spektra byly stejné, jako u předchozí sekvence. Segment je určen rozmezím 9905604-9907958 bp.

Pro vykreslení CpG ostrůvků bylo výjimečně potřeba upravit základní skript *hlavni_1234.m* tak, že byl krok normalizace hodnot RGB umístěn mimo *for* cyklus. Hodnoty z předchozího kroku (mapování do RGB prostoru) se tedy nenormalizují průběžně z krátkých vektorů určených krokem ***o*** ve *for* cyklu, ale až z celkových vektorů R, G, B po ukončení *for* cyklu. Tato úprava byla již použita u vykreslování umělých rozptýlených repetic a vznikl tak skript *FT_normalizace_mimo_for.m*.



Obr.40 Spektrogram segmentu sekvence lidského chr. 22 s CpG ostrůvky



Obr.41 Spektrogram segmentu sekvence lidského chr. 21 s CpG ostrůvky

5.2. Autoregresní model

Hlavní funkce, které volají podfunkci *p_burg.m* jsou pojmenované *hlavni_armodel_norm_mimo_for_cyklus.m* a *hlavni_armodel_norm_ve_for_cyklu.m*. Obsahují stejné kroky jako hlavni skript pro FT s jedinou odlišností – se dvěma normalizacemi, jak už bylo zmíněno na konci úvodu kap. 5. Proto se podfunkce *p_burg.m* nevolá stejně jako *DFT.m* realizující Fourierovu transformaci hlavním skriptem *hlavni_1234.m*.

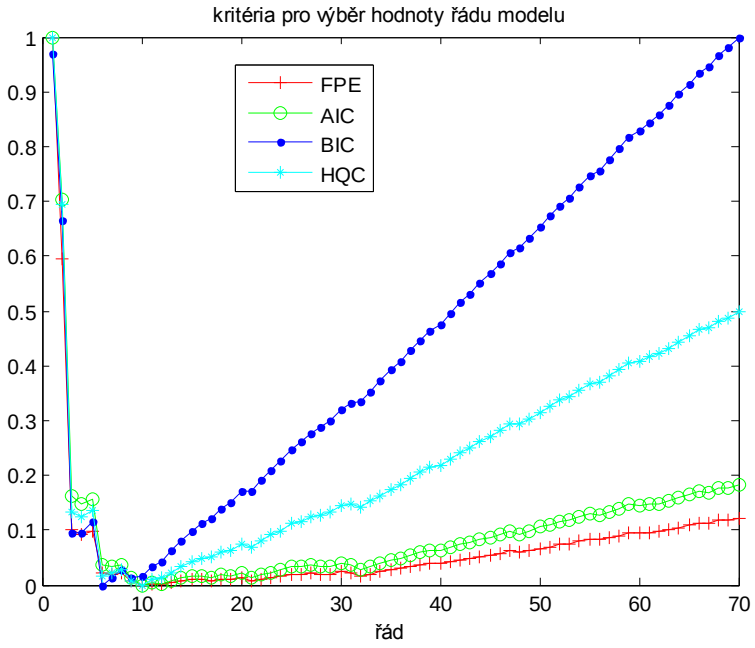
Funkce *p_burg.m* obsahuje tři Matlabovské příkazy pro výpočet autoregresního modelu. První z nich je *'arburg'*, který provádí odhad parametrů ***a*** AR modelu Burgovou metodou. Hlavním principem Burgova algoritmu je odhad rozdílu mezi signálem a ideálním modelem signálu, který je formulován rovnicemi 14. Vstupem pro funkci *'arburg'* je vektor hodnot ***x*** z numerické reprezentace v hlavním skriptu určený sloupci *u* a řádky *v* (podobně jako u FT v kap. 5). Vypočtené parametry ***a*** AR modelu jsou potom spolu s vektorem ***x*** vstupem pro další příkaz *'filter'*. Tento filtruje data z vektoru ***x*** za použití digitálního filtru v čitateli s vektorem jedniček a ve jmenovateli právě s vektorem parametrů ***a***. Výstup ***y*** spolu s hodnotou řádu modelu ***p*** jsou potom vstupními proměnnými pro poslední příkaz *'pburg'*. Tato Matlabovská funkce provádí odhad výkonnostního spektra vektoru ***y*** a výsledek ukládá do proměnné ***spektrum***, což je výstup pro hlavní funkci AR modelu.

```
a = arburg(x,p);
y = filter(1,a,x);
spektrum = pburg(y,p);
```

Dalším vstupním parametrem pro první příkaz *'arburg'* musí být řád modelu ***p***. Jeho hodnota se zadává v hlavním skriptu a odvíjí se od výsledku algoritmu ve skriptu *kriteria.m*. Tento skript realizuje výpočet kritérií FPE, AIC, BIC a HQC pro výběr hodnoty řádu modelu (viz rovnice 16-20). Nejprve se načte pomocí příkazu *fastaread* stejná sekvence, pro kterou chceme nechat vykreslit hlavním skriptem spektrogram. Potom se sekvence musí konvertovat do numerické reprezentace. V tomto případě je lepší pro další výpočty použít EIIP numerickou reprezentaci (viz kap. 3.2.1). Vytvoříme si tak jeden vektor místo čtyř u 4D binární Vossovy reprezentace. Vzorce pro výpočet kritérií jsou uvnitř *for* cyklu, který je propočítá pro hodnoty od 1 do ***q***, což je předem nastavená hodnota maximálního řádu (odhadem). Prvně ale *for* cyklus spočítá Matlabovskou fcí *'arburg'* odhad parametrů ***a*** AR modelu a hodnoty ztrátové funkce ***Vp***, které jsou součástí rovnic pro výpočet kritérií. Ztrátová funkce je vlastně normalizovaný součet druhých mocnin náhodných chyb *e(n)* = rozdíl mezi odhadovaným a vstupním signálem.

```
for i = 1:p
    [a,Vp] = arburg(num_repr,i);
    FPE(i) = Vp*((L+i)/(L-i));
      % finální predikce chyby
    AIC(i) = log(Vp)+((2*i)/L);
      % Akaikeho informační kritérium
    BIC(i) = log(Vp)+((i*log(L))/L);
      % Bayesovo informační kritérium
    HQC(i) = log(Vp)+((2*i*log(log(L)))/L);
      % Hannan-Quinnovo kritérium
end
```

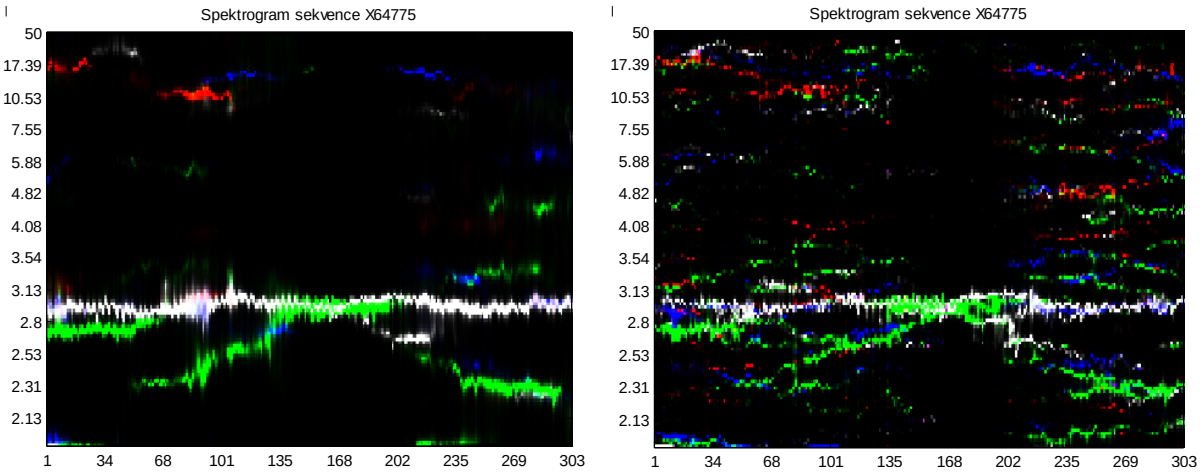
Před vykreslením všech kriteriálních křivek do jednoho grafu se ještě provede normalizace hodnot kritérií na rozsah 0-1 podle rovnice 25. Výsledek algoritmu vidíme na *Obr.42*. Testována byla sekvence O. sativa *X64775.1* [22].



Obr.42 Kritéria pro výběr hodnoty řádu AR modelu sekvence X64775.1

Tento skript *kriteria.m* tedy vznikl proto, abychom mohli určit správnou hodnotu řádu AR modelu a nastavit ji v hlavním skriptu. Ta se podle lit. 2 určí jako nejlepší, když kriteriální funkce dosáhnou minima. Potom korespondující hodnota řádu vyčtená z vodorovné osy je stanovena jako vhodný řád AR modelu. Zde pro sekvenci O. sativa můžeme z grafu stanovit řád p = 10.

Spektrogram sekvence O. sativa s řádem nastaveným na hodnotu 10, oknem o velikosti 50, posunem okna 1 a mapovacími vektory ze vztahu 24 vidíme na *Obr.43a)*. Zvýšíme-li hodnotu řádu, může spektrogram obsahovat falešné vzory, jako na *Obr.43b)*. Příliš nízká hodnota řádu by měla za následek přehnaně vyhlazený spektrální odhad.



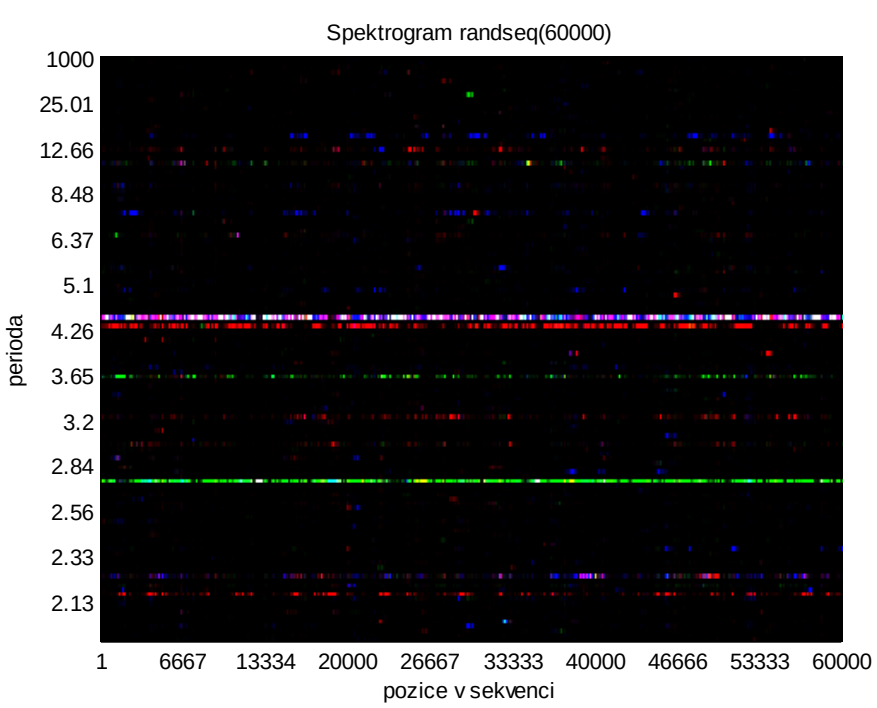
a) *b)*

Obr.43 Spektrogram sekvence O. sativa vytvořený AR modelem s řádem nastaveným na hodnotu a) p=10, b) p=30

5.2.1. Detekce vzorů ve spektrogramech pomocí AR modelu

Dva hlavní skripty pro vytvoření spektrogramů pomocí AR modelu *hlavni_armodel_norm_mimo_for_cyklus.m* a *hlavni_armodel_norm_ve_for_cyklu.m* se mezi sebou liší v umístění kroku normalizace. Už podle názvu se jedná o umístění ve *for* cyklu nebo mimo *for* cyklus. Cílem této kapitoly bude vyzkoušet, který skript vykresluje lepší spektrogramy a potom ho použít pro výpočet stejných sekvencí jako u FT. Výsledné spektrogramy z obou metod se dále budou porovnávat v kap. 6, což je hlavním úkolem celé práce.

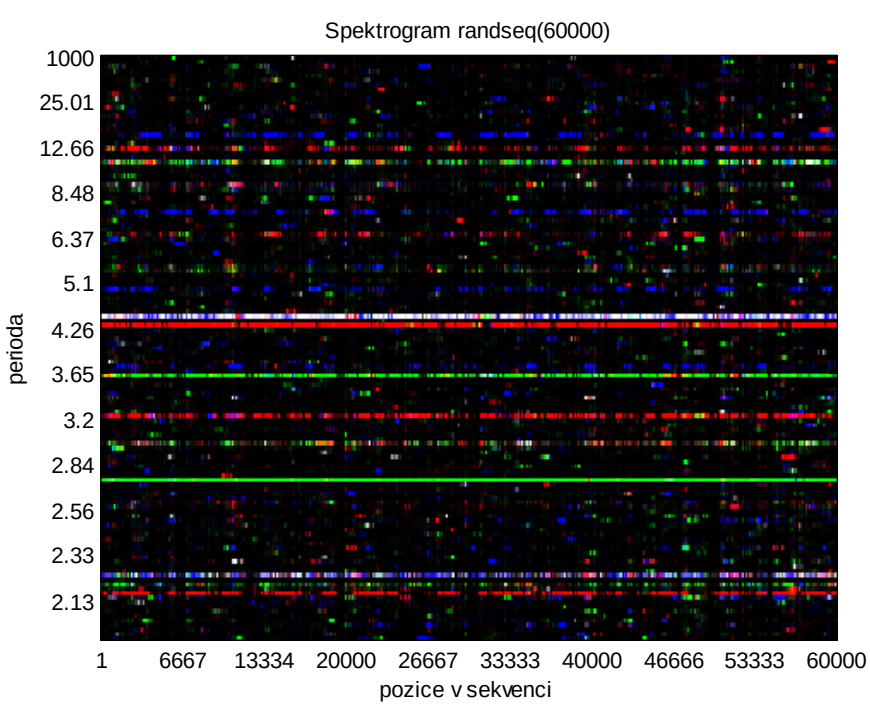
Nejprve pro ověření, zda algoritmus funguje správně byla vykreslována náhodná sekvence o délce 60 bp s bázemi A, T, C a G opakujícími se s periodou 15, 13, 11 a 9, stejná jako byla použita v kap. 5.1.1. Nejlepší parametry byly několikerým testováním zvoleny pro velikost okna $\boldsymbol{w}$ = 1000, pro posun okna $\boldsymbol{o}$ = 100 a pro řád modelu $\boldsymbol{p}$ = 100. Barevné mapovací vektory byly vybrány ze vztahu 24. Výsledný spektrogram je na *Obr.44*. Tento obrázek získáme spuštěním skriptu *hlavni_armodel_norm_ve_for_cyklu.m*, tedy pořadí kroků i umístění kroku normalizace je stejné jako v hlavním skriptu pro FT *hlavni_1234.m*. Normalizace je použita jen jedna podle vztahu 25.



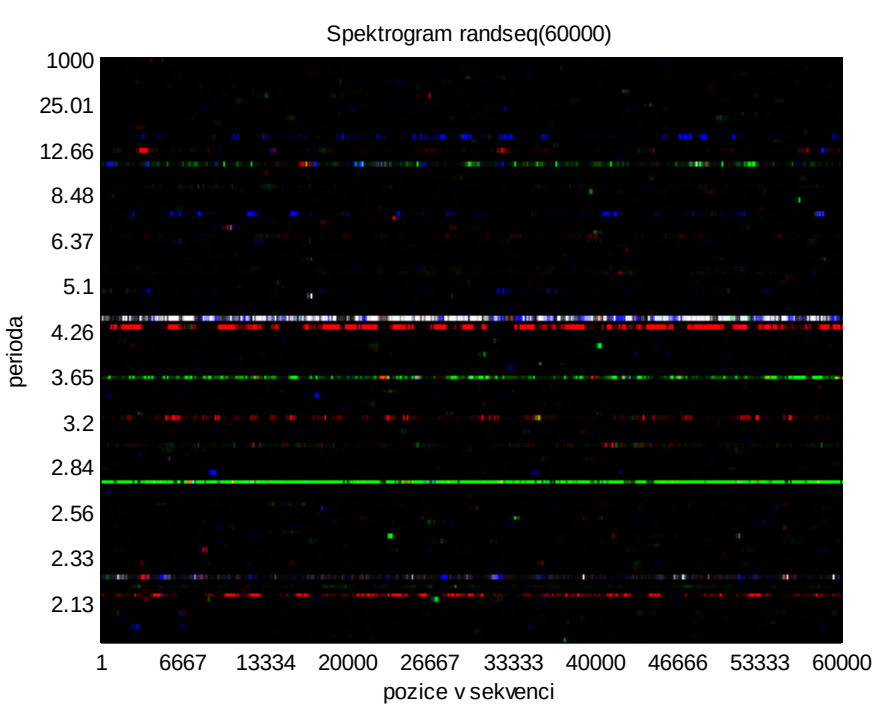
Obr.44 Spektrogram náh. sekvence o délce 60 kbp s periodicky se opak. výskytem nukleotidů získaný pomocí AR modelu s jednou normalizací

Jak už bylo zmíněno v úvodu programové části této práce v kap. 5, pro AR model je potřeba použít dvě normalizace. Je to proto, že pozadí je příliš tmavé a není možno dobře odlišit všechny vzory, které nás zajímají. Spektrogram s oběma normalizacemi je na *Obr.45* a srovnáme-li ho s předchozím obrázkem, můžeme vidět, že právě skript se dvěma normalizacemi rozliší více vodorovných pruhů znázorňujících periodicky se opakující báze. Např. v *Obr.45* jsou navíc linie okolo periody 2,13 a 12,66. Ovšem zdá se, že je to na úkor toho, že celý spektrogram je potom barevnější a některé vzory by mohly být falešné.

Právě proto následovalo vyzkoušení umístění kroku dvou normalizací mimo *for* cyklus. Toto řešení je zapsáno ve skriptu *hlavni_armodel_norm_mimo_for_cyklus.m*. Parametry ***w***, ***o*** a ***p*** pro výpočet spektra, i barevné mapovací vektory, byly ponechány stejné. Výsledný spektrogram je na *Obr.46*. Dalo by se říci, že obrázek je velmi podobný *Obr.44* ze skriptu s jednou normalizací. Pro další závěry je tedy nutné vyzkoušet jiné sekvence a srovnat skripty s normalizací mimo a ve *for* cyklu.



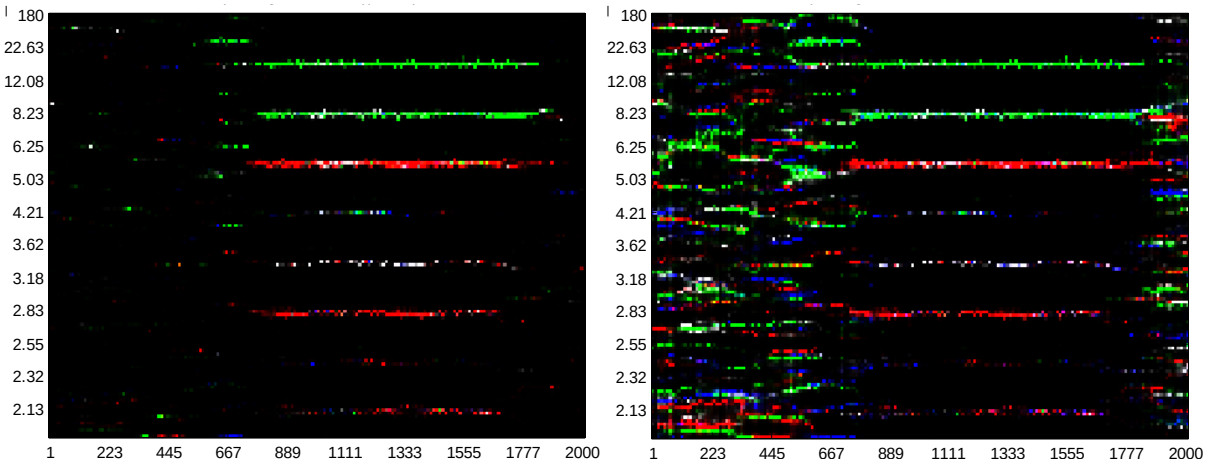
Obr.45 Spektrogram náh. sekvence o délce 60 kbp s periodicky se opak. výskytem nukleotidů získaný pomocí AR modelu se dvěma normalizacemi



Obr.46 Spektrogram náh. sekvence o délce 60 kbp s periodicky se opak. výskytem nukleotidů získaný pomocí AR modelu s normalizacemi mimo for cyklus

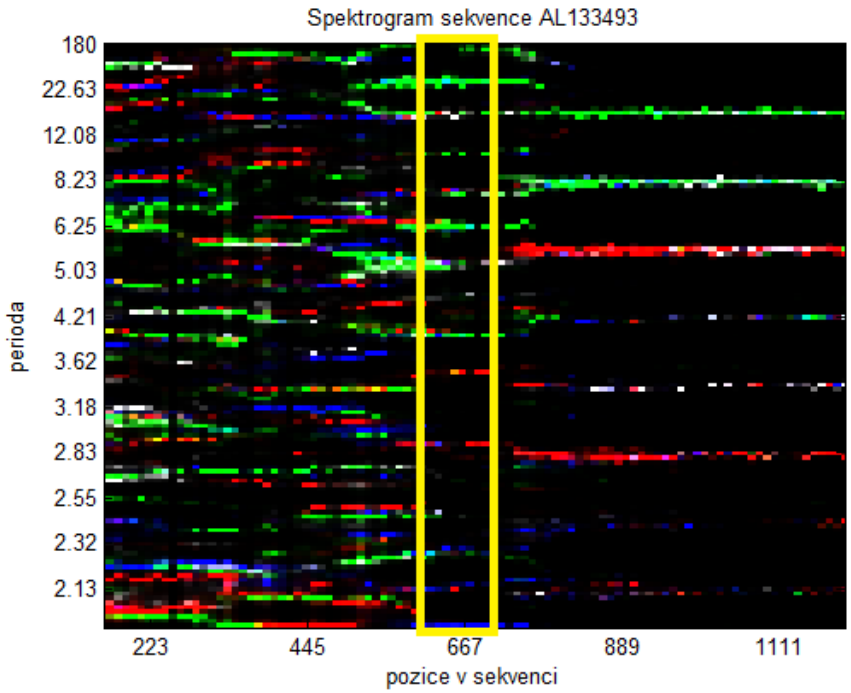
Tandemové repetice

Jako další sekvence pro srovnání skriptů s krokem normalizace umístěným mimo a ve *for* cyklu byla zvolena *AL133493_human_chr21.fasta* s minisatelity. [24] Porovnání výsledků ze dvou skriptů je na obrázcích *Obr.47a)* a *Obr.47b)*. Hodnoty pro výpočet spektra byly: ***w*** = 180, ***o*** = 10 a ***p*** = 50. Barevné mapovací vektory jsou stejné jako pro předchozí náhodnou sekvenci.



Obr.47 Porovnání skriptů s krokem normalizace umístěným: a) mimo for cyklus, b) ve for cyklu pro sekvenci AL133493_human_chr21

Tato sekvence byla zvolena, protože obsahuje tandemové repetice – minisatelity, podle kterých se dají spektrogramy porovnávat. Je žádoucí, aby spektrogram rozlišil dvě různé repetice: první na pozicích 623-737 bp a druhou na pozicích 802-1768 bp. Právě ta první nám umožňuje rozhodnout, který skript je vhodnější. Je jím ten s normalizací ve *for* cyklu (obrázek *b*), protože rozliší víc vodorovných proužků na pozicích 623-737 bp. Mělo by jich být 16 podle vzorce z lit. 14 zmíněného v této práci na straně 45, neboť délka opakujícího se vzoru je 32 bp. Detail je na *Obr.48*, kde vidíme ve žlutém obdélníku právě danou oblast s 16-ti vodorovnými proužky.

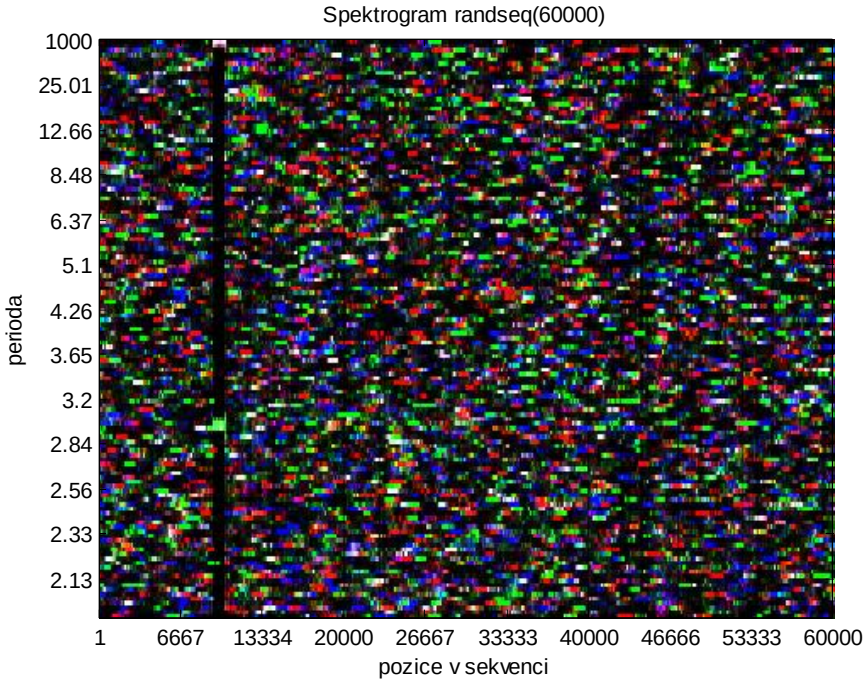


Obr.48 Detail z Obr.47b) ukazující počet vodorovných proužků na poz. 623-737 bp

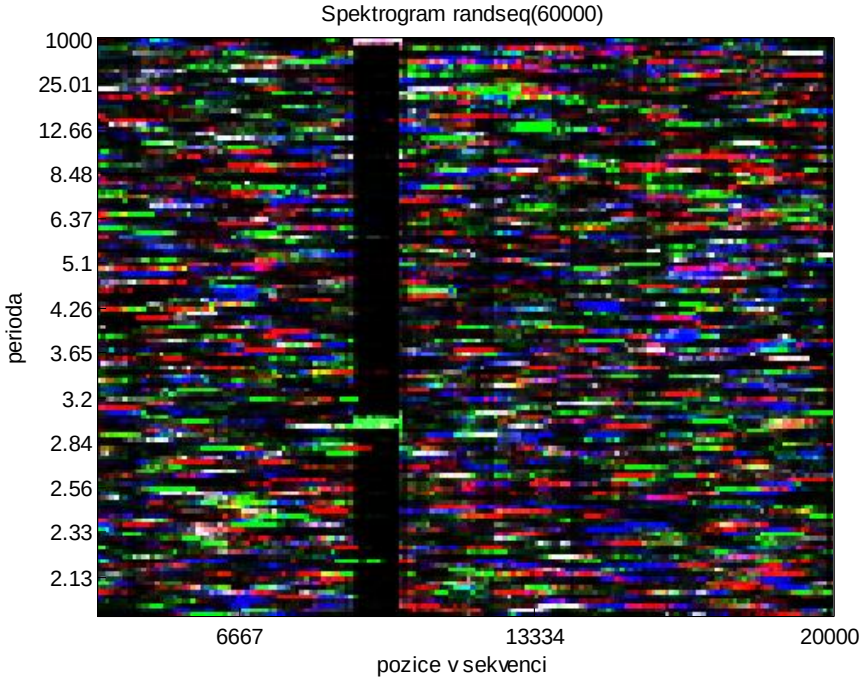
Pro sekvenci *AL133493* s minisatelity i pro náhodnou sekvenci s periodicky se opakujícím výskytem nukleotidů bylo tedy vhodnější použít skript *hlavni_armodel_norm_ve_for_cyklu.m*.

Další sekvencí, kterou můžeme otestovat je náhodná sekvence o délce 60 kbp s tandemovou trinukleotidovou repeticí opakující se v délce 150 bp = mikrosatelitem, tedy stejná jako v kap. 5.1.2. Měl by se okolo pozice 10 kbp objevit tmavý pruh, ve kterém je v místě periody 3 zelená oblast značící přítomnost báze C v repetici. Z *Obr.49* je poznat, že skript *hlavni_armodel_norm_ve_for_cyklu.m* funguje správně.

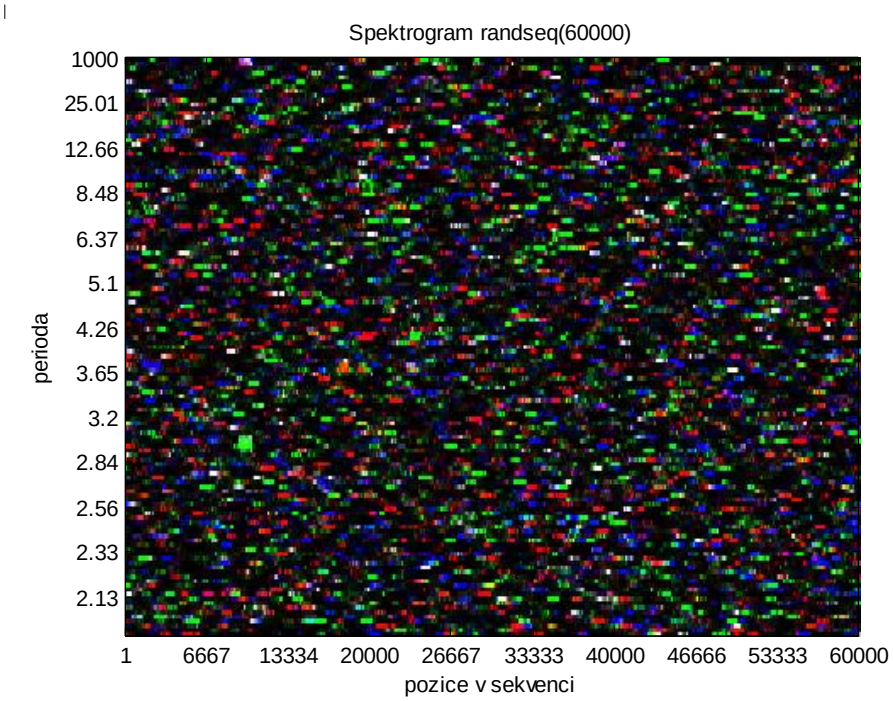
Ponecháme-li ale stejné parametry ***w*** = 1000, ***o*** = 100, ***p*** = 100 a použijeme-li je pro vykreslení stejné sekvence skript *hlavni_armodel_norm_mimo_for_cyklus.m*, získáme spektrogram na *Obr.51*. Repetice už není tak jasně vidět, jako v předchozím obrázku. Opět tedy normalizace umístěná ve *for* cyklu zaručuje lepší výsledek.



Obr.49 Spektrogram náhodné sekvence s umělou repeticí CGG, normalizace ve for cyklu

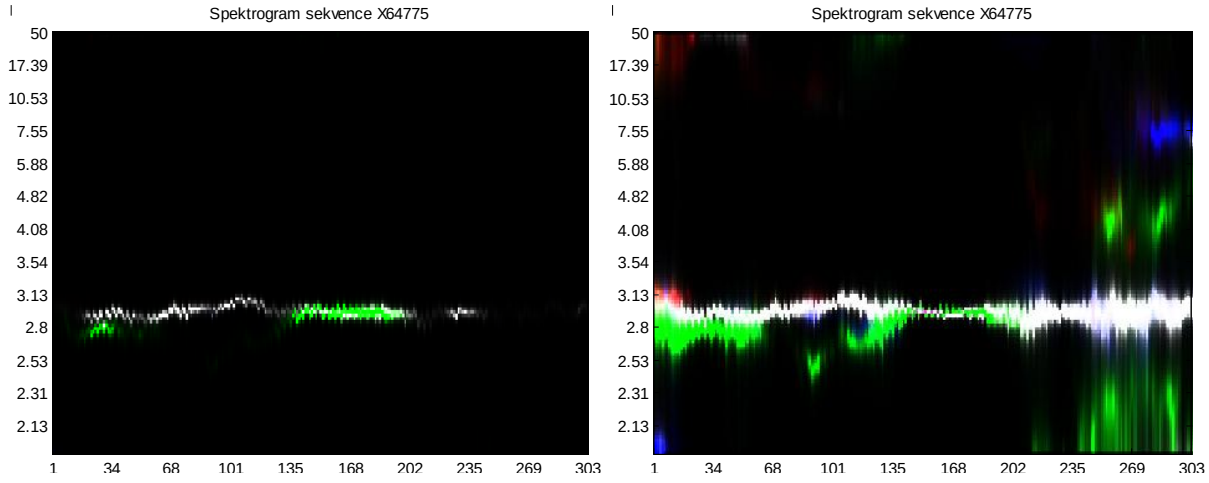


Obr.50 Detail okolo pozice 10 kbp z Obr.49



Obr.51 Spektrogram náh. sekvence s umělou repeticí CGG, normalizace mimo for cyklus

Výjimkou, kdy je lepší použít normalizaci mimo *for* cyklus je sekvence O. sativa, která obsahuje mikrosatelit – trinukleotidovou repetici opakující se ve velmi malé délce 44 bp. V *Obr.52a)* můžeme vidět spektrogram s vodorovným zeleným pruhem mezi pozicemi 135 bp a 202 bp, což při přiblížení odpovídá reálným pozicím repetice (142-186 bp). Zatímco ve spektrogramu vpravo na obr. *b)* je tato oblast těžko rozlišitelná. Bílý vodorovný pruh naznačuje, že délka opakujícího se vzoru je 3 báze, neboť se tato linie nachází na svislé ose v oblasti periody 3. Pro toto měření byly nastaveny hodnoty $\boldsymbol{w} = 50$, $\boldsymbol{o} = 1$, $\boldsymbol{p} = 5$, barevné mapovací vektory ze vztahu 24.

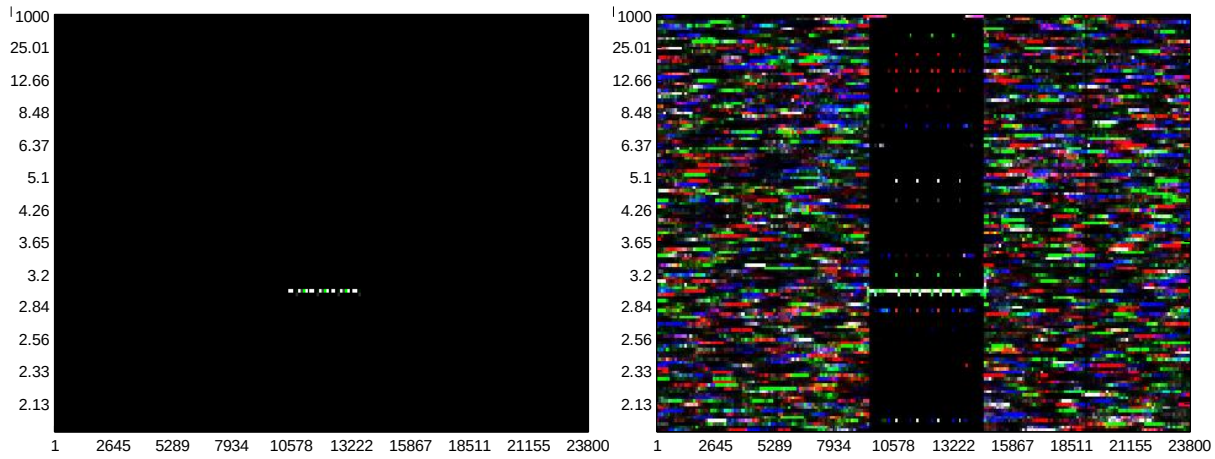


a) *b)*

Obr.52 Spektrogram sekvence O. sativa s mikrosatelity s normalizací a) mimo for cyklus, b) ve for cyklu

Pro sekvence, jejichž délka v poměru k délce obsažené repetitivní části není řádově velmi rozdílné číslo (např. O. sativa je dlouhá 303 bp a mikrosatelity mají 44 bp), je tedy vhodnější použít skript *hlavni_armodel_norm_mimo_for_cyklus.m*. Bylo experimentem ověřeno, že je-li délka sekvence 24 kbp a délka mikrosatelitu 44 bp, spektrogramy vykreslené skripty s normalizací mimo a ve *for* cyklu vypadají podobně. Tento test byl proveden tak, že do náhodné sekvence o délce 24 kbp byla vložena repetice ze sekvence O. sativa na pozici od 10 kbp.

Další experiment slouží pro ověření prvního poznatku o poměru délky sekvence k délce repetitivního vzoru. Nyní byl délka sekvence opět 24 kbp, ale repetice už byla větší = 4 kbp (poměr už je podobný jako u O. sativa). V předchozí umělé sekvenci byla znásobena repetice z O. sativa na délku 4 kbp. Ověřili jsme si, že skript s normalizací mimo *for* cyklus – *Obr.53a),* vykresluje přesně pozici repetice (okolo 10 kbp) a délku opakujícího se vzoru (z periody na ose y). Na rozdíl od toho spektrogram z druhého skriptu na *Obr.53b)* obsahuje nadbytečné informace.



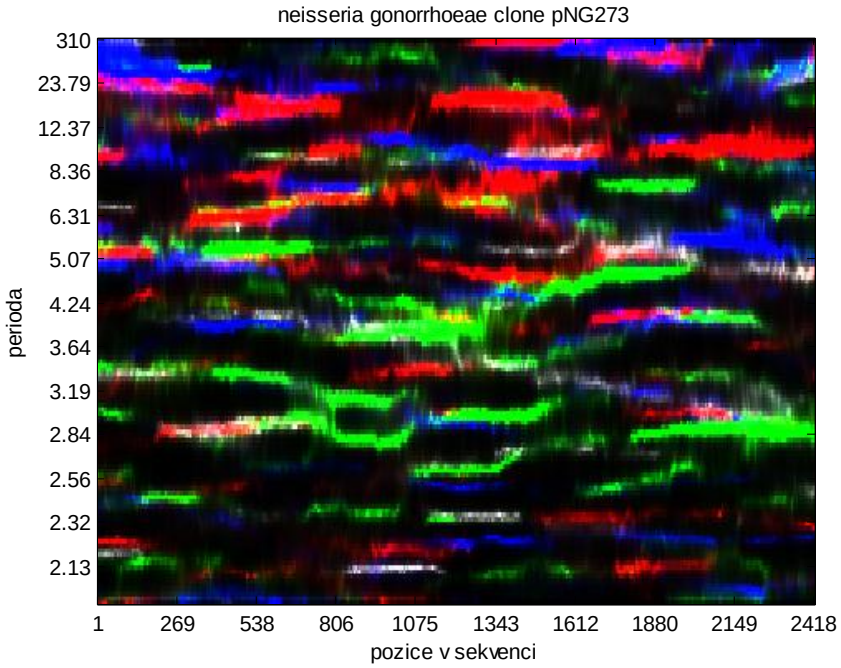
Obr.53 Spektrogram náhodné sekvence s vloženou trinukleotidovou repeticí z O. sativa znásobenou na délku 4 kbp a) norm. mimo for cyklus, b) norm. ve for cyklu

Ovšem podíváme-li se zpět na *Obr.47*, kde je jako lepší zvolena normalizace ve *for* cyklu, přestože poměr délka sekvence / délka vzoru je stejný jako v příkladu na *Obr.53*, odporuje to této teorii. Proto nejlepším řešením bude nechat si vždy vykreslit spektrogramy z obou skriptů.

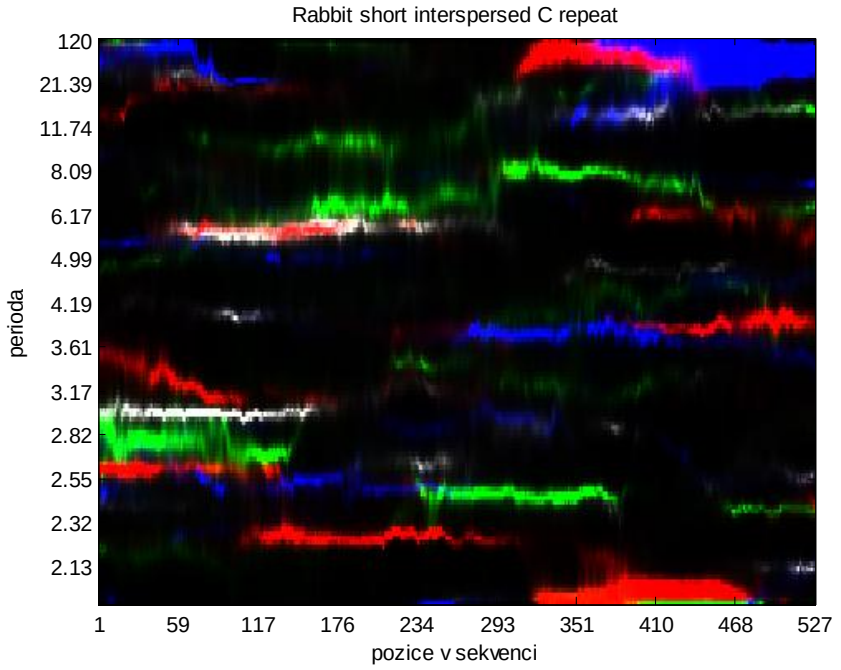
Rozptýlené repetice

Sekvence obsahující rozptýlené repetice mají podle NCBI označení *M19675.1* a *X02216.1* a byly použity v kap. 5.1.2. [33] První z nich – bakterie Neisseria gonorrhoeae zahrnuje rozptýlené repetice dvou typů o délkách 152 bp a 25 bp. Delší z nich se nachází na pozicích 1668-1820 bp a 2257-2408 bp. Kratší je na pozicích 1795-1820 bp a 2383-

2408 bp. Spektrogramy z obou skriptů s normalizací mimo i ve *for* cyklu jsou si velmi podobné, proto je zde uveden pro ukázku jen jeden z nich a to na *Obr.54*. Byl vykreslen s parametry ***w*** = 310, ***o*** = 1, ***p*** = 25. Tak jako v kap. 5.1.2, bohužel z obrázku není repetice detekovatelná. Nabízí se stejný závěr, že repetice se skládá z nukleotidů s rovnoměrně zastoupeným počtem od každého z nich a nemůže tedy být nijak barevně výraznější než ostatní části sekvence.



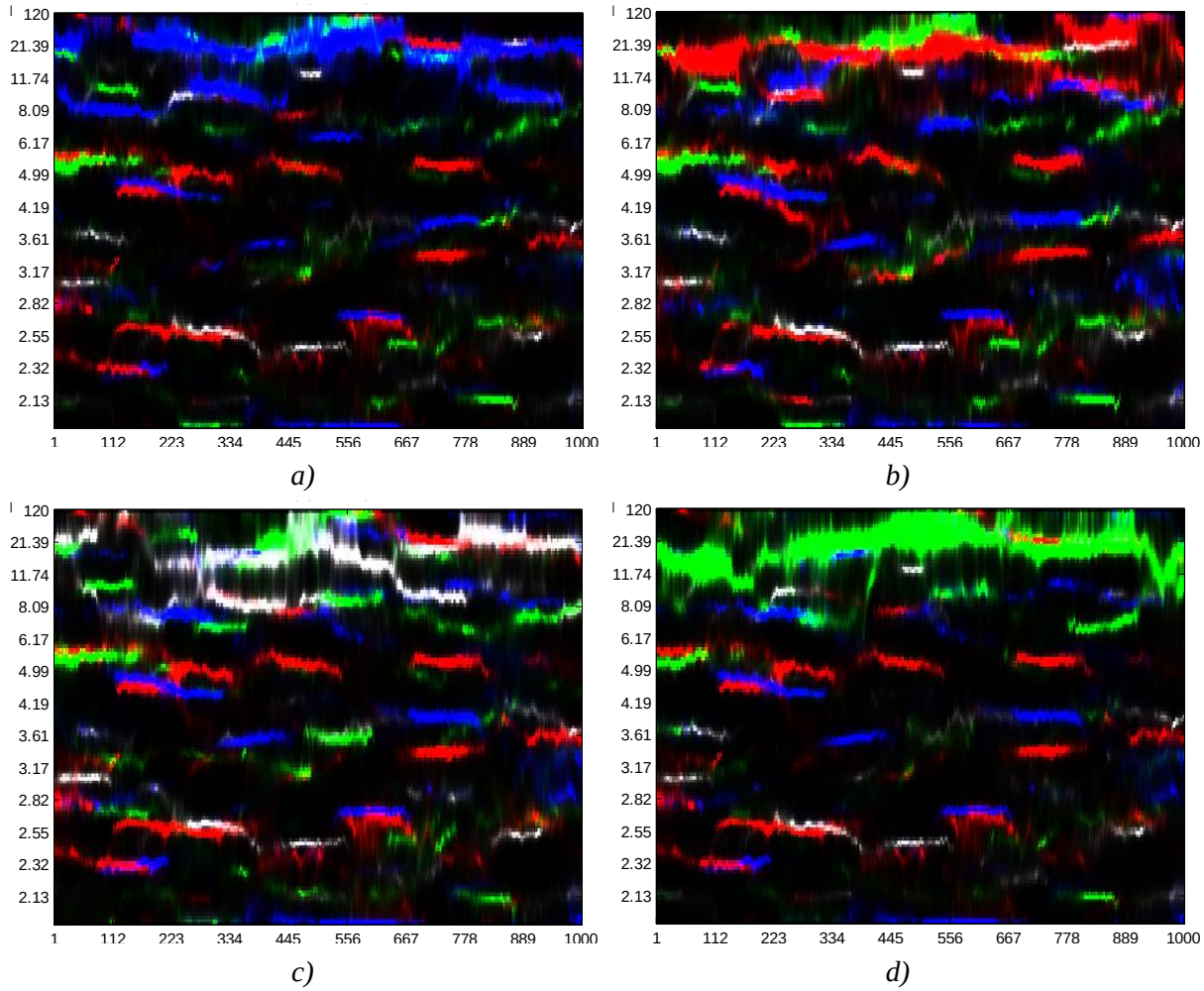
Obr.54 Spektrogram sekvence M19675.1 s rozptýlenými repeticemi



Obr.55 Spektrogram sekvence X02216.1 s rozptýlenými repeticemi

Druhou sekvencí je králičí gen X02216, který má v sobě rozptýlené repetice typu SINE o délce 13 bp na pozicích 64-77 bp a 432-445 bp. V repetici je s nejvyšším zastoupením nukleotid A, měla by být tedy v obrázku odlišena červenou barvou. Tato místa ve spektrogramu na *Obr.55* ale vyčíst nedokážeme. Nejspíše proto, že počet repetic je moc malý.

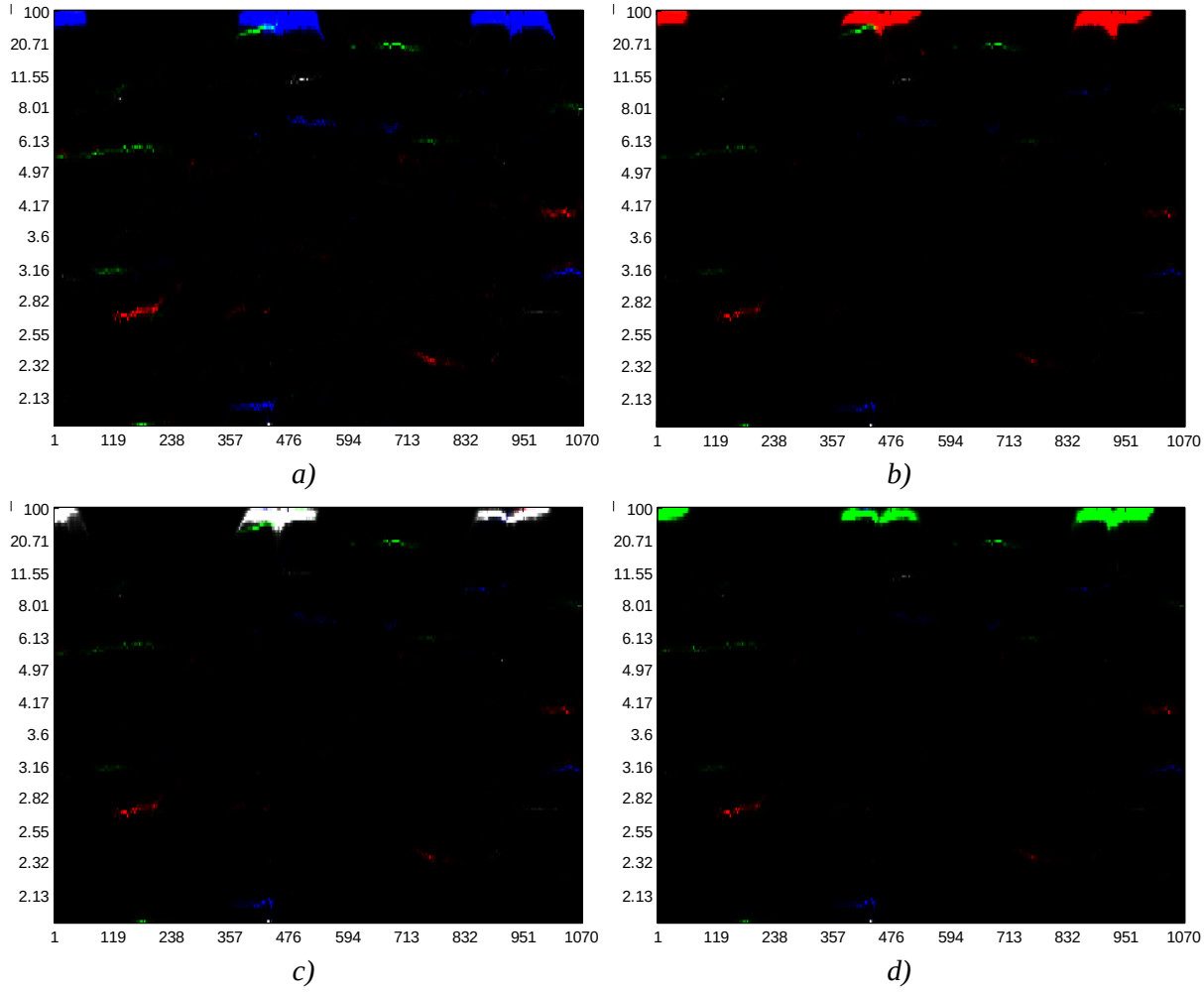
Dále byly otestovány náhodné sekvence s vloženými umělými repeticemi *umele_rozptylene_repetice.fasta* a *umele_rozptylene_repetice_dlouhe.fasta*. Jedna z nich obsahuje pětinukleotidové repetice opakující se v sekvenci 31krát a vidíme ji ve čtyřech provedeních na *Obr.56*. Na obrázku *a)* je sekvence s repeticí AAAAA. Ta se projevuje jako modře zbarvená horní část spektrogramu, protože podle barevných mapovacích vektorů ze vztahu 24 má adenin modrou barvu. Podobně to platí pro ostatní nukleotidy. Obrázek *b)* reprezentuje sekvenci, kde jsou nukleotidy AAAAA nahrazeny nukleotidy TTTTT, v obrázku *c)* jsou to nukleotidy GGGGG a v obrázku *d)* CCCCC.



Obr.56 Spektrogramy náhodné sekvence s umělými krátkými rozptýlenými repeticemi

Druhá sekvence zahrnuje repetici o délce 71 bp, která se v sekvenci objevuje 3krát. Vzor je tvořen vždy stejnými nukleotidy, které se barevně projevují v horních částech

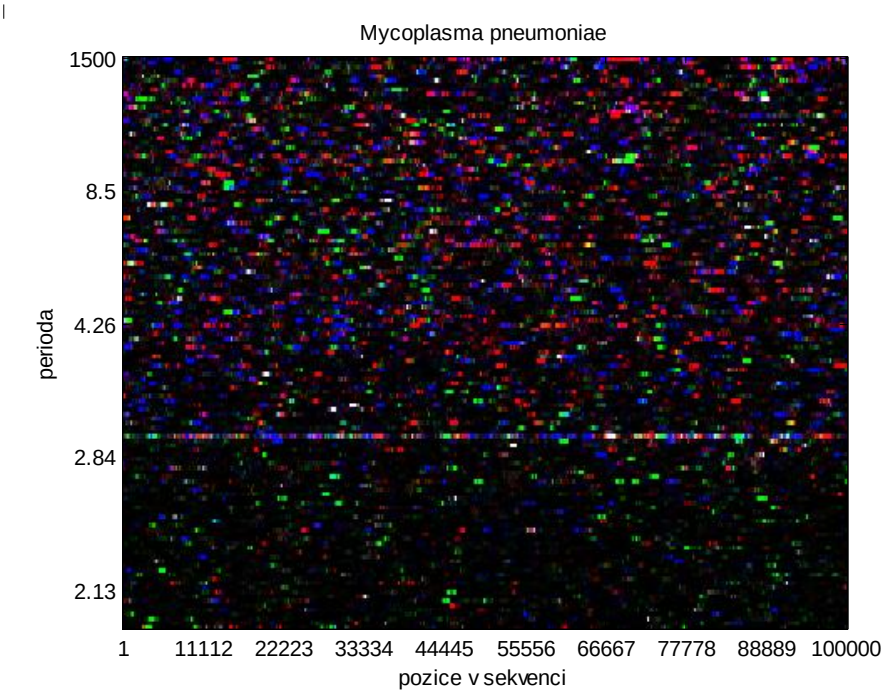
spektrogramů. Barvy odpovídají předchozímu příkladu umělé sekvence. V *Obr.57a)* tedy vidíme oblasti s repeticemi tvořenými bázemi A, protože mají modré zbarvení. V obrázku *b)* je repetice tvořená nukleotidy T, *c)* reprezentují nukleotidy G a *d)* jsou cytosiny (C). V těchto spektrogramech, na rozdíl od předchozích krátkých repetic, můžeme detekovat pozice dlouhých repetic v sekvenci. Nacházejí se v místech 1-71 bp, 427-497 bp a 852-924 bp.



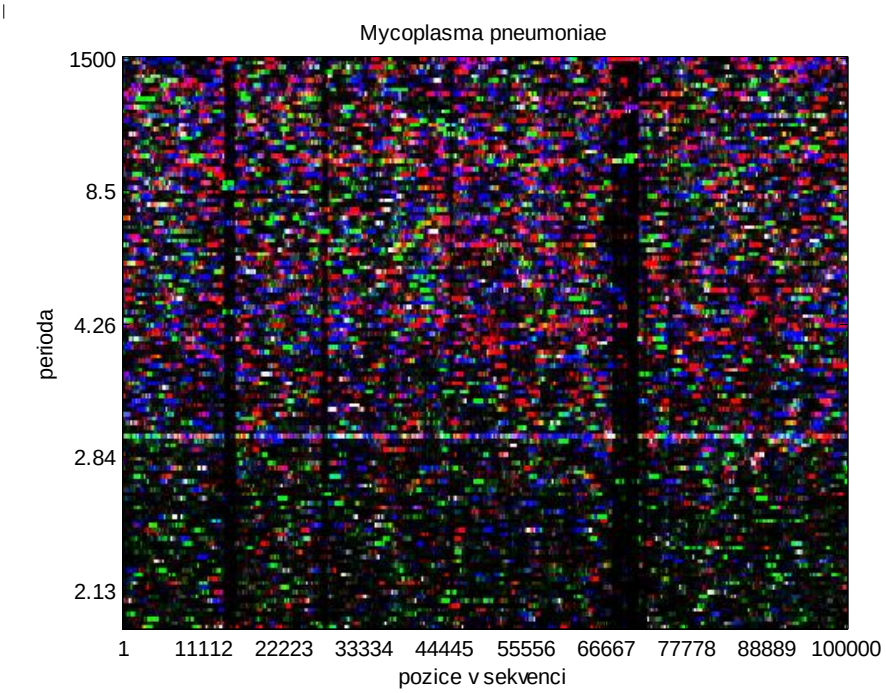
Obr.57 Spektrogramy náhodné sekvence s umělými dlouhými rozptýlenými repeticemi

Kódující sekvence

Testovanou sekvencí s kódujícími regiony byla bakterie Mycoplasma pneumoniae [33] - část jejího genomu, tak jako v kap. 5.1.2. Okno bylo nastaveno na hodnotu 1500, posun okna na 10 a řád byl pomocí skriptu *kriteria.m* určen jako 150. Na *Obr.58* je výsledek ze skriptu s normalizací mimo *for* cyklus, na *Obr.59* je spektrogram z druhého skriptu s normalizací ve *for* cyklu. Z obou je jasně detekovatelná světlá vodorovná linie v místě periody rovné třem, což značí přítomnost kódujících regionů v sekvenci bakterie. Barevné mapovací vektory mají hodnoty podle vztahu 24.



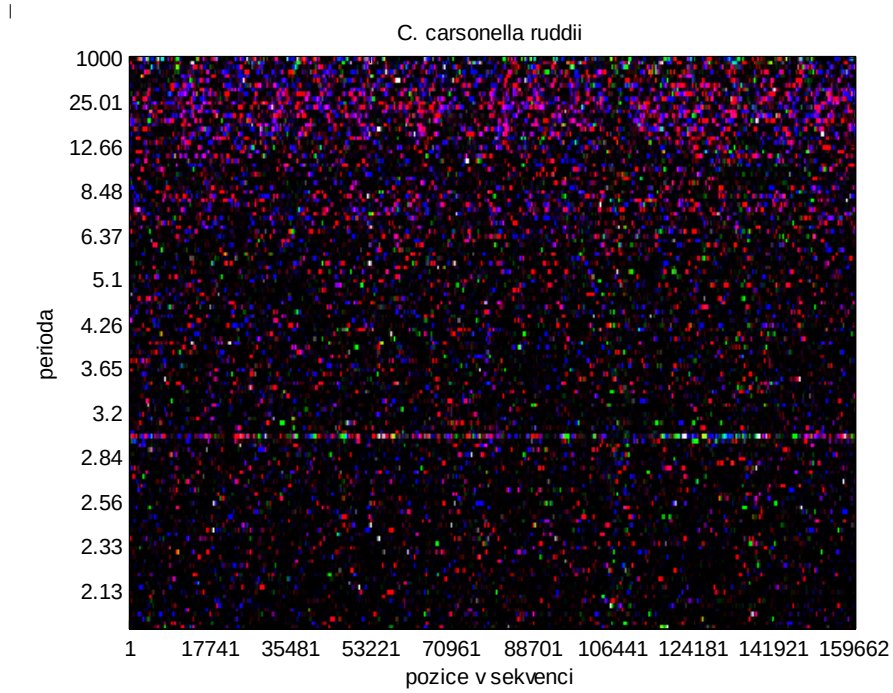
Obr.58 Spektrogram části genomu Mycoplasma pneumoniae, normalizace mimo for cyklus



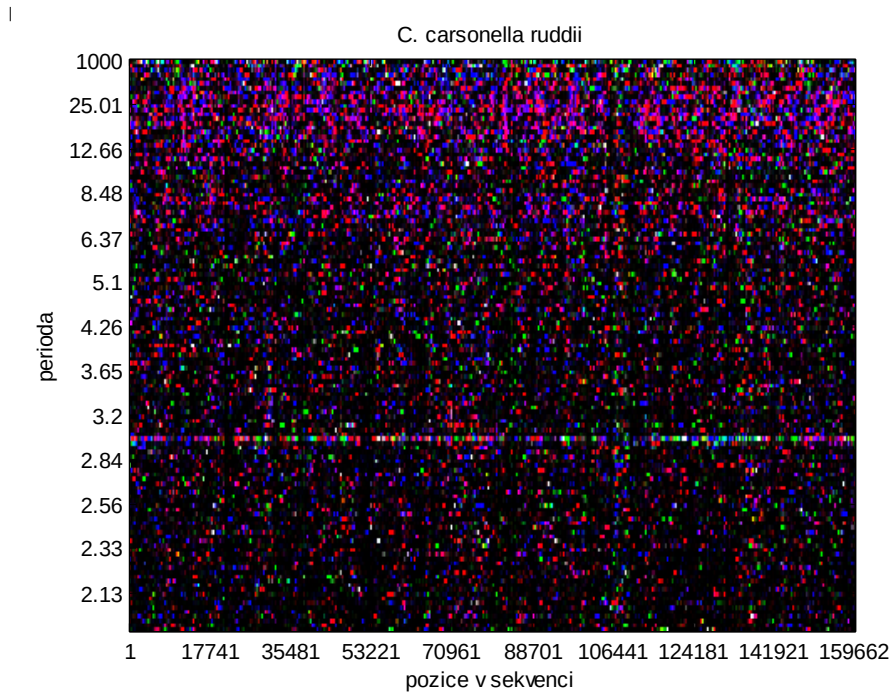
Obr.59 Spektrogram části genomu Mycoplasma pneumoniae, normalizace ve for cyklu

Tak jako v kap. 5.1.2 s FT byla i pro AR model otestována sekvence proteobakterie Candidatus Carsonella ruddii s nejmenším genomem z buněčných organismů. [33] Kódující části v sekvenci jsou z barevných spektrogramů na *Obr.60* a *Obr.61* rozlišitelné podle přítomnosti světlé vodorovné linie v místě periody 3 na svislé ose. Protože je tato linie výrazná, svědčí to o tom, že téměř všechny báze se podílí na kódování proteinů.

Parametry pro vykreslení spektra byly $\boldsymbol{w}$ = 1000, $\boldsymbol{o}$ = 100, $\boldsymbol{p}$ = 150, mapovací vektory stejné jako u předchozí sekvence.



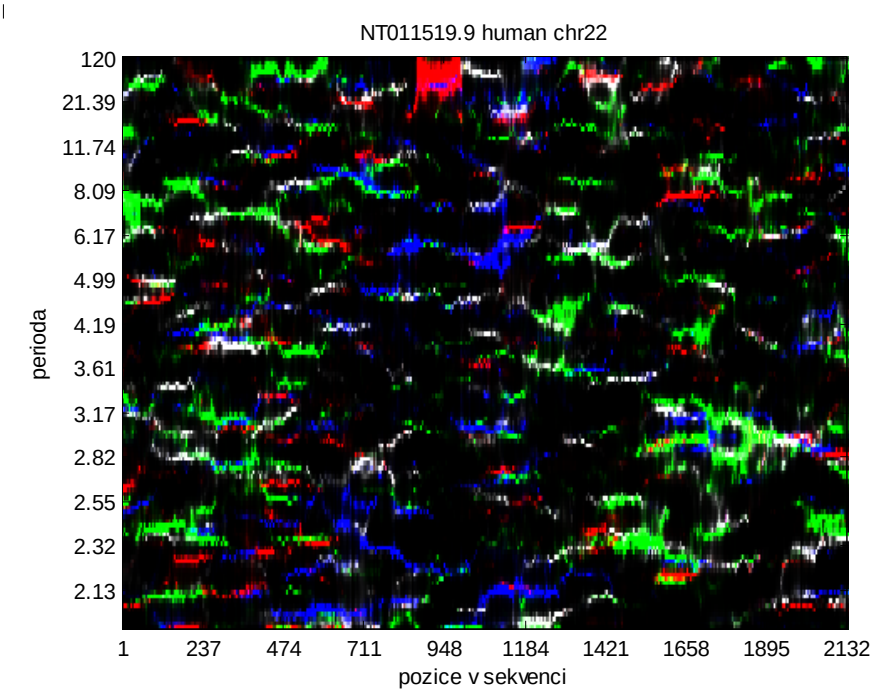
Obr.60 Spektrogram genomu proteobakterie C. Carsonella ruddii, normalizace mimo for cyklus



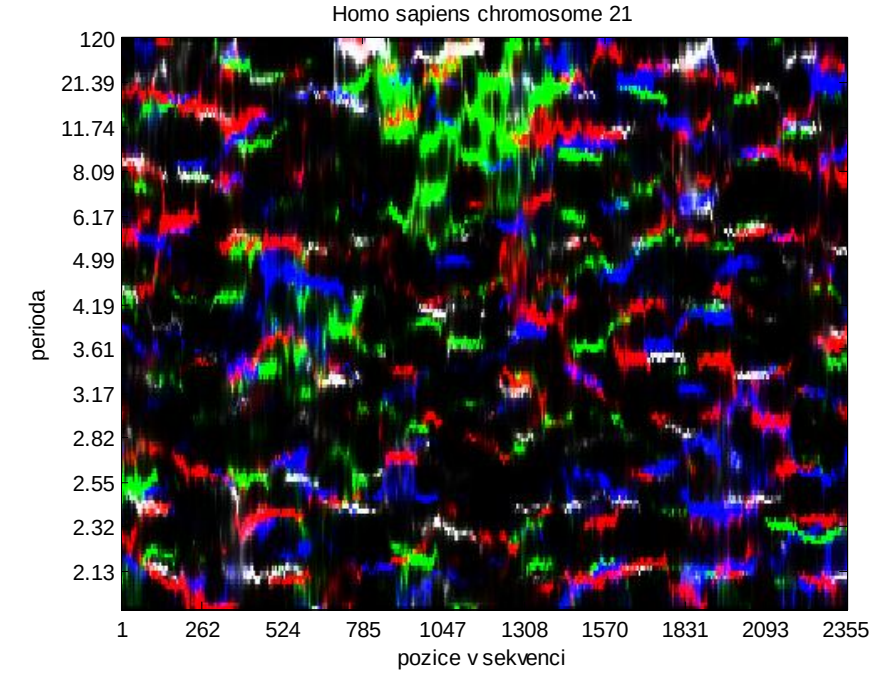
Obr.61 Spektrogram genomu proteobakterie C. Carsonella ruddii, normalizace ve for cyklu

CpG ostrůvky

CpG ostrůvky lépe vykresluje skript *hlavni_armodel_norm_ve_for_cyklu.m*, proto jsou zde uvedeny jen dva obrázky ze segmentů sekvencí lidského chromozomu 21 a 22 z tohoto skriptu.



Obr.62 Spektrogram segmentu sekvence lidského chromozomu 22 s CpG ostrůvky



Obr.63 Spektrogram segmentu sekvence lidského chromozomu 21 s CpG ostrůvky

Obr.62 reprezentuje segment sekvence chromozomu 22 určený pozicemi 2894684-2896815 bp. CpG ostrůvky jsou zde zelené oblasti v levé a pravé časti spektrogramu. V *Obr.63* vidíme segment sekvence chromozomu 21 mezi pozicemi 9905604-9907958 bp. Zde je naopak zelená oblast CpG ostrůvku jedna uprostřed spektrogramu mezi pozicemi 785-1570 bp na ose x.

Zelená barva je dána přítomností báze C mapované podle vztahu 24. Parametry pro vykreslení frekvenčního spektra byly u obou sekvencí zvoleny pro velikost okna $\boldsymbol{w}$ = 120, pro posun okna $\boldsymbol{o}$ = 1 a řád modelu $\boldsymbol{p}$ = 20.

5.2.2. Záměna pořadí kroků pro vykreslení spektrogramu

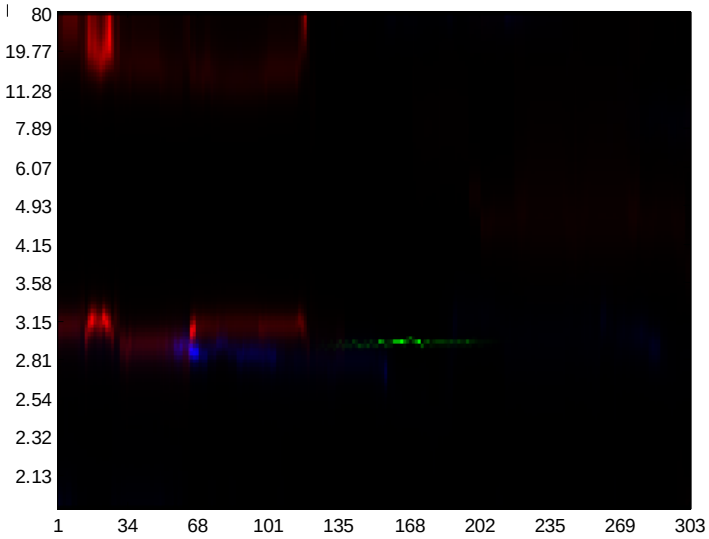
Tak jako u skriptů s FT v kap. 5.1.1 bylo vyzkoušeno, zda ovlivní výsledný spektrogram záměna pořadí kroků 1) – 4). Ze schématu na *Obr.20* vidíme další dvě možnosti pro pořadí kroků ve skriptu kromě základního 1,2,3,4. Skripty se základním pořadím kroků pro výpočet spektra *hlavni_armodel_norm_mimo_for_cyklus.m* a *hlavni_armodel_norm_ve_for_cyklu.m* byly tedy předělány na následující čtyři:

- *ar_model_mimo_for_1324.m*
- *ar_model_mimo_for_1243.m*
- *ar_model_ve_for_1324.m*
- *ar_model_ve_for_1243.m*

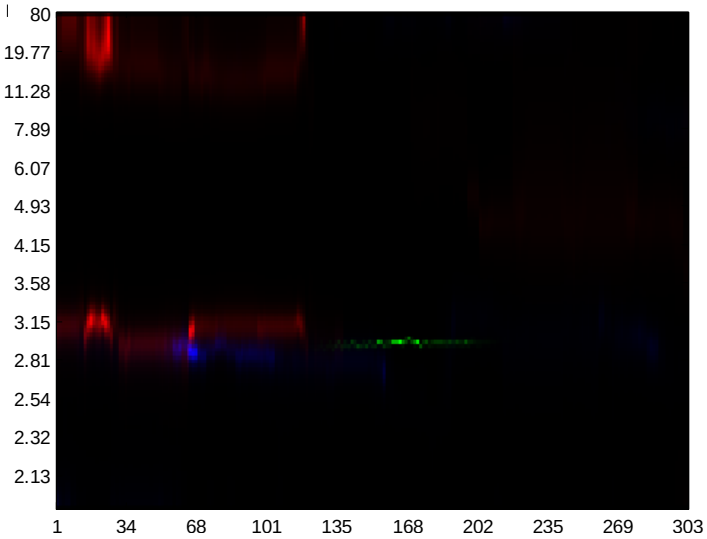
Výsledné spektrogramy s normalizací mimo *for* cyklus jsou na *Obr.64*. Testovanou sekvencí byla Oryza sativa [22] a parametry pro vykreslení spektra byly $\boldsymbol{w}$ = 80, $\boldsymbol{o}$ = 1, $\boldsymbol{p}$ = 5. Barevné mapovací vektory byly nastaveny jako: A – modrá, T – červená, C – zelená, G – černá. Z obrázků *a) – c)* je patrné, že záměna pořadí kroků neovlivní vzhled výsledného spektrogramu. Tedy stejný závěr jako v kap. 5.1.1.

Stejně tak spektrogramy s normalizací ve *for* cyklu na *Obr.65* jsou identické pro všechny možnosti pořadí kroků. Parametry byly ponechány stejné jako pro skripty s normalizací mimo *for* cyklus a testovanou sekvencí byla opět O. sativa.

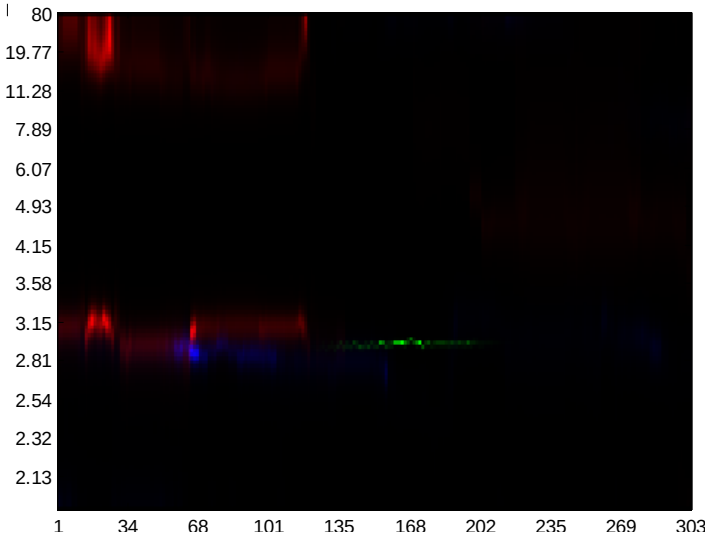
V literatuře je nejčastěji dodržováno pořadí kroků 1,2,3,4, proto v celé kap. 5.2.1 jsou používány skripty s tímto pořadím = *hlavni_armodel_norm_mimo_for_cyklus.m* a *hlavni_armodel_norm_ve_for_cyklu.m*.



a) 1,2,3,4

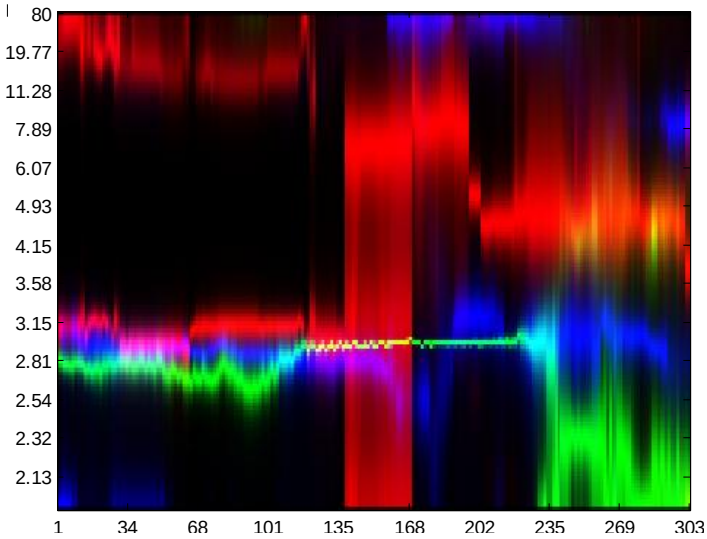


b) 1,2,4,3

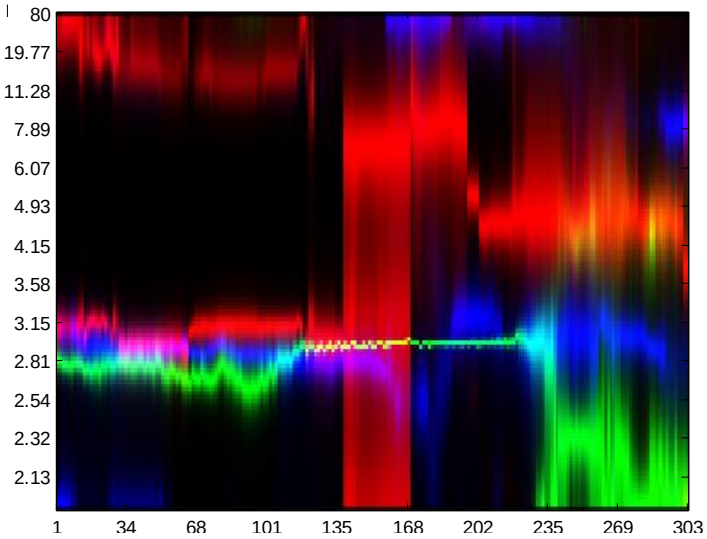


c) 1,3,2,4

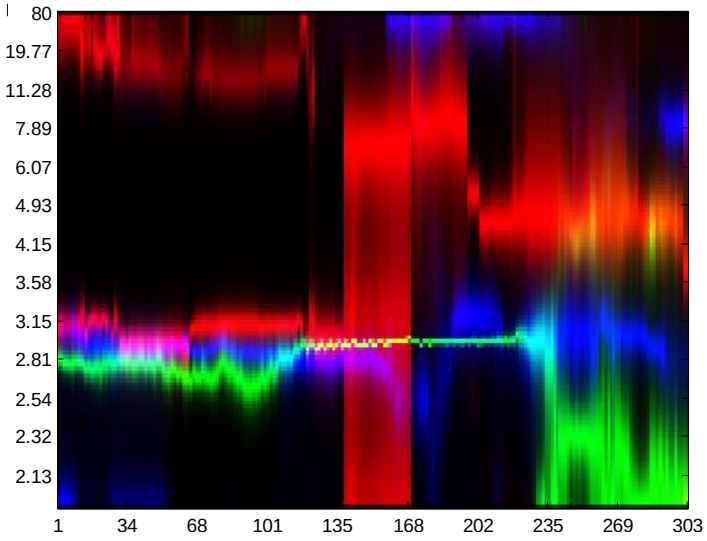
Obr.64 Spektrogramy se zaměněným pořadím kroků pro AR model s normalizací mimo for cyklus, sekvence O. sativa



a) 1,2,3,4



b) 1,2,4,3



c) 1,3,2,4

Obr.65 Spektrogramy se zaměněným pořadím kroků pro AR model s normalizací ve for cyklu, sekvence O. sativa

6. Srovnání spektrogramů získaných pomocí FT a AR modelu

Závěrečná kapitola se věnuje hlavnímu cíli této práce, tedy porovnání metod pro konstrukci DNA spektrogramů. Tyto metody jsou dvě a byly optimalizovány v předchozích kapitolách 5.1 a 5.2 pro použití na vykreslování tandemových a rozptýlených repetic, kódujících míst a CpG ostrůvků v sekvencích DNA. Pro všechny typy vzorů byly u obou metod nalezeny nejlepší parametry pro vykreslování spektrogramů.

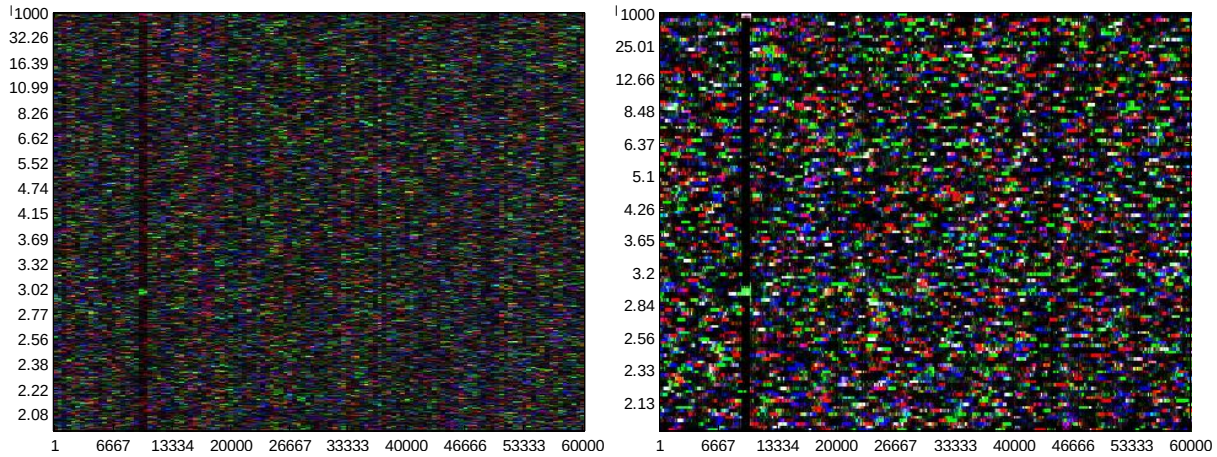
Dále se tedy budou nejlepší výsledné obrázky z těchto metod srovnávat mezi sebou, aby bylo možné finálně zhodnotit, která z nich je pro vykreslení daných vzorů lepší. Půjde hlavně o porovnání vizuální detekcí vzorů. Algoritmy z obou metod se budou posuzovat i z hlediska časové náročnosti výpočtu.

6.1. Porovnání z hlediska vizuální detekce vzorů

Mezi vzory, které můžeme oběma metodami detekovat patří tandemové repetice, rozptýlené repetice, CpG ostrůvky a kódující regiony sekvencí. Pro všechny typy vzorů byly vybrány sekvence, které propočítaly algoritmy z obou metod. Nejlepší skripty pro danou metodu byly vybrány podle kap. 5.1 a 5.2.

6.1.1. Tandemové repetice

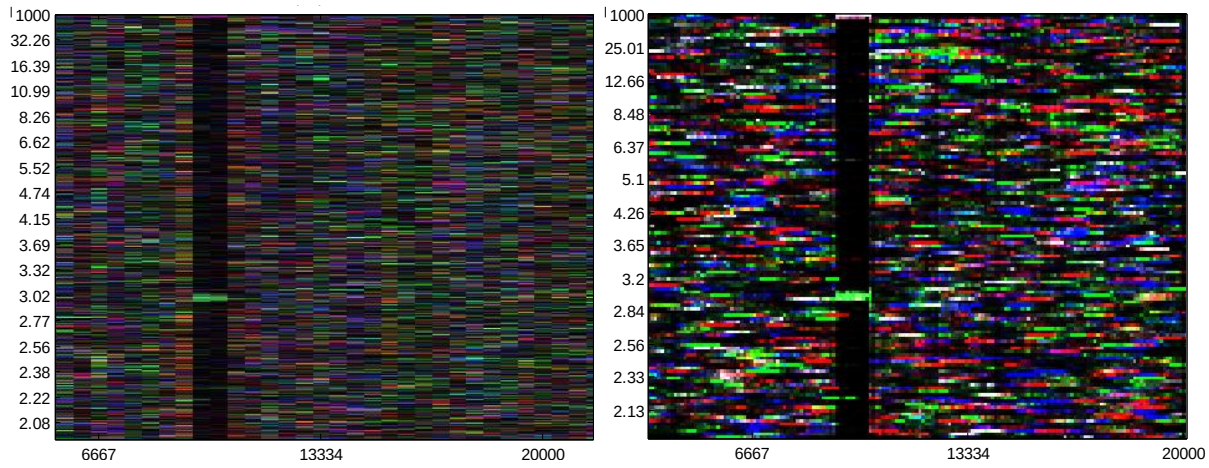
Pro srovnání spektrogramů z obou metod byla jako vhodná sekvence s tandemovou repeticí zvolena náhodná sekvence o délce 60 kbp generovaná v Matlabu příkazem *randseq*. Tato sekvence obsahuje uměle vloženou repetici CGG (mikrosatelit) opakující se v délce 150 bp začínající na pozici 10 kbp.



a) FT *b) AR model*

Obr.66 Spektrogramy náhodné sekvence s umělými tandemovými repeticemi

Parametry pro vykreslení frekvenčního spektra byly pro obě sekvence zvoleny stejné: ***w*** = 1000, ***o*** = 100, ***p*** = 100 (AR model), mapovací vektory ze vztahu 24. Výsledné spektrogramy jsou na *Obr.66a)* pro FT a *b)* pro AR model. Repetice je z obrázků z obou metod stejně dobře detekovatelná jako tmavý pruh na pozici okolo 10 kbp s výrazným zeleným místem okolo periody 3 (podrobnější popis viz kap. 5.1.2). I z detailů repetic na *Obr.67* by se dalo říci, že obě metody vykreslují spektrogramy této sekvence stejně kvalitně.



a) FT *b) AR model*

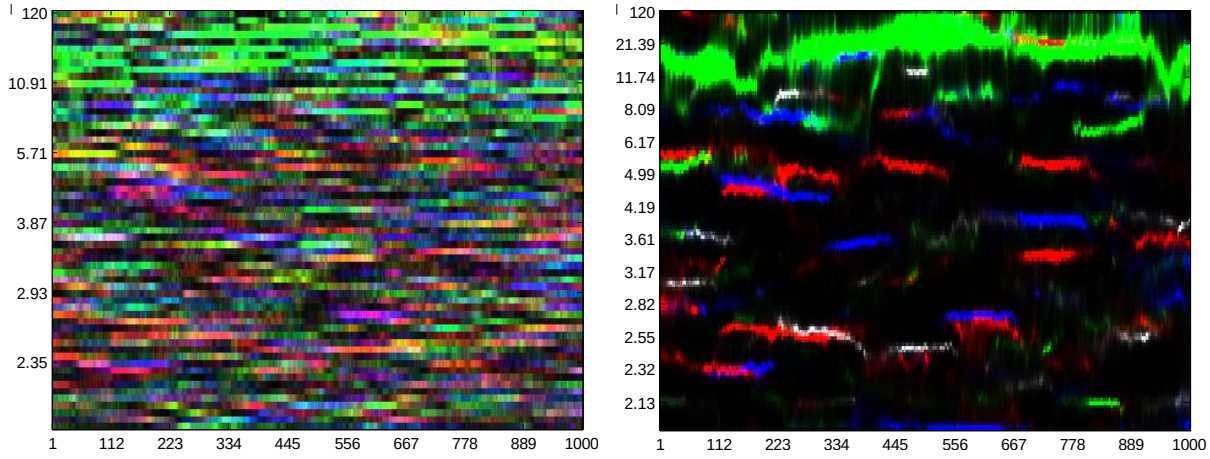
Obr.67 Detailní záběr na repetice z Obr.66

Srovnání spektrogramů z obou metod s reálnými sekvencemi je v kap. 6.2.

6.1.2. Rozptýlené repetice

Sekvence s umělými rozptýlenými repeticemi jsou nejvhodnější pro porovnání spektrogramů z obou metod. Vybrány byly soubory *umele_rozptylene_repetice.fasta* a *umele_rozptylene_repetice_dlouhe.fasta*. Obsahují v prvním případě krátké repetice CCCCC opakující se v sekvenci 31krát. Tyto vzory byly vloženy do náhodné sekvence o délce 1000 bp. V druhém případě se jedná o dlouhé 71 bp dlouhé repetice tvořené nukleotidem C opakující se v sekvenci 3krát. Tyto byly vloženy také do náhodné sekvence o délce 1070 bp.

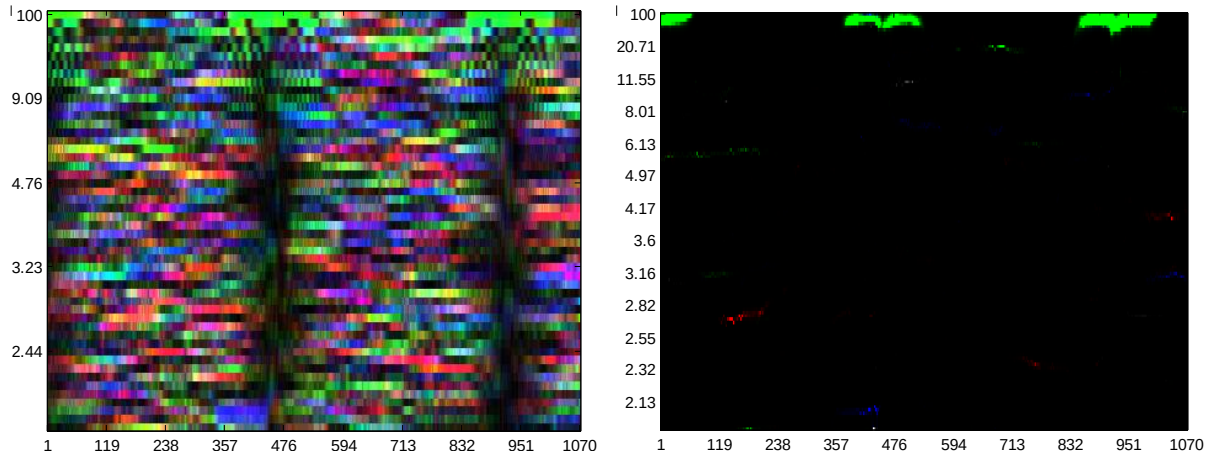
Na *Obr.68* jsou spektrogramy první zmíněné sekvence. Pětinukleotidová repetice CCCCC se projevuje jako zelená oblast v horní části obrázků. Je vidět, že metoda s AR modelem dává lepší výsledek. Oblast repetice je lépe rozlišená od pozadí spektrogramu než u metody s FT. Pro metody byly nastaveny parametry ***w*** = 120, ***o*** = 1, ***p*** = 15.



a) FT *b) AR model*

Obr.68 Spektrogramy náhodné sekvence s umělými krátkými rozptýlenými repeticemi

Na *Obr.69* jsou spektrogramy druhé sekvence s dlouhými rozptýlenými repeticemi. Vzory jsou tvořeny nukleotidy C mapovanými zelenou barvou, proto jsou v obrázcích detekovatelné jako zelené pruhy v jejich horní části. Nachází se v sekvenci 3krát v místech 1-71 bp, 427-497 bp a 852-924 bp. V *Obr.a)* z metody s FT jsou kromě zelených oblastí vidět i tmavé pruhy pod nimi, které směřují směrem dolů k ose x, ze které se pak dají lépe detekovat pozice repetic než u *Obr.b)* z metody s AR modelem. Ale u AR modelu jsou zase zachyceny jen tyto repetice a žádné ostatní redundantní informace. Není možno úplně jednoznačně posoudit, která metoda je pro vykreslování dlouhých rozptýlených repetic lepší.

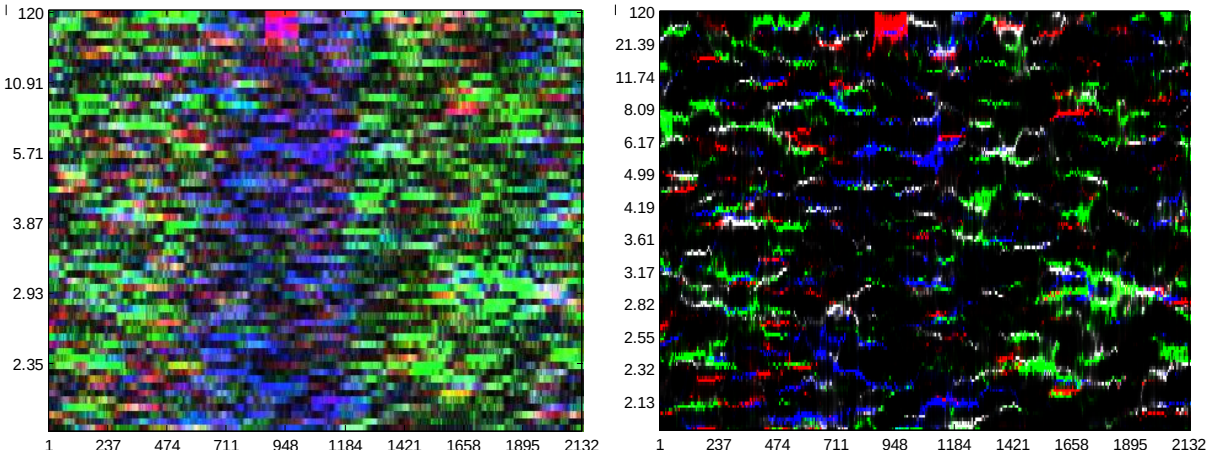


a) FT *b) AR model*

Obr.69 Spektrogramy náhodné sekvence s umělými dlouhými rozptýlenými repeticemi

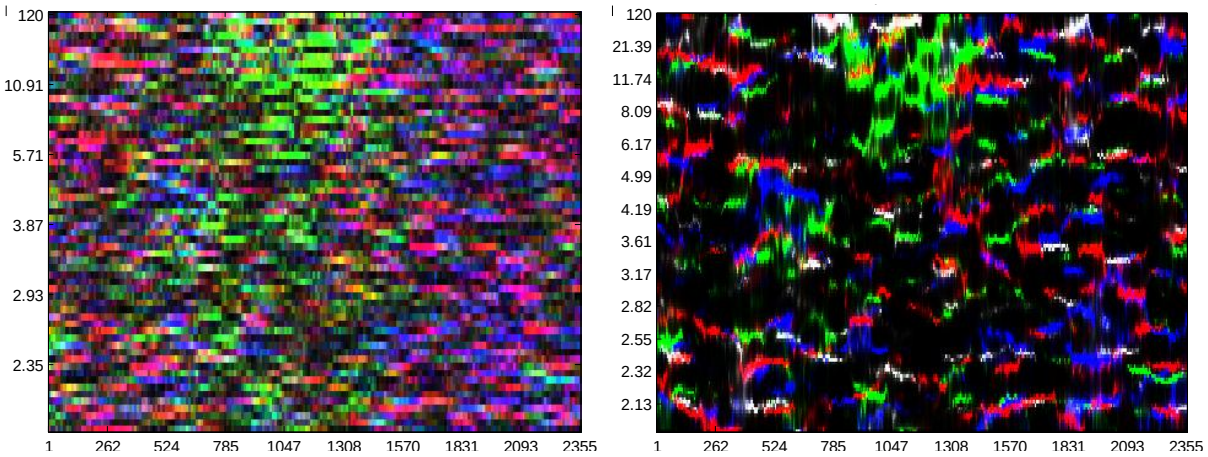
6.1.3. CpG ostrůvky

Oblasti CpG ostrůvků jsou lépe rozlišeny ve spektrogramech vytvořených metodou s FT. Můžeme se o tom přesvědčit z *Obr.70a)* a *Obr.71a).* Jedná se o segmenty sekvencí lidských chromozomů 22 a 21. CpG ostrůvky jsou ty zelené části spektrogramů. Právě u metody s FT jsou tyto části souvisleji ohraničeny a můžeme rozhodnout, na kterých pozicích v sekvenci se CpG ostrůvky nacházejí. U sekvence z chromozomu 22 (*Obr.70*) detekujeme jejich pozice cca mezi 1-500 bp a 1180-2132 bp. U sekvence z chromozomu 21 na *Obr.71* jsou to místa cca okolo 700-1400 bp.



a) FT *b) AR model*

Obr.70 Spektrogramy segmentu sekvence lidského chromozomu 22 s CpG ostrůvky



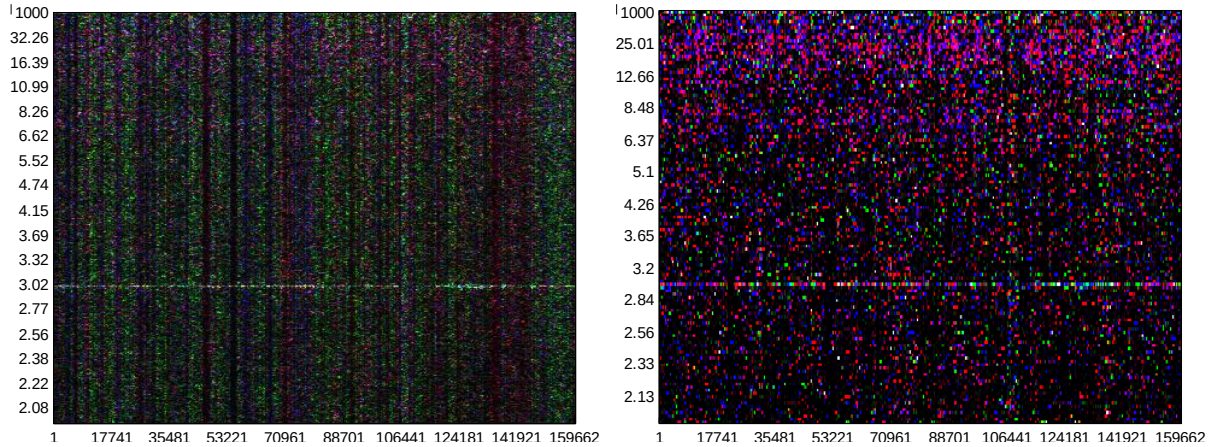
a) FT *b) AR model*

Obr.71 Spektrogramy segmentu sekvence lidského chromozomu 21 s CpG ostrůvky

6.1.4. Kódující regiony

Jako vhodná testovací sekvence s kódujícími regiony byla zvolena *C. carsonella ruddii.fasta*. Jedná se o celý genom proteobakterie Candidatus Carsonella ruddii, ve kterém je dobře detekovatelná linie okolo periody 3 naznačující přítomnost kódujících částí

sekvence. Podrobněji je detekci těchto míst věnována kap. 5.1.2. Spektrogramy z obou metod zobrazují v sekvenci dostatečně kvalitně oblasti zájmu. V *Obr.72a)* z metody s FT jsou navíc rozlišitelná místa s větším výskytem nukleotidů C (zelená barva) či T (červená barva), ale to není při detekci kódujících regionů potřeba rozlišovat.



a) FT *b) AR model*

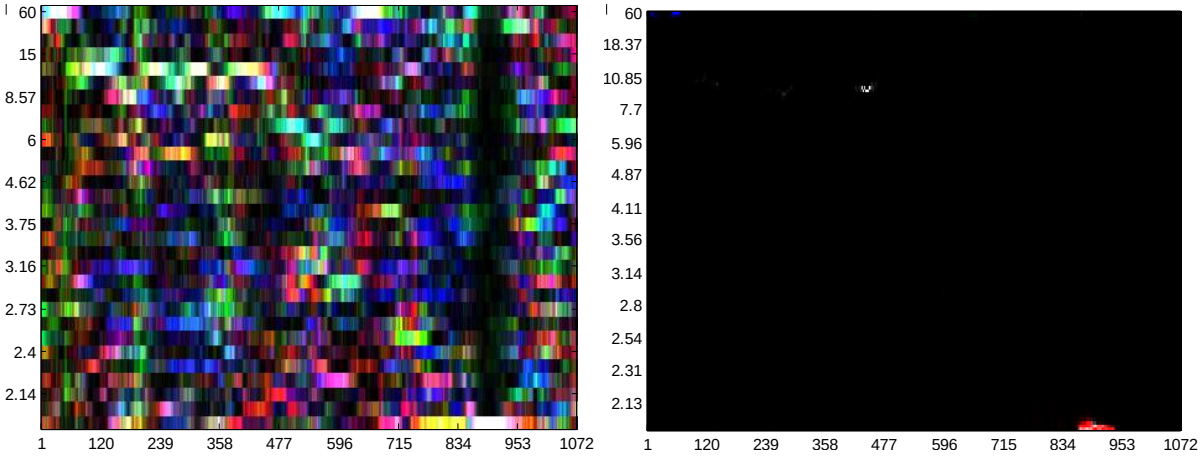
Obr.72 Spektrogramy genomu proteobakterie C. Carsonella ruddii

6.2. Porovnání přesnosti určení tandemových repetic u obou algoritmů

Přesnost obou algoritmů byla ověřena na sekvencích obsahujících různé druhy tandemových repetic z databáze NCBI [33] uvedených v tab. 2. Tři sekvence obsahují mikrosatelity a dvě sekvence minisatelity. Porovnáním s údaji v tab. 2, která vznikla s pomocí databáze TRDB [23], se ověří, zda spektrogramy dobře rozlišují vzory s danými charakteristikami.

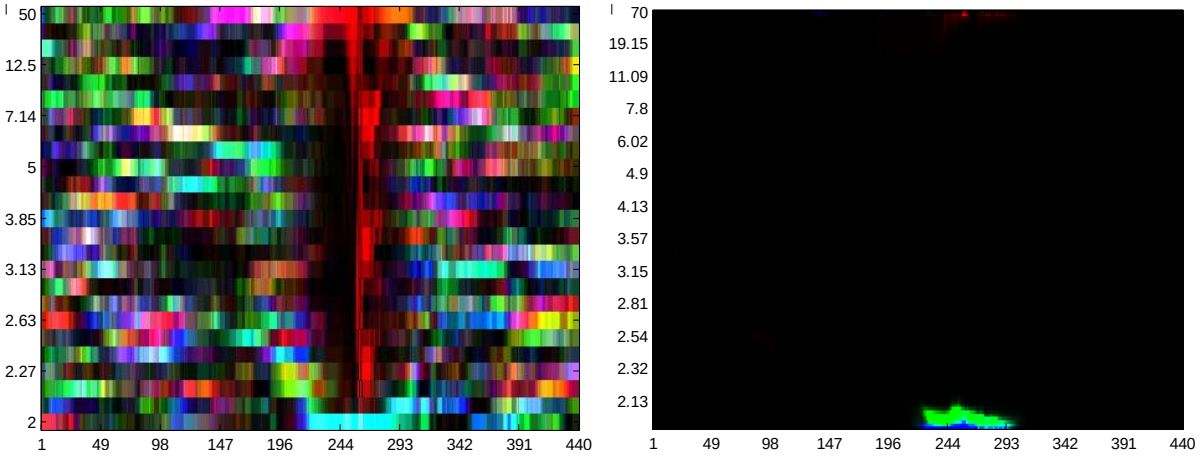
TABULKA2

Sekvence	*NCBI označení*	*Délka repetice*	*Typ repetice*	*Pozice repetice*
lidský gen KLK1	M65145.1	2 bp	GT	860-895 bp
lidský gen D5S253	M96445.1	2 bp	AC	230-282 bp
rýže setá, gen X64775	X64775.1	3 bp	GGC	145-188 bp
lidský chr.21, gen AL133493	AL133493.3	32 bp; 17 bp	GACAGGGTTCCC CCCACCCCGCTC CCAATCCA; CCCCTGGGATGC GGGTG	623-737 bp; 802-1768 bp
houba Leptosphaeria maculans, gen AJ621804	AJ621804.1	16 bp	CCCAGCCACTCA ACCG	98-204 bp



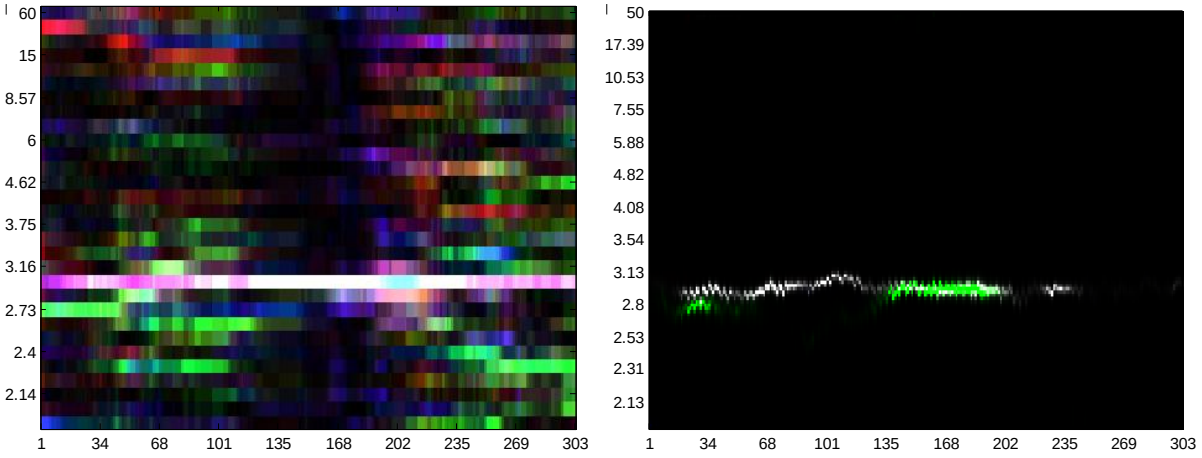
a) FT *b) AR model*

Obr.73 Spektrogramy sekvence M65145.1



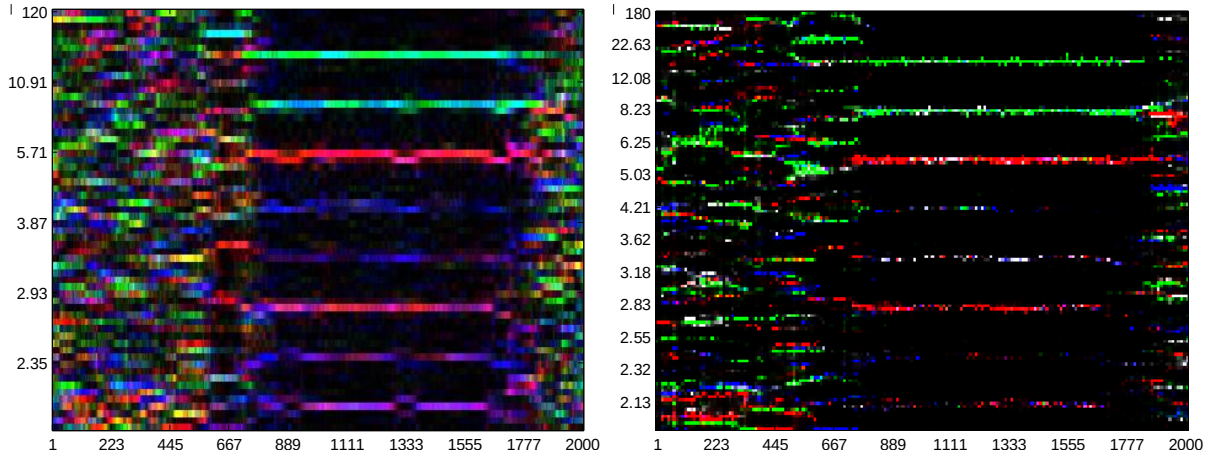
a) FT *b) AR model*

Obr.74 Spektrogramy sekvence M96445.1



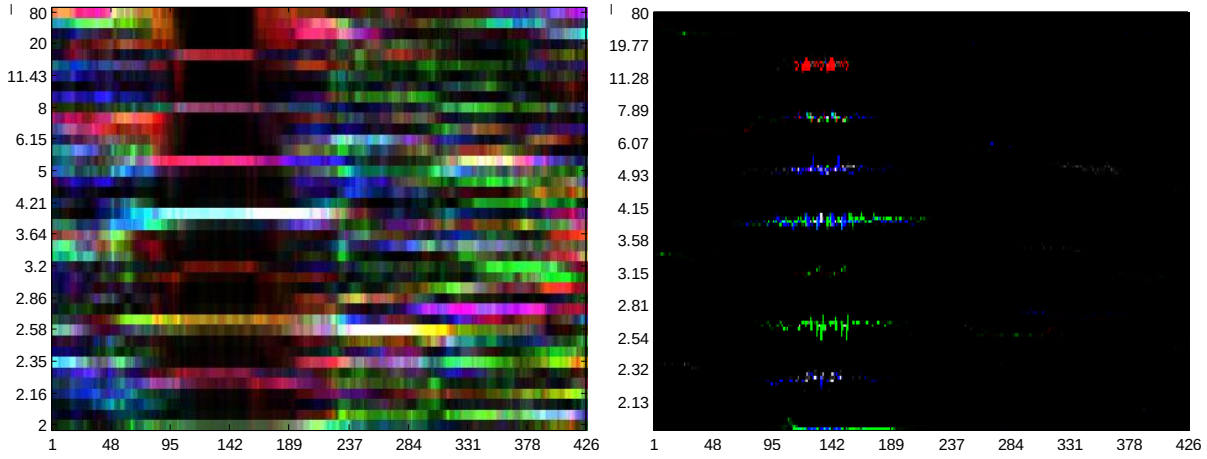
a) FT *b) AR model*

Obr.75 Spektrogramy sekvence X64775.1



a) FT *b) AR model*

Obr.76 Spektrogramy sekvence AL133493.3



a) FT *b) AR model*

Obr.77 Spektrogramy sekvence AJ621804.1

Na první pohled je patrné, že spektrogramy vytvořené metodou s AR modelem nemají barevné pozadí jako spektrogramy z druhé metody. Jejich podklad je černý a barevně zvýrazněny jsou pouze tandemové repetice. Přestože metoda s FT dává správné výsledky jako metoda s AR modelem a je možné barevně odlišit vzory v sekvenci, dalo by se říci, že poskytuje i nadbytečné informace ve formě právě barevného pozadí. Pro vyhledávání tandemových repetic nás informace o rozložení nukleotidů s danými barvami v ostatních částech sekvence nezajímá.

Repetice v sekvenci *M65145.1* je v *Obr.73a)* detekovatelná jako bílý proužek na pozicích 860-895 bp v oblasti periody 2, protože délka vzoru jsou 2 bp. Ve spektrogramu je navíc i tmavý svislý pruh o délce celé osy y. Pro obrázky získané metodou s FT je ale takové zobrazení tmavých pruhů s barevnými proužky uvnitř znázorňujícími přesné místo repetice, potřebné. Kdyby se v obrázcích projevovaly jen ty vodorovné barevné proužky jako u metody s AR modelem, nebylo by možné je odlišit od různobarevných pixelů v pozadí. U metody s AR modelem se žádné svislé pruhy obsahující repetice nezobrazují.

Není to potřeba vzhledem k tomu, že pozadí je černé a barevné vzory jsou tvořeny pouze červenou, modrou, zelenou a šedou (barvy se nemíchají). Z *Obr.73b)* z metody s AR modelem můžeme dokonce odhadnout jaké báze tvoří repetici podle barvy proužku. Je červeno-šedá, tak se dá předpokládat, že jedna z bází by mohla být thymin a druhá guanin (podle mapovacích vektorů ze vztahu 24).

U *Obr.74* sekvence *M96445.1* se nabízí stejný popis. Algoritmus z metody s AR modelem opět vykreslil repetici jako proužek, ze kterého odvodíme dvojici bází A (modrá barva) a C (zelená barva). Jinak informace o délce vzoru a o pozici v sekvenci jsou dobře čitelné z obou obrázků z obou metod.

Na *Obr.75* jsou spektrogramy sekvence *X64775.1*. Sekvence obsahuje hlavní repetici na pozicích 145-188 bp, která je zachycena v *Obr.b)* jako zeleno-šedý proužek. Je tvořena třemi bázemi GGC. Protože se v sekvenci nacházejí i další trinukleotidové repetice tvořené převážně bázemi G, které se ale opakují v menší délce než ta hlavní (viz kap. 4.4), je ve spektrogramu na *Obr.b)* ještě vidět vodorovný šedý proužek téměř v celé délce osy x. Ve spektrogramu z algoritmu s FT na *Obr.a)* je také světlá linie v celé délce vodorovné osy. Hlavní repetici zde odlišíme díky tmavému svislému pruhu na pozicích 145-188 bp.

Další sekvence na *Obr.76* a *Obr.77* obsahují delší tandemové repetice – minisatelity (32, 17 a 16 bp). Takto dlouhé repetice se zobrazují ve formě více svislých proužků nad sebou. Jejich délka se pak odvozuje z počtu nad sebou vyobrazených proužků. Je-li počet vodorovných proužků L sudý, vypočteme délku vzoru jako 2L. Je-li počet L lichý, potom délka vzoru bude 2L+1 [14]. O sekvenci *AL133493.3* je podrobněji psáno v kap. 5.1.2 a 5.2.1.

Sekvence *AJ621804.1* na *Obr.77* má repetici na pozicích 98-204 bp. Její délka je 16 bp, což se dá vyhledat i ve spektrogramech, sečteme-li vodorovné proužky nad sebou. Je jich 8, tedy sudý počet, takže podle vzorce 2L je délka repetice opravdu 16 bp. Z obrázků z obou metod jsou tyto charakteristiky dobře patrné, ale opět *Obr.b)* AR model by se dal označit za přesnější.

Jestliže bychom měli vybrat tu lepší metodu pro vykreslování spektrogramů sekvencí s tandemovými repeticemi, měla by to být metoda využívající k odhadu frekvenčního spektra AR model. Obrázky jsou přehlednější a poskytují informace o typech bází ve vzorech.

6.3. Porovnání z hlediska časové náročnosti algoritmů

Bylo vybráno deset sekvencí s různou délkou, které jsou seřazeny v tab. 3 podle délky vzestupně. Tyto sekvence byly testovány hlavními algoritmy obou metod pro vykreslení spektrogramů s FT a AR modelem v části výpočtu frekvenčního spektra = *hlavni_1234.m* a *hlavni_armodel_norm_ve_for_cyklu.m*. Algoritmy mají kroky

normalizace umístěné ve *for* cyklu, s tím, že pro AR model je potřeba dvou normalizací, aby dával výsledky porovnatelné se spektrogramy z FT.

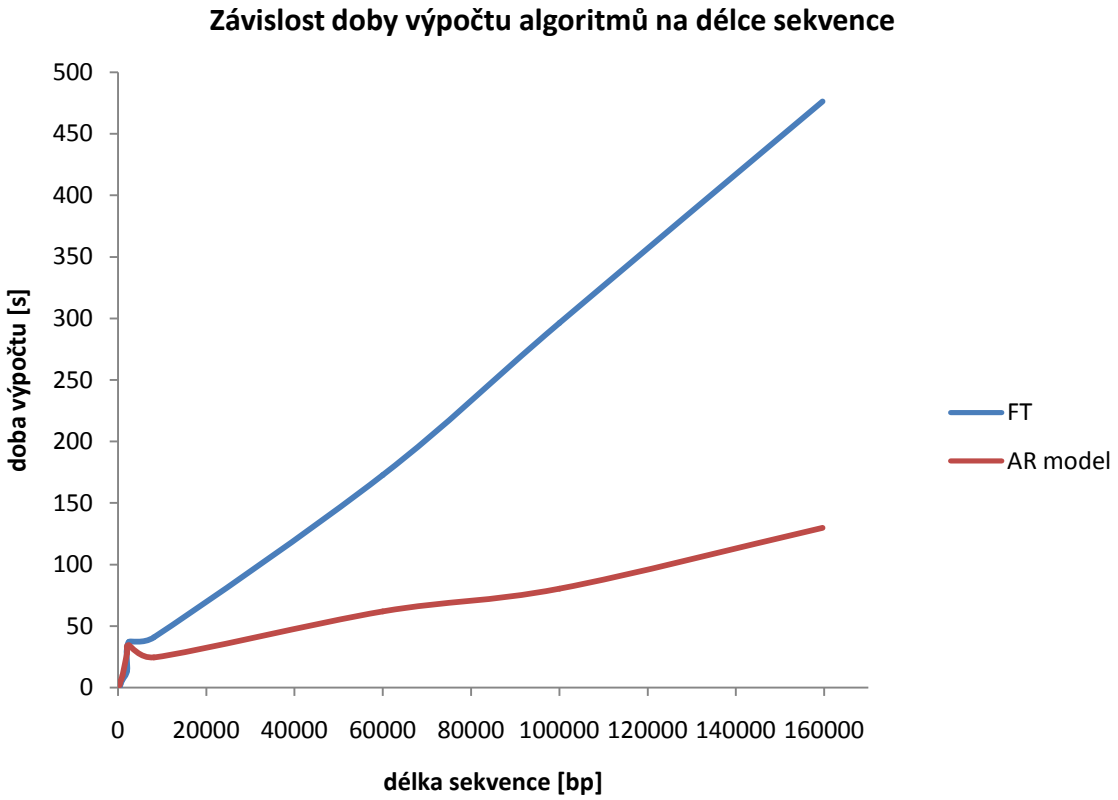
TABULKA3

Sekvence	*Název fasta souboru*	*Délka sekvence*	*Délka okna*	*Posun okna*	*Výpočet pomocí FT [s]*	*Výpočet pomocí AR modelu [s]*
rýže setá, gen X64775	X64775	303 bp	60	1	0,73	1,75
houba Leptosphaeria maculans, gen AJ621804	Leptosphaeria maculans	426 bp	80	1	1,4	3,83
králičí gen X02216	Rabbit short interspersed C repeat	527 bp	120	1	2,9	5,41
náhodná sekvence s umělými repeticemi	umele_rozptylene_repetice	1000 bp	120	1	6,4	9,91
lidský chr.21, gen AL133493	AL133493_human_chr21	2000 bp	120	1	13,51	27,85
bakterie Neisseria gonorrhoeae	neisseria gonorrhoeae clone pNG273	2418 bp	200	1	37,08	34,38
gen F56F11 Caenorhabditis elegans	Caenorhabditis elegans Cosmid F56F11	8100 bp	200	10	40,69	24,51
náhodná sekvence s minisatelity	randomseq_minisatelit	60000 bp	1000	100	173,55	61,92
bakterie Mycoplasma pneumoniae	Mycoplasma pneumoniae	100000 bp	1000	100	296,12	80,4
proteobakterie C. Carsonella ruddii	C. carsonella ruddii	159662 bp	1000	100	476,13	129,74

Výsledná časová závislost výpočtu obou algoritmů na délce sekvence je vidět v grafu na *Obr.78*. Byla vytvořena z hodnot ve třetím, šestém a sedmém sloupci tab. 3. Závislost je lineární u obou metod, tedy doba výpočtu se prodlužuje se stoupající délkou sekvence.

Křivka pro algoritmus využívající pro výpočet spektra FT (modrá barva) má strmější průběh než křivka druhá. Pro dlouhé sekvence je u této metody doba výpočtu vysoká. Dalo by se tedy říci, že pro delší sekvence je lepší použít algoritmus využívající k odhadu spektra AR model (méně strmá červená křivka). Podíváme-li se na hodnoty do tab. 3, kde rozdíl mezi dobou výpočtů u obou metod je u kratších sekvencí malé číslo a u delších

sekvencí (od délky 8100 bp) se hodnota rozdílu mezi časy výpočtů zvyšuje, tento odhad si potvrdíme.



Obr.78 Graf závislosti doby výpočtu algoritmů na délce sekvence

7. Závěr

Cílem této práce bylo provést literární rešerši o numerických reprezentacích DNA sekvencí, o možnostech konstrukce barevných DNA spektrogramů a o vzorech detekovatelných ze spektrogramů.

Metody numerické reprezentace byly klasifikovány do dvou hlavních skupin – metody s pevně stanoveným mapováním a metody založené na chemicko-fyzikálních vlastnostech DNA molekul. Každá z popsaných reprezentací má své výhody i nevýhody. Častěji se však používají metody s pevně stanoveným mapováním a nejvíce z nich 4D binární reprezentace, která byla použita i v této práci pro ověření funkčnosti vytvořených algoritmů.

Po konvertování DNA sekvence do numerické reprezentace se nám otevírá možnost použití technik číslicového zpracování signálů. K těm patří také spektrální analýza, jejímž úkolem je transformovat vybranou numerickou reprezentaci do frekvenční oblasti. Analýza zahrnuje posloupnost kroků vedoucích k vytvoření barevného spektrogramu, jako jsou: výpočet frekvenčního spektra, mapování DFT hodnot do RGB prostoru a normalizace hodnot pixelů.

V této práci jsou popsány dvě metody výpočtu frekvenčního spektra – Fourierova transformace a odhad spektra pomocí autoregresního modelu. Literatura uvádí, že vypočtená spektra pomocí Fourierovy transformace jsou nepřesná, mohou obsahovat falešné píky a mají slabé rozlišení, naproti tomu AR model tento problém nemá. Tato teorie byla v rámci diplomové práce otestována pomocí algoritmů vytvořených v programovém prostředí Matlab vykreslujících barevné spektrogramy. Z výsledků zde uvedených se určitě nedá tvrdit, že autoregresní model je vždy přesnější než Fourierova transformace. Existují typy vzorů v sekvenci DNA, které vykresluje algoritmus s FT mnohem přesněji, např. jsou to CpG ostrůvky.

Co se týká tandemových repetic, je pravdou, že AR model dává přehlednější spektrogramy znázorňující jen místa s repeticemi bez redundantních informací, za které můžeme považovat barevné pixely v pozadí obrázků z metody s FT. Navíc podle barvy oblastí s repeticemi se dá u AR modelu usuzovat, z jakých je tvořena bází, na rozdíl od metody s FT, kdy barvy úplně vždy neodpovídají. U detekce rozptýlených repetic se hodí pro popis výsledků podobný závěr.

Posledním typem hledaného vzoru v sekvenci DNA byly kódující regiony. Tyto jsou při použití AR modelu i FT v obrázcích stejně dobře rozpoznatelné.

Před vykreslením spektrogramů oběma metodami bylo potřeba nastavit vhodně délku okna a posun oken, pro které se počítalo spektrum pomocí FT nebo AR modelem. U druhé metody bylo navíc potřeba účelně nastavit hodnotu řádu modelu. Se zvyšující se délkou sekvence se často nastavuje větší délka okna. Pro velmi dlouhé výpočetně náročné sekvence se ještě zvyšuje i hodnota posunu oken, aby byl algoritmus rychlejší a aby se

zlepšilo rozlišení jednotlivých pozic bází v dlouhé sekvenci. Ale vždy záleží hlavně na pozorovaném vzoru v sekvenci, který má určité vlastnosti, jež má spektrogram odlišit. Proto bylo potřeba parametry několikrát pozměnit a vyzkoušet, se kterými dával algoritmus výsledky s požadovaným rozlišením.

Shrneme-li to, obě metody pro konstrukci barevných DNA spektrogramů byly nejprve optimalizovány pro použití na sekvence s různými vzory, poté byly porovnány z hlediska účinnosti detekce daných typů vzorů.

8. Seznam použité literatury

[1] DIMITROVA, Nevenka, Yee Him CHEUNG a Michael ZHANG. Analysis and Visualization of DNA Spectrograms: Open Possibilities for the Genome Research. In: *Proceedings of the 14th annual ACM international conference on Multimedia*: MULTIMEDIA '06. New York, NY, USA: ACM, 2006, s. 1017-1024. ISBN 1-59593-447-2. DOI: 10.1145/1180639.1180861. Dostupné z: http://doi.acm.org/10.1145/1180639.1180861

[2] ZHOU, Hongxia, Liping DU a Hong YAN. Detection of Tandem Repeats in DNA Sequences Based on Parametric Spectral Estimation. *IEEE Transactions on Information Technology in Biomedicine.* 2009, vol. 13, no. 5, s. 747-755. ISSN 1089-7771.

[3] ANASTASSIOU, Dimitris. Frequency-domain Analysis of Biomolecular Sequences. *Bioinformatics.* 2000, vol. 16, no. 12, s. 1073-1081. DOI: 10.1093/bioinformatics/16.12.1073.

[4] KWAN, Hon Keung a Swarna Bai ARNIKER. Numerical Representation of DNA Sequences. *Electro/Information Technology, 2009. eit '09. IEEE International Conference on.* 2009, s. 307-310. DOI: 10.1109/EIT.2009.5189632.

[5] KWAN, Hon Keung a Swarna Bai ARNIKER. Advanced Numerical Representation of DNA Sequences. *2012 International Conference on Bioscience, Biochemistry and Bioinformatics.* 2012, vol. 31. Dostupné z: www.ipcbee.com/vol31/001-ICBBB2012-T003.pdf

[6] ZÖLZER, Friedo. JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH Zdravotně sociální fakulta. *Radiobiologie buňky.* České Budějovice, 2007. Dostupné z: http://www.zsf.jcu.cz/structure/departments/kra/informace-pro-studenty/ucebni_texty/ochrana-obyvatelstva-se-zamerenim-na-cbrne-aplikovana-radiobiologie-a-toxikologie-krizova-radiobiologie-a-toxikologie/radiobiologie-bunky.doc/view

[7] DNA. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001 [cit. 2012-11-23]. Dostupné z: http://en.wikipedia.org/wiki/DNA

[8] VUT BRNO, FEKT, studijní materiály k předmětu *Pokročilá analýza biologických sekvencí* (2012), garant předmětu: PROVAZNÍK, I.

[9] RAMACHANDRAN, Parameswaran a Andreas ANTONIOU. *Genomic Digital Signal Processing.* Victoria, BC, Canada: University of Victoria, 2004. Dostupné z: http://www.ece.uvic.ca/~andreas/RLectures/GenomicDSP04-Paramesh-Pres.pdf

[10] EURASIP BOOK SERIES ON SIGNAL PROCESSING AND COMMUNICATIONS. *Genomic Signal Processing and Statistics* [online]. New York, NY 10022, USA: Hindawi Publishing Corporation, 2005 [cit. 2012-11-24]. ISBN 977-5945-07-0. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.2778&rep=rep1 &type=pdf

[11] YAU, Stephen S.-T., J. WANG, A. NIKNEJAD, Ch. LU, N. JIN a Y. HO. DNA sequence representation without degeneracy. *Nucleic Acids Research.* 2003, vol. 31, no. 12, s. 3078-3080. DOI: 10.1093/nar/gkg432.

[12] DNA walk method. In: *UNIL / Université de Lausanne: Comparative Genometrics* [online]. 2001 [cit. 2012-11-24]. Dostupné z: http://www2.unil.ch/comparativegenometrics/DNA_walk.html

[13] AKHTAR, Mahmood, Julien EPPS a Eliathamby AMBIKAIRAJAH. On DNA Numerical Representations for Period-3 Based Exon Prediction. *Genomic Signal Processing and Statistics, 2007. GENSIPS 2007. IEEE International Workshop on.* 2007, s. 1-4. DOI: 10.1109/GENSIPS.2007.4365821.

[14] SUSSILLO, David, Anshul KUNDAJE a Dimitris ANASTASSIOU. Spectrogram Analysis of Genomes. *EURASIP Journal on Applied Signal Processing.* 2004, s. 29-42. DOI: 10.1155/S1110865704310048. Dostupné z: http://asp.eurasipjournals.com/content/2004/1/ 790248

[15] NAIR, Achuthsankar S. a Sivarama P. SREENADHAN. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation.* 2006, roč. 1, č. 6, s. 197-202. ISSN 0973-2063.

[16] JAN, Jiří. *Číslicová filtrace, analýza a restaurace signálů.* 2. upravené a rozšířené vydání, brož. Brno: VUTIUM, 2002. ISBN 80-214-2911-9.

[17] Genetické haraburdí - repetitivní DNA. In: *Aktuální genetika: Multimediální učebnice lékařské biologie, genetiky a genomiky* [online]. Ústav biologie a lékařské genetiky 1.LF UK a VFN. 2005-2006 [cit. 2012-11-24]. Dostupné z: http://biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm

[18] ABO-ZAHHAD, Mohammed, Sabah M. AHMED a Shimaa A. ABD-ELRAHMAN. Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques. *I.J. Information Technology and Computer Science.* 2012, č. 8, s. 22-36. DOI: 10.5815/ijitcs.2012.08.03.

[19] DEATON, Aimée M. a Adrian BIRD. CpG islands and the regulation of transcription. In: *Genes & Development 25* [online]. Cold Spring Harbor Laboratory Press, 2011 [cit. 2012-11-24]. ISSN 0890-9369/11. DOI: 10.1101/gad.2037511. Dostupné z: http://www.genesdev.org/cgi/doi/10.1101/gad.2037511

[20] AMBIKAIRAJAH, Eliathamby, Julien EPPS a Mahmood AKHTAR. Gene and exon prediction using time domain algorithms. In: *Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on*. Sydney, Australia, 2005, s. 199-202. ISBN 0-7803-9243-4. DOI: 10.1109/ISSPA.2005.1580230.

[21] HOLDEN, Todd, A. FLAMHOLZ, T. D. CHEUNG a kol. ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes. *SPIE Proceedings* [online]. 2007, roč. 6694 [cit. 2012-12-08]. DOI: 10.1117/12.732283.

[22] NCBI NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. O. sativa short highly repeated, interspersed DNA [online]. [cit. 2012-12-09]. Dostupné z: http://www.ncbi.nlm.nih.gov/nuccore/X64775.1

[23] BENSON, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research [online]. 1999, roč. 27, č. 2, s. 573-580 [cit. 2013-04-17]. Dostupné z: http://tandem.bu.edu/trf/trf.html

[24] NCBI NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Homo sapiens chromosome 21sequence from BAC GS1-53I10 [online]. [cit. 2013-04-17]. Dostupné z: http://www.ncbi.nlm.nih.gov/nuccore/AL133493

[25] VAIDYANATHAN, P.P. a Byung-jun YOON. The role of signal-processing concepts in genomics and proteomics. Journal of the Franklin Institute: special issue on Genomics. 2004, č. 341, s. 111-135. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.72.6984

[26] PALANIAPPAN, R., P. RAVEENDRAN, S. NISHIDA a N. SAIWAKI. Autoregressive spectral analysis and model order selection criteria for EEG signals. In: *IEEE 2000 Tencon Proceedings, Vols I-Iii: Intelligent Systems And Technologies For The New Millennium*. New York, NY 10017 USA: IEEE, 2000, A126-A129. ISBN 0-7803-6355-8 ISSN 0886-1420.

[27] ROTH, K., I. KAUPPINEN, P.A.A. ESQUEF a V. VALIMAKI. Frequency warped Burg's method for AR-modeling. In: Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on. 2003, s. 5-8. ISBN 0-7803-7850-4. DOI: 10.1109/ASPAA.2003.1285795.

[28] DU, L., H. ZHOU a H. YAN. OMWSA: detection of DNA repeats using moving window spectral analysis. *Bioinformatics*. 2007, roč. 23, č. 5, s. 631-633. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btm008.

[29] CHAKRAVARTHY, N., A. SPANIAS, LD IASEMIDIS a K. TSAKALIS. Autoregressive modeling and feature analysis of DNA sequences. *Eurasip Journal On Applied Signal Processing*. 2004, roč. 2004, č. 1, s. 13-28. ISSN 1110-8657. DOI: 10.1155/S111086570430925X.

[30] CRISTEA, P.D. Conversion of nucleotides sequences into genomic signals. *Journal Of Cellular And Molecular Medicine*. 2002, roč. 6, č. 2, s. 279-303. ISSN 1582-1838. DOI: 10.1111/j.1582-4934.2002.tb00196.x.

[31] ZHOU, Hongxia a Hong YAN. Autoregressive models for spectral analysis of short tandem repeats in DNA sequences. In: *IEEE International Conference On Systems, Man, And Cybernetics, Vols 1-6, Proceedings*. New York, NY 10017 USA: IEEE, 2006, s. 1286-1290. ISBN 978-1-4244-0099-7 ISSN 1062-922X. DOI: 10.1109/ICSMC.2006.384892.

[32] SANTO, Evan a Nevenka DIMITROVA. Improvement of spectral analysis as a genomic analysis tool. In: *2007 IEEE International Workshop On Genomic Signal*

Processing And Statistics. s. 46-49. ISBN 978-1-4244-0998-3. Dostupné z: http://ieeexplore.ieee.org/stamp/ stamp.jsp?tp=&arnumber=4365819

[33] NCBI NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION [online]. 8600 Rockville Pike, Bethesda MD, 20894 USA: U.S. National Library of Medicine [cit. 2013-05-08]. Dostupné z: http://www.ncbi.nlm.nih.gov/

[34] PRCHAL, Josef a Boris ŠIMÁK. Digitální zpracování signálů v telekomunikacích. ČVUT, 2001. ISBN 80-01-02149. Dostupné z: http://www.comtel.cz/files/ download.php?id=3370. Skriptum. ČVUT.

9. Seznam obrázků

10. Seznam zkratek a symbolů

A	adenin
AIC	Akaike information criterion = Akaikeho informační kritérium
AICc	corrected Akaike Information Criterion = korigované Akaikeho informační kritérium
AMFD	Average Magnitude Difference Function
AR	autoregresní
BIC	Bayesian Information Criterion = Bayesovo informační kritérium
bp	pár bází
C	cytosin
CA	dinukleotid cytosin-adenin
CpG	cytosin-fosfátová vazba-guanin
DFT	diskrétní Fourierova transformace
DNA	deoxyribonukleová kyselina
EIIP	Electron-Ion Interaction Potential = distribuce energií volných elektronů
FFT	Fast Fourier Transformation = rychlá Fourierova transformace
FPE	Final Prediction Error = finální predikce chyby
FT	Fourierova transformace
G	guanin
HQC	Hannan-Quinn Criterion = Hannan-Quinnovo kritérium
LINE	Long Interspersed Nuclear Elements = dlouhé rozptýlené jaderné elementy
LTR	Long Terminal Repeats = dlouhé terminální repetice
MDL	Minimum Description Length = minimální popis délky
NCBI	National Center for Biotechnology Information
o	posun okna
p	řád modelu
PSD	Power Spectral Density = výkonová spektrální hustota (výkonnostní spektrum)
RNA	ribonukleová kyselina
SINE	Short Interspersed Nuclear Elements = krátké rozptýlené jaderné elementy
STFT	Short-Time Fourier transformation = krátkodobá diskrétní Fourierova transformace
T	thymin
TDP	Time Domain Periodogram
TG	dinukleotid thymin-guanin
TRDB	Tandem Repeats Database
VNTR	Variable Number of Tandem Repeats = variabilní množství tandemových repetic
w	délka okna

11. Obsah přiloženého CD

- Textová část práce *Tereza_Reichlová_DP.pdf*
- Programová část práce *Tereza_Reichlová_DP_přílohy* obsahující:
  - kódy v Matlabu
  - DNA sekvence ve **.fasta* formátu
  - spektrogramy v plné kvalitě ve formátu **.emf*