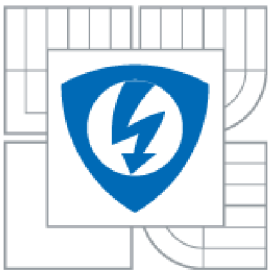
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

FOURIEROVA TRANSFORMACE A SPEKTROGRAMY V ANALÝZE DNA SEKVENCÍ

FOURIER TRANSFORMATION AND SPECTROGRAM ANALYSIS OF DNA SEQUENCES

DIPLOMOVÁ PRÁCE
MASTER´S THESIS

AUTOR PRÁCE — Ing. MICHAL KREJČÍ
AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ
SUPERVISOR

BRNO 2011

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské a ekologické inženýrství

Student: Ing. Michal Krejčí
Ročník: 2

ID: 83585
Akademický rok: 2010/11

NÁZEV TÉMATU:

Fourierova transformace a spektrogramy v analýze DNA sekvencí

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s krátkou diskrétní Fourierovou transformací a tvorbou spektrogramů. Zpracujte rešerši použití Fourierovy transformace a spektrogramů pro analýzu DNA sekvencí. V Matlabu vytvořte grafickou uživatelskou aplikaci pro tvorbu spektrogramů ze sekvencí DNA. Pomocí aplikace proveďte analýzu vybraného souboru sekvencí DNA z databáze NCBI nebo CBOL. Výsledky analýzy zhodnoťte.

DOPORUČENÁ LITERATURA:

[1] JAN, J. Číslicová filtrace, analýza a restaurace signálů. VUTIUM, Brno, 2002.
[2] DIMITROVA, N., CHEUNG, Y.H., ZHANG, M. Analysis and Visualization of DNA Spectrograms: Open Possibilities for the Genome Research. Proceedings of the 14th annual ACM international conference on Multimedia 2006.

Termín zadání: 15.10.2010 ***Termín odevzdání:*** 20.5.2011

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti diplomové práce:

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

UPOZORNĚNÍ:

ABSTRAKT

V této diplomové práci jsou v teoretické části popsány metody úprav DNA sekvencí pro frekvenční analýzu a základní vlastnosti DNA. Využitím krátkodobé Fourierovy transformace jsou vytvořeny barevné spektrogramy, pomocí kterých můžeme rozpoznávat některé charakteristické vzory v DNA. V praktické části práce je popsán program sloužící k vytvoření spektrogramů a k následné analýze. Dále je vytvořena analýza vybraných úseků genomu *C. elegans*. Nalezené vzory jsou porovnány s daty z databáze NCBI. Je zde poukázáno na vztah vytvořených spektrogramů a oblastí kódujících proteiny. Jsou zde uvedeny spektrogramy dobře rozeznatelných vzorů tvořených tandemovými repeticemi složenými ze satelitů, mikrosatelitů a minisatelitů.

ABSTRACT

Various methods of DNA sequences modifications for frequency analysis and basic characteristics of DNA are described in the theoretical part of this thesis. Tricolor spectrograms, created by short time Fourier transform help us to recognize some characteristic patterns in DNA sequences. Practical part of this work deals with developed programme which generates spectrograms and analyse them. Last part deals with the analysis of selected sequences of *C. elegans* genome. Some patterns are related to data of public databases such as NCBI. Various patterns are explained from the biological nature, which relates to chromosome structure and protein coding regions. Another well recognised patterns, tandem repetitions composed of satellites, microsatellites and minisatelites are described by spectrograms as well.

KLÍČOVÁ SLOVA

Fourierova transformace, spektrogram, analýza DNA sekvencí, frekvenční analýza.

KEY WORDS

Fourier transform, spectrogram, analysis of DNA sequences, frequency-domain analysis.

Bibliografická citace

KREJČÍ, M. *Fourierova transformace a spektrogramy v analýze DNA sekvencí*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 63 s. Vedoucí diplomové práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma „Fourierova transformace a spektrogramy v analýze DNA sekvencí“ vypracoval samostatně pod vedením vedoucího diplomové práce s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne: 20. května 2011 .....

podpis autora

Poděkování

Děkuji vedoucí diplomové práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne: 20. května 2011 .....

podpis autora

Obsah

SEZNAM OBRÁZKŮ

1. ÚVOD

DNA tvoří základní předpis stavby živočišných a rostlinných organismů na zemi. S pokroky v oblasti sekvenování DNA již dnes máme dostatek materiálů (sekvencí nukleotidových prvků A, C, T, G) jednotlivých živočichů. Na základě DNA můžeme například: sestavovat evoluční stromy, vyhledávat podobnosti mezi jednotlivými druhy, diagnostikovat choroby a včasně k nim určit náchylnost, modifikovat geneticky plodiny, pomoci při určení pachatele zločinu nebo určení otcovství (paternity). Cílem vědců je porozumění DNA, hlavně funkci jednotlivých genů, což by mělo obrovský dopad na budoucí vývoj lidské populace. Pomocí genové terapie bychom mohli z lidské DNA odstranit "škodlivý kód" jako je například předpoklad k dědičným chorobám a mnohem více. Manipulací s genomem hospodářských plodin lze vyšlechtit odolné odrůdy, které budou mít zásadní vliv na uživení stále rostoucí populace země.

Pro porovnávání DNA sekvencí je možné využít korelaci, ale pro zobrazení a představu struktury dlouhých řetězců je zapotřebí jiných metod. Zpravidla se řetězce nukleotidových prvků převádí pro další zpracování na číslo, kde se využívá sofistikovanějších metod, jako je zobrazení ve frekvenční oblasti. Schopnost extrahování nových poznatků závisí na zobrazení. Člověk zpracovává téměř 80 % poznatků z okolního světa pomocí očí, proto se nabízí vhodné využití barevných spektrogramů.

2. TEORETICKÝ ROZBOR BIOLOGICKÁ ČÁST

2.1 DEOXYRIBONUKLEOVÁ KYSELINA (DNA)[2][4]

DNA je nukleová kyselina, která je nositelkou genetické informace všech organismů s výjimkou některých nebuněčných, u nichž hraje tuto úlohu RNA (např. RNA viry). DNA je tedy pro život nezbytnou látkou, která ve své struktuře kóduje a buňkám zadává jejich program a tím předurčuje vývoj a vlastnosti celého organismu.

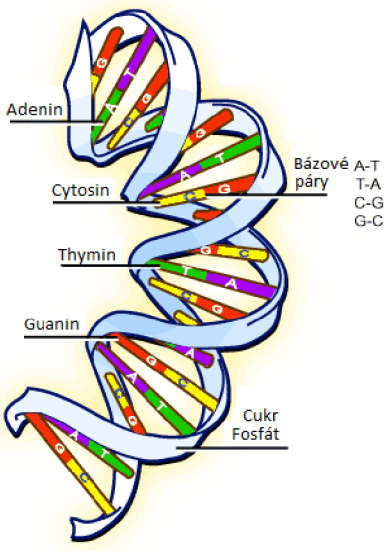
U eukarotických organizmů (jako jsou např. rostliny a živočichové) je DNA uložena vždy uvnitř buněčného jádra, zatímco u prokyryot (např. bakterie) se DNA nachází volně v cytoplazmě. Genová výbava člověka obsahuje přibližně 3,2 x 10^9 vazebných párů (3,2 Gbp).

DNA je biologická makromolekula tvořená dvoušroubovicí se dvěma řetězci nukleotidů v obou vláknech (viz. Obr. 1). Nukleotidy se skládají z heterocyklické dusíkaté báze, které jsou vzájemně propojeny pomocí vodíkových můstků. Spojení jednotlivých bází není dáno náhodně, ale je učeno snahou zaujmout energeticky nejvýhodnější konformaci v rámci dvoušroubovice (tzv. komplementarita bází). Situace propojení bází je znázorněna pomocí strukturních vzorců na Obr. 2. Mezi sousedními bázemi působí van der Waalsovy síly, které pomáhají k celkové stabilitě molekuly. V každém vláknu je tatáž informace, pouze s tím rozdílem, že jde o vzájemný „negativ“ (viz. Obr. 1):

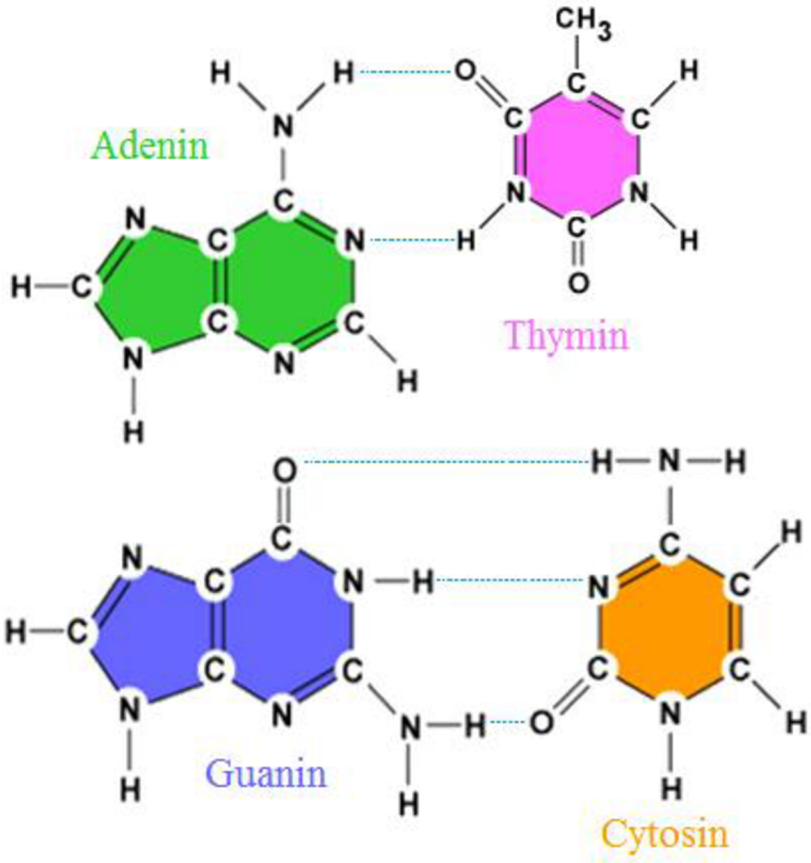
- fosfát (vazebný zbytek kyseliny fosforečné)
- deoxyribóza (pětiuhlíkový cukr - pentóza)
- nukleové báze
    - purinové (adenin A a guanin G)
    - pyrimidinové (thymin T a cytosin C).

Jednotlivé dusíkaté báze se mezi sebou spojují podle jednoduchého klíče:

- A ↔ T + T ↔ A *(vzájemně jsou spojeny dvěma vodíkovými vazbami)*
- C ↔ G + G ↔ C *(vzájemně jsou spojeny třemi vodíkovými vazbami)*



Obr. 1: Struktura DNA[6].

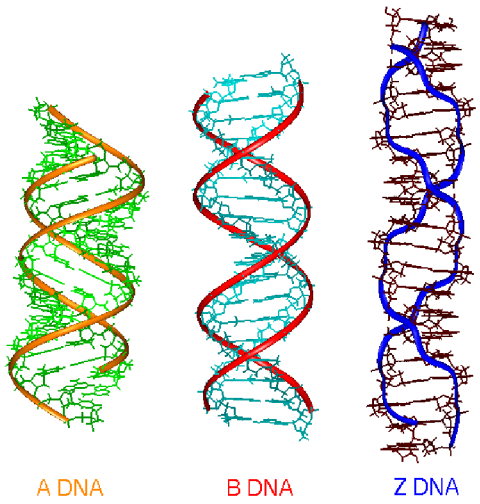


Obr. 2: Strukturní vzorce bázových prvků s naznačenými vodíkovými vazbami.

2.1.1 Úrovně struktury DNA

DNA lze rozdělit na tři základní struktury primární, sekundární a vyšší úrovně. Primární struktura je dána pořadím nukleotidů a přímo určuje genetickou informaci (např. *atcgtagctacg*). Sekundární struktura udává formu stočení dvoušroubovice (helix), jejíž základní model lze popsat následujícími znaky: sestává se ze dvou komplementárních antiparalelně orientovaných polydeoxyribo-nukleotidových řetězců ovíjejících společnou osu uspořádaných tak, že páry bází směřují dovnitř šroubovice a oporná deoxyribózofosfátová kostra směřuje napovrch. DNA může na sekundární úrovni vytvářet různé struktury v závislosti na konkrétní sekvenci a vlastnostech prostředí (vlhkost, iontová síla apod.). Rozeznáváme 3 typy konformací (viz. Obr. 3):

- Typ A (pravotočivá, 10 bp na otáčku, průměr vlákna je 2,3 nm)
- Typ B (pravotočivá, 11 bp na otáčku, průměr vlákna je 1,9 nm)
- Typ Z (levotočivá, 12 bp na otáčku, průměr vlákna je 1,8 nm)

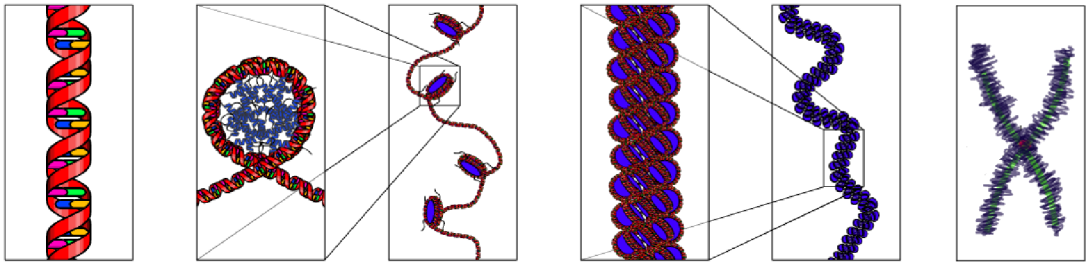


Obr. 3: Druhy sekundárních typů stočení šroubovice DNA. [4]

Typ B se v živých buňkách vyskytuje nejčastěji a představuje konformaci popsanou Watsonem a Crickem. Typ A se v buňkách vyskytuje za nižších vlhkostí (ve sporách mikroorganismů). Typ Z se uplatňuje při procesu rekombinace některých regulacích genové exprese (např. při aktivaci určitých genů). Sekundární struktura DNA může přecházet ve strukturu primární a naopak, takovéto děje nazýváme denaturace (tepelná, resp. chemická) a

renaturace. Reverzibilita přechodů se uplatňuje v základních molekulárně-genetických procesech, jakými jsou replikace a transkripce DNA.

Vyššími úrovněmi struktury se rozumí tvar stočení sekundární struktury v prostoru. Dochází ke vzniku nadšroubovice (*superhelixu*), která bývá označována jako terciární struktura DNA. Jednotlivé úrovně struktury DNA jsou zobrazeny na Obr. 4.



Obr. 4: Vyšší úrovně skládání molekuly DNA až do Chromatinu. [4]

2.1.2 Chromozomy

Jsou schopné samostatné funkce při přenosu informací. Základní stavební jednotkou chromozomů jsou tzv. *nukleosomy*. Jejich spiralizací vznikají chromatinová vlákna, jejichž následná spiralizace tvoří vlastní chromozom.

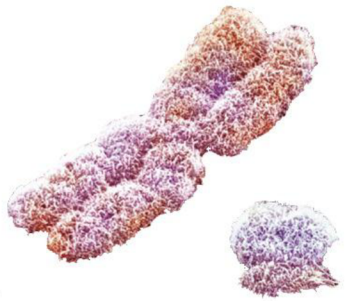
Struktura chromozomu je nejčastěji v podobě dvou ramének, mezi kterými je ztenčená oblast tzv. *centromera*. Chromozomy se liší velikostí ramének (krátké raménko se označuje jako *p* a dlouhé jako *q*). Soubor chromozomů se označuje jako *karyotyp,* u člověka se skládá z 23 párů. Z toho 22 párů jsou *autosomy* (tvoří homologní páry) a poslední pár je heterologní (ženy mají heterologní chromozom XX a muži XY, nazýváme je pohlavní chromozomy (*gonosomy*)). Podle uložení centromery dělíme chromozomy na:

- Telocentrické (jedno raménko)
- Metacentrické (dvě stejně dlouhá raménka)
- Submetacentrické (jedno raménko je mírně kratší)
- Akrocentrické (jedno raménko je značně kratší)

Podle nejnovějších studií je mužský chromozom Y nejrychleji se vyvíjejícím chromozomem vzhledem k tomu, že není párový a nemá tak kopii, pomocí které by prováděl samoopravu. Jednotlivé geny v něm obsažené jsou však zrcadlově zdvojeny. Toto zjištění

bylo učiněno na základě porovnání s chromozomem Y šimpanze, kde bylo docíleno odlišnosti o 30 %, zatímco ostatní chromozomy se liší o 1,5 - 2 %.

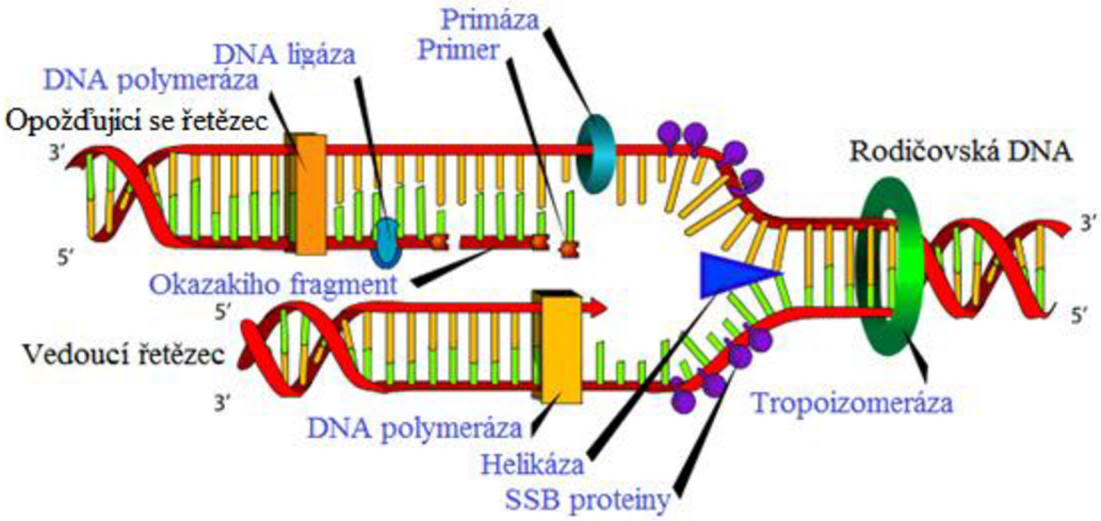
Koncové části jaderných chromozomů eukaryot se nazývají *telomery*. U člověka jsou tvořeny až 2000 opakováními sekvencí „5'-TTAGGG-3'“. Během replikace dochází ke zkracování telomer o 50 - 200 párů bází. Délka telomer je tedy markerem aplikativního stáří buněk a určuje, kolikrát se ještě můžou rozdělit. Dle studie [12] je délka telomer výrazně ovlivněna stresem.



Obr. 5: Chromozom X (vlevo) a chromozom Y (vpravo) – H. sapiens sapiens. [25]

2.1.3 Replikace DNA

Replikace DNA je proces zdvojení DNA obsažené v buněčném genomu. Vzhledem k odlišnostem mezi buňkami prokaryotickými a eukaryotickými existují i odlišnosti v procesu jejich replikace. Nicméně některé znaky jsou společné. Nejdříve je vytvořena replikační vidlice působením enzymu helikázy. Proces replikace je semidiskontinuální, neboť jeden řetězec (tzv. vedoucí) je syntetizován kontinuálně, zatímco druhý (tzv. opožďující se nebo váznoucí) je syntetizován diskontinuálně v úsecích označovaných jako Okazakiho fragmenty. Každá replika původní mateřské DNA se sestává z části původní molekuly a jednoho nově syntetizovaného řetězce, proto nazýváme tento děj semikonzervativní. Proces replikace DNA je popsán na Obr. 6.



Obr. 6: Replikace DNA u eukaryot. [4]

2.2 GEN

Základní jednotkou genetické informace je gen, tvořený lineárním uspořádáním nukleotidů. Lze ho definovat jako určitý úsek DNA (u RNA- virů úsek RNA), který obsahuje informaci o struktuře určitého proteinu nebo o vazbě specifických molekul proteinů k molekule DNA. Geny, které kódují primární strukturu nějakého proteinu, se označují jako *geny strukturní*. Geny, které nepodléhají translaci, nazýváme *funkční RNA* (tRNA a rRNA). *Geny regulační* obsahují informaci nutnou pro rozpoznání specifickým proteinem.

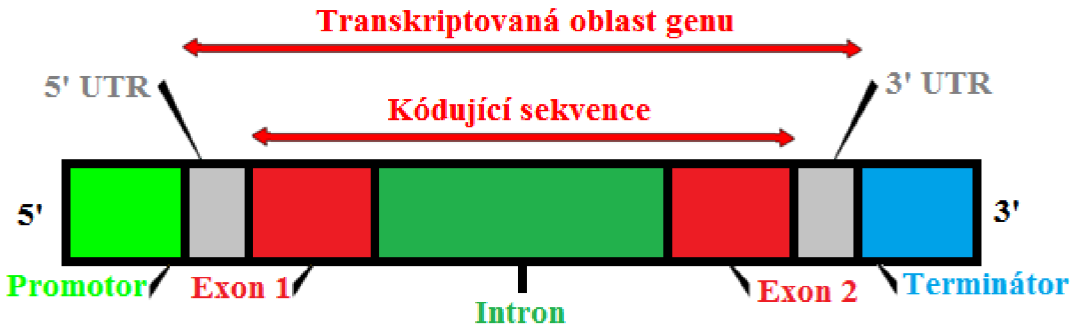
Funkcí genů je vytvoření konkrétního znaku (např. barva očí). Determinace jednoho znaku jedním genem se vyskytuje vzácně. Mnohem častěji je realizace znaku způsobena působením většího počtu genů. Říkáme pak, že znak je závislý na genových interakcích.

Převod genetické informace uložené v DNA je složitý proces označovaný jako *exprese genu*. Na molekulární úrovni probíhá ve dvou stupních:

- Transkripce (přepis) – obdobné replikaci DNA, dojde k oddělení vláken a dle principu komplementarity je vytvořeno mRNA vlákno
- Translace (překlad) – z pořadí nukleotidů mRNA se provede překlad do pořadí aminokyselin v peptidovém řetězci (tj. do primární struktury bílkoviny)

Struktura genu je znázorněna na Obr. 7. Promotor je část molekuly DNA nutný pro spuštění transkripce genu (např. *CAAT, TATA*). Nepřeložené oblasti mRNA (5' UTR a 3'

UTR) uvozují začátek a konec kódující sekvence, jejich rolí je zvýšení stability mRNA, lokalizace mRNA a řízení translační účinnosti. Kódující sekvence se skládá z *intronů* a *exonů*. Introny se nepřekládají do proteinu a jsou „vystřihnuty“ během tvorby mRNA mechanizmem zvaným *splicing*. Exony tvoří kódující oblast, podle níž obvykle v procesu translace vzniká bílkovina. Zajímavou vlastností exonů je 3 bázová perioda. Nalezení 3 bázové periody je jedním ze základních předpokladů pro nalezení kódující oblasti. Metody lokalizace exonů jsou popsány například v literatuře [8], [22].



Obr. 7: Struktura genu.

2.3 ZAJÍMAVÉ OBLASTI V DNA

2.3.1 CpG ostrovy

Jedná se o oblasti DNA bohaté na CG nukleotidy. Tyto úseky se ze 70 % nacházejí v oblasti promotorů, tudíž jsou dobrým ukazatelem začátku genu. CpG značí cytosin (C), fosfátovou vazbu (p) a guanin (G).

Existují dvě různé definice pro určení CpG oblasti. První je z literatury [10] a říká, že za CpG oblast lze považovat úsek minimálně 200 bp, kde je obsah CG větší než 50 % a poměr pozorovaných/očekávaných CpG je větší než 0,6 (tento poměr se vypočítá jako Počet(C)*Počet(G)/Délka segmentu). Druhá definice z literatury [11] uvádí za CpG oblast sekvenci delší než 500 bp, kde je obsah CG větší než 55 % a poměr pozorovaných/očekávaných CpG je větší než 0,65. Tyto oblasti jsou v barevném spektrogramu lehce rozeznatelné, jak je uvedeno například na Obr. 24.

2.3.2 Repetitivní DNA

Kódující i nekódující oblasti DNA mohou být unikátní, anebo se můžou nacházet v genomu ve více identických nebo podobných sekvencích. Sekvence s vysokým počtem kopií nazýváme repetitivní sekvence. Jestliže se kopie motivu vyskytují v blocích za sebou, hovoříme o tzv. *tandemových repeticích*. Pokud jsou repetitivní sekvence rozptýlené v genomu označujeme je jako rozptýlené repetice. Rozptýlené repetice se dále dělí na *transpozony* (přesouvají se bez nutnosti replikace) a *retrotranspozony* (tvoří až 45 % lidského genomu, množí se, označují se „junk DNA“).

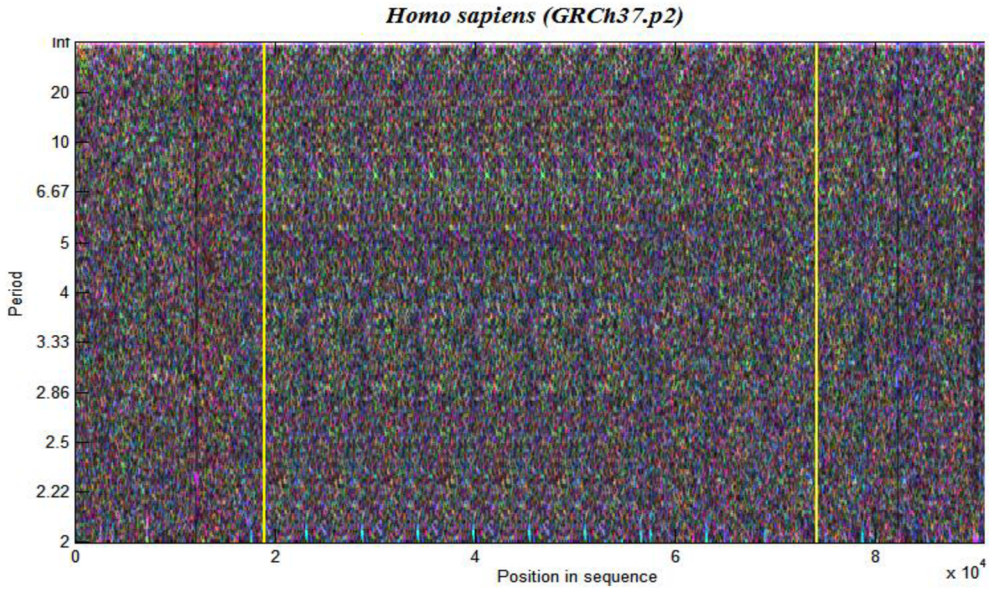
Tandemové repetice se dále dělí dle délky repetice. Nejdelší repetice jsou nazývány **satelity**. Satelitní DNA je hojná v oblasti centromer a konstitutivního heterochromatinu. Opakovaný vzor se pohybuje od 1 bp až po blok několika Mbp. **Minisatelity** jsou kratší tandemové repetice v rozsahu kbp, které se více vyskytují v subtelometrických oblastech chromozomů. Vzor minisatelitu nabývá délek od 6 do 100 bp. Obvykle jsou vysoce polymorfní co do počtu opakování jednotky repetice a používají se jako genetické markery (VNTR). Genetické markery jsou oblasti DNA, které mohou být jednoduše identifikovány a používají se při popisu variace druhů. V tomto případě se jedná o proměnný počet tandemových repetic na dané pozici. **Mikrosatelity** jsou zpravidla tvořeny opakováním 2 až 6 bp s množstvím repetic zřídka překračujícím stovky. Mikrosatelity jsou v genomu velice časté a vysoce polymorfní, jsou proto hojně používány jako genetické markery [13], [14] a [24].

Tandemová repetice tvořená minisatelity je zobrazena na Obr. 8. Repetice dosahuje délky 5,5k bp a je na spektrogramu označena žlutými horizontálními čarami. Nejsilnější opakovaný vzor počíná na 19 kbp a končí na 56 kbp (délka 37 kbp). Minisatelit je zde dlouhý 5 kbp a v tandemové repetici jich můžeme napočítat až 11. Spektrogram zobrazuje gen LPA kódující apolipoprotein, jehož koncentrace je spojovaná s rizikem kardiovaskulárních onemocnění, nachází například na 6 chromozomu *Homo sapiens* (GRCh37.p2 - oblast dlouhá 134.89 kbp od 160952515 do 161087407 bp). Spektrogram byl pořízen s nastavením velikosti okna 500 bp, překrytím oken 450 bp a Hannovým typem okna.

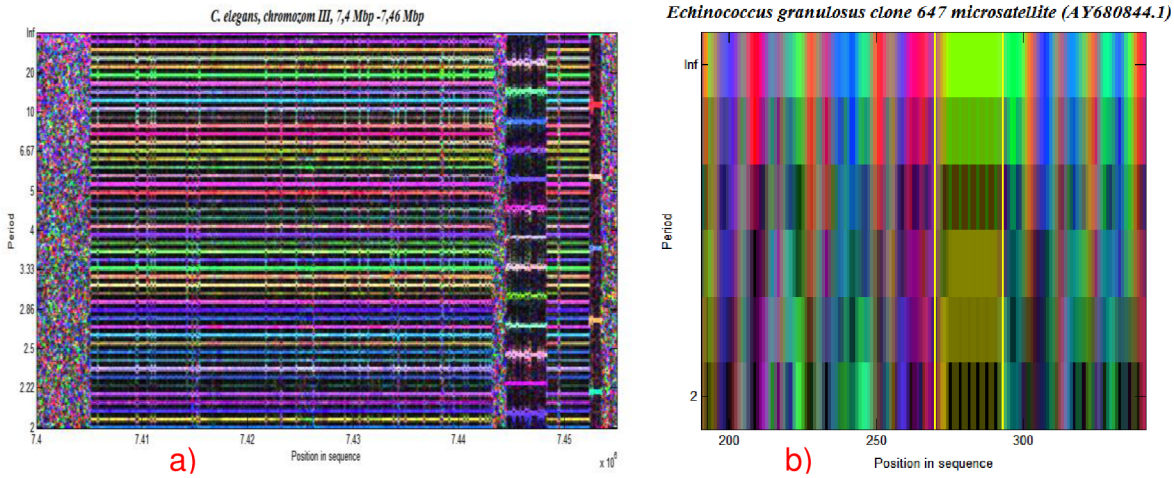
Obr. 9 a) ukazuje jiný typ tandemové repetice s délkou minisatelitu 95 bp a délkou repetice 5 kbp. Jedná se o III chromozom *C. elegans* v rozmezí 7,40 - 7,45 Mbp. Můžeme si zde povšimnout, že tandemová repetice je přerušena sekvencí bez opakování a jiným minisatelitem. Opakovaným minisatelitem s mírnými odlišnostmi tu je

"TTTCCCATTCATTTGTCTACATAGGGCATCGAAAAGCACCCAATATTTAGAGAACAGAA GATTTTGAGAATTACTGCCTCCAGAAATTGATGATT".

Mikrosatelity se pomocí spektrogramu hledají hůře, jelikož se jedná o malé oblasti. Ukázkou mikrosatelitu zobrazeného pomocí spektrogramu může být sekvence *Echinococcus granulosus clone 647 microsatellite* (AY680844.1) na Obr. 9 b). Získaná s nastavením délky okna 10 bp a překrytím 9 bp. Mikrosatelit je zde opět vyznačen pomocí vertikálních žlutých čar a nachází se na pozici 270 - 290 bp.



Obr. 8: Ukázka tandemové repetice tvořené minisatelity.



Obr. 9 a) Ukázka tandemové repetice tvořené minisatelity, b) Ukázka tandemové repetice tvořené mikrosatelity.

2.4 DNA A ČLOVĚK (H. SAPIENS SAPIENS)

Lidské tělo obsahuje okolo 100 trilionů buněk, které navzájem spolupracují. Výjimkou jsou bezjaderné červené krvinky, které neobsahují DNA. Všechny ostatní buňky obsahují lidský genom, což je řetězec 3,3 miliard nukleotidů.

Každá jaderná buňka tedy obsahuje stejnou DNA. Produkce nových buněk je závislá na schopnosti "zapínat" a "vypínat" různé oblasti informace (DNA), které obsahuje. Jádro buňky je odděleno pomocí membrány od okolí buňky a slouží jako řídící středisko regulující svůj růst, metabolismus a reprodukci. Srdcem jádra buňky je lidský genom složený z dvou setů 23 chromozomů (celkem tedy 46 chromozomů). Každý rodič přispívá tedy 23 chromozomy. Okolo 97 % lidského genomu nekóduje proteiny a nemá známou funkci, takovéto části DNA nazýváme odpadní DNA („junk DNA“). Odhaduje se, že člověk je vybaven okolo 25 000 geny. Na základě genů je kontrolováno téměř vše, počínaje růstem buněk a jejich interakcí až k inteligenci a psychologii daného jedince. Gen může být kódován různou délkou bází od několika 100 bází až po několik miliónů. Dva různí lidé na zemi se liší v 1 bázi na 1000 bází, což při počtu 3,3 Gbp znamená, že se od sebe odlišujeme na základě DNA pouze o 0,1%. Zatímco od šimpanze (*Pan troglodytes*), našeho nejbližšího příbuzného dle komparativní genomiky, se lišíme o celá 2 % (66 Mbp). [21]

Za vývoj jednotlivých druhů vděčíme evoluci, jež je způsobena mutacemi a zákony přírody (přežívá a rozmnožuje se jen ten nejsilnější). Většina mutací probíhajících v genomu je negativních a nevede tak k pozitivnímu výsledku pro organizmus. Pokud se negativní mutace uplatní již u embrya, končí většinou takovéto početí spontánním potratem, aniž by si něčeho žena všimla. Některé studie uvádějí, že až 25-50 % početí končí spontánním potratem. Statisticky je dáno, že u každého třetího člověka dojde během života k nebezpečným mutacím, které jsou souhrnně označovány jako rakovina.

Jako krátký příklad genetické mutace v evoluci bych uvedl nedávnou mutaci umožňující zpracování *laktózy* (mléčného cukru) v dospělosti. Všichni lidé jsou schopni zpracovávat laktózu jako malé děti. Lidské mateřské mléko obsahuje nejvyšší podíl laktózy ze všech živočichů (asi 8 %) a je tedy důležité pro správný vývoj lidských mláďat. Laktóza je však stravitelná jen v případě, že buňky stěn tenkého střeva vyrábí bílkovinu *laktáza*, která rozkládá laktózu na jednoduché cukry: *galaktózu* a *glukózu*. Tyto jednoduché cukry pak už mohou procházet přes stěnu střeva do krevního oběhu jako zdroj energie pro buňky všech

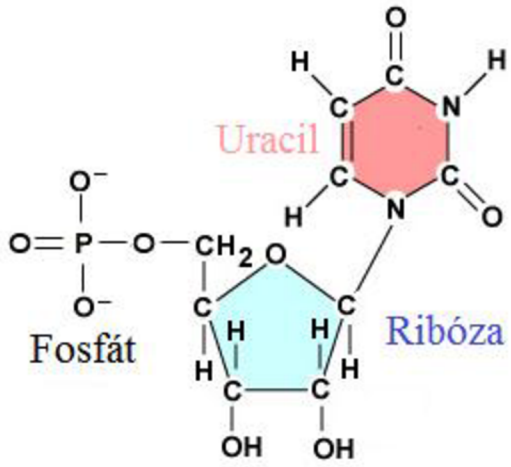
orgánů. Lidé s touto mutací na druhém chromozómu DNA, genu sloužícím pro tvorbu enzymu laktázy mohou lépe zužitkovat okolní zdroje pro své přežití než lidé bez této mutace. V období, kdy nebylo tolik potravy, byli lidé schopni zpracovávat laktózu zvýhodnění, a proto došlo k jejich vyšší reprodukci. Schopnost rozkládat laktózu se vyskytuje především v oblastech, kde se krávy chovají pro výrobu mléka. Gen kódující laktózu zcela chybí u původních amerických indiánů, z velké části se neobjevuje u Eskymáků a obyvatelů střední a jihovýchodní Asie. [15]

Vývoj v oblasti genetiky a bioinformatiky jde neustále vpřed. Uplynulo 10 let od získání celého genomu člověka rozumného (*H. sapiens sapiens*) a vědci po celém světě se neustále snaží porozumět všem informacím v něm skrytým. Nebude to trvat dlouho a přijde doba, kdy se naučíme manipulovat a obohacovat lidský genom a zušlechťovat tak lidskou rasu. Schopnost manipulování s DNA s sebou nese spoustu etických otázek. Ale při správném použití bychom mohli být schopni vymýtit civilizační choroby.

2.5 RIBONUKLEOVÁ KYSELINA (RNA)

Obdobně jako u DNA se jedná o makromolekulu tvořenou vzájemně spojenými nukleotidy. Společnými rysy jsou střídající se fosfátové a cukrové složky a dusíkaté báze adenin (A), cytosin (C), guanin (G). Komplementární bází pro adenin je uracil (U). RNA je obvykle jednovláknová a cukr v řetězci není deoxyribóza (jak tomu je u DNA) nýbrž ribóza. Schopností RNA je nést genetickou informaci a zároveň působit jako katalyzátor biologických reakci. RNA se může vyskytovat ve třech strukturních úrovních: primární, sekundární a terciální. Základní rozdíly ve struktuře RNA vzhledem ke struktuře DNA jsou uvedeny na Obr. 10. Základné dělení RNA:

- Mediátorová mRNA – překládána přímo z genové sekvence DNA
- Nekódující RNA – nenese informaci o struktuře proteinu
    - tRNA - zajišťuje transport aminokyselin k ribozomu
    - rRNA - podílí se na stavbě ribozomů a spoluúčastní se procesů, které se na nich realizují (proteosyntéza)
    - miRNA - regulace genové exprese některých genů

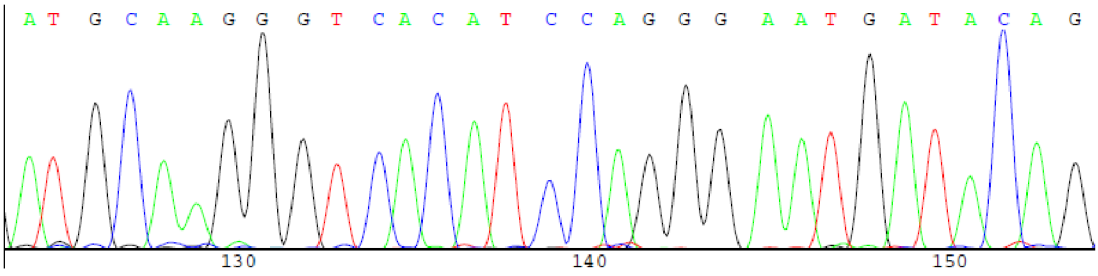


Obr. 10: Odlišnost RNA (ribóza, uracil) od DNA.

2.6 SEKVENOVÁNÍ DNA

Sekvenování DNA je souhrnný termín pro biochemické metody, jimiž se zjišťuje pořadí nukleotidových bází (A, C, G, T) v sekvencích DNA. Tyto sekvence jsou součástí dědičné informace v jádru.

Dnes je známo obrovské množství metod sekvenování DNA. Od sedmdesátých let 20. století je používána zejména metoda Fredericka Sangera, která využívá v klasické podobě dideoxynukleotidů a následné elektroforézy. V poslední době se do popředí dostává hlavně pyrosekvenování. Cílem je nalézt metodu dostatečně rychlou a levnou. Sekvenování DNA je užitečné nejen v základním výzkumu biologických procesů, ale i v aplikovaných oborech, jimiž je diagnostika nemocí či forenzní medicína nebo fylogenetika. Obr. 11 zobrazuje získanou sekvenci DNA.



Obr. 11: Sekvence DNA. [7]

2.6.1 Metody sekvenování [4][7]

Jelikož problematika sekvenování je složitá záležitost a vysvětlení principů není cílem práce, uvedu zde jen chronologický seznam způsobu sekvenování:

- Maxam–Gilbertova metoda
- Sangerova metoda
- Pyrosekvenování
- Nové metody (Whole-genome shotgun , Clone-by-clone)

2.7 HÁĎÁTKO OBECNÉ (CAENORHABDITIS ELEGANS)

C. elegans je volně žijící nepatogenní půdní helmint z kmene hlístic. Žije v půdě po celém světě a je významným modelovým organizmem, jehož výzkum započal v roce 1974. Jde o transparentní mikroskopický organizmus, 1 mm dlouhý, živící se bakteriemi z rozkládajících se materiálů. Vyskytuje se v dvou pohlavích, mužské (obsahující jeden chromozom X) a jako hermafrodit (obsahující dva chromozomy X). Obě pohlaví mají pět párů autosomatických chromozomů.

Jedná se o první mnohobuněčný organizmus, u něhož byl osekvenován kompletní genom (r. 1998 – 97 Mbp). Na základě studie *C. elegans* byla popsána apoptóza (programovatelná buněčná smrt). Skenováním genomu *C. elegans* vedlo k vývoji nových technologií sloužících k oskenování lidského genomu. Výhodou tohoto modelového organizmu je jeho nenáročnost, lehká adaptivita na laboratorní prostředí, krátký životní cyklus (přibližně 2 týdny) a vysoký počet potomků (300 za první 4 dny dospělosti). Celkový počet somatických buněk během postembryonálního vývoje stoupne u hermafroditů na 959 a na 1031 u samečků.

Anatomie těla *C. elegans* je jednoduchá a jednotlivé tkáně a buňky se dají snadno pozorovat (viz. Obr. 12). Zažívací trakt tvoří ústa, jícen, vlastní střevo a konečník. Svalové buňky jsou organizovány ve čtyřech podélných řadách běžících subdorsálně a subventrálně. Koordinovanými kontrakcemi svalových buněk je způsoben sinusoidní pohyb organizmu. Provazcová nervová soustava je tvořena 302 nervovými buňkami, které se nacházejí v okolí jícnu, v hlavové oblasti a v oblasti ocasu. [26]

Obr. 12: a) Hermafrodit C. elegans s červeným obarvením buněk [17], b) nezabarvený hermafrodit [18], c) modře zbarvený hermafrodit. [19]

V genomu hlístice *C.elegans* bylo identifikováno 18841 genů kódujících různé proteiny. Funkce 12000 těchto proteinů je však neznámá a její objasnění čeká na další biochemickou práci mnoha laboratoří. Srovnání genů nalezených v *C. elegans* s geny člověka ukazuje, že 74% dosud nalezených lidských genů má příbuzné geny v *C. elegans*. Přibližně čtvrtina genů *C. elegans* má homology v genomu kvasinky *Saccharomyces cerevisiae*, což je již také poměrně pokročilá forma života. S bakterií *Escherichia coli* má *C. elegans* společných jen 9% genů. [20]

Genom *C. elegans* je vyjádřený v číslech v tabulce č. 1, zdrojem genomu je databáze NCBI. Pomocí genetického experimentování vedoucího k formování různých tkání se vědcům daří pochopit děje odehrávající se ve více komplexních organizmech.

Tabulka č. 1: Genom *C. elegans*

Číslo chromozomu	Počet [Mbp]	Velikost soub. [MB]	Rozložení nukleotidů [Mbp](%)			
			A	T	C	G
I	15,07	14,5	4,8 (32)	4,8 (32)	2,7 (18)	2,7 (18)
II	15,28	15,1	4,8 (32)	4,8 (32)	2,7 (18)	2,7 (18)
III	13,78	13,6	4,4 (32)	4,4 (32)	2,4 (18)	2,4 (18)
IV	17,49	17,3	5,7 (32)	5,7 (32)	3,0 (17)	3,0 (17)
V	20,92	20,7	6,7 (32)	6,7 (32)	3,7 (18)	3,7 (18)
X	17,71	17,8	5,7 (32)	5,7 (32)	3,1 (18)	3,1 (18)
mtDNA	0,013	0,014	(31)	(45)	(9)	(15)

3. TEORETICKÝ ROZBOR TECHNICKÁ ČÁST

3.1 DISKRÉTNÍ FOURIEROVA TRANSFORMACE (DFT) [1]

Diskrétní periodické signály lze popsat diskrétní Fourierovu řadou. Diskrétní aperiodické signály lze popsat Fourierovu transformací diskrétního signálu (DTFT). Spektrum diskrétního signálu získaného pomocí DTFT je spojité, což není žádoucí pro počítačové zpracování. Proto je zavedena Diskrétní Fourierova transformace (DFT), která vzorkům časového průběhu (posloupnost konečné délky N) přiřazuje opět posloupnost konečné (stejné) délky (čárové frekvenční spektrum). Diskrétní Fourierova transformace je dána vztahem:

$$F(m) = \sum_{k=0}^{N-1} f(k) e^{-jm\frac{2\pi}{N}k} = \sum_{k=0}^{N-1} f(k) W^{km} \qquad (1)$$

$$f(m) = \frac{1}{N} \sum_{k=0}^{N-1} F(m) e^{jm\frac{2\pi}{N}k} = \frac{1}{N} \sum_{k=0}^{N-1} F(m) W^{-km} \qquad (2)$$

, kde koeficient *m = 0, 1... N-1* a určuje řád harmonické složky

k = 0, 1... N-1 a určuje pořadí odebraného vzorku v časové oblasti

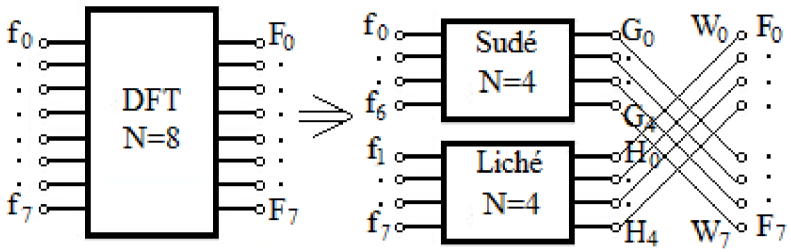
N značí počet odebraných vzorků.

3.1.1 Rychlá Fourierova transformace [1]

Pro výpočet DFT je třeba provést N^2P operací, kde N je počet vzorků a P je jedno komplexní násobení a sčítání. Kvadratická závislost pracnosti výpočtu na délce transformovaných dat je nepříjemná a způsobuje velké výpočetní nároky.

Vzhledem k praktické důležitosti DFT byly v průběhu času vyvinuty metody, které počty operací redukují. Například Goertzův algoritmus založený na teorii lineárních filtrů, který dosahuje úspory až 75% avšak kvadratická závislost zůstává nezměněna. Tento algoritmus je výhodný zejména tam, kde potřebujeme jen určité spektrální koeficienty.

Zásadní inovací, která drasticky snížila nároky na výpočet DFT, byla metoda z roku 1965 od pánů Cooley a Turkey. Algoritmus rychlého výpočtu DFT se označuje jako FFT (*fast Fourier transform*). Základním principem je rozklad vstupní posloupnosti v časové nebo frekvenční oblasti. Celková pracnost je pak dána $P*Nlog_2N$, tento vztah má při vysokých N téměř lineární průběh. Vyčíslením lze ukázat, že pro N = 8 je úspora asi 60 %, zatímco pro N = 131072 dokonce 99,99 %. Obr. 13 ukazuje postupný rozklad vstupní posloupnosti v časové oblasti.



Obr. 13: Rozklad DFT v originální oblasti.

3.2 NUMERICKÉ MAPOVÁNÍ

Jak již bylo řečeno, sekvence DNA nebo RNA je jednorozměrný signál tvořený bázovými prvky A, C, T, G nebo U. Pro potřeby numerického zpracování a následnou analýzu není toto vyjádření vhodné. Proto se práce se symboly nahrazuje reprezentací pomocí čísel, pro které je možné definovat velké množství operací.

Nevýhodou numerického mapování může být ztráta informace způsobená podstatou, že báze nesou informaci o chemických vlastnostech. Další způsoby reprezentace sekvencí jsou grafické metody, jako je reprezentace čtyřstěnem, Liaova metoda, PNN křivka, chaos game reprezentation (CGR). Grafická reprezentace se často používá pro vizuální porovnávání sekvencí.

3.2.1 Binární reprezentace 4D [3][5][8]

Jedná se o nejpoužívanější metodu pro reprezentaci DNA sekvencí. Používá se zejména při zpracování sekvencí pomocí Fourierovy transformace. Je tvořena pomocí čtyř

vektorů ($u_A(n)$, $u_C(n)$, $u_T(n)$, $u_G(n)$), které indikují přítomnost (log 1) nebo nepřítomnost (log 0) dané báze na pozici n. Toto vyjádření můžeme napsat vzorcem 3.

$$u_X[n] = 1\ jestliže\ s[n] = X\ jinak\ u_X[n] = 0 \quad (3)$$

, kde $u_X[n]$ jsou jednotlivé vektory

x je prvek báze (A, C, T, G nebo U)

$S[n]$ je symbolická sekvence

Příklad převodu: Mějme sekvenci bázových prvků *GACTGAGAT*, jednotlivé vektory potom jsou: $u_A = 010001010$

$u_C = 001000000$

$u_T = 000100001$

$u_G = 100010100$

3.2.2 Numerická reprezentace získaná redukcí 4D [8]

Binární reprezentace 4D je redundantní a lze ji tedy redukovat bez ztráty informace. Redukce je provedena pomocí přiřazením 3D jednotkového vektoru směřujícího ze středu do jednoho ze čtyř vrcholů pravidelného čtyřstěnu. DNA sekvence je pak vyjádřena pomocí tří numerických sekvencí:

$$x_r = \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]) \quad (4)$$

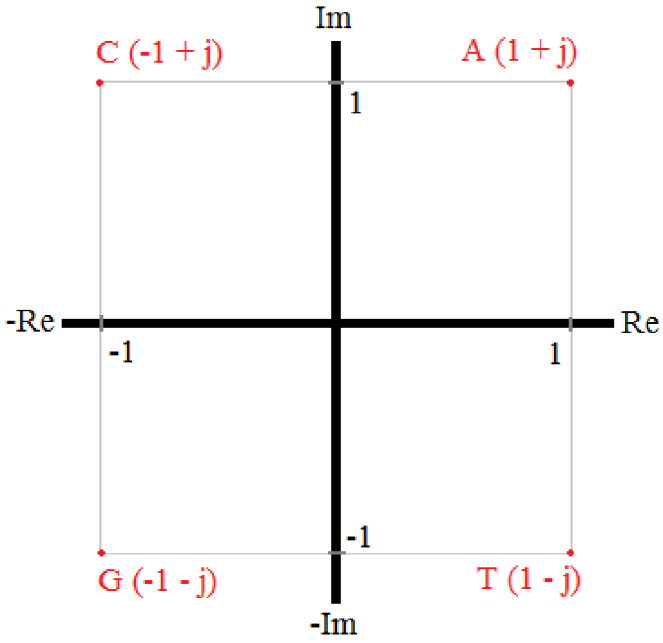
$$x_g = \frac{\sqrt{6}}{3}(u_C[n] - u_G[n]) \quad (5)$$

$$x_b = \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n]) \quad (6)$$

3.2.3 Reprezentace komplexními čísly [9]

Při této transformaci je jednotlivým nukleotidům přiřazeno komplexní číslo, což je výhodné, protože zůstává zachováno stejné množství informace jako v symbolickém zápisu. Purinové nukleotidy mají shodná znaménka pro reálnou a imaginární část, zatímco

pyrimidové mají různá znaménka. Nukleotidy se slabou vazbou mají kladnou reálnou část. Kladná imaginární část znamená, že se jedná o nukleotidy s amino skupinou a záporná imaginární část reprezentuje nukleotidy s keto skupinou. Na Obr. 14 je znázorněno rozložení nukleotidů v komplexní rovině.



Obr. 14: Reprezentace komplexními čísly.

3.3 SPEKTROGRAM

Spektrogram je nástrojem spektrální analýzy, která zobrazuje vývoj spekter signálu v čase. Časově-frekvenční analýza spočívá ve zjišťování spekter signálu z jeho krátkých segmentů a formuluje tak spektrum jako dvourozměrnou funkci, závislou nejen na frekvenci, ale i na pozici v čase. Praktická analýza vychází z konečných úseků signálu získaných pomocí použitého okna. Pokud má okno vhodnou délku N a je definováno jako klouzavé na časové ose, může být použito pro časově frekvenční analýzu. Pozorovací interval je vždy kompromisem mezi požadavky na dostatečnou časovou a frekvenční rozlišovací schopnost neboť rozlišitelná diference frekvencí je nepřímo úměrná délce okna, zatímco minimální rozeznatelný časový úsek je délce okna úměrný. Časovou rozlišovací schopnost lze zvýšit tím, že dílčí okna mají zvolený přesah, např. o polovinu své délky. Pak dostaneme podél

časové osy přiměřeně více spekter a lze lépe sledovat případný rychlý vývoj zejména na straně vysokých kmitočtů.

Spektrogram nejčastěji zobrazujeme jako dvojrozměrný obraz, v němž jedna souřadnice odpovídá frekvenci, druhá času a barva nebo úroveň jasu je přímo úměrná amplitudě odpovídajících koeficientů spekter.

3.3.1 Spektrogram pro DNA sekvence [3][5]

Hlavní výhodou spektrogramů je zobrazení celých chromozomů. Například lidský chromozom 1 má 150 Mbp. Pohled na 150 MB dlouhou sekvenci A, G, T, C v lineárním řazení nám nedovolí extrahovat základní strukturu a skryté informace. Nicméně pomocí spektrogramu jsme schopni zobrazit lidský chromozom 1 v jediném obrázku. Spektrogramy mohou být s různým rozlišením a různou velikostí okna. Představují efektivní způsob sloužící k důkladnému hledání všech typů speciálních vzorů a charakteristik v DNA sekvenci.

DNA spektrální analýza je nový způsob jak se vypořádat s řadou problémů v bioinformatice. Může být použita například k predikci proteinově kódovaných oblastí. Nicméně plné využití této techniky je zatím ve fázi vývoje. Základní myšlenkou je považovat výskyty každé nukleotidové báze v DNA sekvenci jako individuální binární signál (rovnice 3) a poté každý transformovat do frekvenční oblasti (viz. rovnice 6). Amplituda jednotlivých frekvenčních komponentů potom určí, jak silný je určitý vzor bázového prvku opakovaný na dané frekvenci. Vyšší hodnota často signalizuje přítomnost opakování. Pro lepší čitelnost výsledku je každá báze reprezentována vlastní barvou.

Barevný obrázek v podobě spektrogramu může sdělit daleko více informací o vlastnostech DNA sekvence v porovnání s původními neupravenými daty. V podstatě sytost v určité oblasti odráží celkové nukleotidové složení a světlé čáry jsou místa, kde se objevují opakující se vzory. Algoritmus pro vytvoření DNA spektrogramu můžeme definovat dle literatury [3] v pěti krocích:

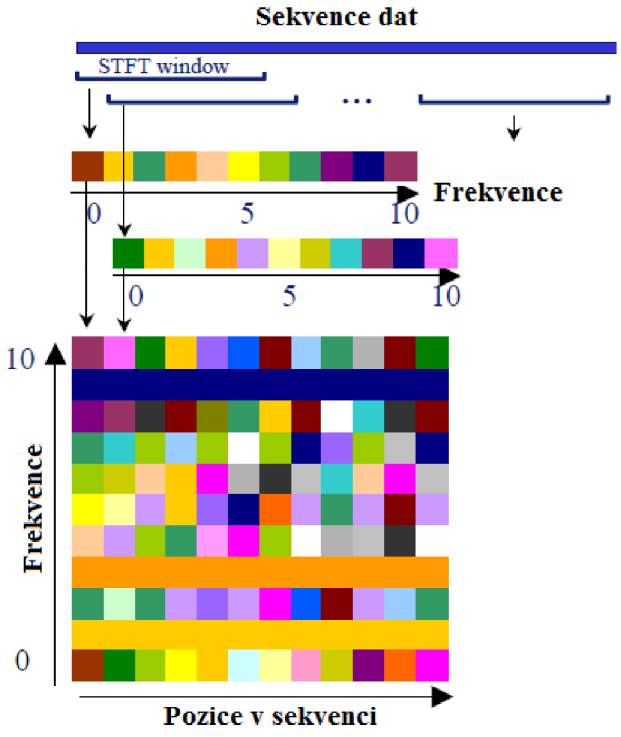
1) Převod DNA sekvence do binární podoby (viz. kap. 3.2.1)
2) STFT jednotlivých složek u_A, u_T, u_C, u_G
3) Mapování DFT hodnot do RGB barev
4) Normalizování velikosti pixelu na rozsah 0 - 1
5) Úprava obrazu (filtrování, hranování, úprava jasu, sytosti apod.)

Takto získaný spektrogram může být dále podroben dalším úpravám, které poslouží k lepší čitelnosti. Takovými úpravami lze docílit odstranění „šumu“, například pomocí morfologických operací (otevření následované uzavřením) nebo vytvořením histogramů a jejich prahováním.

3.4 KRÁTKODOBÁ FOURIEROVA TRANSFORMACE (STFT)

Slouží k vytvoření spektrogramu a je tvořena oknem, které se pohybuje po sekvenci dat. Tím řeší problém souběžného určení času i frekvence, na kterých je rozmístěna energie. Tato transformace tedy provádí časově-frekvenční analýzu pro vybranou část vstupních dat. Mnohokrát se tak opakuje proces popsaný v kap. 3.3.1. Situace popisující vznik spektrogramu je popsána na Obr. 15.

Vzhled výsledného spektrogramu je ovlivněn zvolenou velikostí okna a délkou přesahu oken. Velikost okna ovlivňuje množství frekvencí v jednom okně (frekvenční rozlišovací schopnost) a přesah oken určuje jemnost přechodu mezi jednotlivými okny.



Obr. 15: Vznik spektrogramu z DNA sekvence. [3]

3.4.1 DFT binárních nukleotidových bází [3]

Frekvenční spektrum jednotlivých nukleotidových bází je vytvořeno z jejich binární reprezentace získané pomocí rovnice 3.

$$U_X[k] = \sum_{n=0}^{N-1} u_x[n] e^{-j\frac{2\pi}{N}kn} \tag{7}$$

, kde $k = 0, 1, \ldots, [N/2]+1$

$x = A, T, C, G$

3.4.2 Mapování DFT spekter na RGB [3]

Čtyři DFT sekvence získané pomocí rovnice 7, jsou nyní redukovány na 3 sekvence mapováním do RGB prostoru pomocí následujících lineárních rovnic:

$$X_r[k] = a_r|U_A[k]| + t_r|U_T[k]| + c_r|U_C[k]| + g_r|U_G[k]| \tag{8}$$

$$X_g[k] = a_g|U_A[k]| + t_g|U_T[k]| + c_g|U_C[k]| + g_g|U_G[k]| \tag{9}$$

$$X_b[k] = a_b|U_A[k]| + t_b|U_T[k]| + c_b|U_C[k]| + g_b|U_G[k]| \tag{10}$$

, kde $a_{r,g,b}$, $t_{r,g,b}$, $c_{r,g,b}$, $g_{r,g,b}$ jsou vektory barev pro báze A, T, C, G

$X_{r,g,b}[k]$ je výsledný pixel složený z váhovaných vektorů barev

4. REALIZACE VYBRANÝCH METOD V MATLABU

4.1 VLASTNOSTI PROGRAMOVACÍHO JAZYKA

Programový systém MATLAB vyvinula firma MATHWORKS. Název je odvozen z anglického výrazu *MATrix LABoratory*. Jedná se o velice výkonný jazyk pro vědecké a technické výpočty, zejména v maticových aplikacích. MATLAB byl implementován na všech významných platformách, jako jsou Windows, Linux, Solaris, Mac. MATLAB obsahuje velké množství knihoven, které pokrývají prakticky všechny oblasti lidské činnosti a díky otevřené architektuře je uživateli umožněno vytvářet funkce dle své potřeby. Tyto knihovny jsou neustále vyvíjeny a rozšiřovány dle vývoje vědních a technických oborů. Dalším znakem MATLABu je návaznost na jiné programovací jazyky, jako jsou například C, Java a Fortan. MATLAB také podporuje tvorbu grafických uživatelských rozhraní pomocí programové nadstavby GUIDE. Od verze 7.3 je MATLAB rozšířen o kompilátor, který dokáže vytvořit spustitelnou aplikaci bez nutnosti instalace produktu MATLAB. Další výhodou MATLABu je jednoduchá syntaxe kódu (není nutné definovat proměnné, alokovat paměť apod.), možnosti zobrazení jednotlivých proměnných a trasování programu.

Vzhledem k zmíněným vlastnostem lze odvodit hlavní nevýhodu MATLABu oproti nativním jazykům a tou je jeho rychlost. Proto je vhodné pro praktické nasazení přepsat algoritmus do nativního jazyka (ANSI C, C++ apod.).

4.2 STRUKTURA PROGRAMU

Program pro frekvenční analýzu pomocí spektrogramů se skládá z grafického uživatelského rozhranní využívající jednotlivé funkce. Hlavním spouštěcím programem je *DNAspect*. Program obsahuje více grafických uživatelských rozhraní sloužících pro další analýzu dat. Tato rozhraní se spouštějí z hlavní aplikace a tvoří poté samostatnou aplikaci. Důvodem pro vytvoření dalších oken byl zejména fakt, že hlavním výstupem je spektrogram a pro kvalitní čtení jeho obsahu je zapotřebí, aby jeho plocha byla co největší.

Nejprve je nutné získat data pro analýzu, která můžeme načíst ze souborů ve formátu **.txt* nebo **.fasta*, popřípadě využít funkci na konvertování vstupních dat do zvoleného formátu. Druhou možností získání dat je generování sekvence. Grafické rozhraní pro

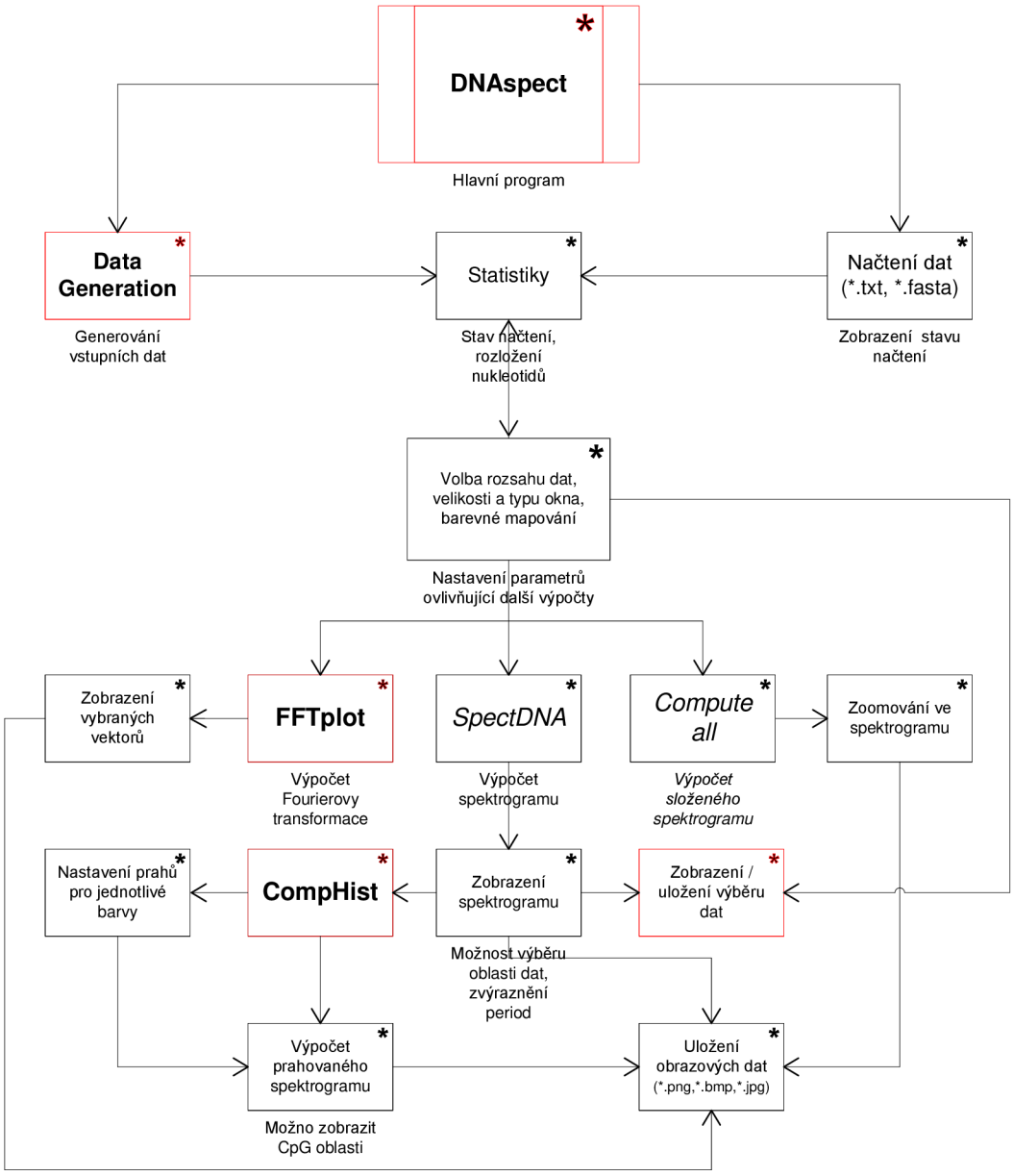
generování sekvence *DataGeneration* umožňuje zadat počet nukleotidů, periodu opakování nukleotidů a vložení náhodného šumu s možností řízení jeho četnosti.

Jakmile máme data načtena, je zobrazena informace o názvu souboru počtu nukleotidů a rozložení nukleotidů. Nyní si můžeme vybrat oblast našeho zájmu (rozsah nukleotidů) a vypočítat FFT nebo spektrogram. Jakmile máme vypočítaný spektrogram, můžeme přistoupit k zvýraznění period, výběru zajímavé oblasti pro zobrazení v lepším rozlišení, zobrazení vybraného rozsahu nukleotidů nebo k prahování spektrogramu. Pokud se jedná o velmi dlouhou sekvenci nukleotidů (>3,5 Mbp) a budeme požadovat například zobrazení celého chromozomu se 21 Mbp (*C. elegans, chromozom V*) je zde funkce *Compute all*, která rozloží vstupní sekvenci a dílčí výsledky spojí do výsledného spektrogramu. V takto rozsáhlém spektrogramu se pak můžeme pohybovat pomocí posuvníku po velikostech okna definovaném pod tlačítkem *Compute all.* Tyto omezení jsou z důvodu velkých nároků na paměť při zpracování takového množství dat.

Program nabízí řadu možností pro uložení výsledků analýzy. Je možnost uložit obrázek celé obrazovky se všemi nastaveními, samotný spektrogram ve formátu obrázku (**.png , *.bmp , *.jpg*) s osami nebo uložení výběru dat (nalezené zajímavé oblasti) do souboru ve formátu *****.txt* nebo **.fasta.*

Vzhled spektrogramu lze značně ovlivnit volbou velikostí okna a překrytí, zvoleným typem algoritmu pro výpočet, typem barevné normalizace, typem okna pro výpočet STFT nebo typem barevného mapování.

Struktura programu je blokově popsána na Obr. 16.



Obr. 16: Bloková struktura programu (červeně nová grafická rozhraní)

4.2.1 Popis a význam funkcí

Nejdůležitější a nejvíce využívané části kódu jsou převedeny do funkcí. Značná část kódu je obsažena u jednotlivých funkčních grafických prvků a slouží zejména k ošetření zadávaných dat, načítání dat, nastavování rozsahů os apod. Zdrojový kód je hojně okomentován, tak aby bylo snadné se v něm orientovat a případně provádět změny.

Každá funkce je okomentována hlavičkou informující k čemu slouží. Zde je seznam funkcí s krátkým popisem jejich významu:

- *CounT* - funkce slouží pro vytvoření grafu s rozložením nukleotidů
- *BinConv* - transformace sekvence nukleotidů do binárních vektorů
- *GenSeq* - generování dat s volbou opakování
- *Convert* - převod načtených dat do zvoleného formátu
- *UimenuFcn* - ovládání kontextového menu na výběr typu okna
- *SpectDNA_I* - výpočet spektrogramu
- *SpectDNA_II* - výpočet spektrogramu
- *SpectDNA_I_all* - spektrogram pro dlouhé sekvence (>3,5 Mbp)
- *SpectDNA_II_all* - spektrogram pro dlouhé sekvence (>3,5 Mbp)
- *Conect_all* - propojení spektrogramů z funkcí *_*all*
- *CpGsearch* - vyhledávání CpG ostrovů (viz. kap. 2.3.1)

Bloky zobrazené červeně na Obr. 15 představují vlastní grafické rozhraní (nové okno) a je možné je spouštět pomocí textových záložek v položce *View,* ikonami pro rychlé spouštění v levé horní části programu nebo pomocí tlačítek ve spodní části programu. Zde je seznam GUI a jejích význam:

- *DNAspect* - hlavní program, ovládání načítání dat a spouštění ostatních GUI, zobrazení rozložení nukleotidů a vlastního spektrogramu, výběr dat ve spektrogramu, zvýrazňování period
- *PlotFFT* - zobrazení Fourierovy transformace binárních vektorů
- *Tresholding* - prahování spektrogramu pro nalezení nejsilnějších vzorů, detekce CpG oblastí
- *Show_sequence* - zobrazení zvoleného rozsahu dat v podobě nukleotidů, možnosti ukládání
- *DataGeneration* - GUI sloužící pro generování dat

4.2.1.1 Funkce pro vlastní výpočet spektrogramu (SpectDNA)

Tato funkce se vyskytuje ve čtyřech variantách, římská číslice (*I* a *II*) určuje typ použitého algoritmu a přítomnost koncovky *all* určuje objem dat pro zpracování.

SpectDNA jsou funkce, do kterých vstupují binární vektory nukleotidů (u_A, u_T, u_C, u_G), délka okna, překrytí oken, typ okna, rozsah dat a typ barevné normalizace. Výstupem funkcí je spektrogram ve formě RGB obrázku a v případě *SpectDNA I a II* i informace o zrušení výpočtu. Základní rozdíl mezi algoritmy *I* a *II* je v pořadí barvení vektorů a v typech barevných vektorů. Program umožňuje volbu algoritmů v záložce *Settings -> Algorithm.* Vývojový diagram pro algoritmus *SpectDNA_ II* je na Obr. 17

SpectDNA_I provádí obarvení čtyř binárních vektorů před výpočtem spektrogramu pomocí tří lineárních rovnic (11, 12, 13).

$$x_r[k] = a_r U_A[k] + t_r U_T[k] + c_r U_C[k] + g_r U_G[k] \quad (11)$$

$$x_g[k] = a_g U_A[k] + t_g U_T[k] + c_g U_C[k] + g_g U_G[k] \quad (12)$$

$$x_b[k] = a_b U_A[k] + t_b U_T[k] + c_b U_C[k] + g_b U_G[k] \quad (13)$$

, kde $a_{r,g,b}$, $t_{r,g,b}$, $c_{r,g,b}$, $g_{r,g,b}$ jsou vektory barev pro báze A, T, C, G , $U_{A,C,T,G}[k]$ jsou binární vektory a $x_{r,g,b}[k]$ jsou vektory pro STFT

Pro barevné mapování je obecně doporučeno, aby barevné vektory byly voleny jako vrcholy pravidelného čtyřstěnu. Následující barevné vektory využité ve funkci *SpectDNA_I* jsou zvoleny dle literatury [5]:

$a_r = 0$	$t_r = 0.911$	$c_r = 0.244$	$g_r = -0.817$
$a_g = 0$	$t_g = -0.244$	$c_g = 0.911$	$g_g = -0.471$
$a_b = 1$	$t_b = -0.333$	$c_b = -0.333$	$g_b = -0.471$

Pro prezentaci dat je důležité provést normalizaci výstupních dat do obrazového formátu (rozsah 0 - 1). Jelikož se jedná o velmi významnou úpravu, která je pro všechny funkce stejná, uvedu jí v samostatné podkapitole.

SpectDNA_II nejdříve vypočítá STFT binárních vektorů a až poté je provedeno obarvení barevnými vektory. Tento typ funkce vznikl za účelem detekce významných složek v obraze. Ukázalo se, že prahováním koeficientů STFT bez obarvení lze dosáhnout lepších výsledků než prahováním ve vytvořeném spektrogramu. Funkce umožňuje změnu barevných vektorů na mapování *A T* jednou barvou (červenou) a *C G* druhou barvou (zelenou) nebo jednobarevné spektrogramy zobrazující rozložení AT nebo CG. Barvení vektorů probíhá pomocí rovnic 8, 9, 10 uvedených v kapitole 3.4.2. Barevné vektory pro čtyři typy vyjádření uvádí tabulka č. 2.

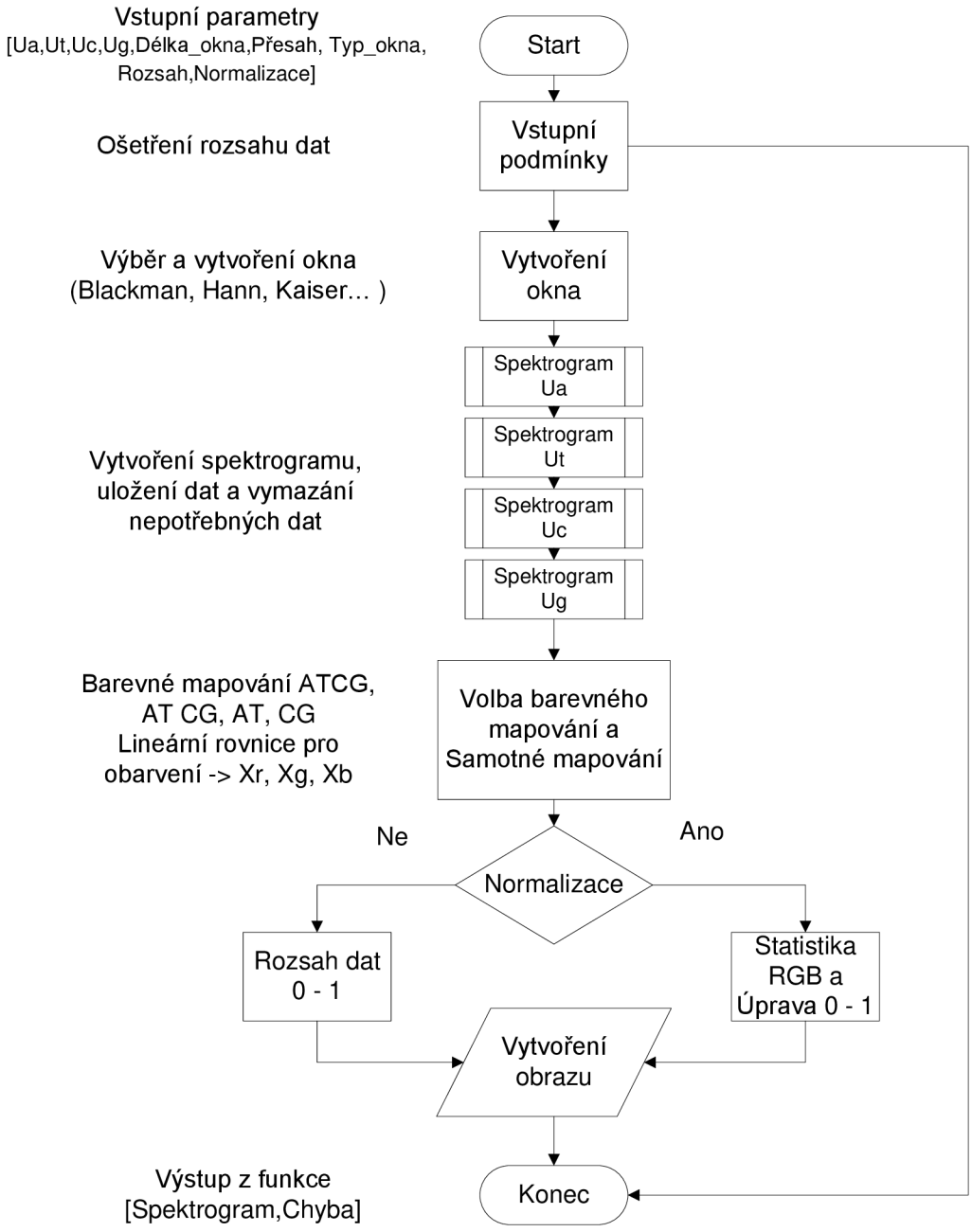
Tabulka č. 2 Barevné vektory

	ATCG				AT CG				AT				CG			
	a	**t**	**c**	**g**	**a**	**t**	**c**	**g**	**a**	**t**	**c**	**g**	**a**	**t**	**c**	**g**
R	0	1	0	1/3	1	1	0	0	1	1	0	0	0	0	0	0
G	0	0	1	1/3	0	0	1	1	0	0	0	0	0	0	1	1
B	1	0	0	1/3	0	0	0	0	0	0	0	0	0	0	0	0

Jelikož se mnohdy jedná o časově náročné výpočty, jsou funkce vybaveny ukazatelem stavu výpočtu a odhadem doby výpočtu. Tento ukazatel taktéž umožňuje zrušit výpočet. Při dlouhých sekvencích dat jsou při výpočtu spektrogramu vysoké nároky na paměť, proto v průběhu výpočtu provádím uložení výsledků do souboru a vymazání nepotřebných dat. Tyto uložené soubory lze později využít například v analýze pomocí histogramu, kde již nemusí být prováděn výpočet spektrogramu, ale pouze obarvení zvolenými barevnými vektory. Veškeré samovolně vytvořené soubory jsou po ukončení aplikace odstraněny.

SpectDNA_*_all jsou upravené předchozí funkce tak, aby automaticky vytvořili spektrogram z velkého množství dat. Jelikož paměťový prostor je omezený provede funkce rozčlenění dat na okna o velikosti zadané uživatelem. V těchto oknech jsou vytvořeny spektrogramy dle zadaných parametrů. Spektrogramy jsou ukládány do souborů a poté jsou spojeny a vymazány. Funkce rovněž umožní se ve vytvořeném spektrogramu pohybovat pomocí posuvníku pod spektrogramem. Rozsah nukleotidů zobrazených najednou na ose X je ovládán pomocí volby délky okna. Tato úprava umožní vytvoření spektrogramu ve velmi

vysokém rozlišení. Časové nároky na takový výpočet jsou značné, proto je funkce opět vybavena ukazatelem stavu výpočtu a předpokládanou dobou výpočtu.

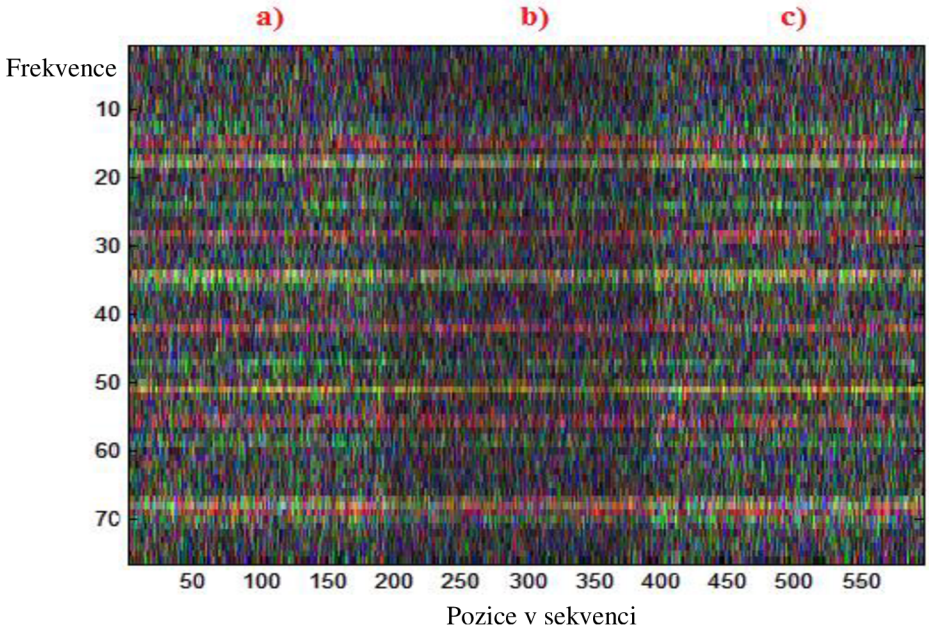


Obr. 17: Vývojový diagram funkce *SpectDNA_II.*

4.2.1.2 Normalizace barev spektrogramu

Jednou z nejdůležitějších úprav je normalizace vytvořených spektrogramů pro barevné složky R, G, B na rozsah 0 - 1. Pokud použijeme klasickou normalizaci pro každou barevnou složku zvlášť pomocí rovnice 14, nejsou jednotlivé vzory dostatečně rozeznatelné. Příkladem může být Obr. 18. Tato základní volba normalizace je zvolena pouze na přání uživatele pokud zruší v *Settings ->Color correction.* Na Obr. 18 jsou zobrazeny 3 typy normalizace tímto způsobem. První typ normalizace (Obr. 18 a) využívá oslabení ostatních barevných vektorů při překročení rozsahu. Druhý typ normalizace (Obr. 18 b) využívá vzorce 14 a poslední typ normalizace (Obr. 18 c) je proveden po jednotlivých STFT oknech.

$$X_{r,g,b}(i,j) = \frac{X_{r,g,b}(i,j) - \min[X_{r,g,b}]}{\max[X_{r,g,b}] - \min[X_{r,g,b}]} \tag{14}$$

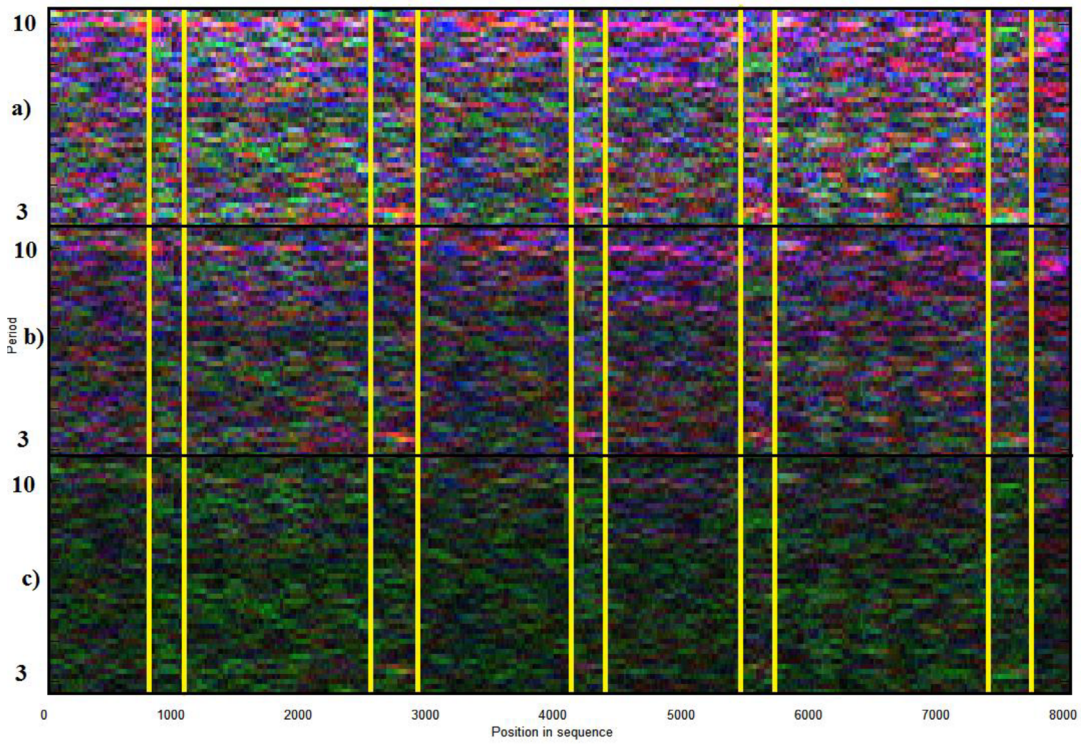


Obr. 18: Normalizace matic barev R, G, B na rozsah 0 - 1.

Druhý typ normalizace vychází ze statistického rozboru hodnot spektrogramu. Pro každou barevnou složku vypočítám disperzi dat pomocí směrodatné odchylky. Pokud je směrodatná odchylka velká, jsou v souboru dat velké odlišnosti. Provedu výpočet střední hodnoty z dat, jež představují šum, pro každou barevnou složku. Normalizační konstantou,

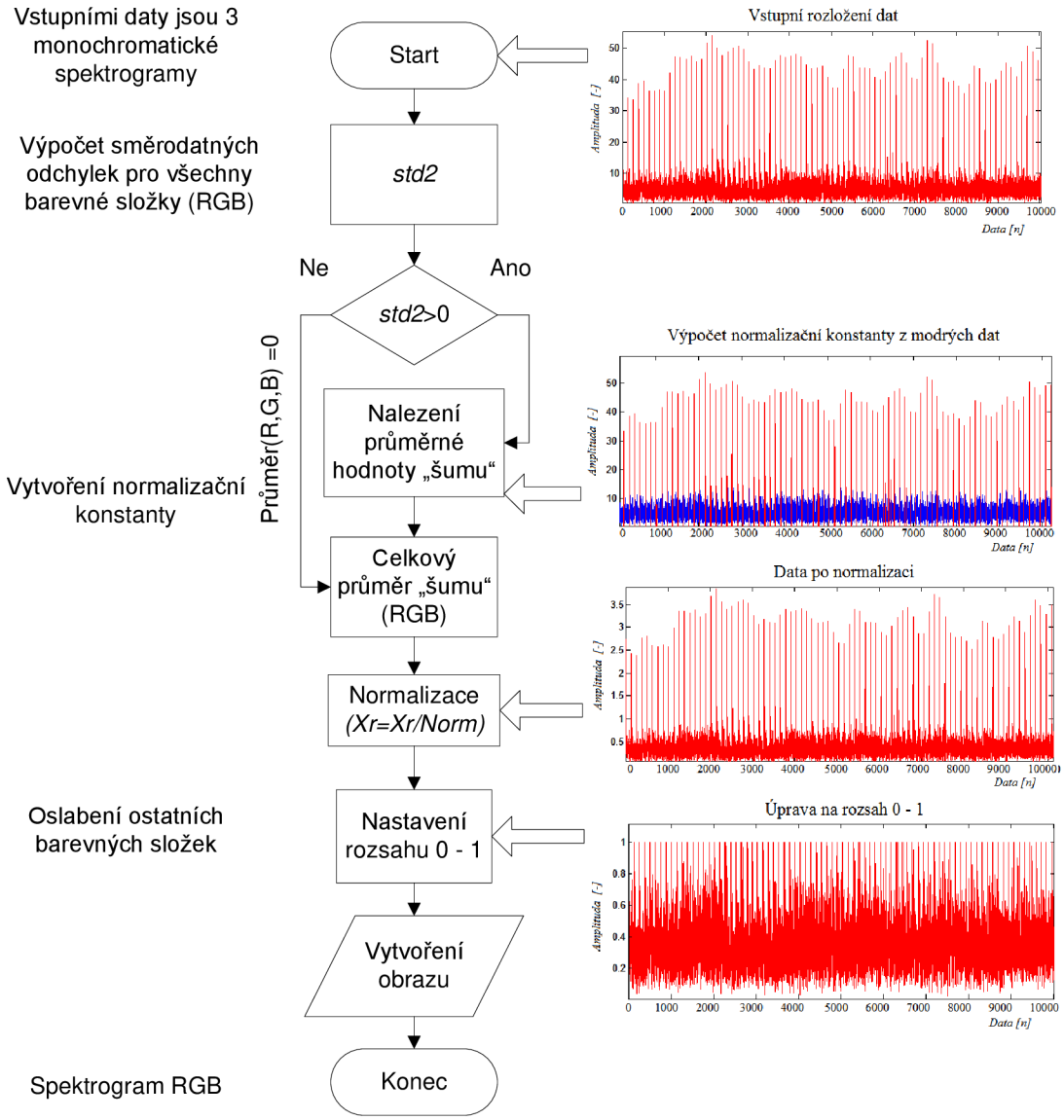
kterou podělím jednotlivé barevné složky, bude průměr středních hodnot šumu barevných složek. Tímto jsem docílil snížení rozptylu hodnot, ale rozsah ještě není upraven pro potřeby zobrazení. Vyhledám tedy hodnoty přesahující maximální povolenou hodnotu „1“ v každé barevné složce a nastavím ji na maximum a současně oslabím amplitudy ostatních složek na daném pixelu. Vývojový diagram algoritmu s ukázkou rozložení dat pro červenou složku spektrogramu je na Obr. 20.

Další způsob normalizace rovněž využívající statistického rozboru, ale je založený na odlišném způsobu výpočtu normalizační konstanty. Normalizační konstanta je zde vypočtena pomocí statistik maximálních hodnot, průměrů a směrodatných odchylek všech barevných vektorů. Výpočetní náročnost tohoto přístupu je vyšší a kontrast je mírně nižší, proto jsem tuto variantu normalizace zavrhl. Porovnání jednotlivých přístupů normalizace je provedeno na proteinu F56F11.4, III chromozomu *C. elegans*. Žlutou barvou jsou zvýrazněny lokace, kde se vyskytují exony (tříbázová perioda). Porovnání je zobrazeno na Obr. 19.



Obr. 19: a) Druhý typ normalizace, b) Druhý typ normalizace s využitím všech statistik, c) První typ normalizace dle rovnice 14.

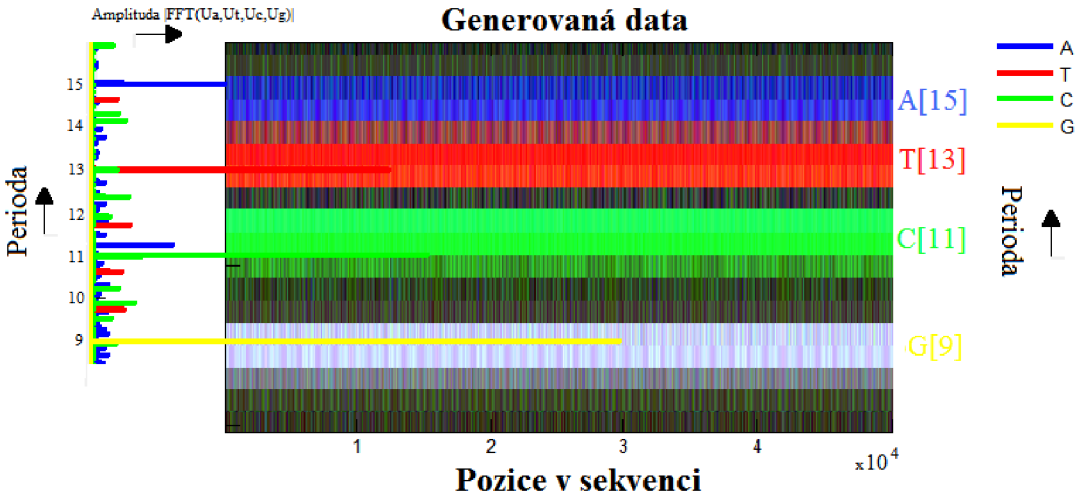
Jelikož všechny exony proteinu F56F11.4 nejsou pomocí spektrogramu dostatečně rozeznatelné (zejména první a poslední exon), nabízí se možnost implementace metod, které by provedly zvýraznění těchto oblastí. Jedním způsobem je prahování spektrogramu pomocí GUI *Tresholding*. Dalšími možnostmi je zobrazení vývoje tříbázového opakování v sekvenci pomocí posuvného okna, avšak tato metoda není zcela spolehlivá a pro kvalitní výsledky vyžaduje kvalitní filtraci. Nalezením exonů pomocí vývoje tříbázového opakování v sekvenci se zabývají například v článcích [8], [22] a [23].



Obr. 20: Vývojový diagram normalizace barevných vektorů.

4.2.1.3 *Vztah mezi FFT a STFT*

Obr. 21 uvádí vztah mezi frekvenčním spektrem získaným z binárních vektorů generované sekvence s periodou opakování A(15), T(13), C(11), G (9) a vytvořeným spektrogramem s délkou okna 200 bp, překrytím oken 200 bp a oknem typu Hann. Jelikož perioda opakování nukleotidů je zachována po celý rozsah dat (50 kbp), jsou přez celý spektrogram patrny vodorovné barevné čáry. Při porovnání se spektrem získaným pomocí GUI *PlotFFT* vidíme, že periody opakování odpovídají nejvyšším složkám amplitud frekvenčního spektra. V tomto případě bylo pro názornost použito na spektrogram přiblížení tak, aby byly zobrazeny pouze periody v rozmezí 6 – 20 bp. Pokud bychom ponechali celý spektrogram byly by patrné i ostatní harmonické složky na násobcích zvolených period. Z obrázku je patrné, že barevné mapování pro A(modrá), T(červená), C(zelená) odpovídá, ale u G(žlutá) se liší a je zde reprezentováno světlým odstínem modré a růžové. Jako legenda pro zapamatování barevného mapování je v programu použito dalšího grafu zobrazujícího zastoupení nukleotidových bází.



Obr. 21: Vztah mezi FFT s STFT.

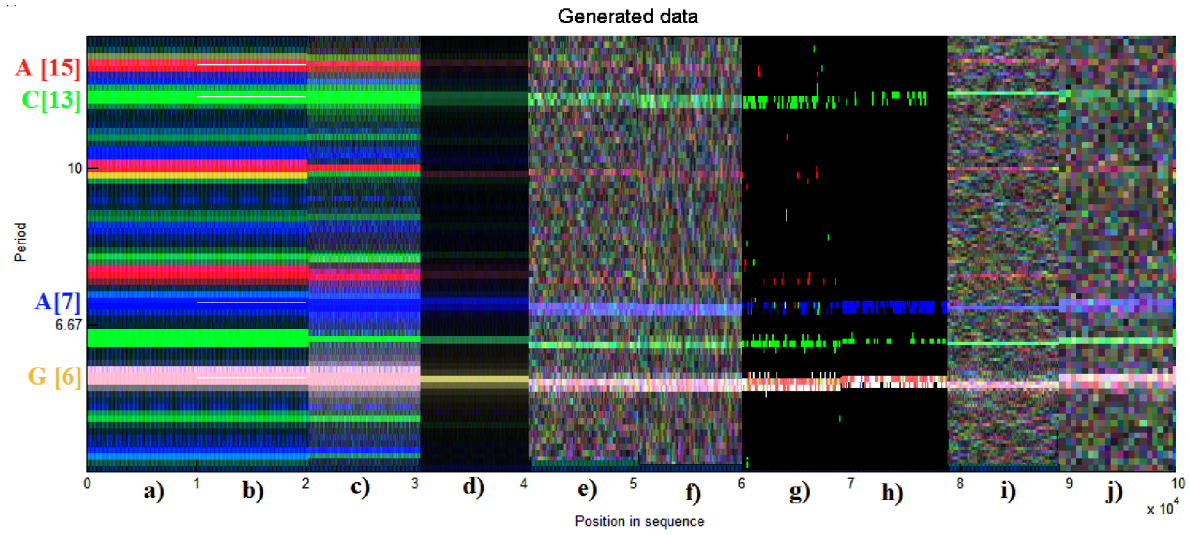
4.2.1.4 *Vlivy parametrů nastavení na výsledný obraz*

Schopnost získávat informace ze spektrogramů vyžaduje pochopení základních principů jeho vzniku a v případě spektrogramů z genetických dat i jistou dávku zkušenosti a informace

o tom, na co bychom se měli při vyhledávání zaměřit. Oblasti patrné pomocí spektrogramů jsou popsány v kapitole 2.3. Při vyhledávání známého vzoru (například známého genu) by bylo vhodné nejdříve vytvořit spektrogram tohoto genu a poté jej na základě porovnávání vyhledávat ve zkoumané sekvenci.

Délka okna by měla být několikrát delší než perioda opakující se sekvence našeho zájmu a menší než oblast, kde se opakující se sekvence vyskytuje. Přesah určuje, kolik bázových prvků bude společných pro dvě po sobě jdoucí okna výpočtu STFT. Čím větší bude přesah, tím bude pozvolnější přechod mezi sousedními okny a větší rozlišení, což vede k lepšímu vizuálnímu dojmu.

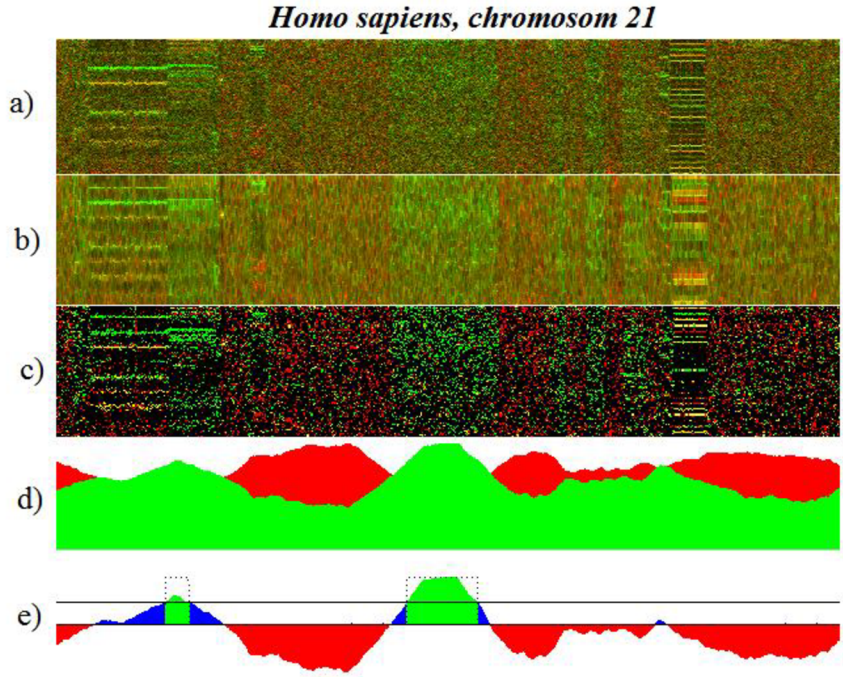
Na Obr. 22 jsou generovaná data s periodou nukleotidů A (7), T (15), C(13), G(6) s celkovým počtem 100 kbp při různých nastaveních vstupních parametrů. Tento složený spektrogram slouží k vizualizaci vlivu jednotlivých parametrů na výstupní zobrazení. Bloky *a)-h)* jsou vytvořeny s velikostí okna 500 bp a překrytím 450 bp. První dva bloky (*a,b*) jsou členěny pomocí Hannova okna a v druhém bloku jsou pomocí bílé čáry zvýrazněny periody 6, 7, 13 a 15. Blok *c* znázorňuje vliv výběru typu okna, zde je vybráno pravoúhlé okno. Blok *d* ukazuje vliv barevných korekcí, v tomto případě je provedeno převedení složek RGB na rozsah 0-1 nezávisle na sobě. Bloky *a)-d)* jsou vytvořeny z dat bez šumu, zatímco do ostatních bloků byl přidán šum s rovnoměrným rozložením. Bloky *e)* a *f)* se liší použitým typem okna (pravoúhlé a Hannovo). Bloky *g)* a h) jsou vytvořeny pomocí prahování, kde v bloku *g)* byl práh nižší než v bloku *h)*. Blok *i)* poukazuje na vliv velikosti okna, zde je vybráno okno s velikostí 1 kbp a překrytím 0,95 kbp. Můžeme si zde povšimnout změny frekvenčního rozlišení (zúžení period opakování). Poslední blok *j)* naznačuje strmost přechodu mezi jednotlivými okny při nepoužití přesahu, velikost okna je zde rovna 0,5 kbp s žádným přesahem.



Obr. 22: Složené spektrogramy.

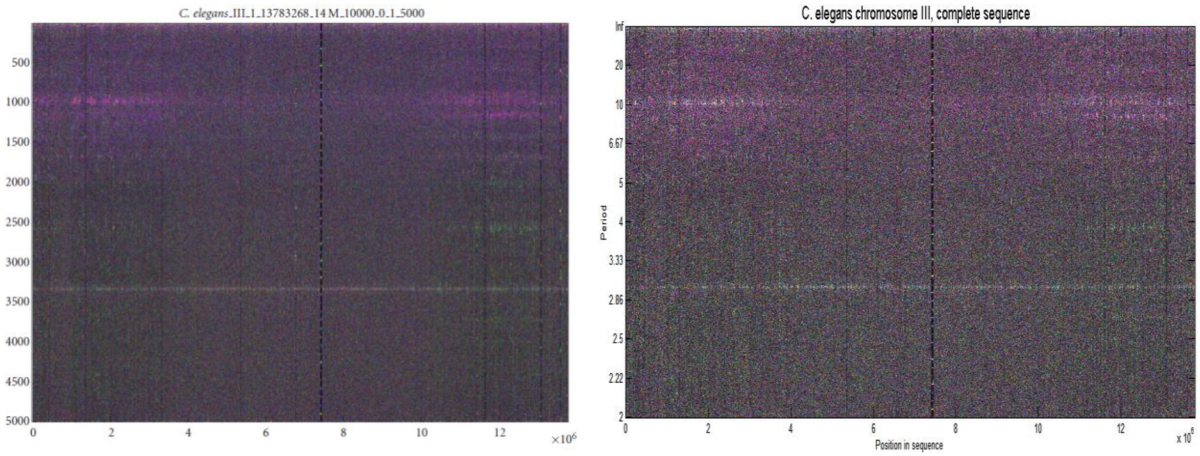
4.3 POROVNÁNÍ VÝSLEDKŮ

V této kapitole provedu stručné porovnání výsledků vytvořeného programu s literaturou [3] a [5], která se zaobírá využitím barevných spektrogramů na biologických datech. Literatura [3] se navíc zabývá automatickou extrakcí a klasifikací dat ze spektrogramu. Uvedené algoritmy pracující s vytvořeným spektrogramem jsem testoval, ale vzhledem k tomu, že se zde vyskytuje příliš mnoho proměnných faktorů, jsem je neimplementoval. Mezi takové faktory můžeme řadit velikost okna, typ použitého okna, volba barevného mapování, způsob normalizace, volba strukturního elementu (slouží pro filtraci pomocí morfologických operací), velikost prahu, metoda detekce prahování (Sobel, Canny, Prewitt atd.). Vytvořený algoritmus nebyl robustní a byl účinný jen na testovaný okruh dat. Algoritmus je obsažen na přiloženém CD jako skript *Xhist.m.* Ukázkou automatického vyhodnocování je Obr. 23, kde jsou nalezeny CpG oblasti. Automatické vyhodnocování dat ze spektrogramu je tedy složitou problematikou, která není vlastním zadáním práce a sama o sobě by svým rozsahem vydala na nové zadání práce.



Obr. 23: a) Vstupní spektrogram, b) morfologické operace, c) prahování a umocnění, d) histogramy četností ve vertikální ose, e) nalezené oblasti.

Porovnání získaných výsledů s literaturou [5] je zobrazeno na Obr. 24. Jedná se o kompletní třetí chromozom *C. elegans* (rozsah 1 – 13,78 Mbp). Tento spektrogram byl pořízen s nastavením velikosti okna 5000 bp, překrytím 0, výpočetním algoritmem *SpectDNA_II_all* a Hannovým typem okna. Dosažené výsledky jsou velmi podobné na obou spektrogramech patrna 3 - bázová perioda, satelit v oblasti 7.4 Mbp s délou 50.9 kbp. Velmi patrné jsou i 10 - bázové periody tvořené zejména nukleotidy A a T a 3.6 - bázová perioda tvořená nukleotidy G v oblasti 12 Mbp. V chromozomu se vyskytuje ještě minimálně 8 minisatelitů, které jsou ovšem o trochu lépe patrny na spektrogramu převzatém z čerpané literatury. Určení periody opakování je snazší z vytvořeného spektrogramu díky frekvenční ose označené v periodách. Následná analýza, kterou můžeme provést na vytvořeném spektrogramu, je velmi užitečným nástrojem. Můžeme například vybrat z dat oblast zájmu a tu zobrazit v lepším rozlišení, zobrazit si rozložení nukleotidů ve zvolené oblasti, procházení spektrogramu po stanoveném rozsahu dat a jiné. Vylepšení zobrazované scény by mohlo být provedeno pomocí filtrace výsledného obrazu nebo například jinými barevnými vektory.

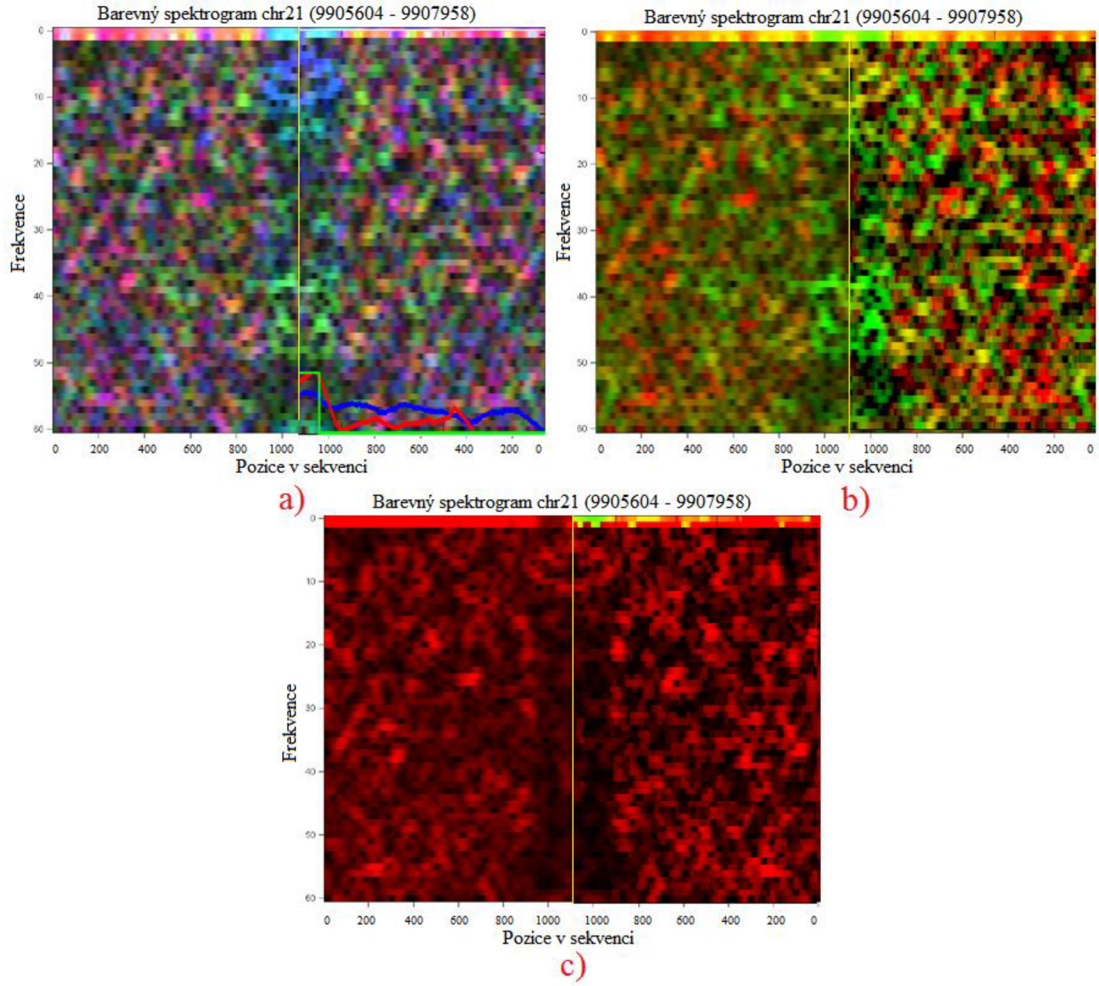


a) *b)*

Obr. 24: a) Spektrogram z literatury [5], b) vytvořený spektrogram v programu *DNAspect.*

Porovnání získaných výsledů s literaturou [3] je zobrazeno na Obr. 25. Porovnání je vytvořeno se stejnými parametry (velikost (120 bp) a typ okna (Hann), překrytí oken (119 bp), barevné mapování) pomocí zrcadlení spektrogramu v rozsahu 0 – 1,1 kbp. První polovina spektrogramu odpovídá čerpané literatuře a druhá je vytvořena programem *DNAspect* při použití funkce *SpectDNA_II.* Kompletní porovnání bez zrcadlení je uvedeno jako příloha 1.

První spektrogram (Obr. 25 a) využívá mapování pomocí čtyř barev (*ATCG*) a je téměř identický. V zrcadlené části je vidět zelenou křivkou nalezená CpG oblast. Druhý spektrogram (Obr. 25 b) je proveden mapováním do dvou barev (*AT_CG*) a v zrcadlené oblasti bylo využito nastavení prahu. Třetí spektrogram (Obr. 25 c) uvádí mapování jedním barevným vektorem (*AT*), kde je opět v zrcadlené oblasti využito prahování (v případě nevyužití prahování by byly spektrogramy identické jak tomu je na Obr. 25 a).



Obr. 25: DNA spektrogram CpG oblasti chromozomu 21 H. sapiens sapiens s různými typy barevného mapování, a) mapování ATCG, b) mapování AT_CG, c) mapování AT.[3]

4.4 GRAFICKÉ UŽIVATELSKÉ ROZHRANÍ

Při vývoji aplikace v MATLABU se postupuje od úrovně m-file k funkcím a poslední fází je provázání funkcí s grafickými objekty. Grafické uživatelské rozhraní (dále jen GUI) slouží pro přehlednější a snazší zadávání vstupních dat a vizualizaci výstupních dat. Zároveň eliminuje neadekvátní nastavení vstupních parametrů a umožňuje spustit jen takové funkce aplikace, které jsou v dané chvíli k dispozici.

Navržené GUI se skládá z hlavní obrazovky a několika dalších provázaných modulů, které se spouštějí ve vlastních oknech. Vzhledem k tomu, že hlavní výstupní veličinou

aplikace je obrazová informace, mnohdy ve vysokém rozlišení, je žádoucí, aby zabírala co největší plochu. Z tohoto důvodu jsou vytvořeny další provázané moduly, které plní následující funkce:

- Generování dat – název okna *Datageneration*
- Analýza dat ve frek. Oblasti - název okna *PlotFFT*
- Analýza spektrogramu - název okna *Tresholding*
- Zobrazení dat - název okna *Show_sequence*

GUI je ošetřeno, aby uživatel nemohl zadávat nesmyslné hodnoty, písmena, záporná čísla apod. Při zadání špatné hodnoty je uživatel upozorněn varovnou zprávou a chybné zadání je červeně zvýrazněno. Během delších výpočtů je vhodné dát uživateli znát, že program pracuje. Proto je vytvořen odhad délky výpočtu zobrazující informace o tom, kolik procent z potřebného času již uplynulo a kolik pravděpodobně zbývá.

Základní GUI *DNASpect* slouží především pro načítání vstupních dat, nastavování parametrů aplikace a výpočtu, zobrazení výsledků a spouštění dalších modulů. Jako jediné obsahuje pro ovládání textové rolovací menu s možnostmi volby pomocí klávesových zkratek. Struktura rolovacího menu se skládá z těchto položek:

- File – *slouží pro ukládání, načítání a převod dat*
  - Open File (*.txt & *.fasta)
  - Convert File
  - Seve Selected Sequence
  - Save Spectrogram(screen)
  - Save Spectrogram(graph)
  - Exit
- Settings – *slouží pro nastavení spektrogramu a chování programu*
  - Type of Window -> Rectangular, Hann, Blackman…
  - Algorithm – *výběr typu výpočtu a obarvení STFT*
    - Colouring -> STFT
    - STFT -> Colouring
  - Color mapping – *druh barevného mapování*
    - ATCG

        - AT_CG
        - AT
        - CG
    - Color correction – *způsob úpravy barev*
    - Programme interaction – *Chování při ukončování GUI's*
- View – *slouží pro řízení zobrazení a spouštění GUI's*
    - Histograms – *spustí modul pro prahování spektrogramu*
    - Data Generation – *spustí modul pro generování dat*
    - FFT - *spustí modul pro FFT binárních vektorů*
    - Char data - *spustí modul pro zobrazení nukleotidů*
        - Selected sequence – *pro vybranou část dat*
        - Whole sequence – *pro všechna načtená data*
    - Data selection – *nástroj pro výběr dat*
    - Base distribution – *zobrazení rozložení nukleotidů*
    - Information about file – *informace o načteném souboru*
- Help
    - Programme help – *spustí nápovědu programu*
    - About programme – *informace o původu programu*

Dalšími ovládacími prvky programu jsou ikony v liště snadného spouštění. Tato lišta slouží ke snadnému a rychlému ovládání základních funkcí programu. Následující seznam uvádí grafickou reprezentaci ikon panelu snadného spouštění a jejich význam:

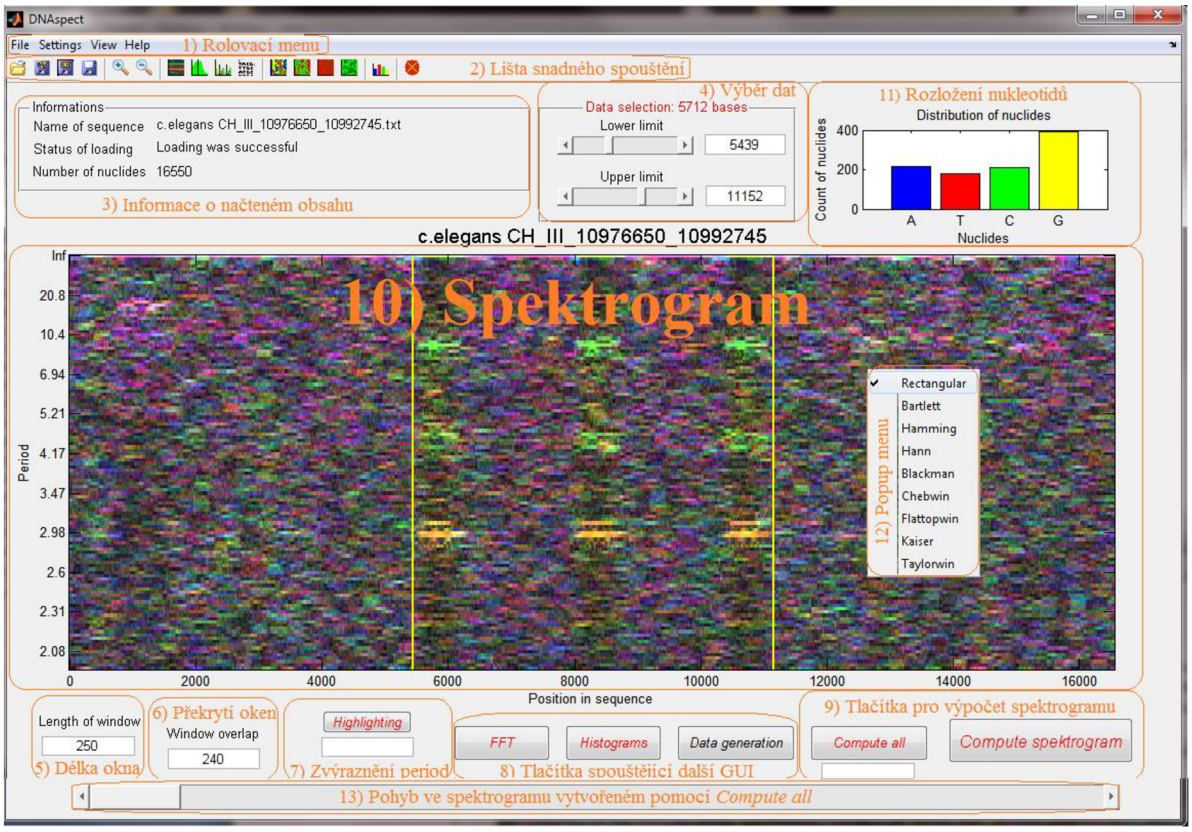
-  Otevření souboru
-  Uložení obrazovky do grafického souboru
-  Uložení spektrogramu s osami do grafického souboru
-  Uložení vybraného rozsahu dat do textového souboru
-  Zoom in/out - přiblížení/oddálení v grafech
-  Otevření modulu pro generování dat
-  Otevření modulu pro prahování spektrogramu
-  Otevření modulu pro zobrazení FFT binárních vektorů

-  Otevření modulu pro zobrazení nukleotidů
-  Nastavení barevného mapování ATCG
-  Nastavení barevného mapování AT_CG
-  Nastavení barevného mapování AT
-  Nastavení barevného mapování CG
-  Výpočet rozložení nukleotidů ve zvolené oblasti dat
-  Konec programu (Ctrl+Q)

Rozložení hlavního GUI *DNAspect* lze vidět na Obr. 26. Hlavní obrazovka je rozdělena do 13 oblastí dle své funkce. Následující číslovaný seznam uvádí stručně význam jednotlivých oblastí hlavního GUI:

1) Textové rolovací menu – slouží pro ovládání programu
2) Lišta snadného spouštění – slouží pro rychlé ovládání programu
3) Informace o načtených datech- název dat, stav načtení a počet nukleotidů
4) Oblast sloužící pro výběr dat podrobovaných analýze
5) Délka okna – nastavovací prvek
6) Překrytí oken – nastavovací prvek
7) Zvýraznění period – lze zapsat i více period ve formátu např. 3 5 11.5 7 nebo 3,5,11.5,7
8) Spouštění dalších GUI pomocí tlačítek
9) Tlačítka pro výpočet spektrogramu - *Compute all* a *Compute spectrogram*
10) Oblast pro zobrazení spektrogramu
11) Graf rozložení nukleotidů
12) Popupmenu sloužící pro rychlý výběr typu okna – aktivace p. tlačítkem myši
13) Prvek sloužící pro pohyb ve spektrogramu vytvořeném pomocí *Compute all*

Popis ostatních modulů GUI, nutných vlastností vstupních a výstupních dat spolu s vlivy a způsoby nastavení aplikace pro tvorbu spektrogramů, jsou uvedeny v samostatné nápovědě programu. Tuto nápovědu lze vyvolat přímo z programu v záložce textového rolovacího menu *Help -> Programme help* nebo přímo spustit z přiloženého CD (*Help.doc*).

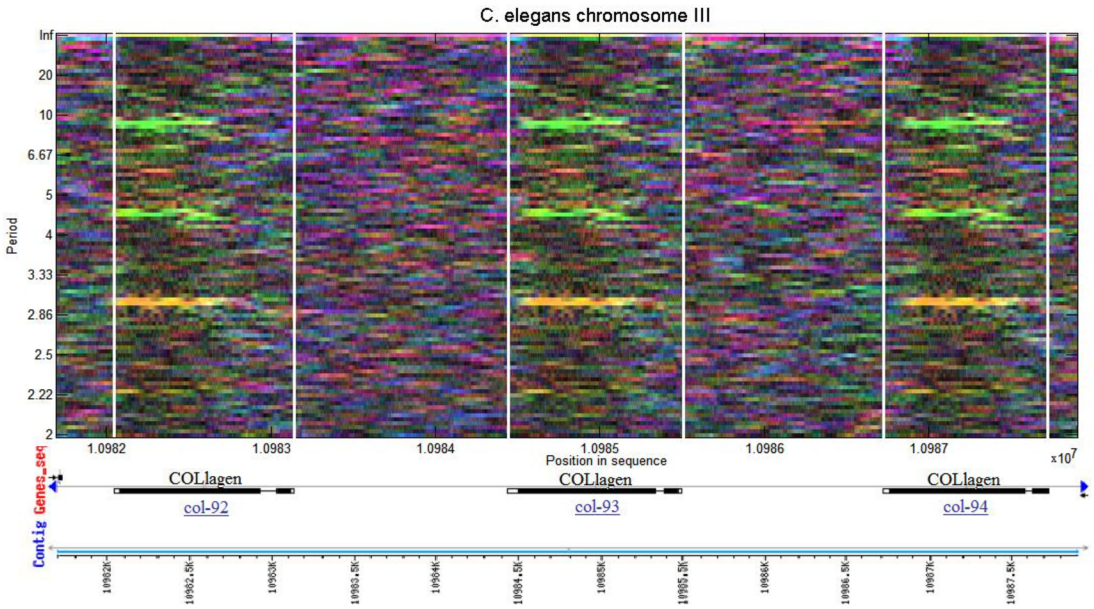


Obr. 26: Hlavní GUI *DNAspect* a jeho rozložení.

5. ANALÝZA VYBRANÝCH ÚSEKŮ DNA C. ELEGANS

V této kapitole jsou uvedeny možnosti využití vytvořeného programu a porovnání informací z NCBI se získanými spektrogramy. Vlastnosti analyzovaného modelového organizmu C. elegans jsou uvedeny v kapitole 2.7.

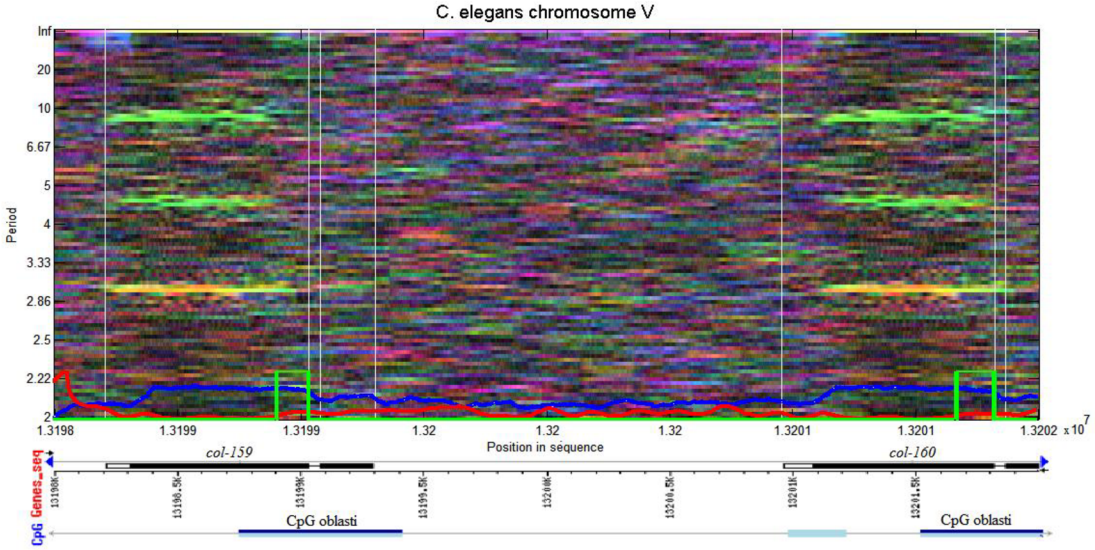
První využití a porovnání již bylo uvedeno na proteinu F56F11.4, III chromozomu *C. elegans* na Obr. 19, kde jsou patrny tříbázové periody v oblasti exonů. V praxi je nepravděpodobné, že bychom našli protein na základě dat ze spektrogramu, jelikož průměrná délka genu u *C. elegans* činí 1,91 kbp, což je pro zobrazení spektrogramu v kvalitním rozlišení nedostatečné. Příkladem mohou být geny col-92, col-93 , col-94 na chromozomu III v oblasti 10981,7 kbp – 10987,9 kbp zobrazené s daty získanými z NCBI na Obr. 27. Tyto geny patří do rodiny COLlagen a podílejí se na strukturálním složení pokožky. Geny lze pomocí spektrogramu snadno rozeznat díky tříbázové periodě. Dalšími periodami, které se v genech vyskytují, jsou 4,5 a 9 tvořené cytosinem. Délka zobrazených genů se pohybuje okolo 900 bp s převážným zastoupením G, C, T a jen malým množstvím A. Na Obr. 27 jsou vztahy mezi geny nalezenými v NCBI a ve spektrogramu vyznačeny bílou svislou čarou.



Obr. 27: Chromozom III, *C. elegans*, geny *col - 92, col - 93, col – 94.*

Při vytváření spektrogramů si můžeme povšimnout, že některé vzory jsou si velice podobné. Většinou se tak děje u rodin genů. Například na předešlém spektrogramu (Obr. 27)

jsou geny z rodiny COLlagen, které se ovšem vyskytují i na jiných chromozomech. Na Obr. 28 je znázorněn spektrogram V chromozomu *C. elegans* v rozmezí od 13198 kbp do 13202 kbp, na kterém se vyskytují geny *col-159* a *col-160*. Tyto geny se rovněž podílejí na strukturálním složení pokožky a mají stejný vzor (stejné periody opakování bp - 3; 4,5; 9). Kódující oblasti genů jsou opět vyznačeny bílými vodorovnými čarami. Pro porovnání bylo do spektrogramu přidáno vyhledávání CpG oblastí s nastavením velikosti okna 100 bp, minimální délkou oblasti 100 bp, minimálním obsahem GC 0,55 (modrá křivka) a minimálním obsahem CpG 0,6 (červená křivka). Můžeme si povšimnout, že nalezené oblasti se překrývají s oblastmi uvedenými v NCBI a jsou lokalizovány na konci prvních exonů genů.

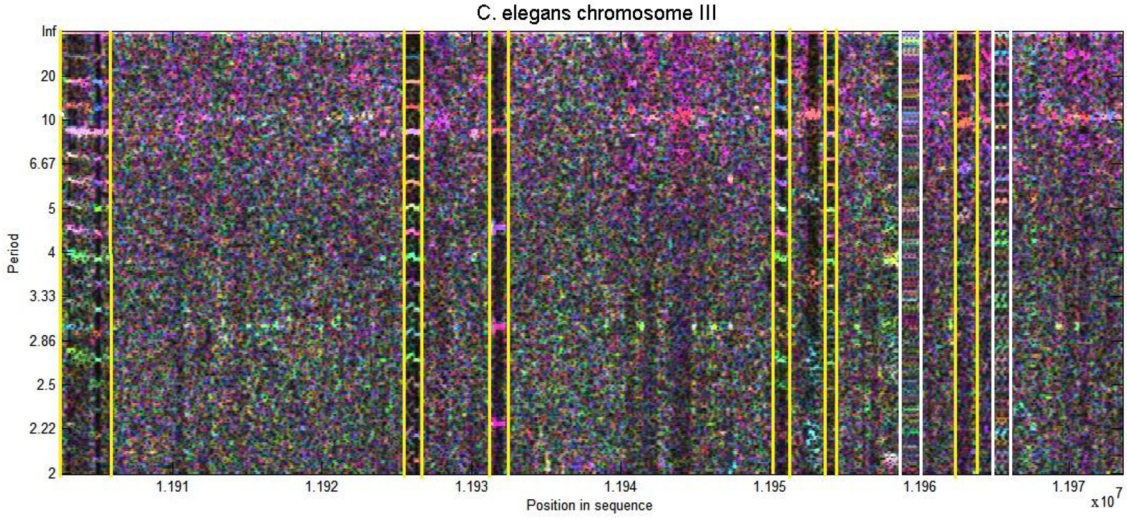


Obr. 28: Chromozom V, *C. elegans*, geny *col – 159* a *col – 160.*

Jak již bylo zmíněno tříbázová perioda je charakteristická pro oblasti kódující protein. Další hojně se vyskytující periodou je 10.5 bázová perioda, která má vztah ke struktuře stočení chromatinu. Pokud jsou v periodě 10.5 zakódovány sekvence „*AA*“, „*TT*“ nebo „*TA*“ je DNA místně posílena. [27]

Na Obr. 29 je znázorněn spektrogram výběru ze III chromozomu *C. elegans* o délce 71,1 kbp (v rozmezí od 11902 kbp do 11973 kbp). Ze spektrogramu jsou na první pohled patrny minisatelity označeny svislými žlutými čarami a tandemové repetice mikrosatelitů označené bílými svislými čarami. Zejména kolem středu spektrogramu (11945 kbp) se střídají

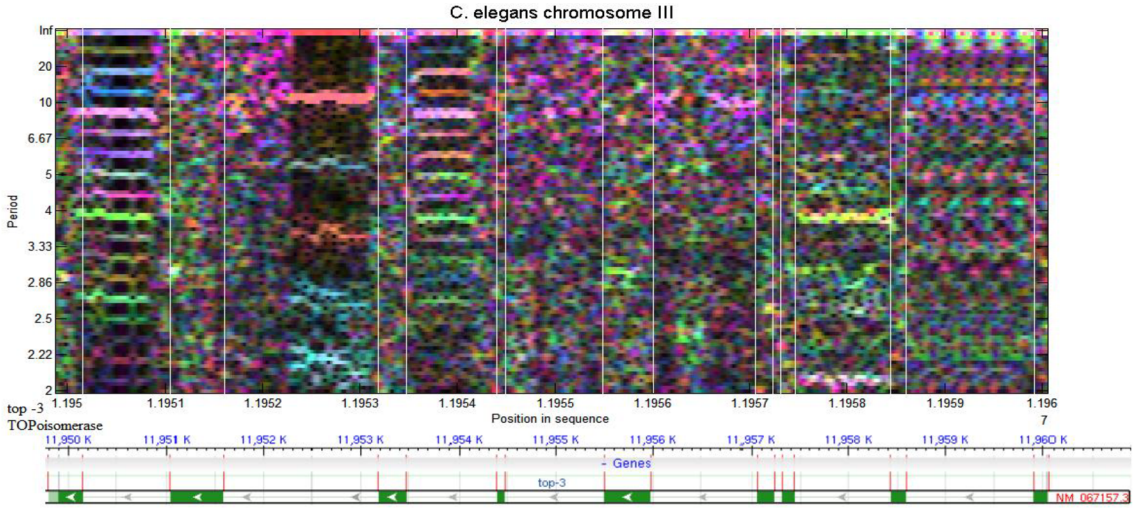
tříbázové periody s desetibázovými. Rovněž je zde patrno rozložení opakujících se nukleotidů (*CG* vyšší frekvence, *AT* nižší frekvence).



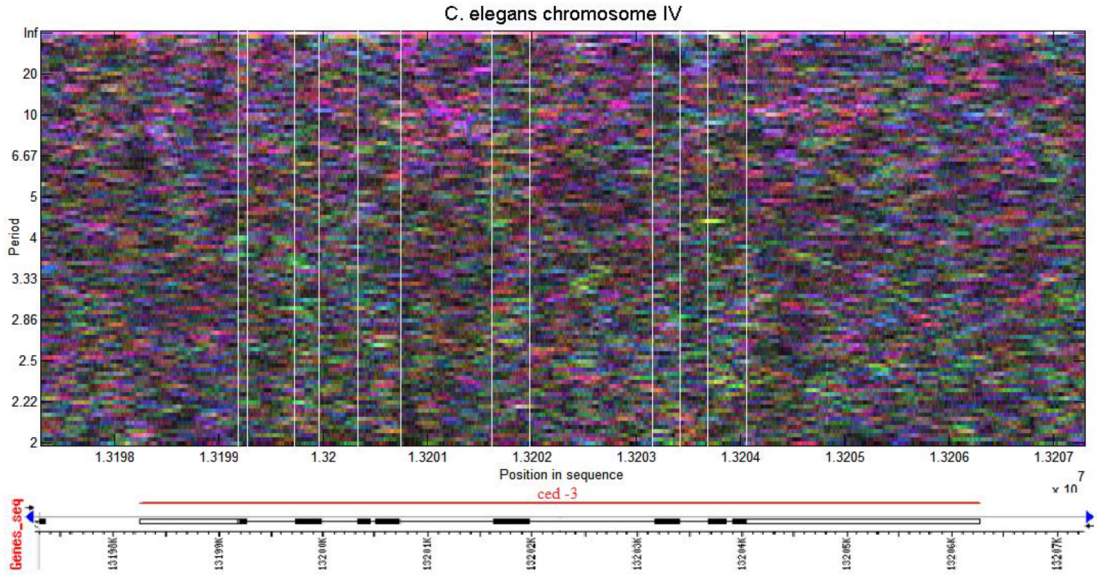
Obr. 29: Chromozom III, *C. elegans*.

Ze spektrogramu na Obr. 29 jsem vytvořil výřez zobrazující gen označený názvem *top- 3*. Zobrazená oblast na Obr. 30 má délku 10,16 kbp a nachází se na pozici 11949,8 kbp – 11960 kbp. Bílými svislými čarami je znázorněn vztah mezi daty (kódujícími oblastmi - exony) získanými z NCBI a vytvořeným spektrogramem s délkou okna 200 bp, překrytím oken 185 bp a Hammingovým typem okna. Můžeme si zde povšimnout, že kódující oblasti se nenachází na pozicích tandemových repetic tvořených minisatelity a mikrosatelity. Tento gen je společný pro řadu organizmů jako je například člověk, šimpanz, pes a uplatňuje se v procesu meiózy a mitózy.

Pomocí spektrogramu nelze s jistotou říci, že se na dané pozici vyskytuje exon nebo intron genu. Tříbázová perioda, která je ve spektrogramech dobře patrna, je pouze jedním z mnoha prediktorů genu. Například na Obr. 31 je vyobrazen spektrogram genu *ced-3* kódující protein CED-3 patřící do skupiny cysteinových proteáz. Tato skupina je důležitá pro apoptósu neboli programovatelnou buněčnou smrt a byla objevena právě u *C. elagans*. Ve spektrogramu si můžeme povšimnout v oblastech kódujících protein zvýšeného obsahu *C* a *G* na vyšších frekvencích a malého výskytu *A* na nízkých frekvencích při porovnání s přilehlými oblastmi.



Obr. 30: Chromozom III, *C. elegans* a gen *top-3.*

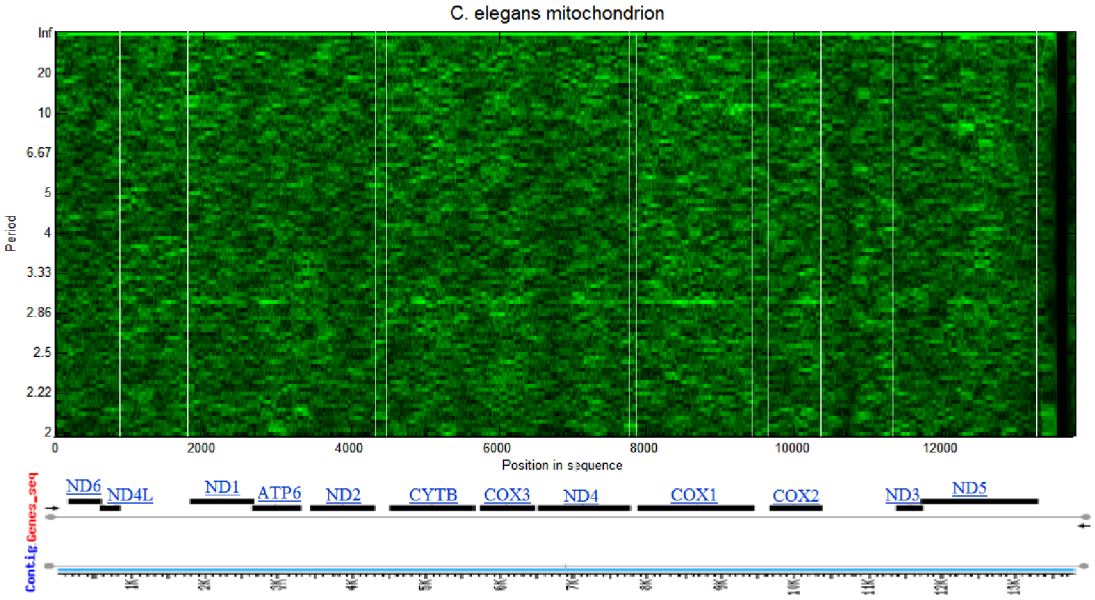


Obr. 31: Chromozom IV (13197,3 kbp – 13207,3 kbp), *C. elegans* a gen *ced-3.*

Mitochondriální DNA neboli mtDNA se nachází v mitochondriích a tvoří tak část mimojaderné genetické informace. Při přenosu této genetické informace dochází v drtivé většině k dědění po matce, hovoříme o tzv. maternální dědičnosti. Mitochondriální DNA se využívá pro různé genetické analýzy, jako jsou například migrace živočichů a fylogenetické stromy.

Při pohledu na spektrogram mtDNA na Obr. 32 je patrné, že data obsahují méně intronů, jak tomu u mtDNA bývá. Velmi patrná je zde tříbázová perioda, která je nejvíce

výrazná v oblastech kódujících geny. Bílou svislou čarou jsou v tomto případě vyznačeny oblasti nekódující geny. Spektrogram byl vytvořen pomocí barevného mapování CG zobrazujícího spektrální rozložení guaninu a cytosinu.



Obr. 32: Mitochondriální DNA *C. elegans.*

6. ZÁVĚR

V této diplomové práci na téma „Fourierova transformace a spektrogramy v analýze DNA sekvencí“ jsou v první části uvedeny základní poznatky z oblasti složení DNA spolu s popisem vzorů v ní se vyskytujících. Další část pojednává o způsobech převodů DNA sekvencí a o problematice spektrogramů. Vlastní část práce se skládá z popisu základních funkcí vytvořeného programu *DNAspect* v prostředí MATLAB a analýzy vybraných úseků modelového organizmu *Caenorhabditis elegans*.

Počáteční náplní práce bylo vytvoření aplikace pro zpracování DNA sekvencí do podoby spektrogramů a její doplnění o užitečné nástroje usnadňující budoucí analýzu. Komplikací při vytvoření spektrogramů z DNA sekvencí jsou velké paměťové nároky, z tohoto důvodu bylo zapotřebí stanovit omezení nastavení parametrů. Výpočetní náročnost algoritmů je přímo úměrná množství dat, proto je nutné uživatele informovat o stavu a předpokládané délce výpočtu, která se může pohybovat v řádu minut.

Tato práce nabízí nový pohled na zpracování DNA sekvencí různých organizmů na základě krátkodobé Fourierovy transformace (STFT). Výstupem algoritmů jsou barevné spektrogramy zobrazující prostorově frekvenční rozložení nukleotidů v sekvenci DNA.

Pomocí spektrogramů lze nalézt biologicky významné vzory, jež mají vztah například k lokalizaci kódujících oblastí proteinů (3 – bázové periody). Dále lze na základě spektrogramů s vhodným nastavením rozpoznat tandemové repetice tvořené satelity, minisatelity a mikrosatelity nebo oblasti bohaté na guanin a cytosin (tzv. CpG ostrovy). Jednotlivé biologicky významné vzory jsou popsány a zobrazeny v kapitole 2.3.2 a 5.

Poslední část zabývající se analýzou vybraných úseků DNA *Caenorhabditis elegans* uvádí vztah mezi obrazovou informací tvořenou spektrogramy a daty získanými z databáze NCBI. Analýzou bylo prokázáno, že podobné vzory na různých chromozomech patří do stejné rodiny genů a plní tak podobnou funkci. Ve většině uvedených případů je patrný vztah mezi spektrogramem a daty z NCBI. Získávání informací ze spektrogramu však není jednoduchou záležitostí a je zapotřebí se tomu naučit.

Další pokračování práce by se mohlo ubírat ve smyslu provázání spektrogramů s jinou databází (např. NCBI) nebo vytvořením algoritmů, které by na základě obrazové informace automaticky predikovaly pozice kódujících oblastí. Popřípadě lze využít biologicky

významný obrazový vzor při vyhledávání podobnosti na jiných chromozomech nebo u jiných živočišných druhů.

7. POUŽITÁ LITERATURA

[1] JAN J., *Číslicová filtrace, analýza a restaurace signálu,* nakladatelství VUTIUM, Brno 2002, ISBN – 80-214-2911-9.

[2] HONZÍKOVÁ N., *Biologie člověka,* skriptum VUT FEKT, Brno 2003

[3] DIMITROVA N., *Analysis and Visualization of DNA spectrograms,* [on-line] [cit. 2010-20-3] Dostupné na internetu < http://portal.acm.org >

[4] WIKIPEDIE, [on-line], [cit. 2010-20-3] Dostupné na internetu: < http://cs.wikipedia.org/wiki/DNA >

[5] SUSSILO D., *Spectrogram analysis of genomes,* Department of Electrical Engineering, Columbia University, NY 10027, USA 2004

[6] *The science creative quaterly*, [on-line] [cit. 2010-20-3] Dostupné na internetu: < http://www.scq.ubc.ca/wp-content/dna.gif >

[7] *Princip sekvencování DNA*, Přírodovědecká fakulta, Masarykovy univerzity, ústav Biochemie, přednášky, [on-line], [cit. 2010-20-3] Dostupné na internetu : < http://orion.sci.muni.cz/kgmb/bioinformat/princip_seq.pdf >

[8] ANASTASSIOU D., *Frequency-domain analysis of biomolecular sequences*, Department of Electical Engeneering, Columbia universty, 2000

[9] CRISTEA P. D., Large *scale features in DNA genomic sinals*, Bio-Medical Engineering Center, University of Bucharest, 2002

[10] GARDINER-GARDEN, FROMMER M., *CpG islands in vertebrate genomes*, J. Mol. Biol. 1987,

[11] TAKAI D., JONES A., *Comprehentive Analysis of CpG Islands in Human Chromosome 21 and 22*, PNAS, Vol. 99, No. 6, březen 2002

[12] EPEL E. S., *In press. Accelerated telomere shortening in response to life stress*. Proceedings of the National Academy of Sciences., [on-line], [cit. 2011-8-3], Science News, Vol. 166, No. 23, Dec. 4, 2004, p. 355, Dostupné na internetu: <http://www.pnas.org/cgi/doi/10.1073/pnas.0407162101 >

[13] ŠEDA O., *Genetické haraburdí-repetitivní DNA,* Ústav biologie a lékařské genetiky 1. LF UK a VFN, [on-line], [cit. 2011-8-3], Dostupné na internetu: < http://biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm >

[14] IPSER J., *Genetika*, skriptum, Univerzita Jana Evangelisty Purkyně – Přírodovědecká fakulta – katedra biologie, Ústí nad Labem 2006, [on-line], [cit. 2011-8-3], Dostupné na internetu:
< http://biology.ujep.cz/vyuka/file.php/1/opory/Genetika.pdf >

[15] VOMELA J., *Zdravotní péče*, přednášky, 2011

[16] ARDUENGO M., *Encyklopedia of Genetics Revised Edition*, Pacific Union College – Depertment of Biology, Salem Press, Inc. ISBN 1-58765-151-3, r. 2004, p. 516

[17] *The Evolution of Self-Fertile Hermaphroditism: The Fog Is Clearing*, Published: December 28, 2004 , [on-line], [cit. 2011-8-3], Dostupné na internetu: < http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0030030 >

[18] The Goldstein Lab , The University of North Carolina at Chapel Hill - Biology Department, [on-line], [cit. 2011-8-3], Dostupné na internetu: < http://www.bio.unc.edu >

[19] *International Genome Team Deciphers Genetic Instructions for a Complete Animal,* Science 282: 2012-2021, 1998, Last Updated October 14, 2010, [on-line], [cit. 2011-8-3], Dostupné na internetu: < http://www.genome.gov/ >

[20] PAČES V., *Genomika – věda pro 21. století*, Ústav molekulární genetiky AVČR a VŠCHT Praha, r. 2000, [on-line], [cit. 2011-8-3], Dostupné na internetu: < http://www.img.cas.cz/paces/Genomika_2000.htm>

[21] VÁCHA M., Od DNA k evoluční psychologii, Masarykova univerzita v Brně - Lékařská fakulta, Ústav lékařské etiky, [on-line], [cit. 2011-8-3], Dostupné na internetu: < http://is.muni.cz/th/98186/lf_d/?lang=cs >

[22] WANG L., Localizing triplet periodicity in DNA and cDNA sequences, Bioinformatics 2010, , [on-line], [cit. 2011-8-3], Dostupné na internetu: <http://www.biomedcentral.com/1471-2105/11/550 >

[23] RUSHDI A., *The filtered spectral station measure*, department of Electrical and Computer Engineering, University of California, Davis, [on-line], [cit. 2011-16-3], Dostupné na internetu:
< http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4176897 >

[24] Manolio T., *Measurement of Genetic Exposure*, Office of Population Genomics, Northwestern University in Chicago, National Human Genome Research Institute, [on-line], [cit. 2011-16-3], Dostupné na internetu:
< http://www.genome.gov/27026645 >

[25] *Whitehead Institute for Biomedical Research*, [on-line], [cit. 2011-16-3], Dostupné na internetu: < *http://www.wi.mit.edu/* >

[26] KOSTROUCHOVÁ M., *Využití modelových organismů pro studium lidských onemocnění (C. elegant)*, [on-line], [cit. 2011-16-3], Dostupné na internetu:
< http://bioprojekty.lf1.cuni.cz/ >

[27] LACHIRI Z., *3D Spectrum Analysis of DNA Sequence: Application to Caenorhabditis elegant Genome*, [on-line], [cit. 2011-15-4], Dostupné na internetu: < http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4375661&tag=1 >

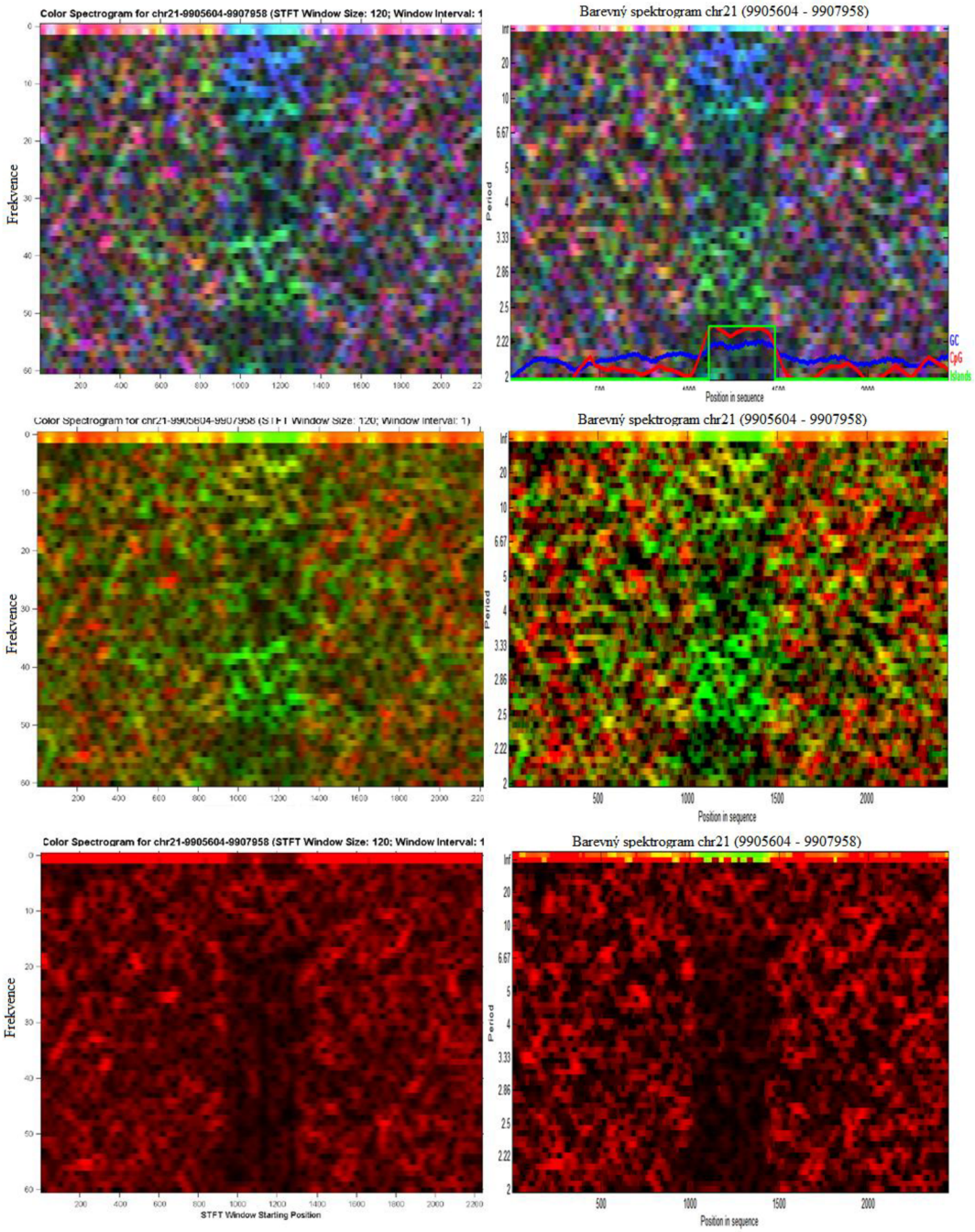
8. SEZNAM POUŽITÝCH ZKRATEK

A	Adenin
bp	Bázové páry (base pair)
C	Cytosin
CpG	Cytosin následovaný Guaninem
DFT	Diskrétní Fourierova transformace (Discrete Fourier Transform)
DNA	Deoxyribonukleová kyselina
FFT	Rychlá Fourierova transformace (Fast Fourier Transform)
G	Guanin
GUI	Grafické uživatelské rozhraní (Graphical user interface)
mRNA	Informační / mediátorová RNA (messenger RNA)
NCBI	Biologická databanka (National Center for Biotechnology Information)
rRNA	Ribozomální RNA (ribosomal RNA)
RGB	3 barevné složky pro vytvoření barevného obrazu (Red, Green, Blue)
RNA	Ribonukleová kyselina
STFT	Krátkodobá Fourierova transformace (Short Time Fourier Transform)
tRNA	Transferová RNA (transfer RNA)
T	Thymin
UTR	Oblast nekódující protein, sloužící k regulaci překladu (Untranslated region)
VNTR	Variabilní množství tandemových repetic (Variable number of tandem repeats)

9. SEZNAM ODBORNÝCH POJMŮ

Alela	Konkrétní forma genu
Centromera	Oblast uprostřed chromozomu, kde se dotýkají obě chromatidy
Exon	Oblast DNA, podle níž se v procesu translace tvoří bílkovina
Eukarota	Všechny jednobuněčné a mnohobuněčné organismy kromě bakterii a archeí
Gen	Úsek DNA se specifickou funkcí
Genom	Veškerá genetická informace uložená v DNA daného organismu
Introny	Oblast DNA, jež se nepřekládá do proteinu
Meióza	Buněčné dělení, během kterého dochází k produkci buněk se zredukovaným počtem chromozómů
Mitóza	Buněčné dělení, jehož úkolem je zajistit rovnoměrné předání nezredukované genetické informace dceřiným buňkám
Nukleosid	Pentosa + báze
Nukleotid	Pentosa + báze + kyselina fosforečná
Prokaryota	Evolučně velmi staré organismy (bakterie a archea)

10. PŘÍLOHA



Porovnání barevných spektrogramů s literaturou [3], popis jednotlivých spektrogramů je uveden v kapitole 4.3

11. OBSAH PŘILOŽENÉHO CD

1) Textová část práce ve formátu *.pdf
2) Vytvořený program *DNASpect* v prostředí MATLAB
3) Návod k programu *DNASpect (help.doc)*
4) Genom *Caenorhabditis elegans*
5) Vědecké články citované v práci
6) Grafické objekty práce spolu se spektrogramy celých chromozomů *Caenorhabditis elegans*