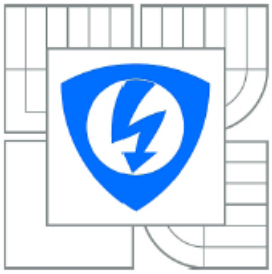
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ KÓDUJÍCÍCH ÚSEKŮ POMOCÍ ANALÝZY DNA SPEKTROGRAMU

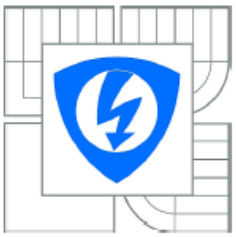
IDENTIFICATION OF CODING REGIONS USING DNA SPECTROGRAM ANALYSIS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — BARBORA SPÍCHALOVÁ
AUTHOR

VEDOUCÍ PRÁCE — Ing. VLADIMÍRA KUBICOVÁ
SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka:	Barbora Spíchalová	***ID:***	136691
Ročník:	3	***Akademický rok:***	2012/2013

NÁZEV TÉMATU:

Vyhledávání kódujících úseků pomocí analýzy DNA spektrogramu

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši o metodách konstrukce spektrogramů z DNA sekvencí a o vzorech detekovatelných ze spektrogramů. Vysvětlete princip vyhledávání kódujících úseků z DNA spektrogramu. 2) Navrhněte metodu, která bude vyhledávat kódující úseky pomocí číslicového zpracování spektrogramu. 3) Navrženou metodu realizujte v programovém prostředí Matlab. Funkčnost metody ověřte na anotovaných sekvencích v NCBI a výsledky detekce porovnejte s pozicemi kódujících úseků uvedenými v této databázi. 4) Proveďte diskusi získaných výsledků a zhodnoťte účinnost a využitelnost řešení.

DOPORUČENÁ LITERATURA:

[1] SUSSILLO, David, Anshul KUNDAJE and Dimitris ANASTASSIOU. Spectrogram Analysis of Genomes. EURASIP Journal of Applied Signal processing, pp. 29-42, 2004. ISSN 1110-8657.
[2] 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.

Termín zadání:	11.2.2013	***Termín odevzdání:***	31.5.2013

Vedoucí práce: Ing. Vladimíra Kubicová
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

ABSTRAKT

Tato bakalářská práce se zabývá vyhledáváním kódujících úseků pomocí analýzy DNA spektrogramu. V teoretické části jsou popsány numerické reprezentace genomických dat, možnosti úprav sekvencí DNA a charakteristika metod pro vyhledávání kódujících úseků. Nejpoužívanější metodou pro zpracování DNA je diskrétní Fourierova transformace, díky které jsme schopni v sekvenci vyhledávat požadované úseky. Dále je uveden teoretický postup pro vytvoření spektrogramu a výčet vzorů z něj detekovatelných. Nabyté teoretické znalosti nám slouží k praktické realizaci konkrétních metod v programovém prostředí MATLAB. Vytvořili jsme program pro detekci kódujících úseků ze spektrogramu a nalezení jejich přesných pozic v sekvenci. Námi dosažené výsledky jsou v závěru porovnány s databází NCBI.

KLÍČOVÁ SLOVA

Sekvence DNA, diskrétní Fourierova transformace, kódující oblasti DNA, spektrogram.

ABSTRACT

The Bachelor's Thesis deals with coding identification of coding regions using DNA spectrogram analysis. The theoretical part describes numerical representations of genomic data and methods for editing DNA sequences. The types of methods used for DNA spektrogram construction and characteristic patterns detected by spectrogram are described. The most used method for data processing is discrete Fourier transformation that enables us to scan sequences for required data. There is also a theoretical part about creating a spectrogram and a list of detected samples. Knowledge of this is used to program specific methods in Matlab. We created a program for detection of coding parts in Spectrogram and defining their accurate positions in the sequence. Acquired results are discussed and compared with the NCBI database at the end of this work.

KEYWORDS

DNA sequence, discrete Fourier transform, coding region of the DNA, spectrogram.

Bibliografická citace práce:

SPÍCHALOVÁ, B. *Vyhledávání kódujících úseků pomocí analýzy DNA spektrogramu.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 54 s. Vedoucí bakalářské práce Ing. Vladimíra Kubicová.

PROHLÁŠENÍ

Prohlašuji, že svoji bakalářskou práci na téma „Vyhledávání kódujících úseků pomocí analýzy DNA spektrogramu“ jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské 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 § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 31. 5. 2013 . . . . . .

podpis autora

PODĚKOVÁNÍ

Děkuji vedoucí bakalářské práce Ing. Vladimíře Kubicové, za její cenné rady, vstřícné jednání, odbornou a hlavně trpělivou pomoc při zpracování mé bakalářské práce.

V Brně dne 31. 5. 2013 . . . . . .

podpis autora

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Bioinformatika je obor, který se zabývá navrhováním metod pro shromaždování, analýzu a vizualizaci biologických dat. Také úzce souvisí s genomikou, která se věnuje především získávání sekvencí DNA z různých organismů. Deoxyribonukleová kyselina je jedinečným profilem každého jedince a proto nachází využití v široké škále odvětví (určení otcovství, původ člověka atd.). Díky přesnosti výsledků a zpracování jsou neustále zkoumány nové metody pro dokonalejší analýzu, separaci a umělou syntézu DNA.

Sekvence DNA u eukaryotních organismů, kterými se zabýváme, je složena z intronů - nekódujících částí a exonů - kódujících částí, jejichž pozice v sekvenci si přejeme v této práci získat. Kódující úseky nesou informace ke stavbě bílkovin, které jsou podstatou všech živých organismů a plní v něm různé funkce. Proto sebemenší zásah do oblasti exonů je naprosto zásadní a vede ke vzniku defektní nebo pozměněné bílkoviny. Náplní práce však není podrobné vysvětlení procesu tvorby bílkoviny ze sekvence DNA a proto není zacházeno v této problematice příliš do detailů. [4] [28]

Tato bakalářská práce se zaměřuje na vyhledávání kódujících úseků pomocí analýzy DNA spektrogramu. Pro zpracování dané problematiky uplatňujeme základní znalosti o vlastnostech a struktuře DNA. Využíváme bioinformatické nemoderované databáze GenBank, ze které získáváme DNA sekvence příslušných organismů. Pro vytvoření spektrogramu a následnou analýzu, je využito programového prostředí MATLAB. Vyhledávání kódujících úseků pomocí spektrogramu je zvoleno z důvodu přehledného zpracování, neboť spektrogramy nám dávají jasný vizuální výsledek o vlastnostech příslušné sekvence. Díky těmto možnostem jsme schopni vyhledávat požadované části a nalézat jejich přesnou polohu.

Práce je rozčleněna do čtyř hlavních kapitol. První kapitola je zaměřena na numerické reprezentace DNA, jejichž znalost je nutná pro metody číslicového vyhledávání kódujících úseků. V úvodu kapitoly je obecně pojednáno o symbolické sekvenci DNA, na které je celá práce založena. V kapitole druhé, se zabýváme obecným popisem metod pro vyhledávání exonů v sekvenci. Metody jsou rozčleněny do dvou kategorií, a to na číslicové vyhledávání a vyhledávání za pomocí znaků. Mezi číslicové vyhledávání, neodmyslitelně patří i zpracování sekvence pomocí spektrogramu. Spektrogramu je věnována samostatná třetí kapitola, jelikož tvoří hlavní složku práce. V této kapitole je teoreticky popsána technika konstrukce spektrogramu a také výčet vzorů z něj detekovatelných. Čtvrtá kapitola se věnuje popisu dílčích kroků při praktické realizaci spektrogramu v programovém prostředí MATLAB. Dále jsou v této kapitole zhodnoceny a porovnány námi dosažené výsledky s výsledky vědeckého zkoumání.

1 NUMERICKÁ REPREZENTACE DNA

V následujících podkapitolách je uveden souhrn metod pro numerickou reprezentaci DNA a jejich charakteristika potřebná pro následující praktickou realizaci. První podkapitola je věnována sekvenci DNA, jejímu složení, vlastnostem a možnostem získání přesného sledu jednotlivých nukleotidů.

Pro zpracování genomických dat lze využít numerického mapování, které nám sekvenci DNA v podobě symbolů převede na reprezentaci čísel. Poté lze považovat řetězec DNA za signál, na kterém můžeme využít metody na zpracovávání signálů - jako je například Fourierova transformace, díky které můžeme hodnotit vlastnosti sekvencí.

Při numerickém mapování obvykle dochází ke ztrátě informací a zároveň zvýraznění některých rysů, které nejsme v symbolických sekvencích schopni rozlišit. V odborné literatuře se můžeme dočíst, že numerické mapování se rozděluje do dvou základních skupin, a to na numerické reprezentace a grafické reprezentace.

Grafická reprezentace genomických dat je navíc od numerické reprezentace zobrazitelná v kartézském souřadném systému a umožňuje tak vizuální informaci o sekvenci. Konkrétní metody grafické reprezentace přináší schopnost sekvence prohlížet, třídit a porovnávat mezi sebou. Nejpoužívanější grafickou reprezentací je reprezentace v 3D kartézském souřadném systému. [5]

1.1 Sekvence DNA

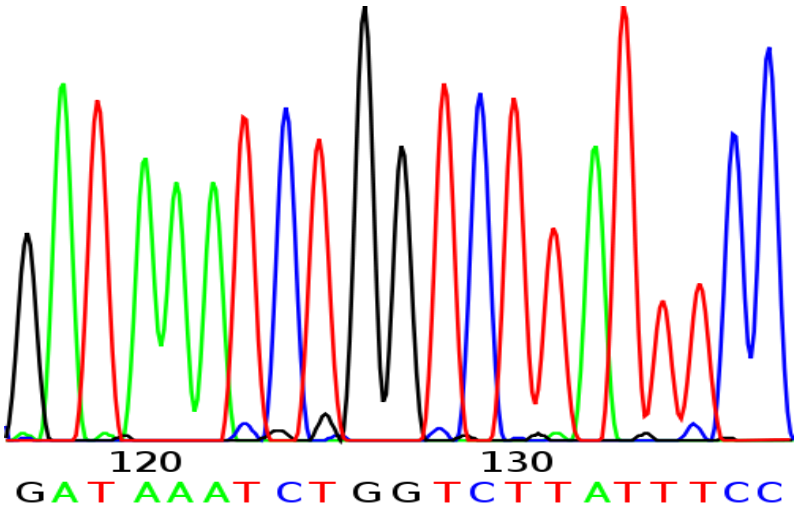
Sekvence DNA neboli genetická sekvence je posloupnost několika set až miliardy symbolů, které představují primární strukturu vlákna DNA.

Symboly *A*, *C*, *G* a *T* používané v sekvenci, reprezentují čtyři nukleotidové báze ve vláknu: adenin (*A*), cytosin (*C*), guanin (*G*) a thymin (*T*).

Jednotlivé nukleotidy se od sebe vzájemně liší v několika biochemických vlastnostech. Odlišují se v molekulární struktuře, kdy cytosin a thymin jsou pyrimidiny, adenin a guanin jsou puriny. V řetězcích dvojité šroubovice DNA jsou pyrimidiny vždy orientovány v jednom řetězci, puriny v druhém. Mezi nukleotidy existují pouze komplementární páry, a to: *T*=*A*, které společně tvoří slabou dvojnou vazbu a pár *C*≡*G*, který vytváří silnou trojnou vazbu. Vodíkové vazby v rámci těchto párů nukleotidů udržují pohromadě oba dva řetězce DNA. Další rozdělení nukleotidů je podle počtu radikálů: *A* a *C* obsahují amino skupinu NH_3, báze *T* a *G* keto skupinu C=O.

Posloupnost libovolného počtu nukleotidů většího než čtyři, lze nazývat sekvencí. Sekvence se obvykle vypisuje bez mezer, např. *ACGTTAGGCCTA*.

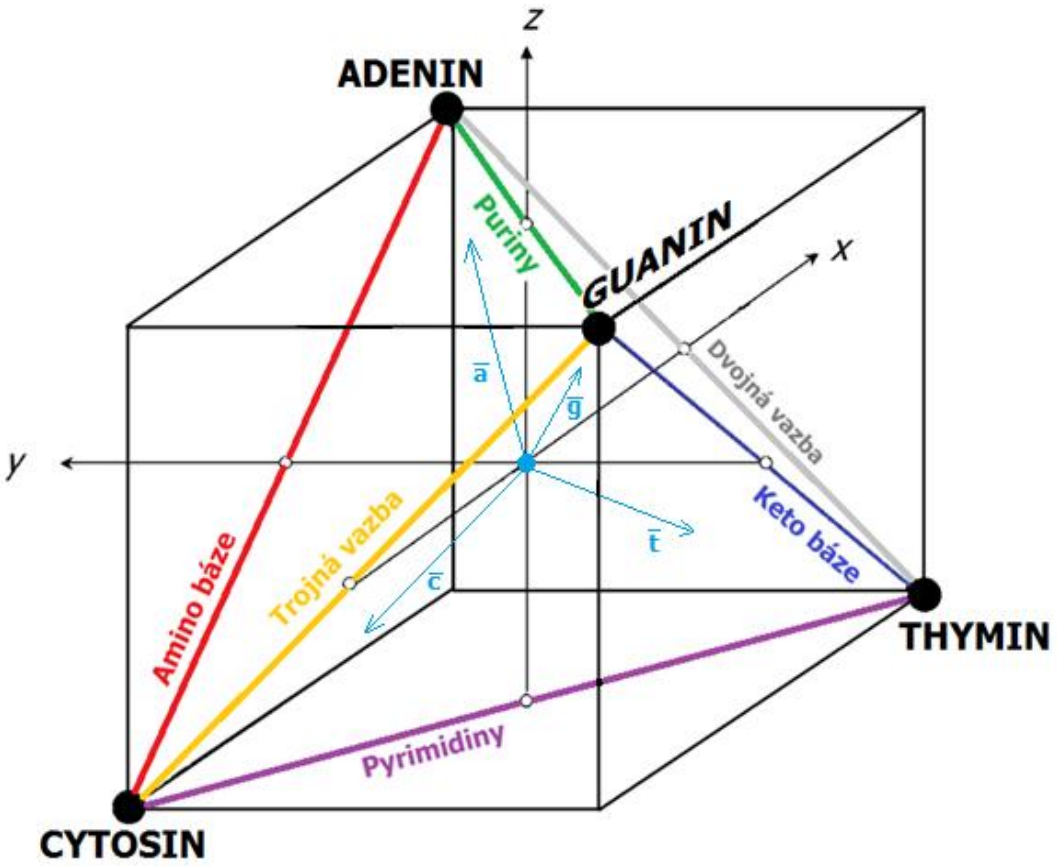
Přesné pořadí nukleotidů lze získat z biologického materiálu dějem nazývaným sekvenování DNA. V podstatě jde o přenesení sekvenční informace do posloupnosti znaků, které lze zapsat na elektronické médium. Mezi nejznámější metody určující pořadí nukleotidů v úseku DNA patří Sangerova a Maxam & Gilbertova metoda, jež jsou založené na gelové elektroforéze. V poslední době existují moderní přístroje, tzv. sekvenátory, které kupříkladu využívají flourescenčních barviv (pro každou nukleotidovou bázi je zvolena konkrétní barva). Obrázek 1 znázorňuje získaný úsek sekvence DNA pomocí sekvenátoru a barviv. [6] [26] [30]



Obrázek 1: Výsledek sekvenace úseku DNA za použití flourescenčních barviv. [26]

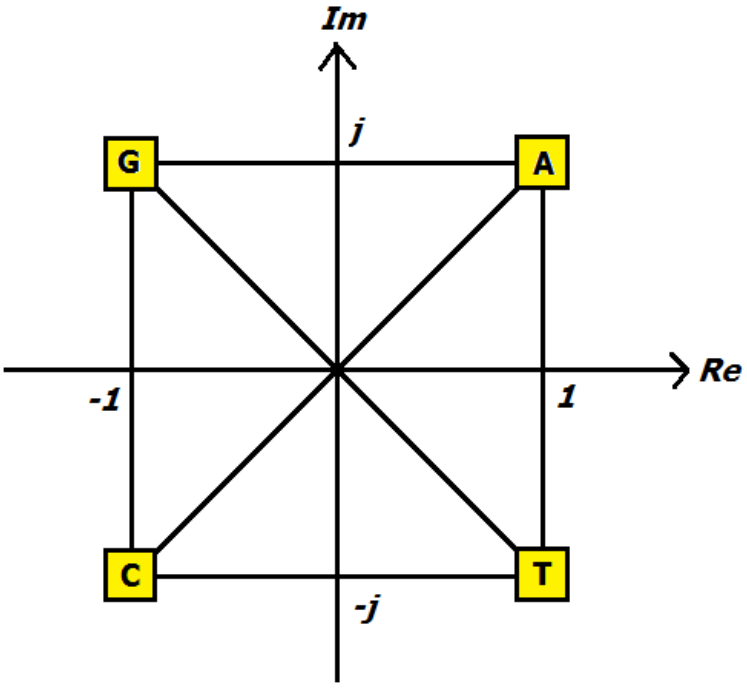
1.2 Komplexní reprezentace

Na základě zmíněné struktury, síly vazby a vázaného radikálu na bázi (viz Kapitola 1.1), můžeme tento klasifikační systém vyjádřit pomocí nukleotidového tetrahedronu neboli čtyřstěnu, který je znázorněn na Obrázku 2. V podstatě se jedná o vyjádření všech vlastností nukleotidů v 3D prostoru. Vrcholy pravidelného čtyřstěnu jsou podskupinou vrcholů krychle, která nám slouží pouze pro lepší orientaci.



Obrázek 2: Nukleotidový čtyřstěn. [5]

Reprezentaci čtyřstěnem lze zredukovat do 2D prostoru pomocí projekce čtyřstěnu do vhodných rovin, kde považujeme zvolenou rovinu za komplexní. Tímto způsobem získáváme komplexní reprezentaci bází. Rovinu zvolíme dle parametrů, které chceme sledovat. My jsme použili následující rozdělení (viz Obrázek 3): dle síly vazby, kde $A=T$ a $C\equiv G$ jsou zobrazeny vzhledem k reálné ose a rozdělení dle molekulární struktury, kdy puriny a pyrimidiny jsou znázorněny vzhledem k imaginární ose.



Obrázek 3: 2D komplexní reprezentace. [5]

Díky této reprezentaci lze každému nukleotidu přiřadit rovnici:

$$\boldsymbol{A} = 1+j, \tag{1}$$

$$\boldsymbol{C} = -1-j, \tag{2}$$

$$\boldsymbol{G} = -1+j, \tag{3}$$

$$\boldsymbol{T} = 1-j. \tag{4}$$

Přiřazení komplexního čísla bázím je prospěšné z toho důvodu, že zůstává zachováno stejné množství informace jako v symbolickém zápisu. Nedochází tím k degeneraci DNA sekvence a můžeme zpětně plně rekonstruovat sekvenci z její grafické podoby. Další výhodou je nepotřebnost složitého grafického nástroje. [5] [16] [34]

1.3 4D binární reprezentace

Binární reprezentace je jednou z nejznámějších a též nejpoužívanějších numerických reprezentací. Metoda je založena na vytvoření indikačního, neboli binárního vektoru pro každý nukleotid: *uA(n), uT(n), uC(n)* a *uG(n)*. Indikační vektor detekuje přítomnost nebo nepřítomnost příslušného nukleotidu na pozici *n* v sekvenci DNA. Vektor *uA* eviduje přítomnost adeninu (*A*), *uC* cytosinu (*C*), *uT* thyminu (*T*) a *uG* guaninu (*G*). Z tohoto tvrzení vyplývá, že po provedení metody získáme celkem čtyři vektory – pro každý nukleotid jeden.

Postup tvorby indikačního vektoru je následující: pokud se na pozici *n*, v úseku sekvence, vyskytuje příslušná báze, zapíše se na danou pozici vektoru číslo *1*. V opačném případě číslo *0*. Toto vyjádření můžeme popsat za pomoci Rovnice 5, která slouží jako ukazatel bází v sekvenci. [5] [16] [34]

$$\boldsymbol{u_x[n] = 1 \quad jestliže \quad s[n] = X \quad jinak \quad u_x[n] = 0,} \tag{5}$$

kde $s[n]$ pro $n = 0, 1,\ldots, N-1$ je symbolická sekvence o délce N,

$u_x[n]$ je indikační vektor pro konkrétní nukleotid,

X je nukleotidová báze – adenin, cytosin, guanin nebo thymin.

1.4 3D numerická reprezentace

Čtyřrozměrná binární reprezentace sekvencí je redundantní a proto ji lze zredukovat na 3D reprezentaci. Čtyři indikační vektory z binární reprezentace jsou nahrazeny pouze třemi vektory takovým způsobem, že nedochází ke ztrátě informací.

Metoda je založena na přiřazení 3D vektoru každému nukleotidu, který směřuje ze středu pravidelného čtyřstěnu (viz Obrázek 2) do jednoho ze čtyř vrcholů. Sekvence DNA je poté zastoupena třemi numerickými sekvencemi x_r, x_g a x_b. [9] [24]

$$\boldsymbol{x_r[n] = \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n])} \tag{6}$$

$$\boldsymbol{x_g[n] = \frac{\sqrt{6}}{3}(u_C[n] - u_G[n])} \tag{7}$$

$$\boldsymbol{x_b[n] = \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n])} \tag{8}$$

3D numerická reprezentace se v praxi využívá pro grafické zobrazování v RGB spektru. Více o RGB spektru se dozvíme v Kapitole 3.2.3.

1.5 EIIP mapování

Pro tento typ mapování, jsou využity biofyzikální a biochemické vlastnosti DNA biomolekul. EIIP (*Electron ion interaction potentials)* neboli vzájemné působení elektron - iontových potenciálů, je metoda, která představuje rozložení energie volných elektronů podél sekvence DNA. Jednotlivým nukleotidům v sekvenci jsou přiřazeny jedinečné číselné EIIP indikátory, a to následovně:

$$\boldsymbol{A} = 0{,}1260,$$

$$\boldsymbol{C} = 0{,}1340,$$

$$\boldsymbol{G} = 0{,}0806,$$

$$\boldsymbol{T} = 0{,}1335.$$

EIIP reprezentace zlepšuje rozlišovací schopnost při vyhledávání genů a snižuje výpočetní zatížení až o 75%. Na druhou stranu také existuje spousta genů, kde tato metoda selhává a není schopna nalézat kódující oblasti. [16]

1.6 Reprezentace pomocí úhlů zkroucení DNA

Jedná se o reprezentaci na základě strukturálních vlastností dvoušroubovice DNA. Metoda je založena na použití dinukleotidového modelu, kdy dva po sobě jdoucí nukleotidy jsou spojeny a převedeny na číselné hodnoty dle úhlu zkroucení.

U dvoušroubovice DNA je úhel jejího zakroucení (*Propeller twist*) měřen ve stupních „°“. Nejvyšší hodnota označuje, že šroubovice je v tomto regionu velmi pevná, zatímco nízká hodnota značí, že region je poměrně flexibilní.

Tabulka 1: Hodnoty úhlů pro jednotlivé dinukleotidy v DNA. [23]

AA	-18,66	**GA**	-13,48
AC	-13,10	**GC**	-11,08
AG	-14,00	**GG**	-8,10
AT	-15,01	**GT**	-13,10
CA	-9,45	**TA**	-11,85
CC	-8,10	**TC**	-13,48
CG	-10,03	**TG**	-9,45
CT	-14,00	**TT**	-18,66

V Tabulce 1 jsou uvedeny hodnoty úhlů pro jednotlivé dinukleotidy dvoušroubovice DNA. Barevně jsou zvýrazněny minimální a maximální hodnoty. Minimální hodnota je -18,66° a vyskytuje se u dinukleotidů *AA* a *TT*. Maximální hodnota je -8,11° a zastupují ji *CC* a *GG*. [23]

1.7 Reprezentace pomocí pevnosti zakřivení DNA

Reprezentace za pomoci pevnosti zakřivení dvoušroubovice DNA, neboli *Bending stiffness,* vychází stejně jako předešlá metoda ze strukturálních vlastností DNA.

Hodnoty pevnosti zakřivení jsou uváděny v *nm* a udávají trvalosti úseků, které jsou odvozeny z experimentálních dat, na základě korelace s anizotropní pružností DNA. Vysoké hodnoty odpovídají tuhým DNA regionům, zatímco nízké hodnoty odpovídají méně pevným DNA oblastem.

Tabulka 2: Tabulka hodnot pevnosti zakřivení jednotlivých dinukleotidů v DNA. [23]

AA	35	**GA**	60
AC	60	**GC**	85
AG	60	**GG**	130
AT	20	**GT**	60
CA	60	**TA**	20
CC	130	**TC**	60
CG	85	**TG**	60
CT	60	**TT**	35

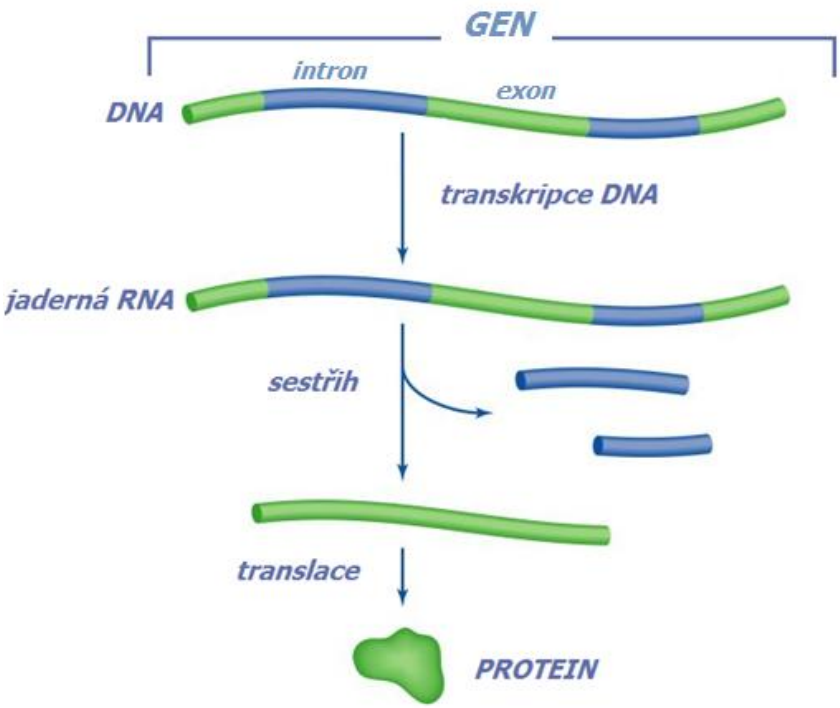
Tabulka 2 zachycuje hodnoty pro pevnosti zakřivení jednotlivých dinukleotidů v sekvenci DNA. Barevně jsou opět zvýrazněny minimální a maximální hodnoty pro lepší názornost a porovnání. Minimální hodnota je 20 *nm* a vyskytuje se u dinukleotidů *AT* a *TA*. Maximální hodnota je 130 *nm* a zastupují ji dinukleotidy *CC* a *GG*. [23]

2 METODY VYHLEDÁVÁNÍ KÓDUJÍCÍCH ÚSEKŮ

Kapitola je zaměřena na identifikaci kódujících oblastí v DNA. Vyhledávání může být realizováno dvěma metodami. Jde nejen o metodu za pomoci číslic, např. převod na binární vektory, zpracování vzniklého signálu Fourierovou transformací a vyhledávání. Nýbrž i o vyhledávání pomocí znaků, kdy se v sekvenci DNA detekují konkrétní kodony, které určují počátek a konec kódujících oblastí.

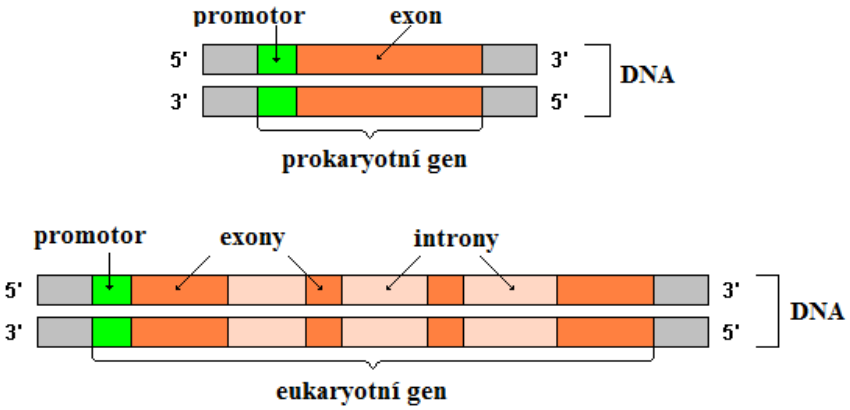
Centrálním dogmatem biologie je, že genetická informace uchovávaná v DNA se transkripcí přenáší do molekul mRNA a translací do proteinů. Toto pravidlo platí pro eukaryotické organismy, kterými se v prácí zabýváme. Po replikaci, neboli zdvojení sekvence DNA a následné transkripci, přepisu, vzniká mediátorová RNA (mRNA), která je reprezentována kodony neboli trojicemi bází. Důležitým tripletem v mRNA je *AUG*, neboli triplet iniciační = START kodon, který určuje počátek genu a začíná zde syntéza proteinu. Triplety *UAA*, *UAG* a *UGA* jsou označovány jako terminační = STOP kodony a ukončují syntézu.

Většina eukaryotických genů obsahuje nekódující úseky - introny, tyto nekódující intronové sekvence se vyštěpují z transkribované DNA v jádře, ještě před jejich přesunem do cytoplasmy, jedná se o mechanizmus zvaný *splicing*. Dnes je známo, že introny přerušují většinu, ale ne všechny eukaryotické geny. Po vystřižení intronů dochází ke spojení exonů a následného vzniku jednotlivých bílkovin tak jak je popsáno na Obrázku 4. Introny přerušují kódující sekvence - exony, které tvoří oblast, podle níž se v procesu translace tvoří bílkovina. Exon je tedy oblast genu, která obsahuje sekvence skutečně kódující proteiny. [22] [25]



Obrázek 4: Postup procesu sestřihu intronů a vzniku proteinu. [15]

Rozdíl mezi prokaryotním genem a eukaryotním je uveden na Obrázku 5. Sekvence prokaryotní obsahuje pouze kódující exony, proto nedochází během procesu transkripce k vyštěpování intronů a tudíž mRNA může obsahovat více genů. Genetická informace u prokaryotické buňky je uložena volně v cytoplazmě v jedné kruhové uzavřené DNA. Naopak DNA u eukaryotické buňky je vláknitá a je obalena jaderným obalem. Na obrázku si můžeme povšimnout zeleného obdélníčku, který značí promotor. Jedná se o specifickou sekvenci DNA, na které se zahajuje transkripce konkrétního genu. [11] [20]



Obrázek 5: Porovnání eukaryotického a prokaryotického genu. [11]

2.1 Číslicové metody vyhledávání kódujících oblastí

Důležitou vlastností exonů je, že vykazují periodu opakování rovnou 3. Tato perioda odpovídá struktuře kodónů a je základním předpokladem pro nalezení kódujících oblastí pomocí metod číslicového zpracování sekvencí DNA. Perioda je pozorována u eukaryotických buněk pouze v exonech, což je pro naši práci velmi přínosné. [5] [10] [31]

Mezi další možnou metodu číslicového vyhledávání kódujících oblastí patří zpracování pomocí spektrogramu, kterému je věnována samostatná Kapitola 3.

2.1.1 Vyhledávání pomocí diskrétní Fourierovy transformace

Nejznámější metodou pro nalezení periody opakování 3 je diskrétní Fourierova transformace (DFT). Jelikož je tato číslicová metoda nejrozšířenější, budeme se jí dále podrobněji zabývat a vysvětlíme si přesný princip.

Jedná se o transformaci získaného signálu do frekvenční oblasti, tj. vstupem do DFT je diskrétní signál a výstupem je diskrétní spektrum tohoto signálu - informace o frekvenčních složkách v něm obsažených.

Jelikož pracujeme se symbolickou sekvencí DNA, je proto důležité, před započetím samotné metody, převést symboly ze sekvence na posloupnost čísel. K převodu nám poslouží numerické reprezentace DNA uvedené v Kapitole 1. Díky této úpravě sekvencí získáme potřebný vstupní diskrétní signál v podobě číslic. DFT následně vytvoří z posloupnosti čísel frekvenční spektrum, ve kterém bychom měli vidět výrazný pík, který odpovídá kódující oblasti.

Rovnice pro výpočet diskrétní Fourierovy transformace je následující: [17]

$$\boldsymbol{U[k] = \sum_{n=0}^{N-1} u(n) e^{-j\frac{2\pi kn}{N}}, \quad 0 \leq k \leq N-1,} \qquad (9)$$

kde $U(k)$ je posloupnost vzorků výsledného spektra,

k je koeficient spektra od 1 do $N/2$,

$u(n)$ je vektor numerických hodnot pro úsek sekvence vymezený délkou okna N, pro který se počítá diskrétní Fourierova transformace,

n je pozice ve vymezeném úseku sekvence.

V případě použití 4D binární reprezentace k převodu symbolů na čísla, získáváme čtyři indikační vektory (viz Kapitola 1.3). DFT je počítána pro každý vektor zvlášť a vzorec pro výpočet musí být mírně upraven, a to následovně:

$$\boldsymbol{U_X[k] = \sum_{n=0}^{N-1} u_X(n) e^{-j\frac{2\pi kn}{N}}, \quad 0 \leq k \leq N-1,} \qquad (10)$$

kde U_X je posloupnost vzorků spektra pro konkrétní nukleotid,

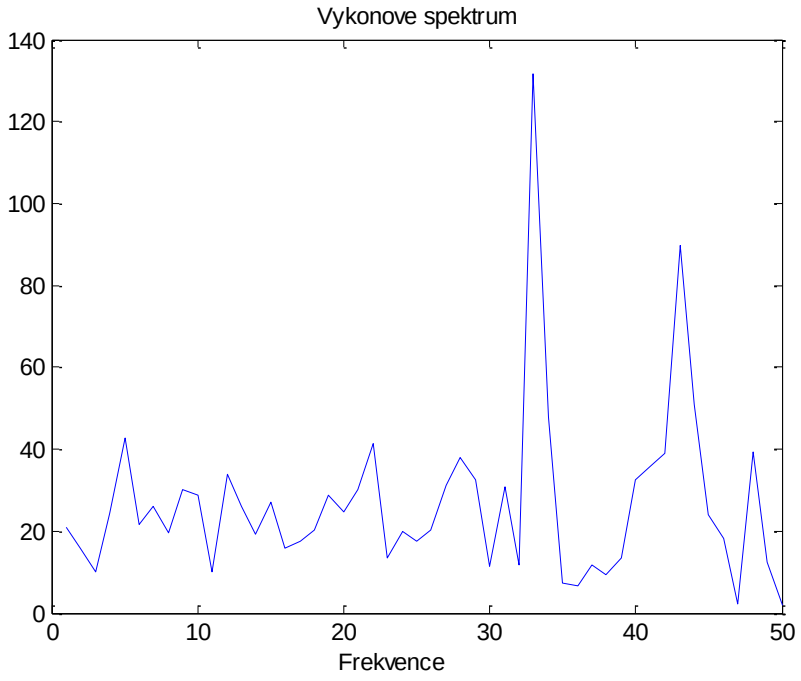
k je koeficient spektra od 1 do $N/2$,

u_X je indikační vektor pro odpovídající nukleotid,

n je pozice v úseku vymezeném délkou okna N, pro kterou se počítá DFT.

Provedením DFT každého segmentu v sekvenci DNA o délce N, získáme frekvenční spektrum úseku. Pro přesnější zjištění převládající frekvence ve spektru a lepší lokalizaci kódujících úseků, použijeme výkonové spektrum. Toto spektrum udává rozložení výkonu signálu podél frekvenční osy. Díky tomu vidíme, která složka signálu je nejvýkonnější.

Výkonové spektrum získáme umocněním frekvenčního spektra na druhou. V tomto spektru následně pozorujeme výrazný vrchol u převládající frekvence. Ve výkonovém spektru pro sekvence DNA, periodě opakování 3 odpovídají vysoké hodnoty pro koeficient $k=N/3$, které korespondují s kódujícími úseky. Na Obrázku 6 je zachyceno výkonové spektrum, za použití okna N=100. Proto vidíme výrazný pík na hodnotě *33,3*, která značí výskyt kódujícího úseku. [17]



Obrázek 6: Výkonové spektrum (X64775).

Pozor si však musíme dát při použití 4D numerické reprezentace, kdy postup získání výkonového spektra je poněkud odlišný. Po provedení DFT, pro každou část indikačních vektorů o velikosti *N*, získáme čtyři dílčí spektra. Každé reprezentuje jeden nukleotid. Následně jsou tyto dílčí spektra umocněna na druhou a sečtena dle Vzorce 11 za vzniku jediného výkonového spektra, ve kterém lze opět detekovat požadované úseky.

$$\boldsymbol{S[k] \triangleq |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2.} \quad (11)$$

Také je důležité vědět, že DFT je symetrická kolem svého středu. Proto stačí vypočíst první polovinu spektra *N/2* a druhá je naprosto identická. [17]

Samotné frekvenční spektrum signálu můžeme vytvořit pomocí několika metod. V základě se jedná o parametrické a neparametrické zpracování. Metody jsou všeobecně popsány v Kapitole 3.1.2, která se věnuje tvorbě frekvenčního spektra, které je dílčím krokem k vytvoření spektrogramu.

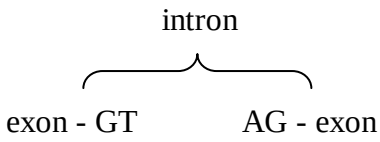
Jelikož vstupem ke spektrální analýze je zpravidla neperiodický signál, je třeba před vlastní DFT provést jeho periodizaci. Periodizace spočívá ve výběru úseku ke zpracování. To se provádí pomocí okénkových funkcí. Nejjednodušší funkcí je obdélníkové okno, které realizuje prostý výřez úseku ze zdrojové posloupnosti bez úpravy jeho amplitudy.

Při výběru úseku mohou nastat dvě situace. Buďto na sebe začátky a konce vybraných úseků spojitě a hladce navazují anebo ne. V prvním případě není s analýzou pomocí DFT problém, protože periodizovaný signál nebude obsahovat žádné strmé přechody. V případě kdy úseky na sebe nenavazují, periodizovaný signál obsahuje nespojitosti. K vyřešení těchto nespojitostí poslouží okno, které kromě výběru úseku navíc vhodně upravuje amplitudu signálu (násobení signálu oknem). Jednou z možností je např. Hammingovo okno. Při jeho použití pak začátek nového i konec předešlého úseku na sebe hladce navazuje, periodizovaný signál neobsahuje žádné nespojitosti. [21]

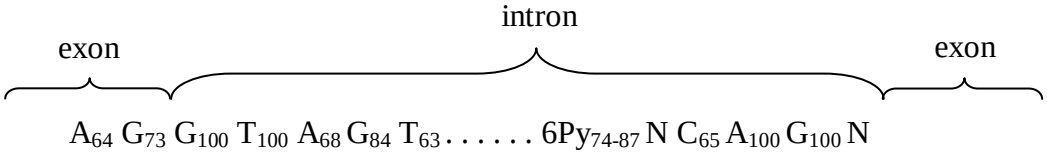
2.2 Znakové metody vyhledávání kódujících oblastí

Jak již bylo řečeno, u eukaryotických genů obsahujících nekódující úseky (introny), dochází během procesu transkripce k jejich vyštěpení z DNA. Mechanizmus sestřihu intronů musí být naprosto přesný, poněvadž exonové sekvence je nutné spojit s přesností jednoho nukleotidu, tak aby bylo zajištěno, že se kodony původně vzdálených exonů budou číst správně.

Proto by se mohlo zdát, že pro zjištění přesnosti takové úrovně je třeba precizních sestřihových signálů. Avšak v primárních transkriptech jaderných genů jsou ukazateli intronů, tzv. dinukleotidové sekvence, které se vyskytují vždy na začátcích a koncích intronů, díky nimž je můžeme od sebe oddělit. Konkrétně:



Navíc existují krátké konzervativní sekvence v místech spojení exonů s introny. Pro jaderné geny vypadají konzervativní spojení takto:



Číselné indexy u písmen ukazují četnosti bází v daném místě v procentech - index *100* tedy znamená, že příslušná báze se na dané pozici v sekvenci vyskytuje vždy. *N* značí, že na příslušné pozici může být kterýkoliv ze čtyř standardních nukleotidů, *Py* značí výskyt pyrimidinů. Na základě této znalosti můžeme definovat algoritmus pro porovnávání znaků, který je schopen rozeznávat exony od inronů. [11] [22] [27]

Další možností pro vyhledávání kódujících úseků je, že tří bázová periodicita v protein-kódujících sekvencích se v nejjednodušší znakové formě vyjadřuje častějším výskytem *G* na prvních pozicích jaderných RNA tripletů, a vyloučením *G* na pozici druhé:

$$\mathbf{(G\ non\text{-}G\ N)_n}.$$

Znakové metody nejsou příliš vhodným a spolehlivým ukazatelem kódujících úseků, jelikož jsou algoritmy pro vyhledávání velmi rozmanité a jsou mezi nimi velké rozdíly v mechanizmu vyhledávání. [27]

3 SPEKTROGRAM

V předchozí kapitole jsme se dozvěděli, že kódující úseky nacházející se v sekvenci DNA vykazují periodu opakování rovnou 3. Tento zásadní poznatek je nesmírně důležitý pro zpracování a hodnocení sekvencí pomocí spektrogramu.

Spektrogram je časový sled krátkodobých spekter signálu, které formulují spektrum jako dvourozměrnou funkci, která je závislá na frekvenci a pozici v čase neboli v našem případě na pozici v sekvenci DNA. Spektrum signálu získáme například výpočtem vektoru Fourierovy transformace omezeného rámce signálu. Též víme, že spektrogramy jsou kvalitní vizuální nástroje pro sekvenční analýzu DNA. Poskytují nám významné informace o sekvenci a slouží tak k podrobnému vyhledávání všech speciálních vzorů a charakteristik.

Hlavní myšlenkou je, považovat výskyt každé nukleotidové báze jako individuální číselný signál a poté ho transformovat do frekvenční oblasti. Amplituda jednotlivých frekvenčních komponentů následně určí, jak silný je určitý vzor bázového prvku opakovaný na dané frekvenci. Vyšší hodnota často signalizuje přítomnost opakování ať již samotných nukleotidů, tak i delších úseků. [24]

3.1 Postup vytvoření spektrogramu

Spektrogramy se mohou vyskytovat v různém rozlišení a s různou velikostí posuvného okna *N*. Skutečná analýza signálu však vychází z konečných úseků, které získáme právě pomocí okna o vhodně zvolené délce. Rozlišitelný rozdíl frekvencí je nepřímo úměrný délce okna, zatímco rozeznatelný poziční úsek je délce okna úměrný. Pokud zvolíme určitý přesah oken, zvýšíme tím poziční rozlišovací schopnost a získáme více spekter, díky nimž jsme schopni lépe sledovat rychlý vývoj u vysokých kmitočtů. Vše však záleží na individuálním nastavení uživatelem.

Díky vizuálnímu pohledu na sekvence, který nám přináší spektrogram, jsme schopni lépe hodnotit vlastnosti a vyhledávat potřebné vzory. Spektrogram je schopný zobrazit celé chromozomy organismů v jednom malém obrázku, což nám přináší výhodu v podobě jednoduššího vyhodnocení vlastností sekvence.

Základní algoritmus pro vytvoření spektrogramu je definován následovně: [7]

- Převod sekvence DNA na numerickou posloupnost.
- Výpočet spektra.
- Sestrojení spektrogramu.
- Normalizace barev.

3.1.1 Převod sekvence DNA na numerickou posloupnost

Nejjednodušší způsob provedení analýzy na symbolické sekvenci je zmapování symbolů na čísla. Podrobné popisy metod pro převod symbolických sekvencí DNA na numerické hodnoty jsou uvedeny a charakterizovány v Kapitole 1. Pro další praktickou realizaci byly zvoleny následující numerické metody pro převod sekvencí:

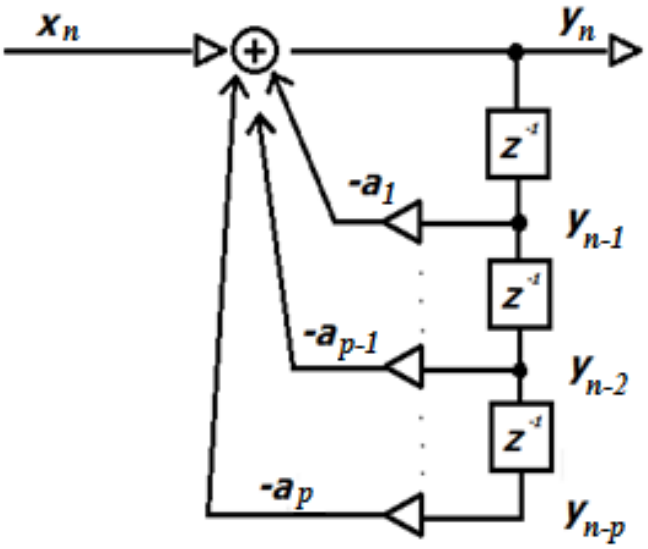
- 4D binární reprezentace,
- EIIP mapování,
- reprezentace pomocí úhlů pro zkroucení dvoušroubovice,
- reprezentace pomocí pevnosti zakřivení dvoušroubovice.

3.1.2 Výpočet spektra

V naší práci jsme si již předem zvolili diskrétní Fourierovu transformaci k výpočtu frekvenčního spektra pro sekvenci DNA. Metodu jsme vybrali z důvodu, že v praxi patří k nejčastějším způsobům zpracování signálů. Podrobně jsme ji charakterizovali v Kapitole 2.1.1. Vytvoření frekvenčních spekter signálů můžeme provést i jinými technikami, než pouze DFT. Rozdělují se na metody parametrické a neparametrické.

Neparametrické metody frekvenční analýzy jsou založeny na pásmových filtrech pro získaný signál, což je pro výpočet spektra ze sekvence DNA nepotřebné. Mezi tyto metody patří již zmíněná Fourierova transformace nebo symbolická autokorelace. [29]

Parametrické metody jsou naproti tomu založeny na výpočtu autoregresního modelu časové řady vzorků. K výpočtu autoregresního (AR) modelu lze využít několika způsobů, např. metoda nejmenších čtverců, řešení soustavy Youle – Walkerových rovnic nebo Levisova rekurzivního algoritmu. Autoregresní model popisuje závislost aktuálního vzorku signálu na předchozích hodnotách. AR model patří k nejpoužívanějším modelům při analýze řeči nebo spektrální analýze. Jeho struktura je uvedena na Obrázku 7. [19] [31]



Obrázek 7: Model autoregresního systému. [19]

Rozdíly mezi těmito metodami jsou v rychlosti zpracování a efektivnosti vyhledávání. Kdy rychlejší je Fourierova transformace a efektivnější je AR model.

Jak bylo řečeno výše, k nalezení periody opakování rovné 3 ze spektra, jsme zvolili Fourierovu transformaci. Konkrétně její modifikaci – diskrétní Fourierovu transformaci. DFT je provedena pro každý ze čtyř vektorů vzniklý numerickou metodou 4D binární reprezentace anebo pro jediný číselný vektor vzniklý metodou EIIP mapování, reprezentací pomocí úhlů zkroucení a pevnosti zakřivení dvoušroubovice DNA.

Vektory u 4D binární reprezentace vytváří matici nul a jedniček, která má čtyři řádky a několik sloupců, kdy počet sloupců odpovídá délce sekvence. Tato primární matice je poté rozdělena na několik dílčích matic – segmentů o zvoleném počtu sloupců, kdy počet sloupců se rovná velikosti zvoleného okna o délce *N*. Následně je na každý segment aplikována DFT. Tímto způsobem získáme 4 dílčí spektra z každého segmentu - pro každý nukleotid jedno.

Obdobným způsobem je provedena DFT i pro ostatní tři zvolené numerické metody, které nám poskytují pouze jeden číselný vektor. Posuvné okno o velikosti *N*, se po vektoru sune a počítá DFT pro každou hodnotu, o to je algoritmus snazší.

Takto získáme frekvenční spektrum sekvence DNA. Dále postupujeme dle kroků popsaných v Kapitole 2.1.1, dokud nezískáme výkonové spektrum potřebné k sestrojení spektrogramu.

3.1.3 Sestrojení spektrogramu

Kroky k vytvoření spektrogramu jsou následující: první získané výkonové spektrum – vektor hodnot, pro první úsek o délce *N*, umístíme ve výsledném spektrogramu jako první sloupec. Druhé spektrum umístíme jako druhý sloupec a takto postupujeme až do posledního spektra. Výsledný spektrogram tedy zobrazuje v každém sloupci dílčí výkonové spektrum z jednotlivých segmentů a jedná se proto o matici hodnot.

Důležitou podmínkou je velikost posunu okna *N* po sekvenci, kdy je možné nastavit libovolnou velikost posunu. Minimální hodnota je *1*, což znamená, že druhé okno je vzato od *2*. pozice v daném segmentu do *N*+2, program je ovšem časově náročný. Pokud je hodnota posunu okna vyšší než *10*, dojde k tomu, že počet dílčích výkonových spekter ve spektrogramu se sníží a výsledek není příliš přesný. Na druhou stranu, u dlouhých sekvencí, dojde při větším posunu oken, ke zkrácení času vypočtu spektrogramu a také k horší rozlišovací schopnosti jeho vlastností. Překrývání oken tedy patří spolu s velikostí posuvného okna k hlavním parametrům, které ovlivňují dobu trvání programu a přesnost výsledného spektrogramu. [9] [13]

Osa *x* značí pozice nukleotidů v sekvenci DNA a osa *y* periodu. Platí, že perioda je podíl *N* a *k*, kde *N* je úsek sekvence a *k* určuje pořadí spektrální složky ($perioda = N / k$). Obě osy jsou uváděny v jednotkách párů bazí [*bp*]. [17]

3.1.4 Normalizace barev

Škálu barev použitou pro vytvoření spektrogramu si volíme dle osobní potřeby. Pro naši práci jsme zvolili 256 barevných úrovní, kdy každá hodnota v matici spektrogramu může mít jeden z 256 barevných odstínů:

0 – odpovídá minimální hodnotě z matice spektrogramu,
255 – odpovídá maximální hodnotě z matice spektrogramu.

Dle tohoto barevného nastavení jsme schopni posuzovat vlastnosti spektrogramů a vyhledávat v nich kódující a jiné úseky. A to proto, že vyšší hodnoty v matici spektrogramu, mají přiřazenu vyšší hodnotu z barevné palety a tak signalizují výskyt kódujících úseků a přítomnost opakování na dané pozici.

3.2 Vzory ve spektrogramu

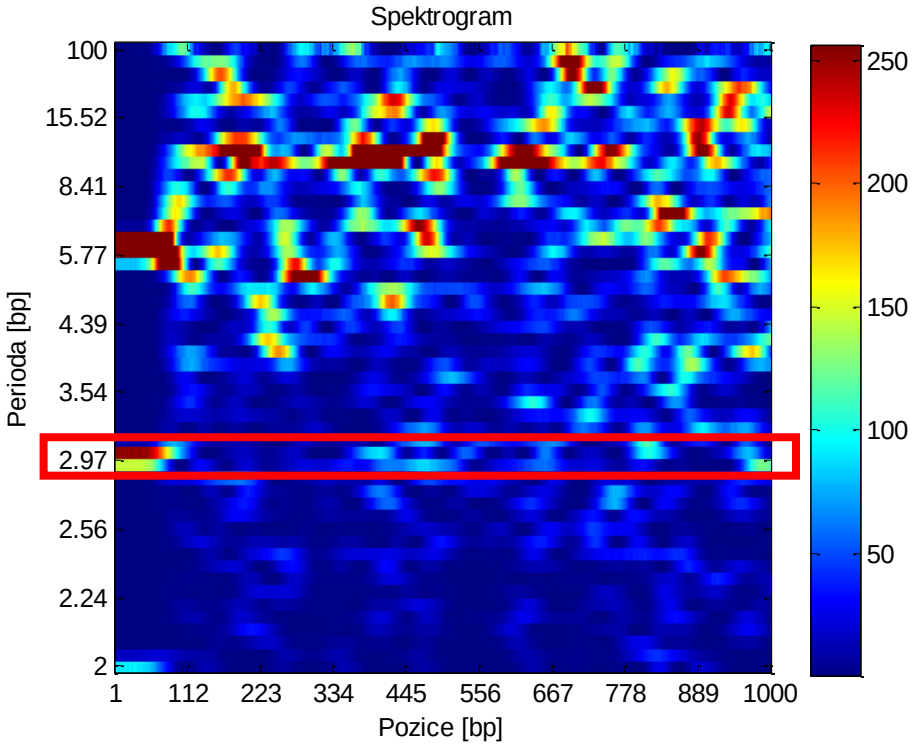
Charakteristické vzory, které můžeme detekovat ve spektrogramu, je možné roztřídit dle jejich velikosti. Vzory složené z milionů párů bazí (*bp*) jsou považovány za velké. Vzory sestávající z několika stovek tisíc nukleotidů jsou považovány za střední a ty, sestávající z několika tisíc jsou pokládány za velmi malé. Typicky větší vzory představují strukturální prvky a menší vzory jsou užitečné pro vizualizaci některých protein-kódujících oblastí, na které se v práci zaměřujeme. [24]

3.2.1 Kódující oblasti

Nejdůležitější vlastností kódujících oblastí ve spektrogramu je silná intenzita píku na periodě opakování 3. Proto nalezení pozice úseku v sekvenci DNA, vykazujícího tuto periodicitu, je výborným ukazatelem genů u většiny eukaryotických organismů.

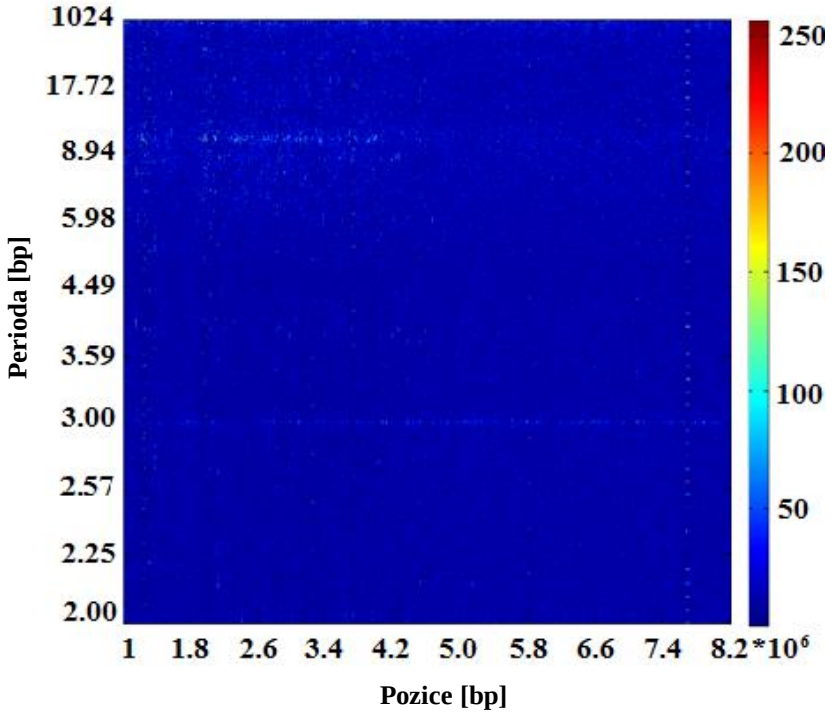
Obrázky 8 a 9 znázorňují spektrogramy pro určité délky sekvence DNA z III. chromozomu Háďátka obecného. Oba spektrogramy byly zpracovány a vytvořeny za pomoci 4D binární reprezentace, diskrétní Fourierovy transformace a dalších technik pro konstrukci spektrogramu v programovém prostředí MATLAB (viz Kapitola 4).

První spektrogram je vytvořen pro úsek sekvence o délce 1-1000*bp*. Je použita délka posuvného okna s velikostí *wl*=100 a posunem okna *posun*=1. Perioda 3 je v obrázku ohraničena červeným obdélníkem pro lepší a rychlejší zhodnocení.



Obrázek 8: Spektrogram III. chrom. (1*kbp*) *C. elagans* (NC_003281).

Druhý spektrogram je vytvořen taktéž z III. chromozomu *C. elegans*, ale o mnohem větší délce sekvence, a to 1-8.2*Mbp*. Je použito posuvné okno s velikostí *wl*=1024 a velikostí posunu okna *posun*=1000.



Obrázek 9: Spektrogram III. chrom. (1-8.2*Mbp*) C. elegans (NC_003281).

Srovnání spektrogramů pro různě dlouhé sekvence DNA provádíme z důvodu, abychom ukázali, že na kratším úseku sekvence se kódující úseky na periodě 3 vyskytují v krátkých blocích a jsou snadněji detekovatelné jejich pozice. Kdežto, při použití mnohonásobně delšího úseku sekvence, nám kódující úseky vytvoří tečky a splynou v jednu vodorovnou, občas přerušovanou linii. Kvůli tomu je jejich vzájemné rozeznání a určení pozic pouhým okem nemožné.

Na Obrázku 9 můžeme ve spektrogramu pozorovat u pravého kraje několik výrazných úseků, o vysoké intenzitě, které jsou naskládány nad sebou. Jedná o oblasti, jež odpovídají tandemovým repeticím, o kterých je pojednáno v následující kapitole.

Další velice výrazné úseky projevující se vysokou intenzitou, se u obou spektrogramů nacházejí na periodě 10 - 11*bp*. Jedná se o zóny v sekvenci DNA s velice silným výskytem *A* a *T* nukleotidů. Tato perioda se může vztahovat na strukturu DNA dvoušroubovice, která má v průměru periodicitu 10,4*bp*. [12]

3.2.2 Repetitivní oblasti

Pouhých 10% genomové DNA u eukaryotních organismů a člověka jsou kódující sekvence organizované v exonech. Tak jako kódující DNA i nekódující může být unikátní anebo se může nacházet v genomu ve více identických nebo podobných kopiích.

U prokaryot většina DNA kóduje proteiny. Naopak u eukaryot jen asi 1,5% DNA kóduje proteiny a zbytek - nekódující DNA je tvořena částečně introny, ale především tzv. repetitivní DNA – oblasti s vysokým množstvím kopií. Repetitivní oblasti se dělí do dvou skupin a to na rozptýlené repetice a tandemové repetice.

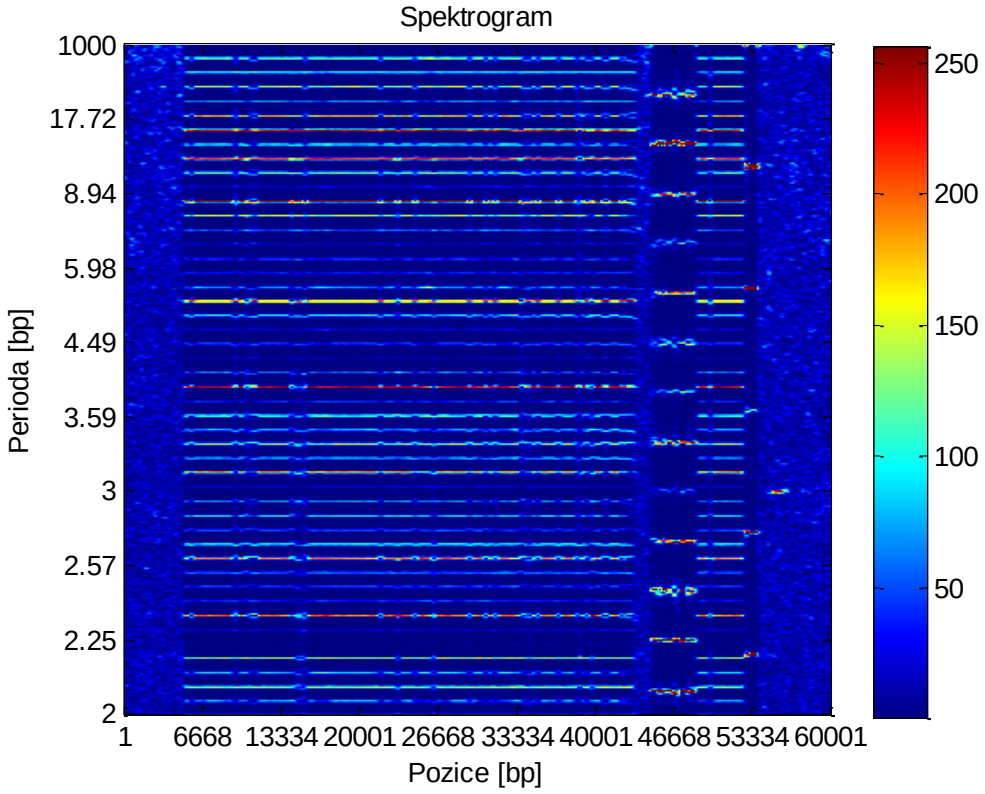
Rozptýlené repetice se nacházejí v sekvenci v samostatných blocích a můžeme je rozdělit na transpozony a retrotranspozony.

Pokud se kopie sekvenčního motivu nacházejí v blocích těsně za sebou, mluvíme o tandemových repeticích. Tandemové repetice jsou sekvence několika nukleotidů, opakované mnohokrát za sebou, např. *GTTACGTTACGTTACGTTACGTTACGTTAC*. Sekvence *GTTAC* se může až několik set tisíckrát opakovat. Díky tomu, že tandemové repetice se opakují tolikrát za sebou, má tato část chromosomu poněkud jiné složení *CG* a *AT* než ostatní části chromosomu a proto jsou dobře viditelné ve spektrogramech.

Tandemové repetice se dělí na tři typy:

- nejdelší části se nazývají satelity, délka sekvence může bát až několik *kbp,* nacházejí se nejčastěji v telomerách,
- kratší úseky jsou minisatelity s délkou sekvence 6 až 100*bp*, používají se jako genetické markery,
- nejkratší pasáže jsou mikrosatelity o délce sekvence 2 až 5*bp,* nejčastějšími mikrosatelity jsou dinukleotidy (*CA*) a trinukleotidy (*CAG*). [24] [30]

Obrázek 10 prezentuje spektrogram III. chromozomu *C. elegans* od délky sekvence $7{,}4.10^6$ až $7{,}46.10^6$*bp*. Tento spektrogram je zobrazen z důvodu lepší představivosti o výskytu a rozložení tandemových repetic v sekvenci. Použitá sekvence DNA je upravena pomocí numerické metody 4D binární reprezentace, DFT a dalších modifikací vedoucích k vytvoření spektrogramu v programovém prostředí MATLAB (viz Kapitola 4). Bylo použito okno o velikosti *wl*=100 a *posun*=1.



Obrázek 10: Spektrogram III. chrom. (7.4 – 7.42.10^6*bp) C. elegans* (NC_003281).

Ve výše představeném spektrogramu jsou jasně viditelné opakující se sekvence, které se projevují vyšší intenzitou barvy. Mírné odchylky v základním opakování se nám jeví jako svislé čáry, které od sebe oddělují jednotlivé úseky. Může se však jednat i o úseky bez opakování.

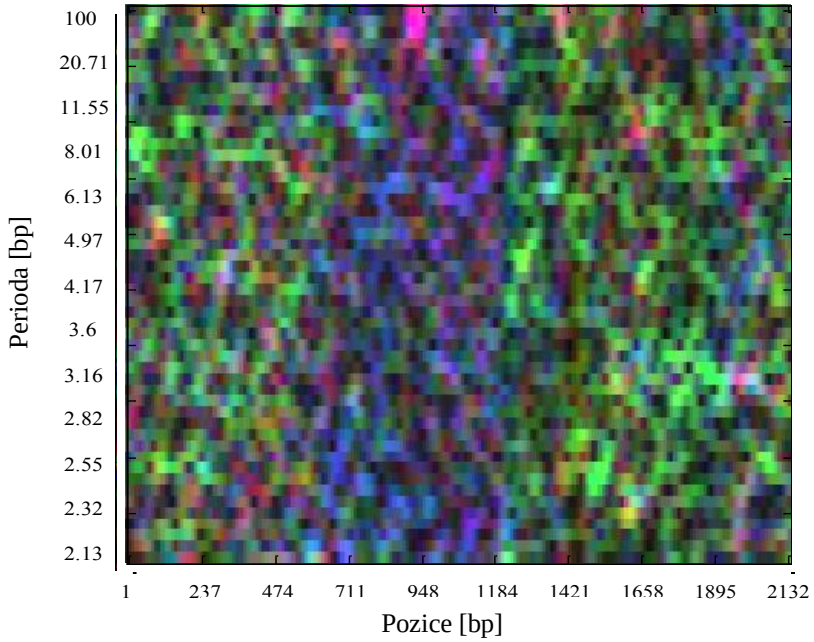
První úsek se vyznačuje velkým počtem opakování vzorů nad sebou a také dlouhou délkou opakujících se vzorů za sebou, čímž vzniká tento ojedinělý jev, který se pozorovateli jeví jako souvislé linie. Podle počtu čar v každém úseku se dá stanovit délka vzoru, který se neustále opakuje.

3.2.3 CpG ostrůvky

Jako *CpG* ostrůvky jsou označovány velmi krátké úseky se zvýšeným výskytem *CG* dinukleotidů než v jiných částech DNA. *CpG* je zkrácené označení pro strukturu cytosin (*C*) – fosfodiesterová vazba (*p*) – guanin (*G*). Základní vlastností těchto ostrovů je jejich minimální délka (1*kbp*) a procentuální zastoupení těchto nukleotidů v genech (60 – 70% v lidském genu). V genomech savců mají *CpG* ostrůvky obvykle délku 300-3000 *bp* a jsou v blízkosti přibližně 40% promotorů, které slouží jako ukazatelé začátků genů a jsou nepostradatelné pro hlavní funkci buňky. V barevných spektrogramech, kterými se zabýváme v naší práci však *CpG* ostrůvky nelze nalézt. Jsou detekovatelné pouze v RGB spektrogramech. [33]

RGB obraz je barevný obraz, který využívá 3 základní vrstvy: červenou (Red), zelenou (Green) a modrou (Blue). Aby byly DNA sekvence čitelnější, mohou být tyto tři vrstvy systému RGB přiřazeny jednotlivým nukleotidům. V důsledku toho, každý nukleotid odpovídá odlišnému odstínu. Dle zabarvení spektrogramu poté určujeme nejčastěji vyskytovanou bázi. V programovém prostředí uživatelskou paletu vytvoříme pomocí matice *M* x 3, kde *M* je počet barev v paletě a hodnoty jsou od 0 do 1, kde každý řádek představuje jednu ze tří barev. Nula znamená 0% podíl dané barvy, jednička 100% podíl dané barvy. [2] [5]

Obrázek 11 prezentuje RGB spektrogram, ve kterém se nacházejí 2 oblasti se zvýšeným výskytem *CpG* ostrůvků - značeny zeleně. Zóny jsou od sebe odděleny sekvencí bohatou na nukleotid *A*, kterému byla přiřazena modrá vrstva. Nukleotidu *T* byla přiřazena vrstva červená. Jakákoli jiná barva ve spektru je složena smícháním těchto tří vrstev. [24]



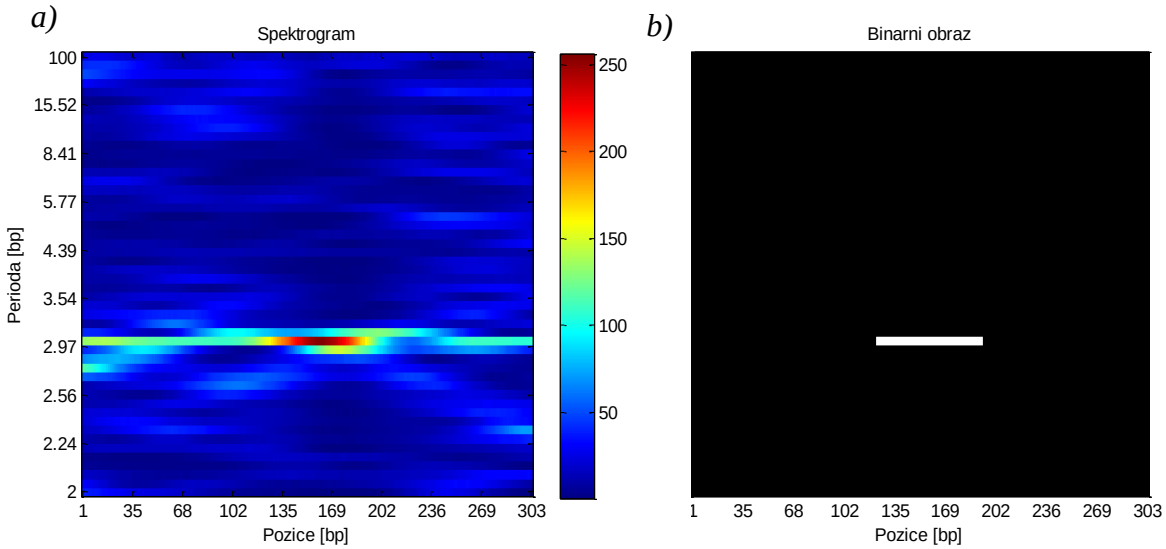
Obrázek 11: RGB spektrogram chr. XXII. Homo sapiens, 2894684 – 2896815*bp*. [24]

3.3 Prahování spektrogramu

Pokud si přejeme zjistit přesné pozice kódujících úseků v sekvenci DNA, které nejsou z barevného spektrogramu příliš jasné, poslouží nám k tomu prahování spektrogramu. prahování je funkce, která převádí barevný obraz na binární – často označovaný jako černobílý, tj. má jen dvě barvy černou a bílou.

Jelikož víme, že vyšším hodnotám v matici spektrogramu jsou přiřazeny vyšší barevné odstíny (viz Kapitola 3.1.4), můžeme v programovém prostředí MATLAB nastavit určitý práh dle barevné palety. Podle tohoto prahu vytvoříme z barevného spektrogramu binární obraz, kdy každý pixel = každá hodnota v matici, má jednu z pouze dvou možných barev. Hodnoty nad zvoleným prahem jsou značeny bíle a představují ve výsledném binárním obraze kódující oblast. Hodnoty pod prahem jsou pro nás nepodstatné, a proto jsou značeny černě. Tímto postupem získáme obraz, ze kterého můžeme přesně lokalizovat pozice kódujících úseků v sekvenci.

Na Obrázku 12 *a)* je znázorněn spektrogram vytvořený v barevné paletě *jet,* na obrázku *b)* je z barevného spektrogramu vytvořen binární obraz, ve kterém jasně vidíme počátek a konec kódujícího úseku. Celá sekvence DNA *O. sativa* byla zpracována pomocí 4D binární reprezentace, DFT v programové prostředí MATLAB. Bylo použito okno o velikosti *wl*=100 a *posun*=1. Práh pak byl nastaven na hodnotu *150.*



Obrázek 12: Barevný spektrogram a binární obraz (X64775).

4 REALIZACE METOD V PROGRAMOVÉM PROSTŘEDÍ MATLAB

Díky výše získaným teoretickým znalostem o vlastnostech kódujících úseků, je v této kapitole popsán postup vytvoření programu pro identifikaci těchto oblastí v DNA sekvencích. Podstatou programu je vytvořit spektrogram, v něm vyhledat kódující úseky, které se projevují vysokou intenzitou na periodě 3 a detekovat jejich pozice v sekvenci.

Program je realizován v programovém prostředí MATLAB a pracuje se sekvencemi, které jsme získali z bioinformatické nemoderované databáze GenBank, kterou provozuje NCBI (The National Center for Biotechnology Information).

Sekvence z databáze jsou uloženy v jednoduchém formátu **.fasta*. Fasta formát obsahuje hlavičku, která je zobrazena na prvním řádku a obecně popisuje sekvenci – název sekvence, anotace a další údaje, které nejsou součástí samotné sekvence. Na druhém řádku pak následuje surová posloupnost v podobě symbolů. [6]

Příklad sekvence DNA ve formátu **.fasta* pro *O. sativa:*

>gi|312289|emb|X64775.1| O.sativa short highly repeated, interspersed DNA

ATGGAGAGCGACTGCCAGTTCTTGGTGGCGCCGCCGCAGCCGCAC…

4.1 Skladba navrženého programu

Z důvodu čtyř metod použitých pro numerické reprezentace DNA (viz Kapitola 3.1.1) a jejich odlišného programového zpracování, jsme vytvořili čtyři hlavní programy pro analýzu DNA sekvencí pomocí spektrogramu (viz příloha na CD). Programy jsme pojmenovali dle použité numerické reprezentace s dodatkem **_spektrogram.m:*

- *Binarni_spektrogram.m,*
- *EIIP_spektrogram.m,*
- *BendingStiffness_spektrogram.m,*
- *PropellerTwist_spektrogram.m.*

Hlavní programy slouží pro načtení sekvence, definování dílčích proměnných, tvorbu spektrogramu a jeho následné úpravy. Dále je utvořeno několik vlastních samostatných funkcí, které jsou důležité a potřebné pro chod programu, numerickou úpravu sekvencí, výpočet DFT a také z důvodu lepší přehlednosti.

V samotném začátku programu je nejprve nutné načíst zvolenou sekvenci DNA, ve formátu*.*fasta*. Toho docílíme pomocí MATLABovské funkce *fastaread*. Tato funkce načte do proměnné *Seq* celou symbolickou sekvenci DNA. Do proměnné *head* jsou uloženy obecné informace o sekvenci. Následuje možnost výběru délky sekvence, nebo oblasti, ze které vyžadujeme sestrojit spektrogram. Např. *Seq=Seq(1:1000),* kdy tento zápis značí, že dále je pracováno pouze se sekvencí v délce 1000 nukleotidů.

Poté co máme sekvenci DNA načtenu, vybránu oblast zájmu, zjištěnu délku sekvence a nastaveny proměnné, můžeme nechat program, aby vytvořil spektrogramu. Po vytvoření spektrogramu následuje několik operací k jeho úpravě, normalizaci a prahování. Podobu spektrogramu lze měnit individuálním nastavením několika parametrů a proměnných, jako je délka zobrazované sekvence, velikost posuvného okna, velikost posunu okna, barevná paleta, normalizace barev či prahování.

4.1.1 Vlastní funkce použité v programu

V tomto oddílu uvádíme funkce nutné pro chod programu a také popisujeme jejich stručný význam. Pro tvorbu funkcí založených na numerických reprezentacích DNA vycházíme z teoretických poznatků o těchto metodách uvedených v Kapitole 1.

- *BinPrevod* – jde o funkci, která zajišťuje převod nukleotidů na základě 4D binární reprezentace (viz Kapitola 1.3). Tato funkce nám vytvoří čtyři indikační vektory. Je součástí programu *Binarni_spektrogram.m.*
- *EIIP_Prevod* – funkce, jenž převádí báze na numerickou posloupnost. Je založena na mapování pomocí elektron – iontového potenciálu DNA (viz Kapitola 1.5). Patří k programu *EIIP_spektrogram.m.*
- *BendingStiffness* – funkce pro transformaci dinukleotidů na numerické hodnoty. Vychází z reprezentace pomocí pevnosti zkroucení dvoušroubovice DNA (viz Kapitola 1.6). Je nedílnou částí programu *BendingStiffness_spektrogram.m.*
- *Propeller_Twist* – jedná se o funkci, sloužící k převodu dinukleotidů na posloupnost čísel. Je založena na reprezentaci pomocí úhlu zakřivení DNA (viz Kapitola 1.7). Řadí se k programu *PropellerTwist_spektrogram.m.*
- *Fourier_transform* – funkce potřebná pro výpočet frekvenčního spektra pomocí diskrétní Fourierovy transformace. Tato funkce je součástí všech 4 programů (*Binarni_spektrogram, EIIP_spektrogram, BendingStiffness_spektrogram, PropellerTwist_spektrogram)*.

4.1.2 Význam funkcí v programu

Nejprve je načtena sekvence DNA. Pokračujeme přivoláním funkcí pro numerickou reprezentaci (*BinPrevod, EIIP_prevod, BendingStiffness* a *PropellerTwist*). Tyto reprezentace převedou symbolickou sekvenci na vektor čísel. Nejsložitější z funkcí je *BinPrevod*. Nejedná se pouze o vytvoření jednoho vektoru hodnot ze sekvence, ale o tvorbu čtyř indikačních vektorů pro nukleotidy. Takto získáme matici hodnot, kterou program *Binarni_spektrogram* musí procházet. Doba pro výpočet spektrogramu se tím prodlužuje.

Dále jsou zavedeny proměnné, jako je délka sekvence *L*, velikost posuvného okna *wl* (v teoretické části práce značeno jako *N*) a velikost posunu okna po sekvenci *posun*. Posun okna udává, o kolik nukleotidů se okno na sekvenci posune vzhledem k předchozí pozici. Čím menší posun okna je, tím získáme větší rozlišení, avšak na úkor delší doby výpočtu. Je také důležité, aby použité okno délky *wl* bylo dostatečně velké (od několika desítek až po několik tisíc) tak, aby efekt periodicity dominoval spektru pozadí. Nicméně dlouhé okno znamená delší dobu výpočtu a také má za následek značné kompromisy v umístění kódujících úseků v sekvenci. V naší práci máme zvolenou velikost okna *wl*=100, pro sekvence v řádech *kbp* a *posun*=1, pro sekvence v řádech desítek *kbp* je zvolena délka okna *wl*=1000 a *posun*=100.

Vytvořením *for* cyklu v programu, docílíme procházení vektoru hodnot a vykonávání operací, které vedou k vytvoření spektrogramu. *For* cyklus prochází vektor hodnot a na každé pozici provádí dané operace. Je nastaven od hodnoty *1* do *L-wl,* s krokem o velikosti proměnné *posun*. V programu *Binarni_spektrogram* je situace složitější a jsou vytvořeny dva *for* cykly, které zajišťují procházení matice hodnot vzniklé 4D binární reprezentací. První *for* cyklus prochází vzniklou matici hodnot po sloupcích. Druhý *for* cyklus prochází matici po řádkách. Cyklus obsahuje proměnnou *seq_1*, jedná se o vektor, jehož počet hodnot odpovídá velikosti okna *wl*. Od každé hodnoty v proměnné *seq_1*, je odečítána střední hodnota, která vznikne jako průměr hodnot, které proměnná obsahuje. Tím zaručíme, že výsledek má střední hodnotu na nule – odstraníme stejnosměrnou složku. Proměnná *seq_1* je následně násobena funkcí *hamming*, která vzniklému signálu zaručí, na základě násobení signálu Hammingovým oknem, jeho periodicitu. Poté úseky vstupují do další části, která z nich počítá frekvenční spektrum, na principu DFT, pomocí přivolané funkce *Fourier_transform*. Následuje umocnění frekvenčního spektra, z daného úseku, na druhou a uložení do proměnné *vykon_spektrum*.

Samotný spektrogram je vytvořen jako transpozice výkonového spektra. Transpozice nám zaručí převod výkonového spektra jednotlivých úseků (řádků) na sloupce ve výsledném spektrogramu.

Pro lepší přehlednost jsou v Tabulce 3 uvedeny další funkce použité v programu. Jedná se o předdefinované MATLABovské funkce.

Tabulka 3: Seznam funkcí použitých v programu.

Význam funkce	Název funkce
sloupec s barevnou paletou	*colorbar*
nastavení barevné palety	*colormap*
exponenciální funkce e^x	*exp*
otevření grafického okna	*figure*
index prvků vektoru	*find*
zaokrouhlení k minus nekonečnu	*floor*
změna hodnot na ose	*gca*
použití Hamming. okna pro periodicitu signálu	*hamming*
zobrazení obrazu	*image*
délka vektoru	*length*
vektor s konečnou aritmetickou posloupností	*linspace*
maximum / minimum	*max / min*
střední hodnota	*mean*
matice obsahující jedničky	*ones*
zaokrouhlení k nejbližšímu celému číslu	*round*
nastavení grafického objektu	*set*
počet řádků a sloupců matice	*size*
součet prvků	*sum*
název obrazu	*title*
převod malých písmen na velká	*upper*
popis osy *x / y*	*xlabel / ylabel*
vynášecí čárky na ose *x / y*	*XTick / YTick*
popis vynášecích čárek na ose *x / y*	*XTickLabel / YTickLabel*
matice obsahující nuly	*zeros*

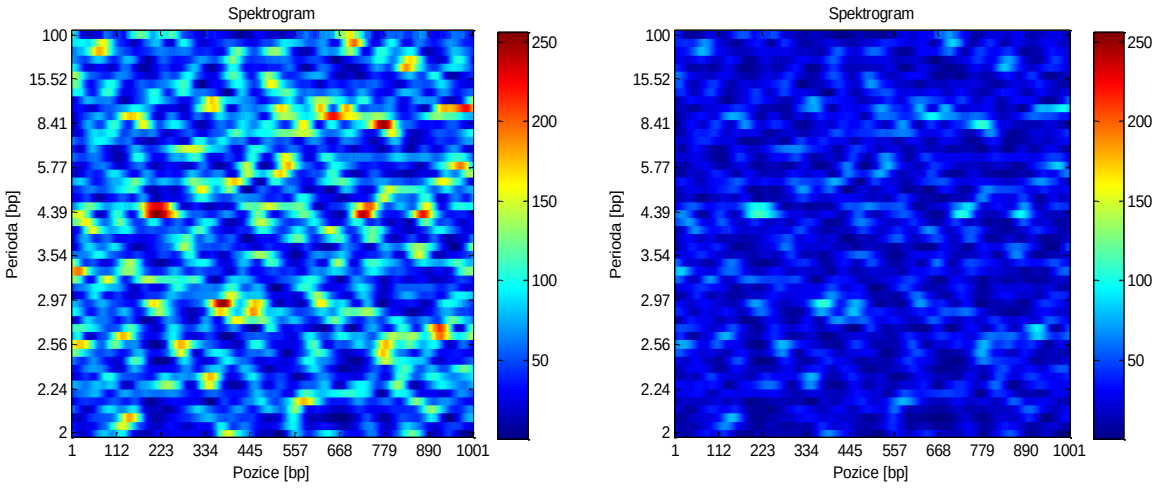
4.1.3 Normalizace barev spektrogramu

Normalizace barevných složek spektrogramu je vypočtena pomocí střední, maximální a minimální hodnoty z matice hodnot spektrogramu. Střední hodnota je zjištěna pomocí funkce *mean* a každá hodnota z výsledného spektrogramu je podělena trojnásobkem této střední hodnoty. Tímto způsobem snížíme rozptyl hodnot v matici spektrogramu.

Posléze je stanovena minimální a maximální hodnota z matice spektrogramu, pomocí funkcí *min* / *max* a dle nich provedena normalizace ostatních hodnot. Zjistíme minimální hodnotu z matice, kterou následně odečteme od všech ostatních hodnot. Takto nastavíme nejmenší hodnoty v matici na *0* – nejnižší intenzita z barevné palety. Poté určíme maximální hodnotu, kterou nejprve podělíme všechny hodnoty v matici a pak vynásobíme číslem *255*, čímž maximálním hodnotám přiřadíme nejvyšší intenzitu z barevné palety. Tak provedeme normalizaci barev ve spektrogramu, jak jsme popsali v Kapitole 3.1.4. Ostatním hodnotám z matice, které se nacházejí mezi maximem a minimem jsou přiděleny barevné složky dle odpovídající barevné palety.

Normalizace hodnot se ovšem nedá rozumně provést pro celý spektrogram. Nastává totiž problém s maximy, která se vyskytují i na jiných periodách a jsou mnohdy vyšší než na periodě 3, kterou vyžadujeme. Tímto způsobem by se nám úseky na třetí periodě potlačili a nebyly by správně detekovatelné. Proto jsme zvolili v programu normalizaci barev spektrogramu pouze podle určitého úseku - kolem třetí periody. Požadovaný úsek hodnot jsme definovali takto: *wl/3-2 : wl/3+2*. Kdy *wl/3* vychází z koeficientu spektra *k=N/3* a odpovídá tak periodě 3 (viz Kapitola 2.1.1). Číslo *2* značí počet řádků nad a pod periodou tři, v matici spektrogramu.

Na Obrázku 13 je znázorněn rozdíl mezi spektrogramy: *a)* s normalizací barev, *b)* bez normalizace. Je patrné, že pokud není provedena normalizace barev, není možné detekovat kódující úseky. Byl použit program *Binar_spektrogram.m*, okno *wl*=100, *posun*=1.



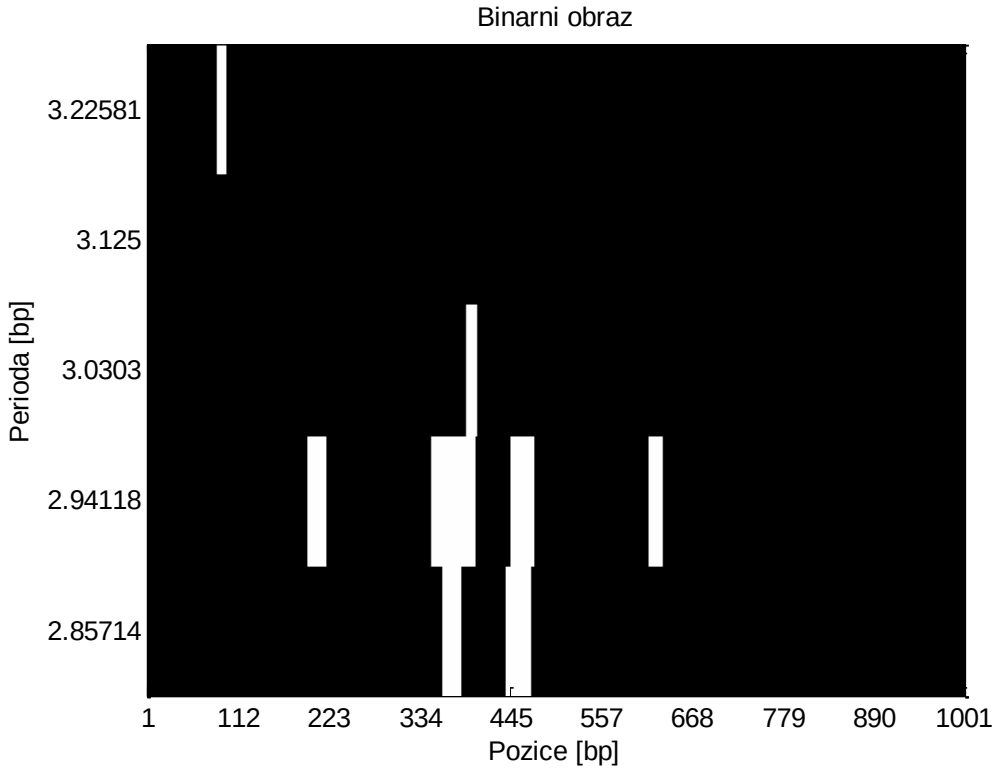
Obrázek 13: Srovnání spektrogramů s / bez normalizace, 1*kbp*-2*kbp*, (NC_003281).

4.1.4 Tvorba binárního obrazu

Pro přesnější zobrazení kódujících úseků a nalezení jejich pozice, využijeme prahování spektrogramu. Prahování je opět provedeno pouze pro úsek kolem třetí periody a výsledek je vykreslen do samostatného obrázku s názvem *Binarni spektrogram*. Výběrem oblasti kolem periody 3 se zbavíme toho, abychom detekovali úseky o vyšší intenzitě, které se nacházejí i v jiných periodách a detekovali pouze kódující oblasti. Jednotlivé barevné složky spektrogramu nám ilustruje škála barev umístěná vedle spektrogramu a vykreslená pomocí funkce *colorbar*. My máme zvolenu 256 úrovňovou barevnou paletu *jet*. Díky této paletě si můžeme nastavit potřebnou prahovou hodnotu, pro identifikaci požadovaných úseků ve spektrogramu, které se projevují určitou barvou.

Např. hodnoty v matici pod 150 jsou uloženy jako číslo 0 a značeny v binárním obraze černě, hodnoty nad 150 jsou uloženy jako číslo 255, značeny bíle a představují kódující oblast (viz Kapitola 3.3). Výsledný binární obraz je vykreslen pomocí funkce *colormap(gray(256))*.

Takto získáme ze spektrogramu binární obraz, ze kterého jsme schopni přesně lokalizovat polohu a výskyt kódujících úseků v sekvenci DNA. Obrázek 14 představuje binární obraz kolem třetí periody pro sekvenci 1*kbp* – 2 *kbp* III. chromozomu *C. elegans*. Bílé oblasti značí kódující úseky.



Obrázek 14: Binární obraz C. elegans, 1*kbp*-2*kbp* (NC_003281).

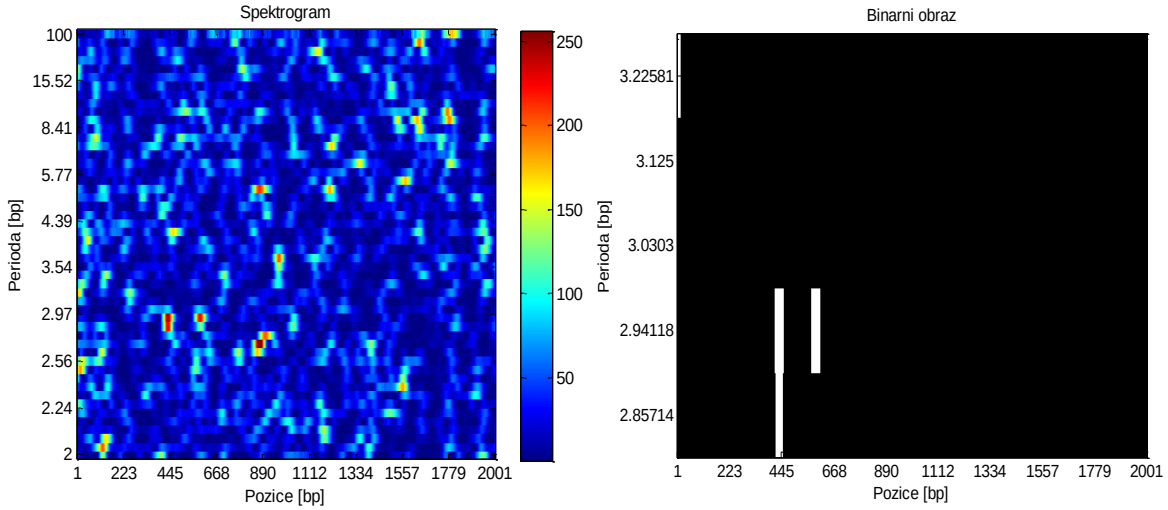
4.1.5 Detekce pozic kódujících úseků

Zjištění polohy kódujících úseků v sekvenci DNA provedeme ze vzniklého binárního obrazu, ve kterém jasně vidíme bíle značené úseky, jejichž pozice požadujeme. Bílé oblasti vyskytující se těsně nad sebou, jsou sečteny a považovány za jeden kódující úsek.

V programovém prostředí je detekce pozic realizována tak, že nejprve zjistíme umístění hodnot, rovných číslu 255, v binárním obraze, tyto hodnoty značí kódující úseky. Následně je zjištěna vzdálenost mezi těmito pozicemi a dle toho stanovena podmínka. Podmínka nám říká, že pokud vzdálenost mezi dvěma sousedními pozicemi je rovna hodnotě 1, jedná se o začátek kódujícího úseku a ten má délku takovou, dokud trvá podmínka. Při překročení vzdálenosti větší než 1, jde o ukončení kódujícího úseku a uložení poslední pozice, která podmínku splňovala. Je však nutné přetransformovat pozice úseků z binárního obrazu na skutečné pozice v sekvenci DNA. K tomu nám posloužila funkce *linspace.*

Výsledky pozic jednotlivých kódujících úseků jsou ukládány do proměnné *pozice_exonu* a vypisovány v okně *Command Window.* Úseky jsou zapisovány do sloupečků, kdy 1. řádek v každém sloupci představuje začátek úseku a 2. řádek značí ukončení daného úseku v sekvenci DNA. Pozice jsou udávány v bazích párů [*bp*].

Obrázek 15 představuje formu vykreslení spektrogramu, binárního obrazu a výpis pozic kódujících úseků v sekvenci DNA pro III. chromozom *C. elegans* (NC_003281) v délce sekvence 1*kbp*-3*kbp*. Sekvence je zpracována programem *EIIP_spektrogram.m.*

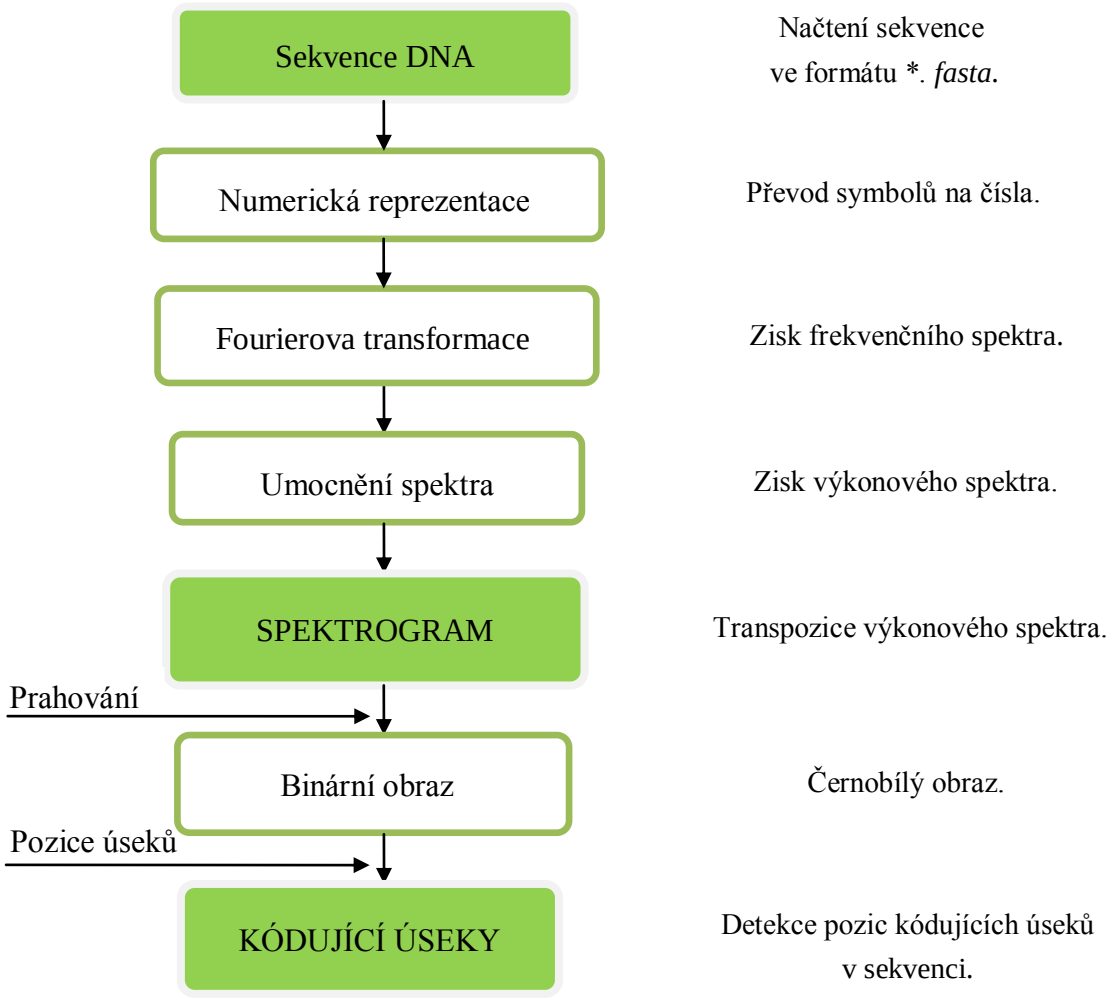



```
pozice_exonu =

     1   417   570
    16   455   607
```

Obrázek 15: Výsledky pro programu *EIIP_spektrogram.m,* 1*kb*-3*kbp,* (NC_003281).

Grafické znázornění postupu tvorby spektrogramu a získání pozic kódujících úseků je naznačeno na Obrázku 16.



Obrázek 16: Blokové schéma vytvoření spektrogramu a získání kódujících úseků.

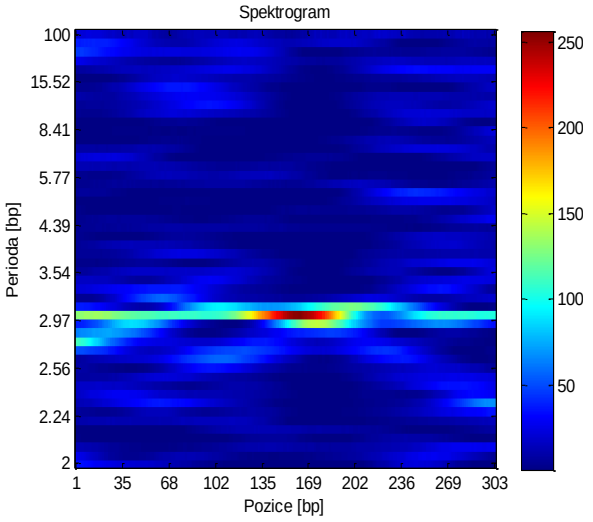
4.2 Výsledky jednotlivých programů

Na níže uvedených obrázcích jsou uvedeny barevné spektrogramy, pro konkrétní sekvence, zpracované pomocí 4 programů (viz Kapitola 4.1). Programy jsou založeny na rozdílných numerických reprezentacích sekvence DNA (viz Kapitola 1). Dále jsou uvedeny výsledky detekce pozic kódujících úseků pomocí námi vytvořených programů a srovnány s výsledky v databázi NCBI.

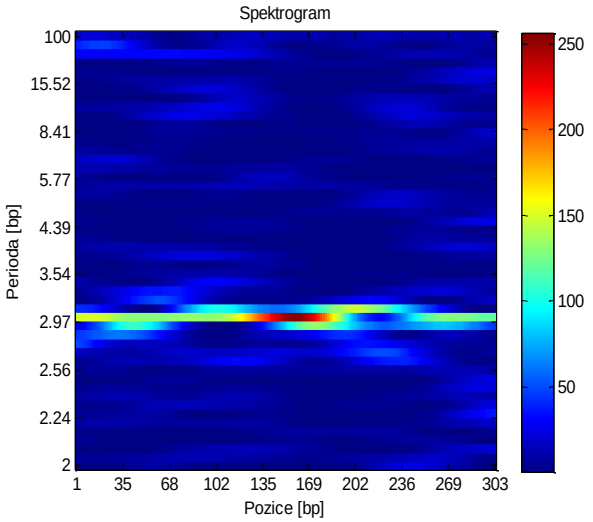
4.2.1 Oryza sativa

Sekvence DNA Rýže seté (*O. sativa*), kóduje v celé délce 303*bp* nejmenovaný protein a navíc se projevuje velkým množstvím opakujících se úseků na periodě 3. Proto je dobrým ukazatelem pro zjištění přesnosti námi navržených metod.

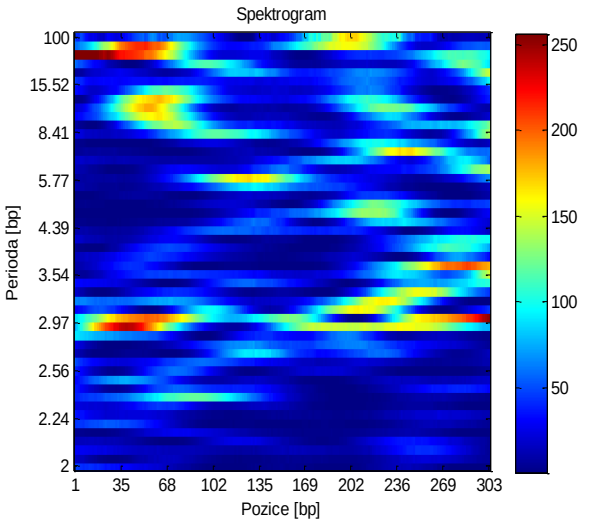
a) 4D numerická reprezentace



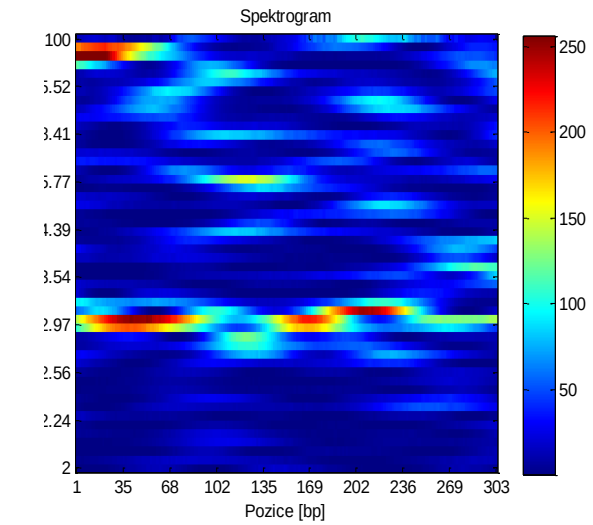
b) EIIP mapování



c) reprezentace pomocí úhlu zkroucení



d) reprezentace pomocí pevnosti zakřivení



Obrázek 17: Barevné spektrogramy pro celou sekvenci *O. sativa* (X64775).

U všech programů byla nastavena délka okna *wl*=100, *posun*=1 a *prah*=170. Tabulka 4 prezentuje získanou pozici opakujícího se úseku v sekvenci DNA pomocí různých programů. Dosažené výsledky jsou porovnány s výsledky v databázi NCBI.

Tabulka 4: Pozice úseků v sekvenci *O. sativa*.

Program	Pozice opakujícího se úseku [*bp*]
Binarni_spektrogram.m.	131… 188
EIIP_spektrogram.m.	124… 183
PropellerTwist_spektrogram.m.	16… 73, 260… 303
BendingStiffness_spektrogram.m.	4… 89, 146… 239
databáze NCBI	142… 186

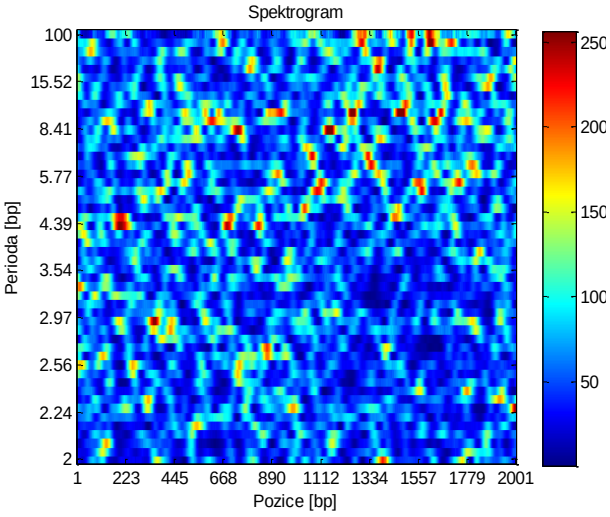
Nejblíže správnému řešení je metoda zpracování sekvence pomocí 4D binární reprezentace, která se liší pouze o několik jednotek pozic. Blízko je též metoda EIIP mapování, ale ostatní dvě metody detekovali úsek naprosto mimo, což je vidět i na první pohled z barevných spektrogramů.

4.2.2 III. chromozom Caenorhabditis elegans

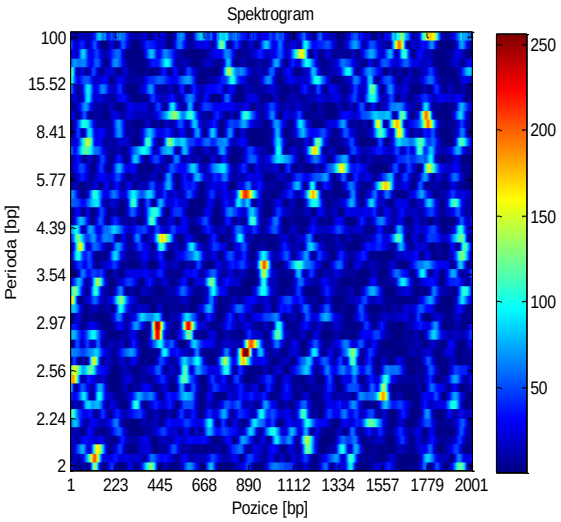
Námi použitá sekvence DNA Háďátka obecného (*C. elegans*), je v praxi hojně využívána. Jelikož bylo u tohoto hlísta identifikováno přes 18 tisíc genů kódujících proteiny, stal se tak dobrým indikátorem pro hodnocení experimentálních výsledků a je významným nástrojem molekulární a vývojové biologie. Jedná se rovněž o první mnohobuněčný organismus, u něhož byl osekvenován kompletní genom. Spolu s octomilkou nebo potkanem je jedním z modelových organismů využívaných při výzkumu. [35]

Pro zpracování sekvence III. chromozomu *C. elegans* v programovém prostředí MATLAB, je vybrán pouze úsek (1*kbp* – 3*kbp*) z celé sekvence tohoto chromozomu, který obsahuje přes 13*Mbp*. Pozice získaných kódujících úseků jsou přepočteny na přesné pozice v celé sekvenci, ne pouze v dané části sekvence. V programech jsme použili délku okna *wl*=100, *posun*=1 a *prah*=170.

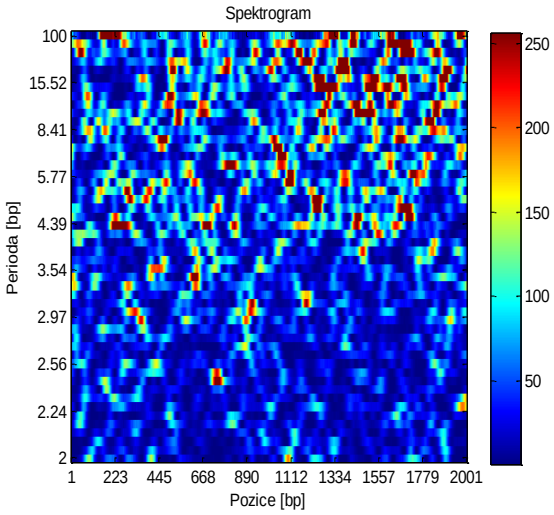
a) 4D numerická reprezentace



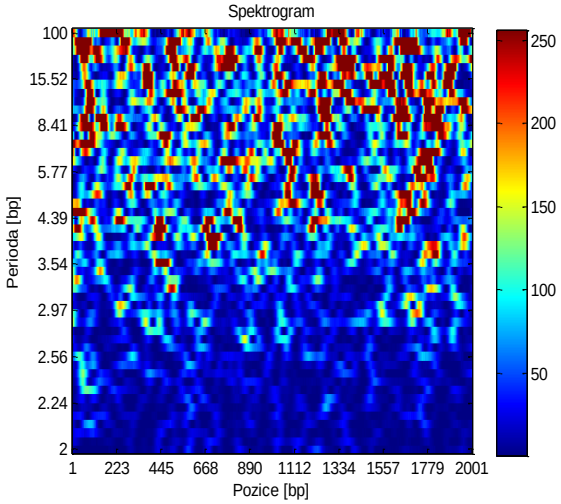
b) EIIP mapování



c) reprezentace pomocí úhlu zkroucení



d) reprezentace pomocí pevnosti zakřivení



Obrázek 18: Barevné spektrogramy pro III. chrom. *C. elegans*, 2k*bp*, (NC_003281).

Tabulka 5 představuje seznam získaných pozic kódujících úseků v sekvenci DNA získaných pomocí 4 programů (viz Kapitola 4.1). Dosažené výsledky jsou porovnány s výsledky v databázi NCBI, kdy správnému řešení se nejvíce přiblížil program založený na numerické metodě zpracování sekvence pomocí reprezentace pevnosti zakřivení dvoušroubovice DNA. Metoda však nedetekovala přesné pozice kódujících úseků, ale na rozdíl od dalších tří metod, se pozice lišili jen nepatrně.

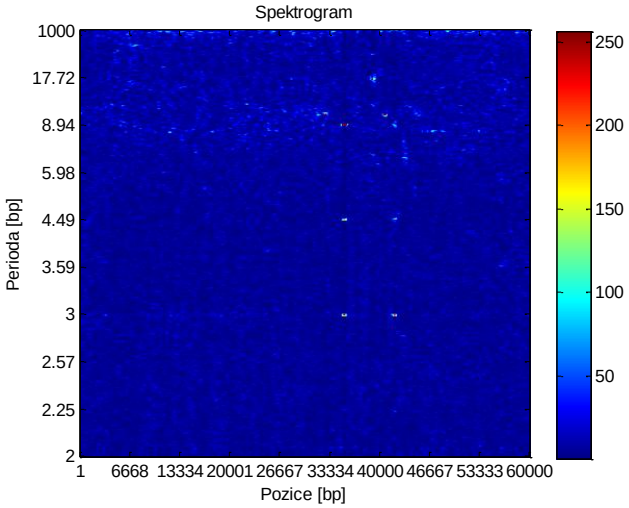
Tabulka 5: Pozice kódujících úseků v sekvenci *C. elegans*, 2k*bp*.

Program	Pozice kódujících úseků [*bp*]
Binarni_spektrogram.m.	1334… 1379, 1421… 1440
EIIP_spektrogram.m.	1002… 1005, 1008… 1008, 1419… 1453, 1573… 1604
PropellerTwist_spektrogram.m.	1297… 1367, 1873… 1929, 2172… 2203, 2909… 2915
BendingStiffness_spektrogram.m.	1001… 1004, 1228… 1259, 1365… 1375, 1477… 1521, 2542… 2555, 2667… 2667, 2669… 2669, 2674… 2743, 2776… 2812
databáze NCBI	1271… 1507, 1558… 2167, 2681… 2764, 2817… 2917

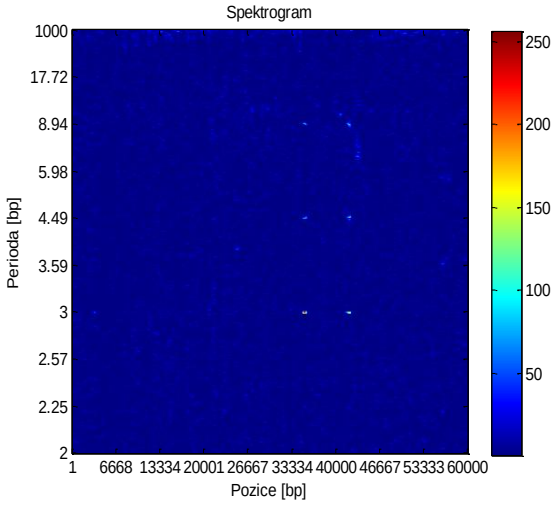
Pro přesnější detekci kódujících pozic ze sekvence, by bylo možné nastavit kratší délku posuvného okna, jelikož dlouhé okno má za následek značné kompromisy v umístění kódujících úseků v sekvenci a také znamená delší dobu výpočtu.

Na další sadě čtyř spektrogramů jsme použili stejnou sekvenci III. chromozomu *C. elegans*, ale zvolili jsme větší délku sekvence, konkrétně 60*kbp*. V programech jsme nastavili také větší délku okna *wl*=1000, jelikož velikost posuvného okna, by měla odpovídat délce sekvence. Pokud je okno příliš krátké, program nenajde požadované kódující úseky a z toho důvodu je spektrogram nevěrohodný. Také se prodloužuje doba výpočtu. Naopak při délce okna moc velké, je spektrogram nepřehledný a špatně se z něj hodnotí vlastnosti dané sekvence DNA. Další parametry v programech jsou nastaveny takto: *posun*=100, *prah*=170.

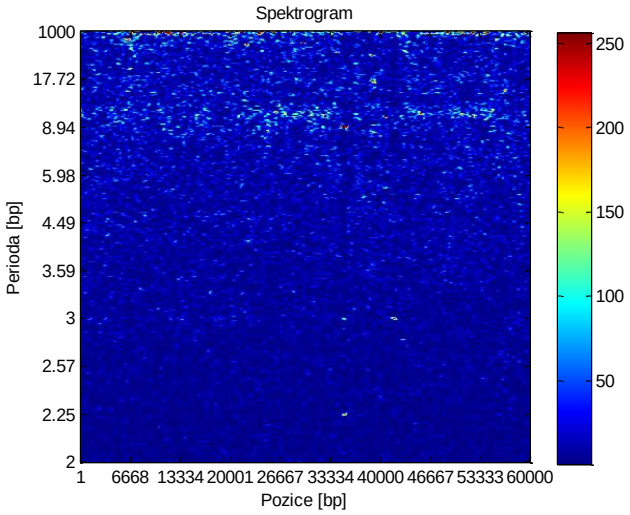
a) 4D numerická reprezentace



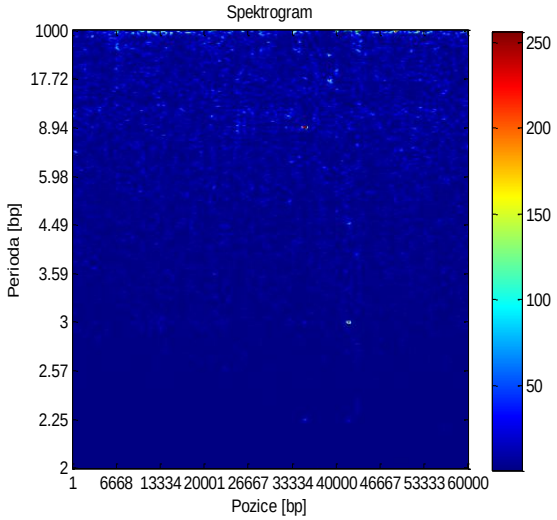
b) EIIP mapování



c) reprezentace pomocí úhlu zkroucení



d) reprezentace pomocí pevnosti zakřivení



Obrázek 19: Barevné spektrogramy pro III. chrom. *C. elegans*, 60k*bp*, (NC_003281).

Tabulka 6: Pozice kódujících úseků v sekvenci *C. elegans*, 60k*bp*.

Program	Pozice kódujících úseků [*bp*]
Binarni_spektrogram.m.	35043… 35348, 41766… 41970
EIIP_spektrogram.m.	35043… 35348
PropellerTwist_spektrogram.m.	41664… 42072
BendingStiffness_spektrogram.m.	41664… 41970

Tabulka s hodnotami pozic uvedených v databázi NCBI není uvedena, jelikož v této délce sekvence se vyskytuje přes 11 krátkých úseků kódujících proteiny. Naše metody je nebyli schopné detekovat a zachytili pouze úseky o větší délce.

Díky spektrogramu vytvořenému z celé sekvence DNA, jsme schopni rozlišit, zda se jedná o tandemovou repetici nebo kódující úsek. Proto sestrojujeme spektrogram z celé sekvence a ne jen z části kolem třetí periody.

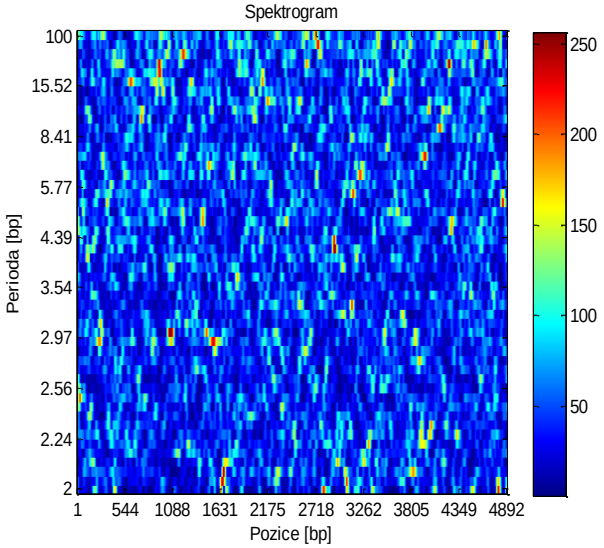
I když v tomto případě, bychom mohli z vytvořených obrázků usoudit, že se jedná o tandemové repetice, ve skutečnosti se na pozici 41063… 41971*bp* opravdu vyskytuje kódující úsek. Proto je nutné si správnost získaných výsledků překontrolovat s výsledky v databázi NCBI.

4.2.3 Interleukin 10

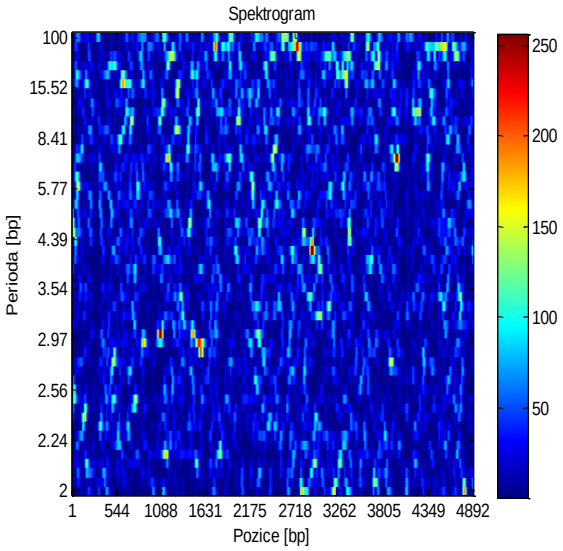
Interleukin 10 (*IL 10*) je gen, vyskytující se v sekvenci *Homo sapiens*. Jeho délka sekvence je 4892*bp*, takže se jedná v podstatě o velmi krátký úsek. Protein kódovaný tímto genem je cytokinin, produkovaný hlavně monocyty a v menší míře také lymfocyty. Tento cytokinin má víceplošné účinky při regulaci imunitních onemocnění a zánětech u člověka. Studie na myších objevila další funkci tohoto cytokininu a tou je, že funguje jako základní imunoregulátor ve střevním traktu. Mutace v tomto genu jsou spojeny se zvýšenou citlivostí na HIV infekci a revmatoidní artritidy.

Pro zpracování sekvence genu *IL10* v programovém prostředí MATLAB, jsme nastavili proměnné následovně: délku okna *wl*=100, *posun*=1 a *prah*=140.

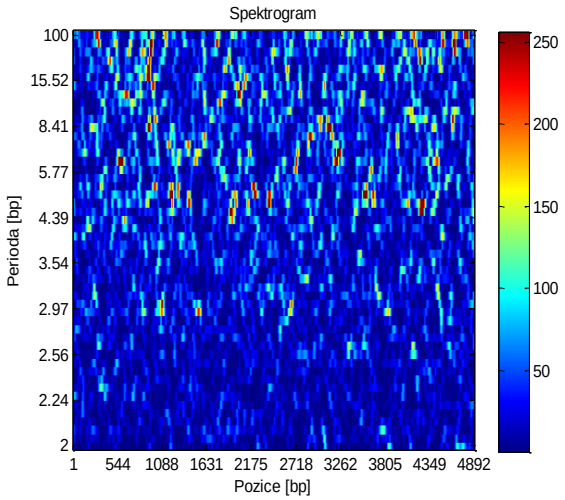
a) 4D numerická reprezentace



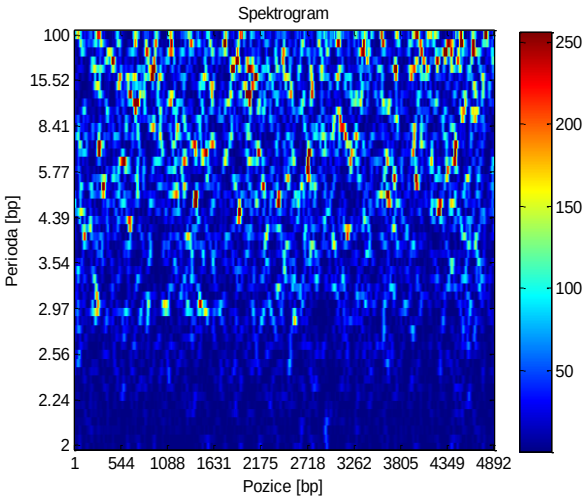
b) EIIP mapování



c) reprezentace pomocí úhlu zkroucení



d) reprezentace pomocí pevnosti zakřivení



Obrázek 20: Barevné spektrogramy pro celý gen *IL 10* (NC_000001.10).

Tabulka 7 obsahuje soubor pozic kódujících úseků v sekvenci daného genu. Pozice jsme získali pomocí 4 programů blíže specifikovaných v Kapitole 4.1. Námi dosažené výsledky jsou porovnány s výsledky v databázi NCBI.

Tabulka 7: Pozice kódujících úseků v sekvenci genu *IL 10*.

Program	**Pozice kódujících úseků [*bp*]**
Binarni_spektrogram.m.	229… 292, 862… 875, 1029… 1100, 1244… 1245, 1453… 1491, 1597… 1597, 1599… 1600, 1614… 1614, 2253… 2262, 2264… 2264, 2582… 2610, 2653… 2657, 3010… 3037, 3114… 3125, 3481… 3505, 3827… 3858
EIIP_spektrogram.m.	858… 890, 1047… 1104, 1456… 1489, 1516… 1590, 2265… 2270
BendingStiffness_spektrogram.m.	250… 295, 864… 872, 1036… 1084, 1086… 1087, 1445… 1492, 1515… 1547, 2372… 2380, 2552… 2581
PropellerTwist_spektrogram.m.	856… 864, 1063… 1106, 1503… 1555, 2629… 2677, 3688… 3700, 3714… 3734, 3814… 3855, 4424… 4454

databáze NCBI	60… 224, 1080… 1139, 1436… 1588, 2601… 2666, 3767… 3859

Nejblíže se experimentálním výsledkům přiblížil program založený na EIIP mapování sekvence. Metoda opět nedetekovala přesné pozice kódujících úseků, ale byla jim velmi blízko. Ostatní metody detekovali i jiné, ve skutečnosti nekódující pozice.

ZÁVĚR

Náplní této bakalářské práce je vyhledat pozice kódujících úseků v sekvenci DNA eukaryotických organismů, pomocí analýzy spektrogramu. V samém začátku práce jsme seznámeni s tematickou oblastí numerického zpracování sekvencí. Následuje výčet metod, pomocí kterých lze vyhledávat požadované úseky ze sekvence. Jedná se o vyhledávání pomocí čísel, anebo pomocí znaků. V teoretické části je také uveden způsob zhotovení spektrogramu ze sekvencí DNA a výčet různých vzorů, které jsme z něj schopni detekovat.

Praktická část práce je popsána v samostatné čtvrté kapitole. Tato kapitola je zaměřena na realizaci konkrétních programů v programovém prostředí MATLAB, vedoucích k sestrojení spektrogramu. Jednotlivé programy jsou založeny na rozdílných numerických reprezentacích sekvence DNA a následného číslicového vyhledávání kódujících úseků.

Vytvořením čtyř odlišných programů, jsme mohli hodnotit jejich schopnosti a spolehlivost při vyhledávání úseků ze sekvencí. Pro porovnání programů jsme používali stejné sekvence a nastavili shodné parametry. Námi získané výsledky jsme následně posuzovali s již známými výsledky uvedenými v databázi NCBI pro každou sekvenci DNA.

Jak jsme zjistili z porovnání výsledků, náš vytvořený algoritmus není zcela přesný pro vyhledávání pozic kódujících úseků. Nejpřesnější výsledky jsme získali pro sekvenci *O. sativa*, kterou jsme použili jako modelovou sekvenci, jelikož obsahuje významnou opakující se oblast, kterou jsme jednotlivými metodami detekovali. V ostatních případech programy nalezli pouze přibližné pozice, lišící se od několika jednotek, přes stovky až k nalezení oblastí, které proteiny vůbec nekódují. Přesnost programů je též dána velikostí okna, což je velmi složité pro nastavení, jelikož delší okno nám přináší jen přibližné výsledky pozic, ale rychlejší vyhotovení spektrogramu. Jako nejvhodnější a nepřesnější program, který jsme navrhli, je založen na numerickém mapování sekvence pomocí EIIP, elektron – iontového potenciálu DNA.

Vypátrali jsme také, že při použití delší sekvence, řádově nad desítky *kbp* a větší velikosti posuvného okna, není efekt tří bázové periodicity dominantní. Proto je lepší vyhledávat kódující úseky z kratších sekvencí a s menší velikostí posuvného okna. Spektrogram a výsledky pozic kódujících úseků jsou poté mnohem věrohodnější.

Díky vytvořeným programům a teoretickému základu pro zpracování a hodnocení sekvencí DNA, získává uživatel možnost samostatně hodnotit různé, libovolně zvolené sekvence, na základě získaného barevného spektrogramu.

Při stanovení dalšího postupu rozšíření této práce, by byla možnost praktické realizace dalších numerických reprezentací DNA, vyhledávání kódujících úseků pomocí znakových metod, rozšíření a propracování samotných, již vytvořených, programů v programovém prostředí MATLAB.

SEZNAM LITERATURY

[1] Aktuální genetika. *Genetické haraburdí - repetitivní DNA* [online]. 2005-2006 [cit. 2012-11-27]. Dostupné z: http://biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm#obr1

[2] Barevné modely: RGB model. *PEKO KARTON* [online]. 2012 [cit. 2012-11-28]. Dostupné z: http://www.pekokarton.cz/technologie/barevne-modely

[3] *Biochemické pojmy - výkladový slovník*. 2004. ISBN 80-7080-551-X. Dostupné z: http://vydavatelstvi.vscht.cz/knihy/uid_es-002_v1/motor/main.anotace.html

[4] Bioinformatika a DNA: Informačné technologie, Medicína. *Posterus* [online]. 2012, roč. 5, č. 8 [cit. 2012-11-27]. ISSN 1338-0087. Dostupné z: http://www.posterus.sk/?p=13501

[5] CRISTEA, Paul Dan. Large scalee features in DNA genomic signals. *Signal processing*. 2003, s. 18.

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

[7] *Časově frekvenční analýza signálů*. Fakulty mechatroniky, informatiky a mezioborových studií, Technické univerzity v Liberci, 2007. Dostupné z: http://www.rss.tul.cz/ftppub/tdg/P6_cas_freq_analyza.pdf

[8] DATTA, Suprakash, Amir ASIF a Haoyuan WANG. Prediction of Protein Coding Regions in DNA sequences Using Fourier Spectral Characteristics. *Datta coding region Fourier transformation*. 2009.

[9] DIMITROVA N., CHEUNG Y., ZHANG M., *Analysis and Visualization of DNA spectrograms*

[10] TRIFONOV E. N., "The pitch of chromatin DNA is reflected in its nucleotide sequence," *Proc. of the Nat. Acad. Sci., USA*, vol. 77, pp. 3816–3820, 1980.

[11] Genetika: Translace. *Translace a proteosyntéza* [online]. 2011 [cit. 2012-11-27]. Dostupné z: http://genetika.wz.cz/transl.htm

[12] HERZEL, H., O. WEISS a E. TRIFONOV. 10 - 11 bp periodicities in complement genomes reflect protein structure and DNA folding. *Bioinformatics*. 1999, s. 187-193.

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

[14] JOHN, Radek. Dědičnost bez DNA? Nenápadný červ popírá zákony genetiky. *Evoluční triky* [online]. 2011 [cit. 2013-03-17]. Dostupné z: http://www.tyden.cz/rubriky/veda/priroda/dedicnost-bez-dna-nenapadny-cerv-popira-zakony-genetiky_220638.html

[15] KODÍČEK, Milan. Exon. *Vydavatelství VŠCHT Praha*. Dostupné z: http://vydavatelstvi.vscht.cz/knihy/uid_es-002_v1/hesla/exon.html

[16] KWAN, Hon. Numerical Representation of DNA Sequences. *Department of Electrical and Computer Engineering University of Windsor*. s. 4

[17] LIN, Shang-Ching. Digital Signal Processing for DNA Sequence Analysis. National Taiwan University, Taipei, Taiwan, s. 1-31.

[18] MIHULKA, Stanislav. *Objective Source E- Learning: Po stopách intronů hub* [online]. 2005[cit. 2013-03-18]. Dostupné z: http://www.osel.cz/index.php?obsah=6&akce=showall&clanek=1243&id_c=2919

[19] Parametrické a adaptivní metody zpracování signálů: Parametrické metody. *České vysoké učení tecnické v Praze: Fakulta elektrotechnická* [online]. s. 14 [cit. 2012-11-28]. Dostupné z: www.comtel.cz/files/download.php?id=3370

[20] *Rozdíly mezi prokaryoty a eukaryoty* [online]. 2009 [cit. 2012-11-28]. Dostupné z: http://www.botanika-puchnerova.estranky.cz/clanky/eukaryota/rozdily-mezi-prokaryoty-a-eukaryoty.html

[21] Rychlá Fourierova transformace (FFT) pro AVR. *Elektronika* [online]. [cit. 2013-03-17]. Dostupné z: http://elektronika.kvalitne.cz/ATMEL/necoteorie/transformation/AVRFFT/AVRFFT.html

[22] SNUSTAD, D. Peter a Michael J. SIMMONS. *Genetika*. 5th ed. Brno: Masarykova univerzita, 2009, 297 - 315. ISBN 978-80-210-4852-2.

[23] SONG, Nancy a Hong YAN. Autoregressive Modeling of DNA Features for Short Exon Recognition. *IEEE International Conference on Bioinformatics and Biomedicine*. 2010.

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

[25] ŠÍPEK, Antonín. Translace. *Genetika - Biologie* [online]. 2012 [cit. 2013-05-27]. Dostupné z: http://www.genetika-biologie.cz/translace

[26] ŠÍPEK, Antonín. Mapování genomů. *Genetika - Biologie* [online]. 2012 [cit. 2012-11-27]. Dostupné z: http://www.genetika-biologie.cz/mapovani-genomu

[27] TRIFONOV, E.N. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A 249*. 1998.

[28] TROUSILOVÁ, Alžběta. Deoxyribonukleová kyselina. *Novinky.cz*. Dostupné z: http://tema.novinky.cz/deoxyribonukleova-kyselina

[29] TŮMA, Jiří. PARAMETRICKÁ METODA VÝPOČTU FREKVENČNÍCH SPEKTER SIGNÁLŮ. *7 th International Scientific - Technica l Conference - PROCESS CONTROL 2006*. 2006, s. 9. Dostupné z: http://homel.vsb.cz/~tum52/publications/TumaFt_CZ.pdf

[30] ÚSTAVU BIOLOGIE A LÉKAŘSKÉ GENETIKY 1. LF UK A VFN. *Analýza DNA*. Praha. Dostupné z: http://biol.lf1.cuni.cz/09AnDNA2.pdf

[31] WANG L., Localizing triplet periodicity in DNA and cDNA sequences, Bioinformatics 2010, [online], [cit. 2013-02-15], Dostupné na internetu: http://www.biomedcentral.com/1471-2105/11/550

[32] ZELINKA, Jiří a Jan KOLÁČEK. Jak pracovat s MATLABem. *Výpočetní matematické systémy* [online]. s. 7-26 [cit. 2013-03-20]. Dostupné z: http://www.math.muni.cz/~kolacek/vyuka/vypsyst/navod.pdf

[33] ZHAO, Zhongming a Leng HAN. CpG islands: algorithms and applications in methylation studies. *CpG island* [online]. 2005-2006 [cit. 2012-11-27]. Dostupné z: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2679166/tivni_dna.htm#obr1

[34] ZHOU, H., DU, L., YAN, H., Detection of tandem repeats in DNA sequences based on parametric spectral estimation, IEEE Transactions on information technology in biomedicine, 2009

[35] ZÖLZER, Friedo. *Radiologie buňky*. České Budějovice, 2007. Doplňkové texty pro posluchače kombinované formy studia studijního programu „Ochrana obyvatelstva". JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH, Zdravotně sociální fakulta.

[36] ZVÁROVÁ, Jana a Ivan MAZURA. *Biomedicínská statistika: STOCHASTICKÁ GENETIKA*. Praha: Karolinum Praha 1, Ovocný trh, 2002. ISBN 80-246-0264-4.

SEZNAM ZKRATEK

A	adenin
AR	autoregresní model
bp	bazí párú
C	cytosin
C*p*G	struktura cytosin – fosfodiesterová vazba – guanin
DFT	diskrétní Fourierova transformace
DNA	deoxyribonukleová kyselina
EIIP	mapování pomocí elektron-iontového potenciálu DNA
G	guanin
mRNA	mediátorová RNA
NCBI	nemoderovaná biologická databáze
RGB	model složený z 3 vrstev (červená, zelená, modrá)
RNA	ribonukleová kyselina
T	thymin