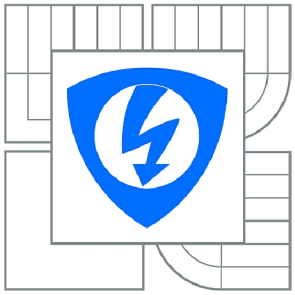
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ TANDEMOVÝCH REPETIC V DNA POMOCÍ NUKLEOTIDOVÝCH DENZITNÍCH VEKTORŮ

TANDEM REPEAT FINDING IN DNA USING NUCLEOTIDE DENSITY VECTORS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

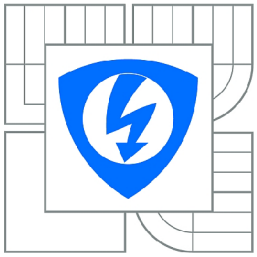
AUTOR PRÁCE — MICHAL HRACHO

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Student:	Michal Hracho	***ID:***	147478
Ročník:	3	***Akademický rok:***	2013/2014

NÁZEV TÉMATU:

Vyhledávání tandemových repetic v DNA pomocí nukleotidových denzitních vektorů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma tandemových repetic v sekvencích DNA. Zaměřte se především na moderní metody vyhledávání. 2) Na vhodně zvoleném souboru dat a uměle vytvořených sekvencích otestujte alespoň 3 volně dostupné vyhledávače. 3) V libovolném programovém prostředí naprogramujte převod sekvencí DNA do nukleotidových denzitních vektorů. 4) Prověřte možnosti vyhledávání tandemových repetic pomocí analýzy nukleotidových denzitních vektorů a proveďte analýzu na souboru dat. 5) Výsledky porovnejte a diskutujte s volně dostupnými vyhledávači.

DOPORUČENÁ LITERATURA:

[1] KRISHNAN, A., TANG, F. Exhaustive whole-genome tandem repeats search. Bioinformatics, 2004, roč. 20, no. 16, s. 2702-2710.
[2] HAUTH, A. M., JOSEPH, D. A. Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinformatics, 2002, roč. 18, suppl. 1, s. S31-S37.

Termín zadání:	10.2.2014	***Termín odevzdání:***	30.5.2014

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

ABSTRAKT

Tato práce se zabývá výskytem tandemových repetic v makromolekule DNA a možnostmi jejich vyhledávání v genomu. Součástí této bakalářské práce je krátké seznámení se strukturou DNA, popis tandemových repetic a jejich definice, členění a jejich vliv a význam v živém organismu. Dále práce obsahuje úvod ke způsobům vyhledávání repetic a popis některých internetových vyhledávačů.

KLÍČOVÁ SLOVA

Tandemové repetice, repetitivní DNA, vyhledávání tandemových repetic, internetové vyhledávače tandemových repetic, denzita nukleotidů

ABSTRACT

This thesis deals with the occurrences of tandem repeats in DNA macromolecule and with possibilities of their search in genome. Part of this work will be mainly short introduction into structure of DNA, a description of tandem repeats and their definition, classification and their impact and importance in living organism. Furthermore introduction to their search, description of internet search engines.

KEYWORDS

Tandem repeats, DNA, repetitive DNA, tandem repeats search, internet tandem repeats search engines, nucleotide density

HRACHO, M. *Vyhledávání tandemových repetic v DNA pomocí nukleotidových denzitních vektorů.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2014. 47 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Vyhledávání tandemových repetic v DNA pomocí nukleotidových denzitních vektorů jsem vypracoval 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 autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji vedoucí své bakalářské práce Ing. Denise Maděránkové za odborné vedení, ochotu další a cenné rady, které mi v průběhu zpracování bakalářské práce věnovala.

V Brně dne ....

(podpis autora)

Obsah

SEZNAM OBRÁZKŮ

ZDROJE OBRÁZKŮ

Obr. 1: http://cronodon.com/images/DNA_Structure.jpg

Obr. 2: http://www.lsa.umich.edu/UMICH/eeb/Home/People/snakes.jpg

Obr. 3: http://upload.wikimedia.org/wikipedia/commons/7/78/D1S80Demo.gif

Obr. 4: http://www.nature.com/scitable/nated/content/33589/10.1073_pnas.0408118101-f3b_thumb_0.jpg

Obr. 5: http://tandem.sci.brooklyn.cuny.edu/Algorithm.do

Obr. 6: http://tandem.bu.edu/trf/figure1.gif

Obr. 7: http://tandem.bu.edu/trf/figure2.gif

SEZNAM TABULEK

ÚVOD

Tato práce se zabývá vyhledáváním tandemových repetic v sekvencích makromolekul DNA.

V současné době prožívá biologická věda genomika, jakožto obor genetiky, který se zabývá především studiem a popisem genomů organismů, exponenciální rozmach, způsobený především rozvojem technologií a prudkým poklesem jejich cen. Genomika zahrnuje sekvenování DNA, anotace genů neboli strukturní a funkční zařazování, a genetické mapování.

Genomika se uplatňuje v mnoha oblastech lékařské diagnostiky, jako jsou DNA fingerprinting, jinak řečeno genetická daktyloskopie, genové inženýrství, fylogenetika atd. Používá spoustu metod získávání informací z různých zdrojů - oblastí DNA a jejich zpracování. Zpracováním informací se zabývá bioinformatika, věda zaměřená na práci s biologickými daty, jako jsou např. sekvence nukleových kyselin v DNA/RNA, sekvence aminokyselin, struktura proteinů, funkce a exprese genů atd.

Díky analýze a anotaci genomu jsme schopni určit význam a funkci jednotlivých genů a význam různých struktur vyskytujících se v DNA. Jednou z těchto významných struktur jsou tandemové repetice. Jedná se o poměrně krátké, opakující se úseky nukleotidů. Tandemových repetic využívá především DNA fingerprinting v kriminalistice a při určování otcovství a jako nástroj účinného způsobu provádění PCR, což je metoda rychlého a snadného rozmnožení požadovaného úseku DNA.

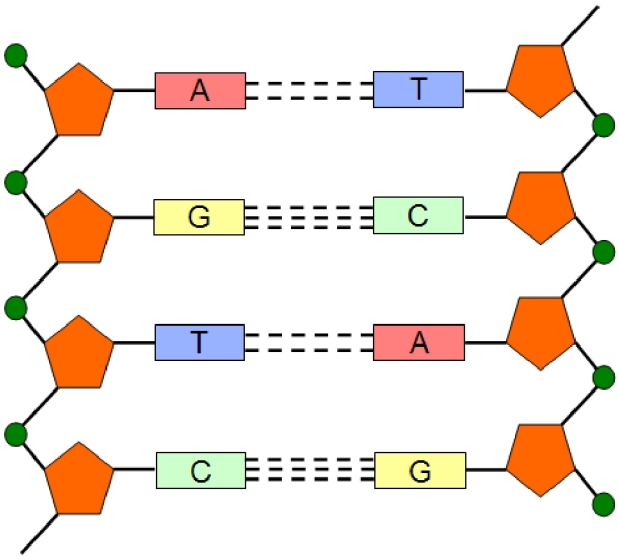
1 DNA

Molekuly DNA, popř. RNA jsou nositelem genetické informace, to znamená, že obsahují informaci o syntéze proteinů všech živých organismů. Jedná se o jakýsi biologický zdrojový kód všech organismů. Každá buňka obsahuje svou vlastní DNA, která expresí vhodných genů řídí její činnost. DNA obsahuje úseky nesoucí informaci o syntéze bílkoviny; tyto úseky se nazývají geny.

Elementárním informačním prvkem DNA je nukleotid. Jedná se o chemickou sloučeninu monosacharidu pentózy, kyseliny fosforečné a specifické dusíkaté báze. Dva nukleotidy jsou mezi sebou vždy spojeny fosfodiesterovými vazbami, které spojují fosfátovou část (5'konec) s cukernatou částí (3'konec) nukleotidu, a tvoří tak kostru jednoho vlákna polynukleotidu.

Molekuly DNA jsou nejčastěji dvouvláknové; opačně orientované polynukleotidy jsou mezi jejich jednotlivými bázemi navzájem pospojovány mnohačetnými vodíkovými vazbami podle pravidla komplementarity, čili A=T, C≡G. Výskyt jiných spojení bývá spíše produktem chyby.

Samotná genetická informace je dána sekvencí bází, která určuje pořadí aminokyselin v bílkovině, jež je produktem exprese této genetické informace, takovým způsobem, že sekvence tří po sobě jdoucích bází určuje danou aminokyselinu. DNA se nachází především v jádrech buněk kondenzována do formy chromatinu. Malé množství se rovněž vyskytuje v mitochondriích. [1] [3]



Obrázek 1: Elementární struktura DNA a způsob spojování nukleotidů

převzato z http://cronodon.com/images/DNA_Structure.jpg

Z hlediska stavby DNA rozlišujeme tyto hlavní struktury:

Primární struktura

Jedná se o posloupnost nukleotidů v jednom vlákně DNA. Má hlavní informační hodnotu, neboť sekvence nukleotidů (bází) přímo definuje primární, sekundární a částečně i terciární strukturu výsledných bílkovin. [1]

Sekundární struktura

Dvouvláknová molekula DNA se nejčastěji formuje do tvaru pravotočivé dvoušroubovice. Mimo to existují i tří- a čtyřvláknové struktury, např. forma G-kvartet. Ty se ovšem in vivo vyskytují jen velmi zřídka. [1] [3]

Terciární struktura

Terciární strukturou se rozumí kondenzace a uložení vlákna DNA do formy chromatinu za použití specializovaných strukturních proteinů, jako jsou histony a tzv. „bílkovinná lešení“. Tato struktura se sama skládá z několika úrovní. Výskyt DNA v konkrétních úrovních kondenzace v daný okamžik vypovídá o momentální míře exprese dané oblasti DNA. [1]

1.1 Exprese genu

Vlastnosti a rysy organismů jsou produktem dlouhé dráhy kauzalitních vztahů DNA, produkovaných bílkovin a vnějších vlivů. Ačkoliv každá buňka mnohobuněčného organismu obsahuje celý genetický kód, exprimuje jen několik určitých genů, a to na základě kombinace regulačních procesů – formou již vyprodukovaných bílkovin – s vlivem především vnějšího prostředí. [1] [2] [4]

Transkripcí a translací genetické informace vznikají molekuly bílkovin se strukturními, enzymovými a signálními funkcemi. Jejich specifickými interakcemi jsou dány všechny konečné vlastnosti buňky, tj. struktura, metabolismus, reaktivita atd. [2]

Geny člověka podílející se na kódování bílkovin zabírají jen asi 1,5% celé DNA. Další zlomek DNA kóduje RNA jiného typu než mRNA. Zbytek délky DNA tvoří introny a především tzv. **repetitivní DNA.** [2] [3]

2 REPETITIVNÍ DNA

DNA eukaryot a také člověka obsahuje značný podíl nekódujících sekvencí. Tak jako kódující DNA i nekódující může být unikátní anebo se může nacházet v genomu ve více identických nebo podobných kopiích. Sekvence DNA s vysokým množstvím kopií se nazývají repetitivní sekvence. Tyto oblasti nejsou transkribovány a jejich přesný význam není úplně znám.

Dříve byla tato oblast DNA považována za tzv. „junk DNA" neboli oblast odpadních a nepotřebných sekvencí. V dnešní době již víme, že jsou tyto oblasti nepostradatelné a pokud je o ně buňka připravena, má to fatální následky. Tyto oblasti se podílejí na formování jaderných struktur, jako jsou telomery, centromery, oblasti navázání histonů atd. Dnes se již mezi repetitivní sekvence řadí i některé funkční geny, například geny pro ribozomální RNA a jiné strukturní, přesněji řečeno enzymové využití RNA.

Pokud jsou kopie sekvenčního motivu v blocích, v řadě za sebou, hovoříme o tandemových repeticích. Od nich odlišujeme repetitivní sekvence rozptýlené v genomu jako jednotlivé kopie. [3] [7]

2.1 Rozptýlené repetice – DNA transpozony a retrotranspozony

Krátké shrnutí podstaty rozptýlených repetic:

Naprostá většina rozptýlených repetic vzniká transpozicí sekvencí DNA. Jedná se o přesun celých sekvenci na jiná místa v DNA. [7]

Transpozibilní sekvence DNA se dělí na:

DNA transpozony

DNA transpozony jsou považovány za pozůstatky a akumulace mutací v průběhu fylogenetického vývoje organismů, a tak můžeme najít pouze jejich evolučně staré zbytky, neboli „fosilie". Nicméně aktivní transpozon odvozený z lidských fosilních elementů může být vyroben s použitím informací získaných z lidského genomu i genomu ostatních obratlovců. Tyto transpozony se v genomu pohybují způsobem „cut-and-paste"; jejich množství se tedy žádným závratným způsobem nemění. [7]

Retrotranspozony

Mnohem hojnějšími transpozibilními elementy v lidském genomu jsou retrotranspozony. Ty tvoří až 45% celé délky genomu, podle některých odhadů i více. Mnohem důležitější vlastností retrotranspozonů je to, že jsou v lidském genomu stále aktivní. Tento druh transpozonů se přesunuje transkripcí do RNA a následnou reversní transkripcí do DNA; původní kopie tak zůstává na svém místě, takže se jedná o „copy-and-paste" způsob přesunu. Tímto způsobem se sekvence může dostat na jakékoliv místo v genomu. Během transkripce a reversní transkripce jsou tyto segmenty náchylné k chybám, a tak většinou proces transpozice končí chybnou kopií, která je následně inaktivována. Na dalším šíření se tak podílí opět už jen původní aktivní originál. Ovšem i tak je tímto způsobem genom zatížen obrovským množstvím fosilních členů dané rodiny sekvencí.

Retrotranspozony můžeme ještě dále dělit na autonomní a neautonomní. Autonomní obsahují kód pro proteiny nezbytné k jejich transpozici (především RNA-polymerázu a reversní transkriptázu), neautonomní nikoliv, a nadto jsou závislé na enzymech okolních transpozonů.

Některé retrotranspozony obsahují na svém konci tzv. dlouhé koncové jednotky (long terminal repeats, odtud jejich název - LTR). Jedná se o další repetice uvnitř retrotranspozonů, jež svým složením připomínají proviry skutečných retrovirů, avšak vždy alespoň jeden z genů, potřebných k sestavení infekčních virů chybí nebo je zmutován. [7]

2.2 Tandemové repetice

Tandemové repetice jsou po sobě jdoucími identickými nebo téměř identickými jednotkami sekvencí, které se opakují. Poprvé byly pozorovány při centrifugaci DNA v hustotním gradientu jako oddělenější a vzdálenější frakce než ostatní DNA; jsou označovány jako „satelitní DNA". Označení „satelitní" bylo později doplněno a rozšířeno a dnes zahrnuje i repetitivní sekvence, které nevytvářejí tyto vzdálené frakce – satelity. Periodicky se opakující jednotky jsou obvykle degenerované s různou mírou nepravidelností – při větší míře se mohou nepravidelnosti opakovat a dokonce tvořit repetitivní podjednotky. To samozřejmě platí pro dostatečně dlouhé jednotky repetic. [7] [8]

Tandemové repetice dělíme podle celkové délky na satelity, minisatelity a

mikrosatelity.

Satelity jsou úseky repetic dosahující celkových délek až v řádech Mpb. Jsou hlavní funkční komponentou centromer a tvoří hlavní strukturní prvek heterochromatinu. Rozeznáváme několik hlavních typů satelitů, např. α-satelit s repeticemi o délce 171 bp a celkové délce až 4 Mbp, který tvoří funkční jádro centromer, protože váže kinetochor při buněčném dělení, nebo β o délce 68 bp který se nachází v centromerách chromozomů 1, 9, 13, 14, 15, 21, 22 a chromozomu Y. [7] [10]

Minisatelity dosahují délek až 20 kbp a délka repetic se pohybuje od 10 do 25 bp, někdy i více. Jejich nejznámější výskyt je ve formě telomer a jejich okolních oblastí. Minisatelity jsou silně polymorfní, díky čemuž jsou hojně využívány v DNA daktyloskopii jako genetické markery například pro populační studie a studium genetických chorob nebo genových vazebných skupin. Také se předpokládá jejich úloha jakožto regulátorů genové exprese. [8] [11]

Mikrosatelity jsou nejčastější formou tandemových repetic. Jedná se o úseky o délce repetic nanejvýš 5 bp a celkové délce okolo 100 bp. Vyskytují se obecně v celém genomu. Vyskytují se jako mononukleotidové repetice – nejčastěji A (T na komplementárním vlákně), dinukleotidové repetice – nejčastěji jako CA (GT na komplementárním vlákně) – ty jsou zároveň nejčastěji se vyskytující formou mikrosatelitů, představují 0,5% celého genomu. Mikrosatelity jsou rovněž vysoce polymorfní, což umožňuje jejich využití jako genetických markerů při konstrukci genetických map, v populačních studiích, při testování příbuzenských vztahů a v klinické a forenzní praxi. Rovněž je možno je využít při studiích genových duplikací nebo delecí nebo určení „open reading frame“ při genové predikci. [12] [13] [14]

2.2.1 Polymorfismus

Již několikrát byl zmíněn pojem polymorfismus. O co se jedná? Genetický polymorfismus je současný výskyt dvou nebo více variant genů na jedné oblasti chromozomu, každý s patrnou frekvencí výskytu, jež by měla být minimálně 1%. Genetický polymorfismus je aktivně a stabilně udržován v populacích pomocí přírodního výběru, čímž se liší od přechodného polymorfismu. Polymorfismy vznikají především mutacemi v zárodečných nebo somatických buňkách organismu. Protože většina délky lidské DNA spadá do nekódujících oblastí, nepůsobí na ně selekční tlak, a z toho důvodu jsou rozdíly v oblastech repetitivní DNA mezi jedinci tak obvyklé. [2] [5] [6]

Obrázek 2: Projev genetického polymorfismu ve fenotypu – převzato z http://www.lsa.umich.edu/UMICH/eeb/Home/People/snakes.jpg

2.2.2 Význam tandemových repetic z biologického hlediska

Většina savců má přibližně stejnou sadu genů, nicméně, jak je evidentní, existuje množství významných rozdílů, které rozlišují různé živočišné druhy.

Studie Fondon a Garnera z roku 2004 naznačily, že významným zdrojem variací a odlišností jsou právě tandemové repetice, sekvence VNTR (z anglického variable number tandem repeats). [5]

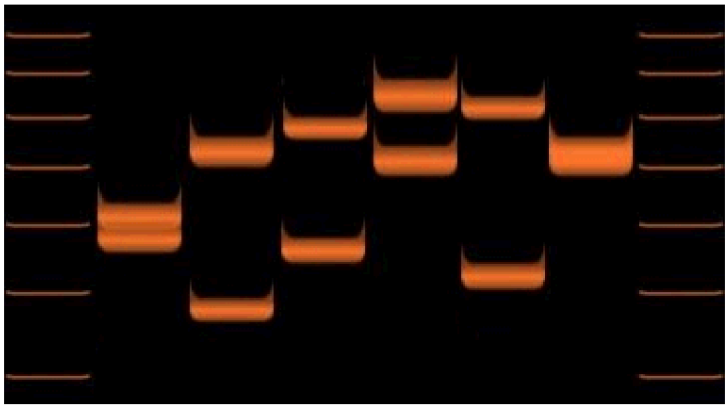
VNTR jsou oblasti genomu organizované jako tandemové repetice, které vykazují variace ve svých délkách mezi různými jednotlivci téhož druhu.

Populace jednoho druhu (včetně lidí) může nést obrovské množství různých variací délek tandemových repetic. Je to způsobeno tím, že tandemové repetice mají velkou tendenci mutovat a frekvence jejich specifických mutací je přibližně 100 000 krát vyšší než frekvence výskytu bodových mutací. [5]

Hlavní typ mutace tandemových repetic je tzv. slipped-strand mispairing, kterýžto typ je zároveň zdrojem variability jejich délek. Je způsoben strukturou tandemových repetic – ty, protože jsou tvořeny krátkými, opakujícími se úseky, se mohou snadněji komplementárně spojit s jinou repeticí na komplementárním vlákně než s tou, která leží „naproti nim“. Takto vznikají na komplementárním vlákně nepřipojené oblasti, a ty se stávají objektem opravných enzymů, které je buď odstřihnou, nebo k nim vytvoří jejich protějšky. V obou případech tak dochází ke změně celkové délky oblasti tandemové

repetice, jinak řečeno satelitu. Mutace slipped-strand mispairing může být omezena bodovými mutacemi, protože ty pak na vlákně tvoří lokální odlišnosti, které způsobují správné spojení vláken, a tím stabilizují danou oblast satelitu. Zde se projevuje pozitivní zpětná vazba, protože stabilizované úseky, které přestávají být náchylné na slipped-strand mispairing akumulují bodové mutace, což dále stabilizuje danou oblast, ale způsobuje ztrátu jejího repetitivního charakteru. Na druhou stranu delece způsobená mutací slipped-strand mispairing může odstranit lokální odlišnost, a tím prohloubit náchylnost oblasti na mutace chybným spojením, čili opět se vyskytuje pozitivní zpětná vazba, která má za následek produkci čistých sekvencí bez bodových změn. Indukcí tak můžeme vyhodnotit, zda v dané oblasti docházelo v průběhu její minulosti k delecím a délkovým variacím. [5] [6] [9] [15] [16]

„Čistota“ tandemových repetic je proto určitým měřítkem míry výskytu selekce v původu daného organismu. Ve studii Fondon a Garnera byly srovnány repetice lidí a psů, přičemž bylo zjištěno, že v 29 z 36 případů jsou psí repetice čistší, čili mající nízké množství bodových mutací. Tento závěr jasně naznačuje, že variace u psů nebyly jen případem náhodných změn, ale spíše produktem nedávné selekce, jakou je šlechtění.



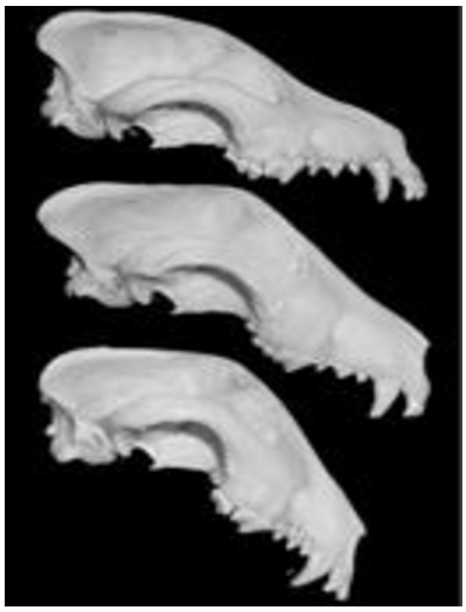
Obrázek 3: Rozdílné VNTR markery šesti jedinců při zobrazení gelovou elektroforézou - převzato z http://upload.wikimedia.org/wikipedia/commons/7/78/D1S80Demo.gif

Vnější projev délky VNTR:

Příklad projevu – gen Runx-2:

U obratlovců slouží gen Runx-2 k regulaci diferenciace osteoblastů (podílí se na formování kostí při jejich růstu). Obsahuje dvě tandemově repetitivní oblasti – poly-Q (kódující 18 až 20 glutaminů) a poly-A (kódující 12 až 17 alaninů). Délka poly-Q oblasti má posilující vliv na transkripci genů, které reguluje, zatímco oblast poly-A má

účinek tlumící. Bylo statisticky zjištěno, že u psů má poměr délek poly-Q a poly-A rozhodující vliv na délku a dorzoventrální zahnutí čumáku. [5] [6]



Obrázek 4: Projev genu Runx-2 – převzato z http://www.nature.com/scitable/nated/content/33589/10.1073_pnas.0408118101-f3b_thumb_0.jpg

2.3 Forma a získávání dat – sekvenování DNA

Sekvenování DNA je proces určování přesného pořadí nukleotidů v sekvencích DNA. Nástup DNA sekvencingu umožnil prudký vývoj oblastí biologického a medicínského výzkumu.

Existuje obrovské množství biochemických metod pro sekvenování. V dnešní době je nejhojněji rozšířena Sangerova metoda, hlavně především pro její jednoduchost a spolehlivost. Díky těmto vlastnostem byla již automatizována a dnes je prováděna pomocí přístrojů zvaných DNA sekvencery.

Výsledkem této metody jsou data obvykle ve formě sekvence písmen, reprezentující jednotlivé nukleotidy na jednom vlákně molekuly DNA. [17]

FASTA formát je běžný textový formát standardně využíván pro zápis bioinformatických dat, především sekvence nukleotidů v DNA, popř. RNA nebo i sekvence aminokyselin v bílkovině. Každý element – nukleotid nebo aminokyselina – je reprezentován jedním písmenem, jež je pro něj vyhraněno. Soubory tohoto formátu mají několik druhů přípon, ovšem obvykle používaná je forma *.fasta*.

Soubor začíná znakem >, jenž je následován řádkem pro stručný popis sekvence. Zbytek těla souboru již obsahuje data dané sekvence, jež jsou zapsána po řádcích, jejichž délka by neměla přesáhnout 80 znaků. [18]

První řádek bývá u spousty internetových vyhledávačů vyžadován, nicméně jeho obsah nemá vliv na samotný postup vyhledávání. Obvykle je detekován jen první znak. Existují však také i vyhledávače, které vyžadují vstupní data jako čistý text i bez znaků „Line feed“, což jsou znaky pro nový řádek.

3 VYHLEDÁVAČE TANDEMOVÝCH REPETIC

V dnešní době, kdy je genomika již značně rozvinutou vědou, lze na internetu nalézt nepřeberné množství volně dostupných databází a nástrojů k analýze a zpracování bioinformatických dat. Jedním z typů takových nástrojů jsou on-line nástroje pro vyhledávání tandemových repetic.

Následující část se zabývá vybranými internetovými vyhledávači. Budou rozebrány algoritmy, které tyto nástroje využívají, a parametry samotných procesů vyhledávání, které lze měnit.

Byly vybrány tři různé vyhledávače, na kterých se pomocí vygenerovaných i reálných sekvencí budou testovat jejich možnosti. Dále budou srovnány jejich výsledky a efektivita. Vybrané vyhledávače používají naprosto odlišné algoritmy, proto byly shledány vhodnými k dalšímu srovnávání.

Zvoleny byly tyto vyhledávače:

1. TRedD – Tandem Repeats Database – Edit Distance Program. Odkaz
2. Tandem Repeats Finder – Boston University. Odkaz
3. Phobos v3.3.12 – Christoph Mayer. Odkaz

3.1 Princip vybraných vyhledávačů

V žádném z vybraných vyhledávačů se nevyskytují jednoduché nebo snadno slovně či matematicky popsatelné algoritmy a žádný z nich nepoužívá výlučně jen jeden nýbrž kombinaci několika algoritmů. Hlubší analýza jednotlivých algoritmů by byla spíše vhodným předmětem samostatné práce. Pro značně velký rozsah těchto algoritmů, je v následující části jejich fungování shrnuto a zkráceno, popřípadě je zmíněn obecný název těchto algoritmů pro případnou potřebu vyhledání bližších informací o nich.

Obecně lze s jistou nadsázkou prohlásit, že s deseti různými vyhledávači přichází dvanáct různých algoritmů.

Hlavním problémem při automatizovaném vyhledávání tandemových repetic v reálných sekvencích jsou mutace v repeticích. Podle typu změny se mutace dělí do dvou kategorií: na bodové mutace měnící jen nukleotid a na tzv. *indely* (název vznikl

spojením slov inzerce, což je vložení nového nukleotidu, a delece, což je odstranění nukleotidu), které způsobují posun celé sekvence. [19]

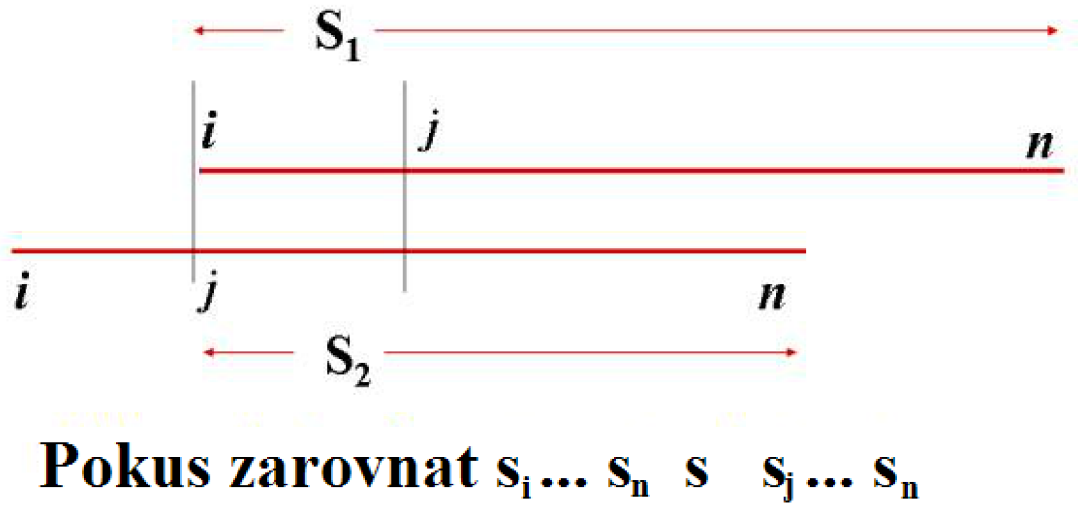
TRED - Edit Database Program:

U bodových mutací vznikají dva druhy chyb: Konsenzuální chyba a evoluční chyba.

Konsenzuální typ chyby je vždy jen jednoduchá změna jednoho nukleotidu oproti originálu. Naproti tomu evoluční chyba je kumulativní mutace nukleotidu, která se přenáší dále a další mutace probíhá na této již zmutované repetici. Tento program proklamuje na rozdíl od ostatních vyhledávačů schopnost detekce evolučních mutací.

K-edit repetice je tandemová repetice, která má maximálně *k* chyb mezi jednotlivými repeticemi v celé své délce. Oblast *vlákna* (*string*) je nazvána *k-edit repeticí* pokud existuje podobnost s jejich *předponou* (předchozí repetice) s *k* nebo méně chybami.

- Definice problému:
  - Vstup:
    - Vlákno S
    - Hodnota k
  - Výstup:
    - Všechny maximální k-edit repetice, které se vyskytují v S
- Přímý algoritmus
  - Zvážit všechny možné páry: $1 \leq i < j \leq n$, kde *i* a *j* jsou indexy pozic a *n* je délka vlákna
  - Sestavit vzdálenostní porovnávání S_1 a S_2, což jsou dvě totožné kopie S, kde $S_1 = s_i \ldots s_n$; $S_2 = s_j \ldots s_n$; (viz Obrázek 5)
  - Pokud se prvních *j-i* znaků S_1 vyskytuje ve spojení s *j-i* v S_2 s *k* nebo méně chybami, pak na daném místě existuje repetice.



Obrázek 5: Schéma srovnávání sekvence programem TRED – převzato z http://tandem.sci.brooklyn.cuny.edu/Algorithm.do

Toto je základní způsob vyhledávání TR. Tento algoritmus se dá samozřejmě dále urychlovat, protože potřebný čas pro výpočet roste se čtvrtou (!) mocninou délky analyzovaného vlákna (n^4). Proto se běžně používá Main-Lorentzův algoritmus, který má účel snížení počtu iterací na $n \cdot \log(n)$, a další algoritmy, jmenovitě Ukkonenův, Landau-Vishkinův, Landau-Myers-Vishkinův, které časovou závislost sníží na $n \cdot k \cdot \log(n) \cdot \log(k)$, kde *n* je délka sekvence a *k* je délka detekované repetice. [19]

Vyhledávač požaduje před každým vyhledáváním zadat parametry, jimiž jsou: *maximální množství chyb* (v jedné repetici), *minimální* a *maximální délka repetice*, *minimální délka repetic*. Tyto parametry představují v praktickém využití značné omezení, jelikož většinou nejsou předem známy, a uživatel je tak odkázán na metodu pokus-omyl. Vyhledávač navíc na výstupu zobrazí jen tabulku číselných údajů o nalezených repeticích, jako jsou: *Začátek* a *konec*, *délka*, *procentuální shoda* atd. Slibně však působí možnost vyhledávání i evolučně zmutovaných repetic.

Tandem Repeats Finder (TRF):

Tento vyhledávač pracuje s pravděpodobnostním modelem tandemových repetic.

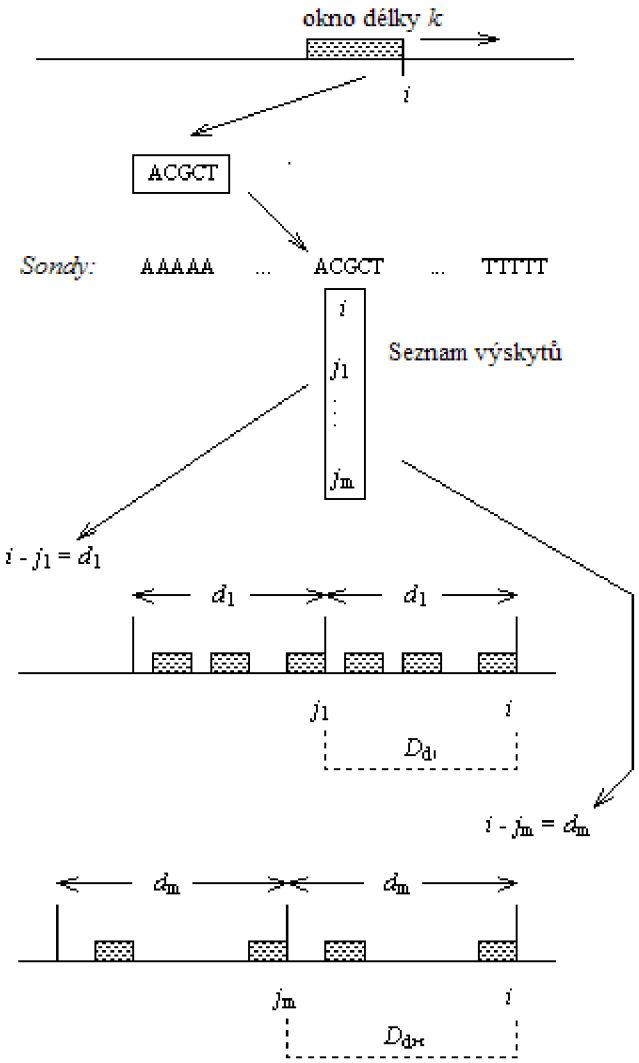
Provádí se Smith-Watermanovo zarovnání, dále skórování dvou tandemových kopií vzorku délky n pomocí sekvencí n nezávislých Bernoulliho pokusů (pravděpodobnostní pokus typu „hod mincí – panna-orel“). Pravděpodobnost shody p_M (matching probability) v jednom elementu zároveň zastupuje průměrnou procentuální identitu mezi kopiemi. Každý element této Bernoulliho sekvence reprezentuje nukleotidy dvou kopií vzorku a nabývá hodnot T - true (souhlasné nukleotidy) nebo F - false (nesouhlasné nukleotidy, popř. indel). Pravděpodobnost p_I (indel probability) zastupuje průměrný procentuální výskyt indelu. Sledování vztahu těchto parametrů (p_M, p_I), resp. jejich vývoj ve vztahu k různým posunům, obecněji řečeno zarovnáním kopií vzorku je hlavním cílem této metody.

```
A G C T C A C T A G T A C A C A C A C T T A C A C C A G A
C G C T C A C T G G T - - A C A C A C T C A C A C C A G -
T H H H H H H H T H H T T H H H H H H H T H H H H H H H T
```

Obrázek 6: Dvě sekvence a Bernoulliho sekvence ve formě H (head - panna), T (tails - orel) – převzato z http://tandem.bu.edu/trf/figure1.gif

Celkový algoritmus se dělí na dvě části:

- Detekční komponenta:
  - Předpoklad existence souhlasných znaků v počáteční vyšetřované oblasti – toto je zajištěno z parametrů p_M, p_I.
  - Sledování souhlasných nukleotidů, oddělených vzdáleností d (v první iteraci není přesně určena).
    - Je určena nebo vybrána počáteční délka repetice k.
    - Vytvoří se všechny možné variace s opakováním délky k (nazvány *sondy*).
    - Každé *sondě* přísluší jeden seznam výskytů ${}^{H}p$.
    - Po sekvenci se posunuje okno délky k a na každé pozici i je určen druh sondy a tato pozice je zaznamenána do sondě příslušejícího seznamu výskytů (${}^{H}p$).
      - Při přidání pozice i do ${}^{H}p$ a existenci předchozích i, nechť se zvou j, existuje vzdálenost d, pro kterou platí $d = i - j$, což je možná velikost detekované repetice.
      - Testovací kritérium: Po sekvenci se zpět posunuje okno délky d a vyhledává ostatní variace na sekvenci mezi i a j. Tyto informace uchovává distanční seznam ${}^{D}p$.



Obrázek 7: Schéma algoritmu vyhledávače TRF – převzato z http://tandem.bu.edu/trf/figure2.gif

- Statistická kritéria jsou založena na binomickém rozdělení Bernoulliho sekvencí, odpovídající detekovaným shodám, uloženými v distančním seznamu.
  - Kritéria jsou založena na čtyřech základních rozděleních podle:
    - Velikosti vzdálenosti – d
    - Pravděpodobnosti shody – p_M
    - Pravděpodobnosti výskytu indelu – p_I
    - Velikosti délky variací – k

- Každé rozdělení je buď vypočteno podle vzorců, nebo aproximováno.
- Nakonec je určena hraniční hodnota, která slouží jako *kritérium*.

➢ Jsou vykonány čtyři další algoritmy vycházející z předchozích *kritérií*:
- *Suma hlav* – algoritmus využívající binomické, resp. normální rozdělení, zde sloužící k určení, kolik shod je potřeba specifickou hodnotu vzdálenosti *d*.
- *Náhodná procházka* – rozdělení, popisující možné změny vzdáleností *d* způsobené výskytem inzercí a delecí.
- *Zdánlivá vzdálenost* – algoritmus pro rozpoznání skutečných tandemových repetic, podle jejich rozložení v intervalu *i-j*.
- *Čekací doba* – neboli forma Poissonova rozdělení sloužící k správnému výběru počáteční velikosti repetic (a také variací) *k* v dalších iteracích tohoto procesu.

❖ Analyzační komponenta:

➢ Pokud informace uložené v distančním listu projdou kritérii testů, úsek tvořen z pozic *j+1 ... i* je vybrán z nukleotidové sekvence a zarovnán se svým okolím. Pokud jsou nalezeny alespoň dvě úspěšně zarovnané kopie úseku se sekvencí, je úsek prohlášen za tandemovou repetici. [20]

Vyhledávač se prezentuje ve třech režimech – *základní*, *středně pokročilý* a *pokročilý*. Tyto režimy se odlišují pouze malým množstvím vstupních parametrů, jejichž změna, jak bude dále zmíněno, má jen malý (a obvykle negativní) význam na výsledek procesu vyhledávání.

Phobos

Phobos je ke stažení a pro akademické účely volně dostupný program k pokročilému vyhledávání tandemových repetic. Nejedná se tedy o online vyhledávač, nicméně jeho možnosti jsou několikanásobně rozsáhlejší než u běžných online vyhledávačů. Značnou nevýhodou je jeho hardwarová náročnost; výpočetní operace jsou složité a velmi náročné na tzv. *CPU resources.*

❖ Základní princip:

Na každé pozici v analyzované sekvenci Phobos kontroluje, zda je tato pozice vhodným počátečním bodem. Dále se od tohoto bodu vyhledává v obou směrech. Díky této strategii není potřeba žádná knihovna repetic. Program také dále znovu hledá na již analyzovaných pozicích s využitím poznatků (parametrů) z později

analyzovaných sekcí, což umožňuje vyhledat skryté nebo překrývající se satelity nebo i spojení rozdílných repetičních oblastí s vysokým skóre, což umožňuje najít příbuzné satelity a ihned hodnotit jejich fylogenetickou vzdálenost.

- Princip skórování
  - Skóre tandemové repetice s daným počátečním a koncovým bodem v subjektivní sekvenci je skóre nejlepšího lokálního zarovnání, které bylo nalezeno v možných po sobě jdoucích kopiích pravděpodobné repetiční jednotky.
  - Každá shoda v zarovnání má skóre 1.
  - Skóre neshod zvlášť ve formě bodových mutací a zvlášť ve formě indelů lze nadefinovat.
  - Toto skóre je využito jako kritérium optimality k rozhodování, zda má být analýza dané repetice rozšířena za pozice neshod. Čili pokud již byla nalezena delší repetice s vyšším skóre, rozšíří se analýza právě analyzované repetice.
  - Pokud jsou nalezeny dva překrývající se satelity, je z analýzy odebrána ta střetávající se repetice, která má nižší skóre.

Program přináší tři základní módy vyhledávání:

- Vyhledávání nedokonalých repetic (*imperfect search*) – základní vyhledávání, umožňující nález nedokonalých repetic. V tomto módu nemusí být dvě repetice pro jejich detekci stejné.
- Vyhledávání dokonalých repetic (*perfect search*) – vyhledává pouze přesně totožné repetice. Pokud se ve *fasta* souboru objeví *N*, lze nastavit, jak s ním bude naloženo.
- Rozšíření přesného vyhledávání (*extend exact search*) – nejprve jsou nalezeny dokonalé repetice a posléze je analýza rozšířena do obou směrů pro vyhledání nedokonalých repetic. Jedná se o nepřesnější mód, avšak má minimální požadavky pro správné fungování: je zapotřebí přítomnosti minimálně 5 bp pro mononukleotidovou, 6 bp pro di-, a 7 bp pro trinukleotidovou repetici. V případě repetic s delšími podjednotkami, musejí být přítomny alespoň dvě dokonalé podjednotky.

Nastavení většiny požadovaných vstupních parametrů vyžaduje poměrně značnou znalost základních principů toho, jak Phobos funguje, proto je tento program poměrně nevhodný pro krátkodobé nebo nepravidelné užívání. Výstupem je soubor s příponou *.phobos*, který je vhodné otevírat v programu *WordPad* a který poskytuje přehledný výpis i popis detekovaných dat. [21]

Jak již bylo zmíněno, je v dnešní době běžný výskyt velkého množství volně

dostupných genomických nástrojů, mimo jiné i vyhledávačů, a díky jejich volné dostupnosti může být jejich kvalita i značně nedostačující, proto byly po pečlivém volném testování a analýze jejich principů zvoleny tyto tři, na kterých bude dále probíhat testování k rozlišení jejich efektivity a účinnosti pro specifické typy vstupních dat ve formě sekvencí DNA s obsahem tandemových repetic.

3.2 Testovací sekvence

Protože budou v následující kapitole testovány vybrané vyhledávače, bylo zvoleno několik testovacích sekvencí, z nichž tři jsou reálné a ostatní uměle vytvořené. Jako reálné sekvence byly vybrány:

- Skot – lokus BOVTGN: poměrně krátký úsek s výskytem minisatelitů.
- Vinná kvasinka – chromozom I: úsek o přibližné délce 230 kbp.
- Člověk – výstřižek chromozomu I: zatěžovací test, úsek o délce přes 9 Mbp.

Dále byl pro vytvoření umělých sekvencí a přímé testování citlivosti vyhledávačů na různé vlastnosti vstupních dat v programovém prostředí Matlab vytvořen algoritmus pro generaci sekvencí s výskytem tandemových repetic s několika nastavitelnými parametry (zdrojový kód).

Princip generátoru:

- Na každém místě je generováno jedno náhodné písmeno {A, C, G, T} jako nerepetitivní „junk DNA“, přičemž existuje šance, že se na dané pozici vyskytne satelit.
- Při výskytu satelitu je zadaná repetice umísťována podle předem zadaného množství, přičemž pro každý její nukleotid existují šance na mutace typu bodová, kdy je nukleotid náhodně změněn na jiný, delece a inzerce opět náhodně vybraného nukleotidu napravo od něj. Na počátku i na konci formování jednoho satelitu jsou do sekvence umístěny speciální znaky, které slouží jako markery pro přesnou alokaci začátku a konce každého vygenerovaného satelitu.
- Taktéž byla implementována možnost výběru přítomnosti evolučních mutací, kdy se struktura dané, množené repetice mění. Evoluční mutace se může vyskytnout jak formou bodových mutací, tak formou indelových mutací; důležité je, že každá vyskytnuvší se mutace se tak přenáší do následujících repetic.

- Program obsahuje čítače na různé výsledné parametry, jako jsou délka výsledné sekvence, celkové množství jednotlivých mutací aj. Data jsou následně vypsána a uložena ve formátu *fasta*.

- Vstupní parametry generátoru:
  - Repetice – lze si nadefinovat přímo danou repetici, jež se bude vyskytovat.
  - Maximální délka sekvence – omezení míry generování horní hranicí.
  - Počet repetic – množství generovaných repetic v jednom satelitu.
  - Pravděpodobnost výskytu satelitu – na každém generovaném místě existuje pravděpodobnost výskytu satelitu, obrácená hodnota této pravděpodobnosti, značí četnost, s jakou se satelity proti nerepetitivním nukleotidům vyskytují.
  - Pravděpodobnost bodové mutace – při tvorbě repetic má každý nukleotid šanci na mutaci.
  - Pravděpodobnost inzerce – napravo od každého nukleotidu se může vyskytnout inzerce.
  - Pravděpodobnost delece.
  - Evoluční mutace – zda budou mutace probíhat touto cestou. Při aktivaci této funkce, je doporučeno snížit celkovou pravděpodobnost vzniku jednotlivých mutací pro dosažení relevantního a měřitelného výsledku.
  - Způsob výpisu dat a zápis do souboru – určení způsobu výpisu, jméno souboru.

- Výstupní hodnoty:
  - Celková délka celé vygenerované sekvence.
  - Četnost jednotlivých bází v celé sekvenci.
  - Celkový počet vygenerovaných satelitů.
  - Počáteční a koncové pozice vygenerovaných satelitů a jejich délka.
  - Celkový počet vygenerovaných repetic.
  - Celkové počty bodových mutací, inzercí a delecí.
  - Pokud je zapnuta evoluční mutace, zobrazí se pod sebou vývoj mutací dané repetice a pozice, na které k mutaci repetice došlo.
  - Výpis dat – podle určení jsou data vypsána s hlavičkou pro *fasta* formát nebo bez ní. Takto zobrazená data jsou zkopírována do schránky. Pokud je zvolen zápis do souboru, je vytvořen soubor ve formátu *fasta* s hlavičkou a daty formátovanými na řádky po sedmdesáti znacích.

Pomocí tohoto generátoru je vytvořeno několik sekvencí s odlišnými parametry, na kterých budou testovány vybrané vyhledávače.

3.3 Testování vybraných vyhledávačů

Každý z vyhledávačů nabízí několik základních nastavitelných parametrů, jež umožňují zpřesnit a v některých případech i zrychlit proces vyhledávání, nicméně vyžadují znalost některých vlastností vstupních dat, což je mnohdy v praxi nereálné, proto bude testování probíhat formou co nejbližší k reálným situacím. Takto bude zajištěno přesné zhodnocení využitelnosti a vhodnosti jednotlivých vyhledávačů pro praktické úlohy.

3.3.1 Testování umělých sekvencí

Tabulka 1: Shrnutí vlastností testovacích sekvencí

Soubor **.fasta*	Délka sekvence	Repetice	Počet repetic v satelitu (průměr)	Počet satelitů	Mutace v celé sekvenci			Popis
					B	I	D	
Test_1	1992	AT	80	4	0	0	0	Jednoduchá testovací sekvence
Test_2	2990	CAGCTT	180	5	40	13	10	S výskytem mutací
Test_3	5132	$(AT)_8C$	340	5	10	1	3	Test zda bude „*C*“ správně vyhodnoceno jako součást repetice
Test_4	4990	ATCGTCCCATC	280	5	469	222	184	Silně zmutovaná sekv.
Test_5	4986	ACGTAC	110	4	4	1	1	Výskyt evolučních mutací

Každá z těchto sekvencí je postupně testována na každém z vybraných vyhledávačů.

Test_1.fasta – dinukleotidové repetice bez mutací

TRED – program nalezl a dokázal téměř dokonale určit všechny hodnoty.

Tabulka 2: Výstupní data programu TRED 1. sekvence

Začátek	Konec	Délka	Perioda TR	Počet repetic	Chyby	%Shoda
173	213	41	2,0	20,5	0	100,00
859	900	42	2,0	21,0	0	100,00
1168	1208	41	2,0	20,5	0	100,00
1922	1962	41	2,0	20,5	0	100,00

TRF – program nalezl všechny satelity a téměř dokonale určil jejich parametry. Vyskytnutá chyba, spočívající v nepřesné detekci začátků a konců některých satelitů, se následně projeví i v nepřesnosti některých jiných výsledných parametrů.

Tandem Repeats Finder Program written by:

Gary Benson
Program in Bioinformatics
Boston University
Version 4.08

Please cite:
G. Benson,
"Tandem repeats finder: a program to analyze DNA sequences"
Nucleic Acid Research(1999)
Vol. 27, No. 2, pp. 573-580.

Sequence: Hlavicka pro fasta soubory|Test_1
Parameters: 2 7 7 80 10 50 30
Length: 1992

Tables: 1

This is table 1 of 1 (4 repeats found)

Click on indices to view alignment

Table Explanation

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
168--213	2	23.0	2	95	0	83	50	2	0	47	1.13
859--900	2	21.0	2	100	0	84	50	0	0	50	1.00
1168--1208	2	20.5	2	100	0	82	48	0	0	51	1.00
1922--1962	2	20.5	2	100	0	82	51	0	0	48	1.00

Tables: 1

The End!

Obrázek 8: Grafický výstup programu TRF

```
Version 4.08

Sequence: Hlavicka pro fasta soubory|Test_1

Parameters: 2 7 7 80 10 50 30

Pmatch=0.80,Pindel=0.10
tuple sizes 0,4,5,7
tuple distances 0, 29, 159, 500

Length: 1992
ACGTcount: A:0.25, C:0.24, G:0.23, T:0.27


Found at i:180 original size:2 final size:2

Alignment explanation


    Indices: 168--213  Score: 83
    Period size: 2  Copynumber: 23.0  Consensus size: 2

        158 GTCGCGAGCC

                  *
        168 TA TA CA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA
          1 TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA TA


        210 TA TA
          1 TA TA

        214 ATTCAGCATA


Statistics
Matches: 42,  Mismatches: 2, Indels: 0
        0.95            0.05        0.00

Matches are distributed among these distances:
   2    42  1.00
```

Obrázek 9: Podrobnosti detekovaných dat po kliknutí na daný rozsah pozic zobrazen v předchozím obrázku

Phobos – v režimu *imperfect search* byly nalezeny všechny satelity správně, nicméně byly navíc detekovány i další, náhodné, krátké výskyty repetitivních oblastí. Protože byly tyto oblasti dokonalé, čili bez chyb repetitivity, režim *perfect search* je rovněž chybně detekoval.

Tabulka 3: Výstupní data programu Phobos v režimu imperfect search

Souřadnice výskytu		Délka repetic
173	213	2
218	228	5
562	569	2
859	900	2
971	977	1
1168	1208	2
1405	1412	2
1922	1962	2

Test_2.fasta – běžná sekvence s výskytem mutací

TRED – Jak je patrné z tabulky, program detekoval značné množství satelitů. Při zaměření se na konkrétní výsledné hodnoty je patrné, že se jednotlivé detekované satelity překrývají. Při vhodně zvolených vstupních parametrech se může počet detekovaných satelitů snížit, obvykle však nejvíce o tři. Správně detekované satelity mají poměrně správné hodnoty; jsou to ty, jejichž hodnota *Period* se pohybuje okolo 6, což ale v praxi není předem známo. Celkově je tedy tento výsledek značně nevhodný.

Tabulka 4: Výsledky druhého testování programu TRED

Začátek	Konec	Délka	Perioda TR	Počet repetic	Chyby	%Shoda
53	231	179	6,0	29,7	15	91,53
494	699	206	6,1	33,7	27	86,83
500	700	201	12,2	16,4	23	88,14
501	617	117	31,0	3,8	10	88,76
1746	1843	98	17,9	5,5	10	87,80
1747	1940	194	6,1	32,0	34	82,74
1773	1930	158	12,1	13,1	20	86,67
1963	2140	178	17,8	10,0	16	90,24
1963	2116	154	11,8	13,0	16	89,12
1963	2141	179	6,0	30,0	20	88,76
2651	2836	186	6,0	31,2	28	84,86
2657	2830	174	18,1	9,6	22	86,25

TRF – program se stejnou relativní přesností jako předchozí vyhledávač detekoval dané pozice a údaje. Nicméně třetí a čtvrtý satelit byly sloučeny do jednoho. To je nepochybně způsobeno skutečností, že byly od sebe oba satelity vzdáleny jen asi 30 bp, kterýžto úsek tedy program interpretoval jako výskyt mutací, což je patrné z přiloženého obrázku. S vyskytnuvšími se mutacemi všech typů si program poradil výborně. Výsledný tvar vypsané repetice není přesný – *TTCAGC*, místo *CAGCTT* – je názorně vidět posun.

Tabulka 5: Výsledky druhého testování programu TRF

Souřadnice	Perioda	Poč. repetic	Perioda konsenzu	Procento shod	Procento indelů	Skóre	A	C	G	T	Entropie (0-2)
53--225	6	28,7	6	91	3	285	17	32	17	32	1,93
506--683	6	29,5	6	90	3	286	18	32	18	30	1,95
1749--2141	6	64,0	6	79	11	388	18	31	18	31	1,95
2655--2830	6	29,3	6	86	4	248	16	32	17	33	1,93

```
Indices: 1749--2141  Score: 388
Period size: 6  Copynumber: 64.0  Consensus size: 6

   1739 TAACAATCTG

                        *          *                    *
   1749 CAGCTTT CAGCTT AAGCTT CAGCGT CAGCTT CTAGCTT CAACTT CAGCTT
      1 CAGC-TT CAGCTT CAGCTT CAGCTT CAGCTT C-AGCTT CAGCTT CAGCTT


   1799 CAGCTT CAGC-T CAGCTT CAGCTT CAGC-T CAGCTT CAGCTT CAGCTT
      1 CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT

                                  *       *         *
   1845 CCAGCTT CAGCTT CAGCTT CAGCCT CAGCTG CAG-TCT AAGCTT CAGCTGT
      1 -CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCT-T CAGCTT CAGCT-T

                *                                    *      * * *  *
   1895 CAGCTT CACCTT CAGCTT CAGCTT CAGCTT CAGCTT GAAG-TAGG AAACGGT
      1 CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT -CAGCT--T CAGC-TT

        *     **          *           *
   1946 AACGTCCC CAG-TGGGAG CAGCTT CAGGTT CAGCTT CAGCTTT CAGCTT
      1 CA-G-CTT CAGCT----T CAGCTT CAGCTT CAGCTT CAGC-TT CAGCTT

         *       *
   1994 CTGCTT CATCTT CAGCTT CAGCTT CAGC-T CAGCTT CAGCTT CAGCTT
      1 CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT

                                              *
   2041 -ATGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGATT CAGCTT CAGCTT
      1 CA-GCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT

                                               *
   2089 CAGCTT CAGCTT CAGC-T CAGCTT CAGCTT CAGCTA CAGCTT CAGCTT
      1 CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT CAGCTT


   2136 CAGCTT
      1 CAGCTT
```

Obrázek 10: Názorná ukázka chybné interpretace oblasti mezi satelity - uprostřed

Phobos – výstupní data v režimu *imperfect search* byla prakticky srovnatelná s předchozím vyhledávačem. Ovšem režim *extend exact search* již byl schopen rozlišit ony dva blízké satelity od sebe. Bohužel se v obou případech vyskytla nová, značná chyba, a tou je chybné vypsání detekované repetice, která byla chybně vypsána ve všech detekovaných případech – *AAGCTG*, místo *CAGCTT*. Můžeme taktéž vidět chybně detekované, náhodné, repetitivní oblasti. Ostatní

údaje jsou v pořádku.

Tabulka 6: Relevantní výsledky testování v programu Phobos

Imperfect search		Extend exact search	
Souřadnice výskytu		Souřadnice výskytu	
53	225	53	225
506	683	506	683
1749	2141	1754	1930
2655	2830	1963	2141
		2655	2830

Test_3.fasta – detekce dlouhé, zdánlivé dinukleotidové repetice

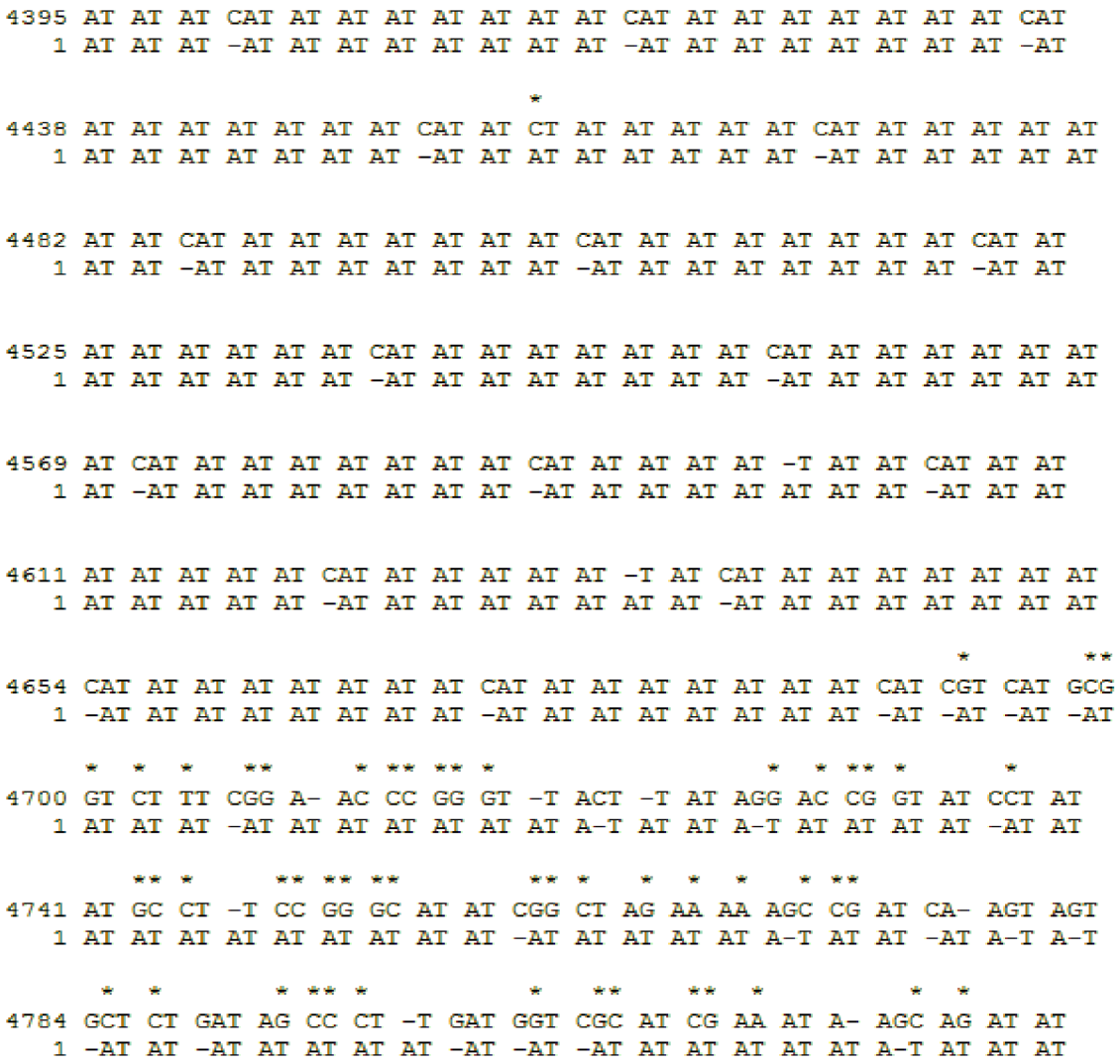
TRED – jelikož formát výstupu tohoto programu neumožňuje zobrazit, zda byla repetice detekována celá, a ne jen jako dinukleotid *AT* s pravidelnou inzercí *C*, není možné hodnotit tuto úlohu. Stejně jako v předchozích případech program detekoval každý satelit několikrát.

TRF – zde se vyskytly očekávané nesprávně detekované dinukleotidy. Jakékoliv pokusy odstranit tuto chybu úpravou vstupních parametrů byly bez výsledku. Pro odstranění chyby tohoto druhu by se hodil parametr *minimální délka periody* (repetice), který ale není přítomen. Mimo tuto chybu byly ostatní údaje změřeny přesně.

Tabulka 7: Výstupní data třetího testování programu TRF

Souřadnice	Perioda	Poč. repetic	Perioda konsenzu	Procento shod	Procento indelů	Skóre	A	C	G	T	Entropie (0-2)
520--860	17	20,1	17	99	0	664	46	6	0	46	1,28
2268--2607	17	20,0	17	99	0	671	46	5	0	47	1,26
3208--3545	2	160,0	2	86	11	462	47	5	0	47	1,28
3207--3546	17	20,1	17	97	0	637	46	6	0	46	1,30
4351--5161	2	381,0	2	78	14	551	43	8	4	43	1,56

4351--5162	17	46,0	17	83	10	947	43	8	4	43	1,56



Obrázek 11: Detail chybných interpretací nukleotidu "C"

Phobos – při implicitním nastavení program v tomto testu selhal a označil všechny nalezené satelity za tandemové repetice *AT* a všechna *C* byla označena za indely. Mimo jiné byly opět nalezeny náhodné, repetitivní oblasti. Při různých změnách parametrů program poskytoval buď již zmíněné výsledky, nebo nedetekoval žádné. Pouze při užší specifikaci délky vyskytující se repetice program správně rozpoznal satelit, ovšem z nepochopitelných důvodů pouze ten první.

Test_4.fasta – silně zmutované repetice

TRED – program nebyl schopen nalézt žádné tandemové repetice.

TRF – při implicitním nastavení parametrů nebyly nalezeny žádné repetice. Až při změnách některých parametrů byl program schopen detekovat dva výskyty TR. Tyto výskyty byly správně lokalizovány v satelitu, nicméně i ty byly detekovány jen ve své zmutované podobě.

Tabulka 8: Výsledek testování zmutovné sekvence programu TRF

Souřadnice	Perioda	Poč. repetic	Perioda konsenzu	Procento shod	Procento indelů	Skóre	A	C	G	T	Entropie (0-2)
346--382	12	2,9	12	76	8	45	21	43	10	24	1,84
2342--2371	11	2,6	12	78	5	40	23	43	10	23	1,83

Phobos – nebyly nalezeny žádné relevantní výsledky.

V tomto testu všechny vyhledávače selhaly, což je pochopitelné, protože míra zmutování repetic testovací sekvence byla skutečně enormní – téměř každý druhý nukleotid repetice prošel mutací.

Test_5.fasta – evoluční mutace

TRED – deklarovaná skutečnost, že tento program dokáže vyhledat i evolučně zmutované repetice a správně je zařadit do satelitů, se prokázala jako pravdivá. Program byl schopen správně zařadit satelit, ve kterém se objevuje více různě zmutovaných repetic, jako jeden. Ostatní hodnoty jsou také relativně přesné.

Tabulka 9: Výsledek detekce evolučních mutací programu TRED

Začátek	Konec	Délka	Perioda TR	Počet repetic	Chyby	%Shoda
2748	2875	128	5,4	23,5	9	93,08
3090	3223	134	4,1	33,0	13	90,44
4482	4619	138	4,6	30,0	18	87,14
4851	4959	109	5,4	20,2	1	99,04

TRF – program podal smíšené výsledky. Satelity byl jen čtyři, nicméně program je

našel několikrát a to podle různých obměn dané repetice. Výskyt krajních bodů některých satelitů částečně koresponduje s výskytem evolučních mutací. Program není evidentně koncipován na detekci tandemových repetic s výskytem evolučních mutací.

Tabulka 10: Srovnávací tabulka reálných pozic počátků a konců satelitu (zeleně) a pozic, na kterých došlo k evoluční mutaci repetic (červeně)

Pozice satelitu	**1**	**2**	**3**	**3**
Počátek	**2752**	**3095** **3095**	**4505**	**4851**
	2836	**3149** **3179**	**4550**	**4911**
Konec	**2872**	**3209**	**4605**	**4959**

Tabulka 11: Tabulka výsledků poskytnutých programem TRF. Srovnej hodnoty s nahoře

Souřadnice	Perioda	Poč. repetic	Perioda konsenzu	Procento shod	Procento indelů	Skóre	A	C	G	T	Entropie (0-2)
2752--2871	6	20,0	6	99	0	186	28	38	16	16	1,91
3095--3176	6	13,7	6	98	0	128	4	50	17	28	1,66
3095--3208	6	20,0	6	97	1	105	9	47	17	25	1,78
3149--3208	5	11,2	5	98	1	84	18	45	18	18	1,86
4505--4605	5	20,2	5	98	0	121	19	29	30	19	1,97
4851--4959	5	20,2	5	99	0	146	18	19	44	18	1,88

Phobos – program poskytuje výsledek srovnatelný s předchozím vyhledávačem.

3.3.2 Testování reálných sekvencí

- Skot – lokus BOVTGN: poměrně krátký úsek s výskytem minisatelitů.

Tabulka 12: Skutečný výskyt repetic v sekvenci BOVTGN

Začátek	Konec	Délka
327	352	25
378	402	24
425	462	37
491	508	17
534	556	22
585	608	23
641	690	49

Výsledky jednotlivých vyhledávačů se zdají být poměrně odlišné, nicméně při bližším pohledu lze spoustu detekovaných úseků označit za podobné. V tomto případě jsou nejlepší výsledky poskytnuty vyhledávačem TRF.

Tabulka 13: Výsledky vyhledávání v reálné sekvenci programu TRED

Začátek	Konec	Délka	Perioda TR	Počet repetic	Chyby	%Shoda
309	451	143	52,2	2,7	9	90,53
313	462	150	2,3	64,0	39	75,32
322	424	103	44,0	2,3	10	84,13
327	556	230	109,1	2,1	6	95,08
378	663	286	52,2	5,5	29	87,97
491	508	18	2,0	9,0	0	100,00
537	556	20	2,0	10,0	0	100,00
585	617	33	2,3	14,3	4	87,10
634	690	57	1,8	32,5	2	96,43

Tabulka 14: Výsledky vyhledávání v reálné sekvenci programu TRF

Souřadnice	Perioda	Poč. repetic	Perioda konsenzu	Procento shod	Procento indelů	Skóre	A	C	G	T	Entropie (0-2)
327--352	2	13,0	2	100	0	52	0	0	50	50	1,00
425--462	2	19,0	2	100	0	76	0	0	50	50	1,00
309--658	48	6,8	50	80	15	376	13	12	36	38	1,82
378--508	62	2,1	62	95	0	235	9	9	38	41	1,70
641--690	2	25,0	2	100	0	100	0	0	50	50	1,00
443--660	52	4,4	51	92	6	337	12	12	36	38	1,81

Tabulka 15: Výsledky vyhledávání v programu Phobos

Souřadnice výskytu		Délka repetic
174	180	1
327	352	2
330	448	48
425	462	2
443	656	48
641	690	2

- Vinná kvasinka – chromozom I: úsek o přibližné délce 230 kbp.

  Pro tuto sekvenci nejsou zobrazeny výsledky v grafické podobě, protože jejich velké množství činí jakýkoliv pokus o grafické srovnání nemožným.

  Po hlubší analýze výsledků se jeví až překvapivá podobnost mezi výsledky jednotlivých vyhledávačů. Při testování této sekvence exceluje program Phobos, a to především svou rychlostí, navzdory jeho vysokým hardwarovým požadavkům (které nebyly splněny při testování), a také absencí rekurentních výsledků. Ovšem opět se u něj objevuje chybové vypsání repetic.

- Člověk – výstřižek chromozomu I: zatěžovací test, úsek o délce přes 9 Mbp.

  Toto testování bylo provedeno jako forma zátěžového testu, zda jednotlivé vyhledávače budou schopny zpracovat takto dlouhé sekvence a poskytovat výsledky. Program TRF se jeví jako nejvhodnější, protože výsledky poskytnul zdaleka nejrychleji. Program TRED poskytnul výsledky výrazně později, což bylo způsobeno i tím, že jsou zobrazovány na jedné internetové stránce, což na druhou stranu spíše napomáhá v přehlednosti. Použití programu Phobos bylo značně omezené, protože při nastavení postupného vyhledávání repetic od délky 1 do 50 se k délce 35 dostal až po sedmi hodinách. Tato skutečnost je bezpochyby způsobená tím, že byly splněny jen jeho minimální hardwarové nároky.

3.4 Shrnutí výsledků testování

Testování prokázalo, že každý z vybraných vyhledávačů má svá využití, druhy dat, pro jejichž zpracování se hodí nejlépe.

Program Tandem Repeats Database – Edit Distance Program se na první pohled jeví vhodným jen pro detekci nezmutovaných repetic. Ani v tomto případě není však nejlepší volbou. Vstupní parametry mohou být v případě některých úloh omezující, jinak pomáhají omezit množství výsledků jen na relevantnější. Jeho největší předností a jedinečností je schopnost detekce evolučně mutovaných repetic, jejichž existence je obvykle při tvorbě nových algoritmů opomíjena.

Program Tandem Repeats Finder lze prohlásit za nejvhodnější pro vyhledávání předem nespecifikovaných typů repetic. Poskytuje průměrně nejlepší výsledky a je schopen si poradit i s některými extrémními situacemi. Skutečnost, že zadávané parametry jen málo ovlivňují výsledek vyhledávání, je paradoxně spíše prospěšná, neboť relevantní výsledky se obvykle dostavují i při implicitním nastavení a změna těchto parametrů je neodstraní. Pravděpodobně jako největším problém by mohlo být nepřesné vypisování nalezených tandemových repetic. Téměř vždy byly vypsány posunuté.

Program Phobos pracuje nejlépe se středně velkými objemy dat a to až do řádově stovek kbp. Poskytuje také nejvíce výsledků v nejpřehlednější formě. Částečnou nevýhodou jsou detekce nevýznamných, krátkých, náhodných sekvencí, jež jsou označeny jako repetice, nicméně ty lze díky přehlednosti výpisu lehce rozpoznat a přejít; kromě toho se jim vhodným nastavení vstupních parametrů můžeme vyhnout, to na druhou stranu vyžaduje předchozí alespoň částečnou znalost vlastností sekvence. Program také neposkytuje rekurentní výsledky a překrývající se satelity. Velké množství vstupních parametrů je výhod v tom, že existuje možnost do značné míry zdokonalit výsledek, ovšem za cenu dlouhé práce na jedné sekvenci. Během testování se vyskytl problém s vypisováním detekovaných repetic. Ačkoliv byla oblast repetic správně detekována a zarovnána, vypsaná repetice se naprosto odlišovala od v zobrazeném zarovnání vyskytnuvší se repetice. Z toho lze usoudit, že se jedná o bug, který by mohl být v budoucích verzích programu odstraněn.

4 METODA ANALÝZY DENZITNÍCH VEKTORŮ

Tato část si klade za cíl otestovat novou metodu vyhledávání tandemových repetic, a to s využitím nukleotidových denzitních vektorů, které budou vytvořeny z několika vzorků DNA s různým obsahem a různou formou tandemových repetic. Deterministický charakter tandemových repetic naznačuje, že by mělo být možné je detekovat touto metodou.

4.1 Indikační a denzitní vektory

Předpokládá se, že vstupní data budou již čistě ve formátu posloupnosti nukleotidů. Z této posloupnosti jsou následně vytvořeny čtyři indikační vektory pro každý typ nukleotidu. Indikační vektor je datová struktura typu „pole“. Pro určitý typ nukleotidu obsahuje hodnotu 1 na všech pozicích, na kterých se v původní sekvenci nachází daný nukleotid. Ostatní pozice jsou vyplněny nulami. Z důvodů dalšího zpracování platí, že všechny čtyři indikační vektory budou stejně dlouhé, a sice tak, jak dlouhá je vstupní sekvence, i přesto, že jejich poslední pozice obsahují pouze nuly. Tento způsob vyjádření se nazývá binární reprezentace nukleotidů.

Indikační vektor tedy zaznamenává pozice výskytu daného nukleotidu. Indikační vektory jsou vypočítány pomocí následujícího vzorce:

$$I_n(x) = 1 \quad \forall \quad S(x) = n; \qquad n = \{A, C, G, T\};$$

kde $I_n(x)$ značí x-tou pozici indikačního vektoru pro n-tý nukleotid a S(x) značí x-tou pozici ve vstupní posloupnosti.

Sekvence „ATG TCG CCG TAT TAA“ bude v binární reprezentaci vyjádřena takto:

Sekvence	A	A	G	T	C	G	C	C	G	T	A	T	T	A	A
I_A	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1
I_C	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
I_G	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0
I_T	0	1	0	1	0	0	0	0	0	1	0	1	1	0	0

V podmínkách programového prostředí Matlab je výhodné tyto indikační vektory umístit do matice. Matice je datová struktura typu „vícerozměrné pole“.

Nejprve je vytvořena matice nulových vektorů o délce vstupní sekvence. Jedná se v podstatě o deklaraci proměnné. Poté jsou pouze na vhodných pozicích nuly nahrazeny jedničkami. Nukleotid je uracil zahrnutý do stejného indikačního vektoru jako thymin. Tato jednoduchá úprava umožní zpracování nejen DNA ale i RNA.

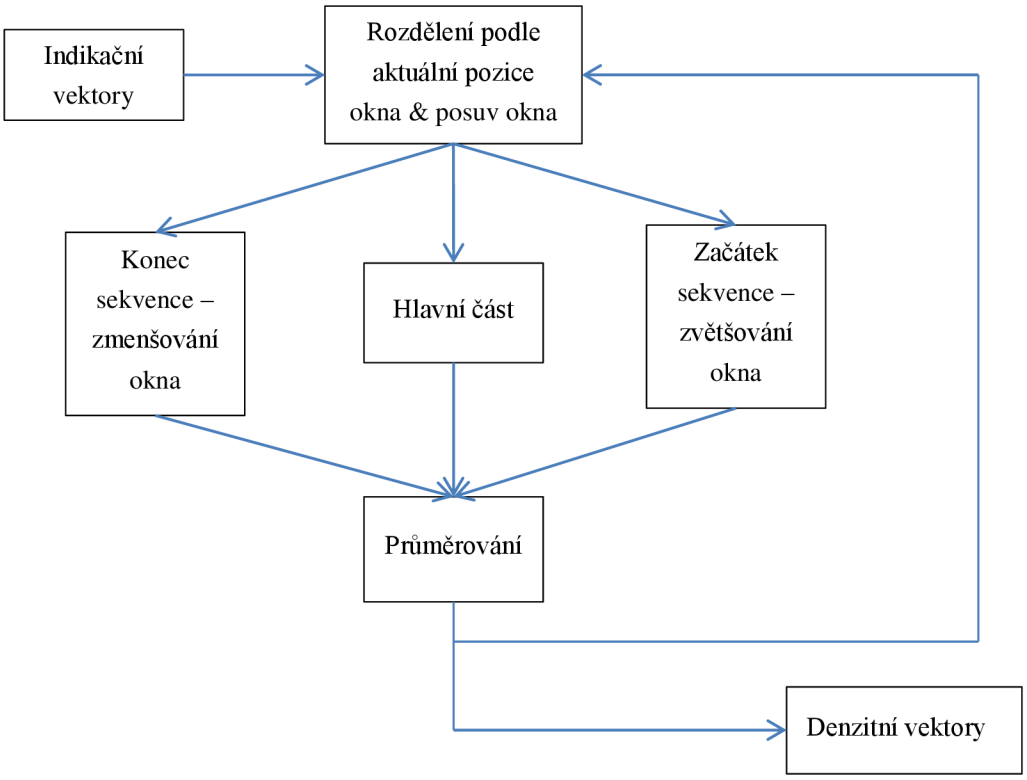
Z každého z těchto indikačních vektorů budou následně vytvořeny denzitní vektory. Stejně jako v případě indikačních vektorů existuje pro každý typ nukleotidu jeden denzitní vektor. Hodnoty každého denzitního vektoru vyjadřují hustotu výskytu daného nukleotidu v jejich okolí. Počítá se tím způsobem, že se každá pozice vyjádří jako podíl součtu všech hodnot indikačního vektoru v rámci určitého okolí (okna) a velikosti samotného okolí. Pro přesné vyjádření hustoty je také zapotřebí vhodně ošetřit začátek a konec vektorů. Zde bylo zvoleno ošetření zkrácením onoho okolí vždy z jedné nebo druhé strany. Tímto způsobem je zachován hustotní charakter krajních hodnot, ačkoliv částečně za cenu přesnosti.

Pro výpočet denzitních vektorů s ošetřenými okraji platí následující matematicky exaktní vzorec:

$$D_n(x) = \begin{cases} \sum_{i=x}^{x+\left\lfloor\frac{O}{2}\right\rfloor} \frac{I_n(i)}{\left\lfloor\frac{O}{2}\right\rfloor + x} \; pro\; x < \frac{O}{2}; \\ \sum_{i=x-\left\lfloor\frac{O}{2}\right\rfloor}^{x+\left\lfloor\frac{O}{2}\right\rfloor} \frac{I_n(i)}{O} \; pro\; x \geq \frac{O}{2}; \\ \sum_{i=x-\left\lfloor\frac{O}{2}\right\rfloor}^{x} \frac{I_n(i)}{\left\lfloor\frac{O}{2}\right\rfloor + \max(x) - x + 1} \; pro \max(x) - x < \frac{O}{2}; \end{cases} ;$$

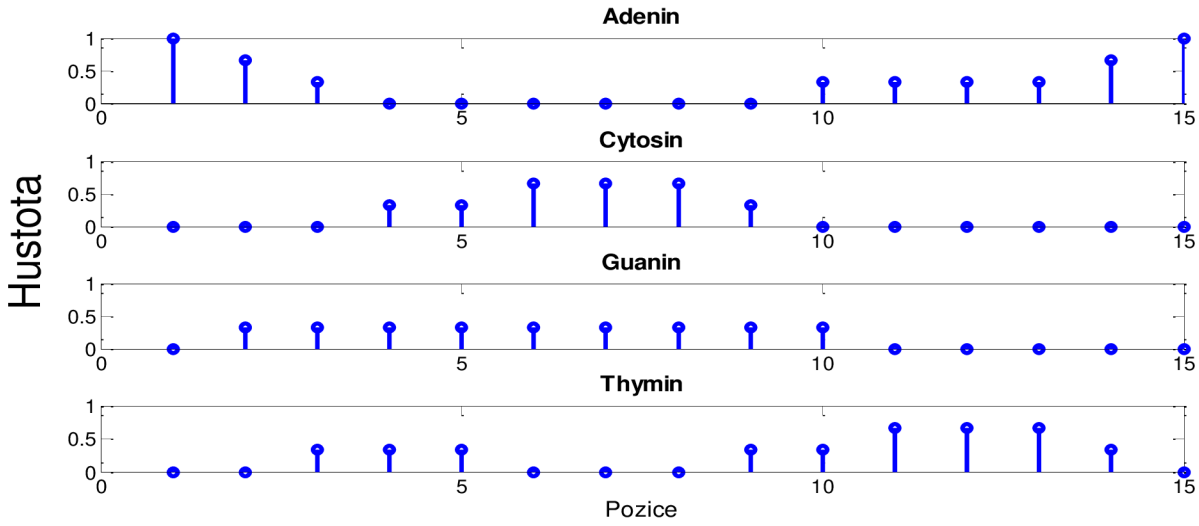
$$pro\; x \in \mathbb{N}, O = (2\mathbb{N} + 1)$$

kde $D_n(x)$ značí x-tou pozici denzitního vektoru n-tého nukleotidu, O je velikost okna neboli okolí, výraz *max(x)* představuje celkovou délku vektoru a I_n je indikační vektor n-tého nukleotidu.



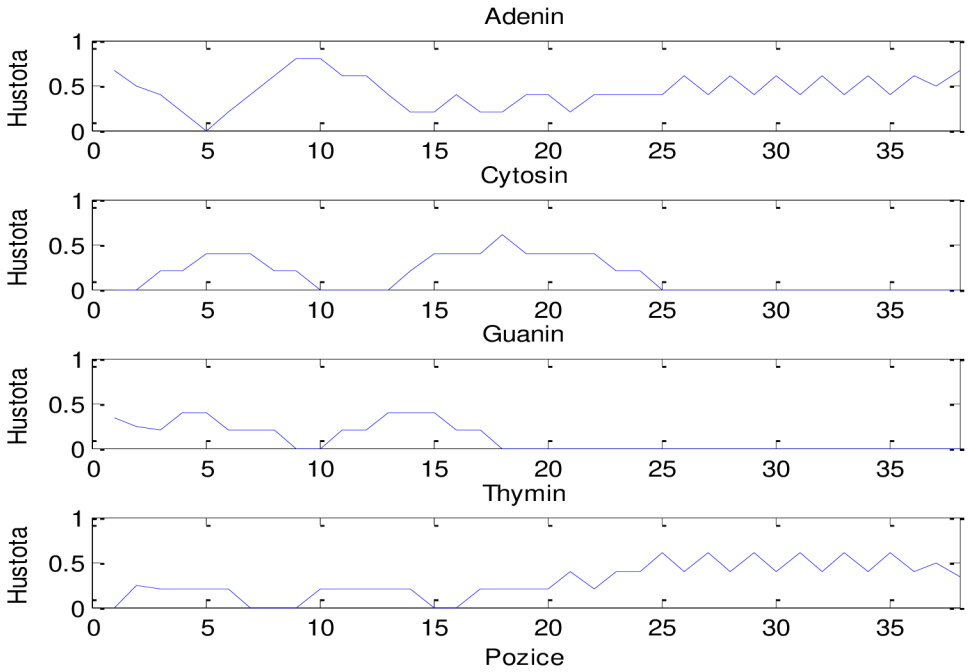
Obrázek 12: Blokové schéma procesu výpočtu denzitních vektorů

A takto již vypadá graficky zobrazený výsledek předchozí vzorové sekvence – matice *denzity* zobrazená po jednotlivých vektorech. Jedná se tedy o čtyři diskrétní funkce hustoty nukleotidů v závislosti na pozici. Pro tato data bylo *okno* délky 3.



Obrázek 13: Denzitní vektory vzorové sekvence

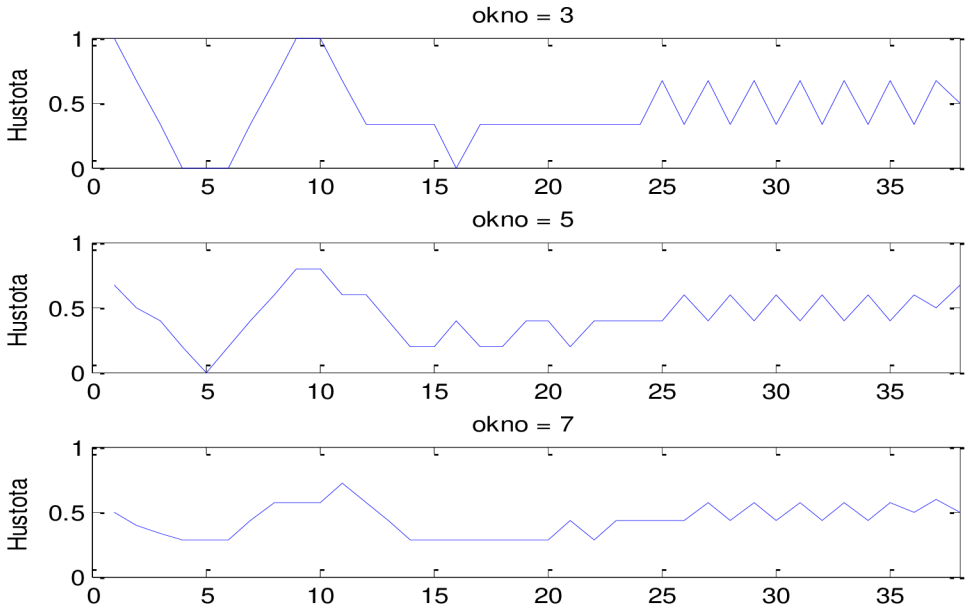
Pro posouzení schopnosti detekce byla vytvořena o něco delší umělá sekvence s výskytem dinukleotidových repetic. Takto vypadá výsledek při zpracování oknem o velikosti 5:



Obrázek 14: Denzitní vektory sekvence s dinukleotidovým motivem

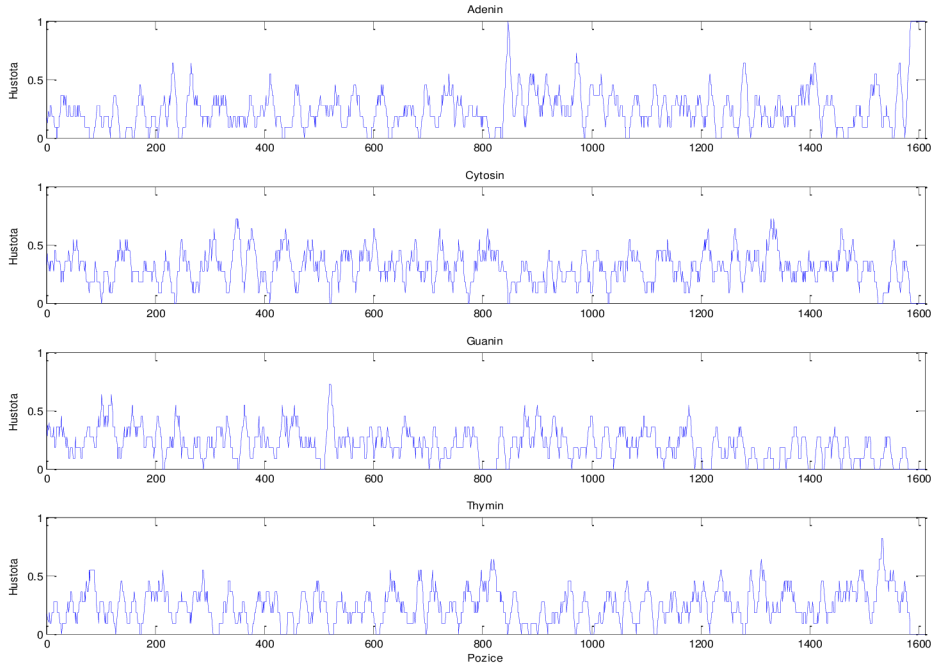
Již z obrázku je jasně patrné, že ona dinukleotidová repetice měla formu „AT“.

Na dalším obrázku jsou zobrazeny denzitní vektory pro adenin stejné sekvence, každý zpracován různou délkou okna. Z těchto výsledků lze interpretovat, že pro vícenukleotidové repetice je zapotřebí delší výpočetní okno.

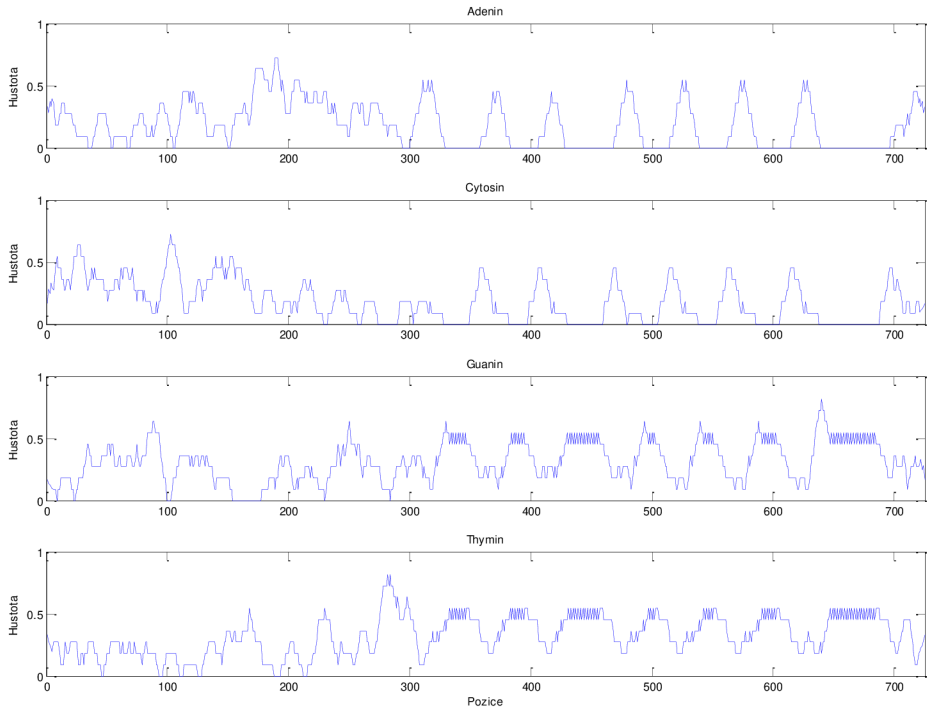


Obrázek 15: Denzitní venktory pro adenin vypočítány různou délkou okna

Dále jsou zpracovány dvě testovací reálné sekvence o stejné délce, jedna bez výskytu tandemových repetic a jedna s výskytem přirozených TR (viz tabulka 11). Obě jsou zpracovány oknem délky 11.



Obrázek 16: Denzitní vektory sekvence bez výskytu TR



Obrázek 17: Denzitní vektory genu BOVTGN

Zatímco v případě první testovací sekvence nejsou na grafech denzitních vektorů patrné žádné významné motivy, v případě druhé lze jasně pozorovat opakující se oblasti značně podobného charakteru. Při vhodně zvoleném okně je výskyt TR obvykle charakterizován významně delší oblastí malých oscilací. Ty jsou patrné na druhém obrázku.

4.2 Fourierova transformace indikačních vektorů

Zatímco výpočet denzitních vektorů umožnil poměrně přesnou alokaci oblastí s výskytem tandemových repetic, ke specifikaci struktury repetic se díky jejich periodickému tvaru jeví jako vhodná Fourierova transformace indikačních vektorů.

$$F_n(x) = \sum_{n=0}^{N-1} I_n(n) \cdot e^{\frac{-2 \cdot \pi \cdot i \cdot x \cdot n}{N}},$$

kde F_n je funkce spektra n-tého nukleotidu, indikačního vektoru I_n a N je délka indikačního vektoru. x ve funkci F_n značí spektrální koeficient.

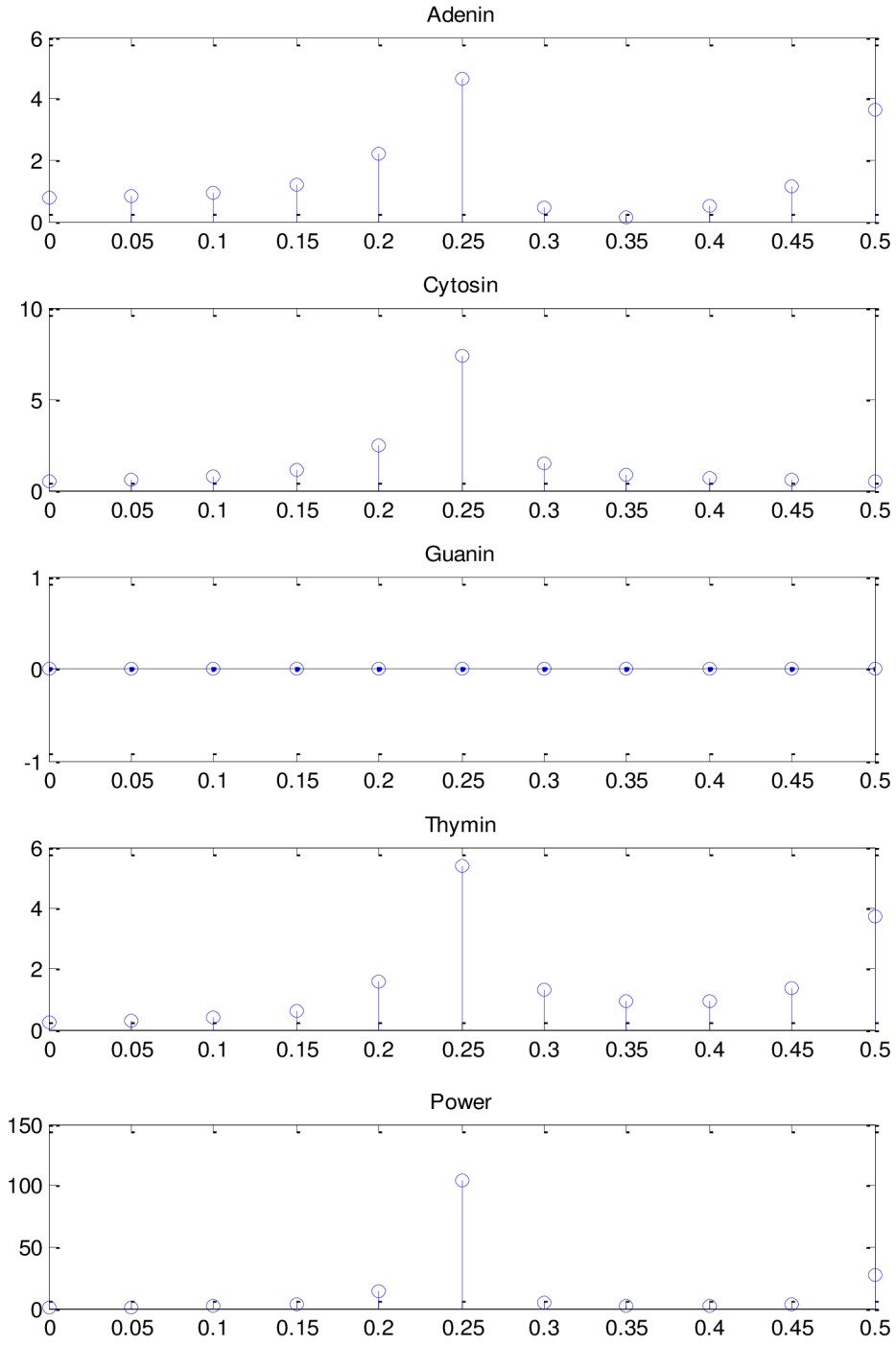
Fourierovou transformací lze získat spektrum jako vektor spektrálních

koeficientů, z nichž relevantní je první polovina. Spektrální koeficienty označují periodicitu o určité délce. Spektrální koeficient s hodnotou $\frac{N}{4}$ značí opakování čtyřprvkového motivu, koeficient $\frac{N}{9}$ analogicky opakování devítiprvkového motivu. Koeficient 0 je určen výškou nulové izolinie. Protože je transformace počítána pouze z hodnot 0 nebo 1, tedy z funkce ekvivalentní obdélníkovým průběhům, jsou spektra zatížena určitou mírou šumu. Je tedy výhodné provést výpočet výkonového spektra, a z jeho tvaru následně odhadovat výskyt repetičních struktur. Výpočet se provede podle následujícího vzorce:

$$V(x) = |F_A|^2 + |F_C|^2 + |F_G|^2 + |F_T|^2,$$

kde V je výsledné výkonové spektrum a F_n jsou spektra jednotlivých nukleotidů.

Na dalším obrázku můžeme vidět spektra pro jednotlivé nukleotidy a výkonové spektrum krátké sekvence s výskytem tetranukleotidové repetice. Hodnota spektrálního koeficientu 0,25 je zvláště vysoká ve výkonovém spektru; z míry jeho zastoupení v ostatních spektrech lze vyvodit, že se jedná o repetici složenou z nukleotidů A, C a T, kde C se vyskytuje vícekrát, zatímco A a T jsou rozloženy stejně. Malá míra nestejnosti, pozorovatelná na obrázku mezi A a T je způsobena absencí posledního nukleotidu A v sekvenci. Jedná se tedy o zanedbatelný šum.



Obrázek 18: Spektrální koeficienty krátké sekvence s výskytem TR ve formě "TCCA"

4.3 Program pro grafické vyhledávání tandemových repetic

V programovém prostředí Matlab byl vytvořen program pro vyhledávání specifikovaných repetic v sekvencích nukleotidů. Program je ovládán pomocí uživatelského grafického rozhraní. Je zapotřebí také bioinformatický toolbox.

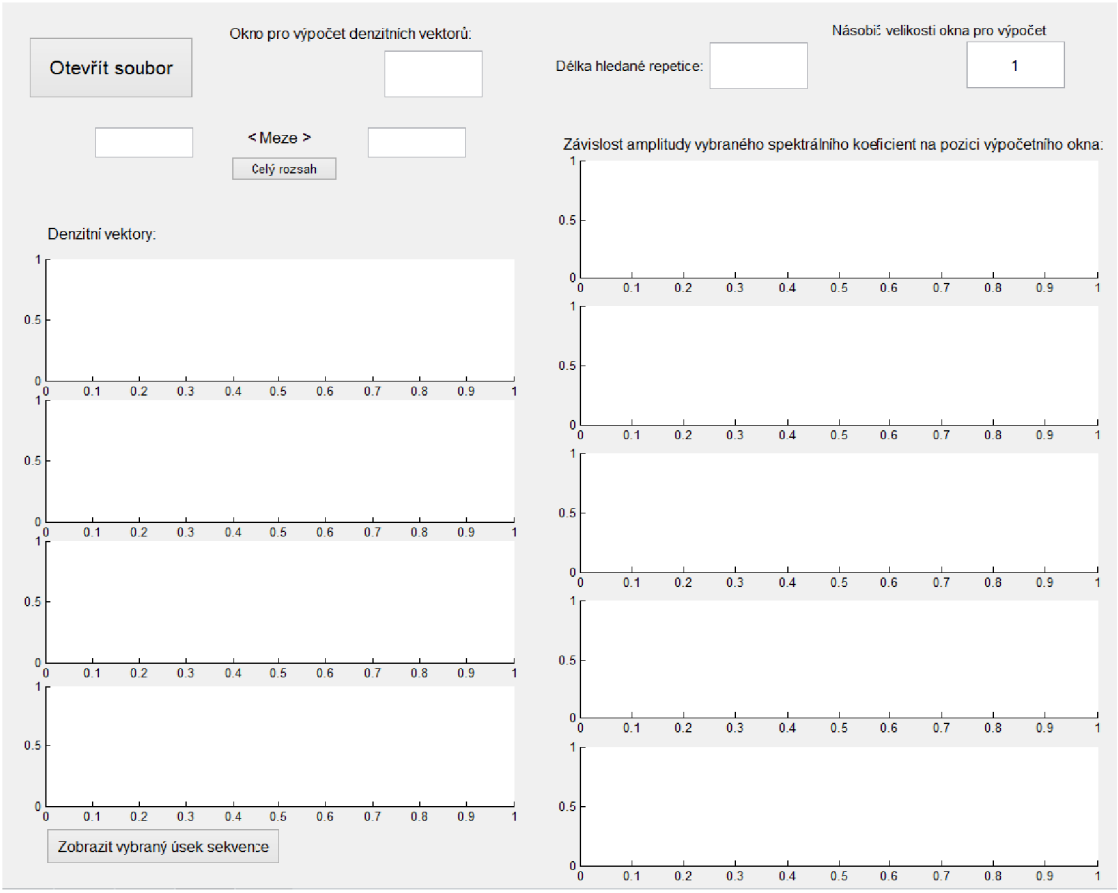
Pomocí programové funkce lze pracovat se soubory ve formátu *fasta* a *genbank*, které jsou standardními formáty, ve kterých lze získat sekvence DNA z databáze GenBank.

Po načtení souboru se specifikuje velikost okna pro výpočet denzitních vektorů. Čtyři grafy vlevo zobrazují vypočtené denzitní vektory. Je možné nastavit meze zobrazení, jedná se v podstatě o zoom v x-souřadnicích.

Vyhledávání repetic není plně automatické, je nutné nejprve nastavit velikost hledané repetice. Program podle toho sám určí velikost okna pro Fourierovu transformaci. Toto okno je dále možné zvětšit pomocí pole násobiče; délka okna je zvětšena jako násobek zadané hodnoty. Délka okna je určena podle vzorce

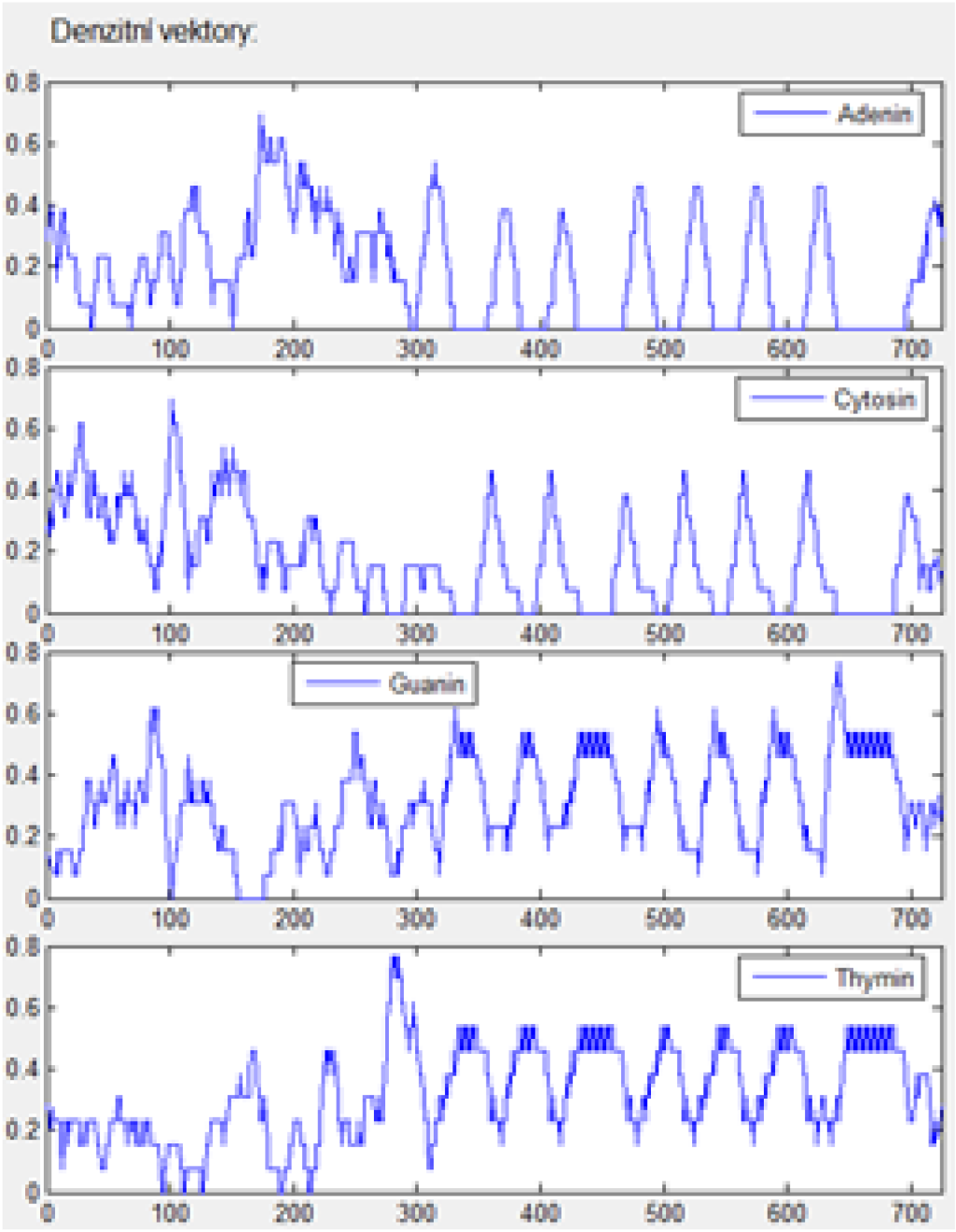
$O = 2 \cdot [(n + 1) + k \cdot n]$, kde n je velikost hledané repetice a k je násobič.

Větší délka okna způsobuje snížení vlivu šumu za cenu rozostření okrajů detekovaných repetic. Grafy zobrazují závislost spektrálního koeficientu související s hledanou repeticí na posunu okna. Meze zobrazení platí i pro tyto grafy. Oblasti s výskytem hledané repetice budou vykazovat vysoké hodnoty spektrálního koeficientu pro danou repetici.

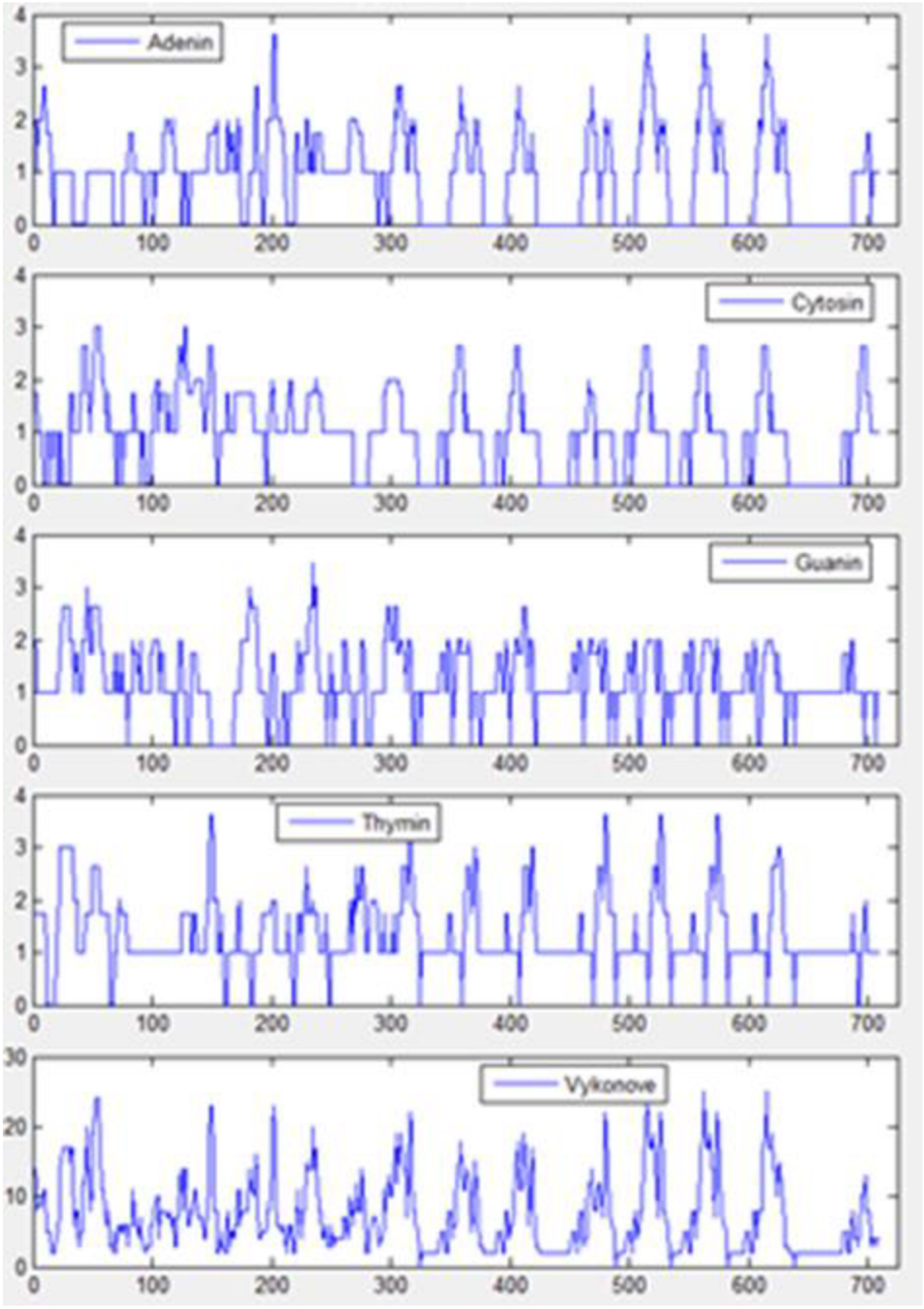


Obrázek 19: Prázdné uživatelské rozhraní programu

Ve spodní části je umístěno tlačítko, které slouží zobrazení mezemi vyhrazené sekvence ve formě nástroje *seqviewer*.



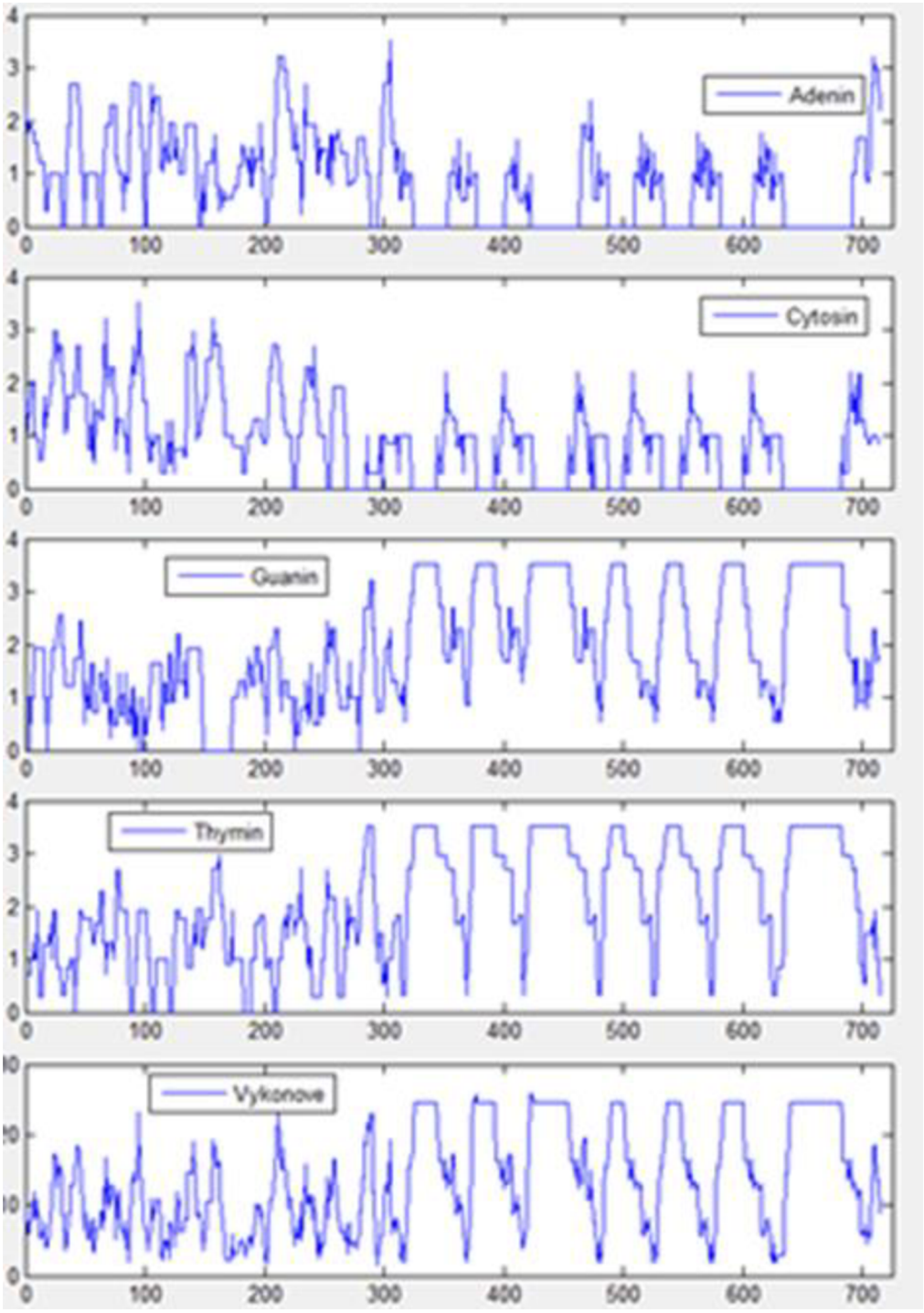
Obrázek 20: Denzity genu BOVTGN - výskyt dinukleotidových TR



Obrázek 21: Závislost vývoje amplitudy spektrálního koeficientu 1/3 na pozici výpočetního okna v sekvenci

Obrázky výše zobrazují typický výstup programu. Z grafů denzitních vektorů jsou patrné oblasti oscilací značící výskyt TR. Grafy amplitud spektrálních koeficientů sice mají periodický charakter, ovšem nenaznačují přímý výskyt satelitů s vybranou délkou repetice 3 nukleotidy. Pokud je nastaveno hledání repetic o délce 2, tedy hledání repetic,

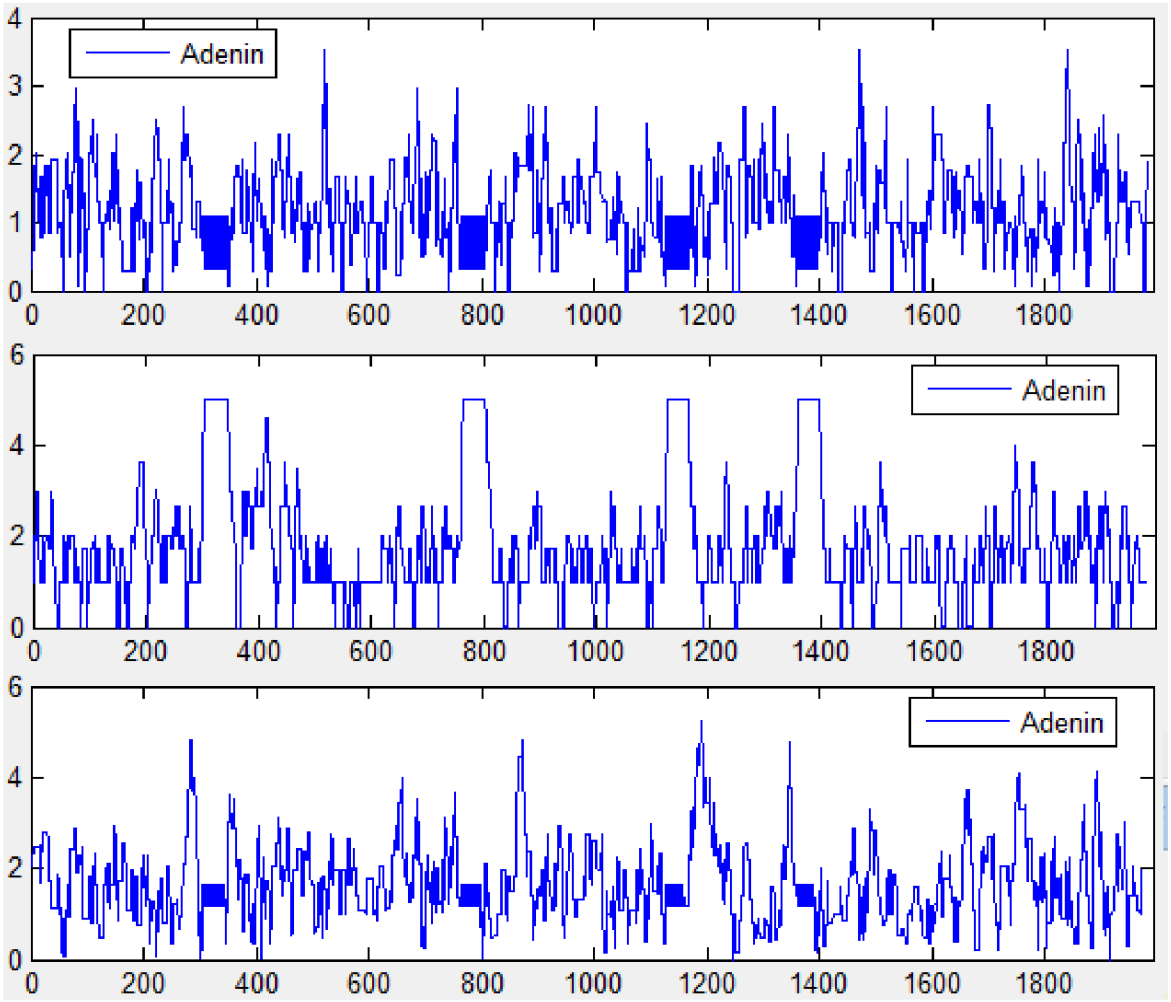
které jsou obsaženy v testovací sekvenci, je výsledek následující:



Obrázek 22: Závislost vývoje amplitudy spektrálního koeficientu 1/2 na pozici výpočetního okna v sekvenci

Z grafů je na první pohled jasný výskyt repetic složených z nukleotidů G a T. Náběžná hrana těchto obdélníků je obvykle ostřejší a tím pádem přesnější, než sestupná

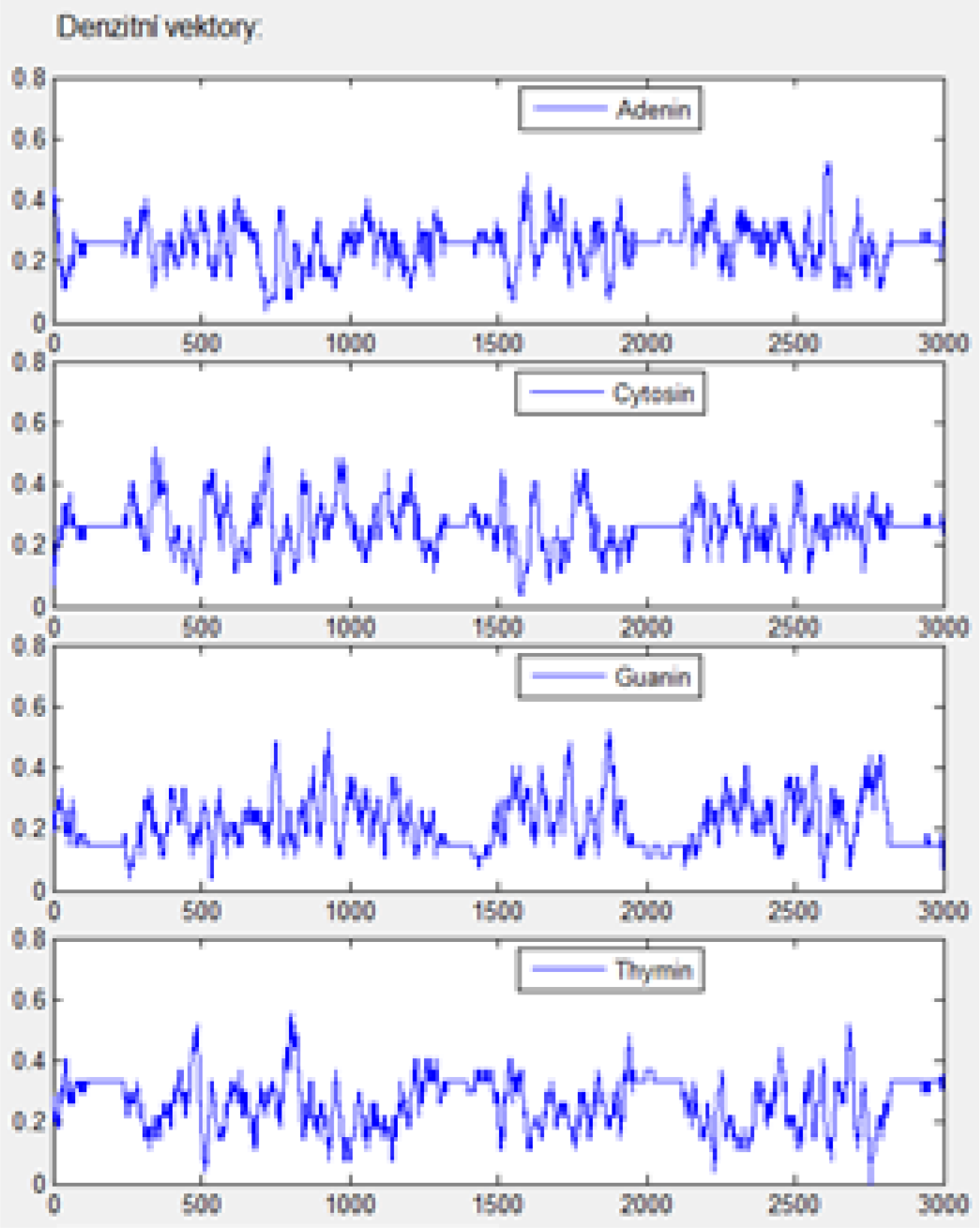
hrana. To je dáno charakterem výpočtu okna Fourierovy transformace.



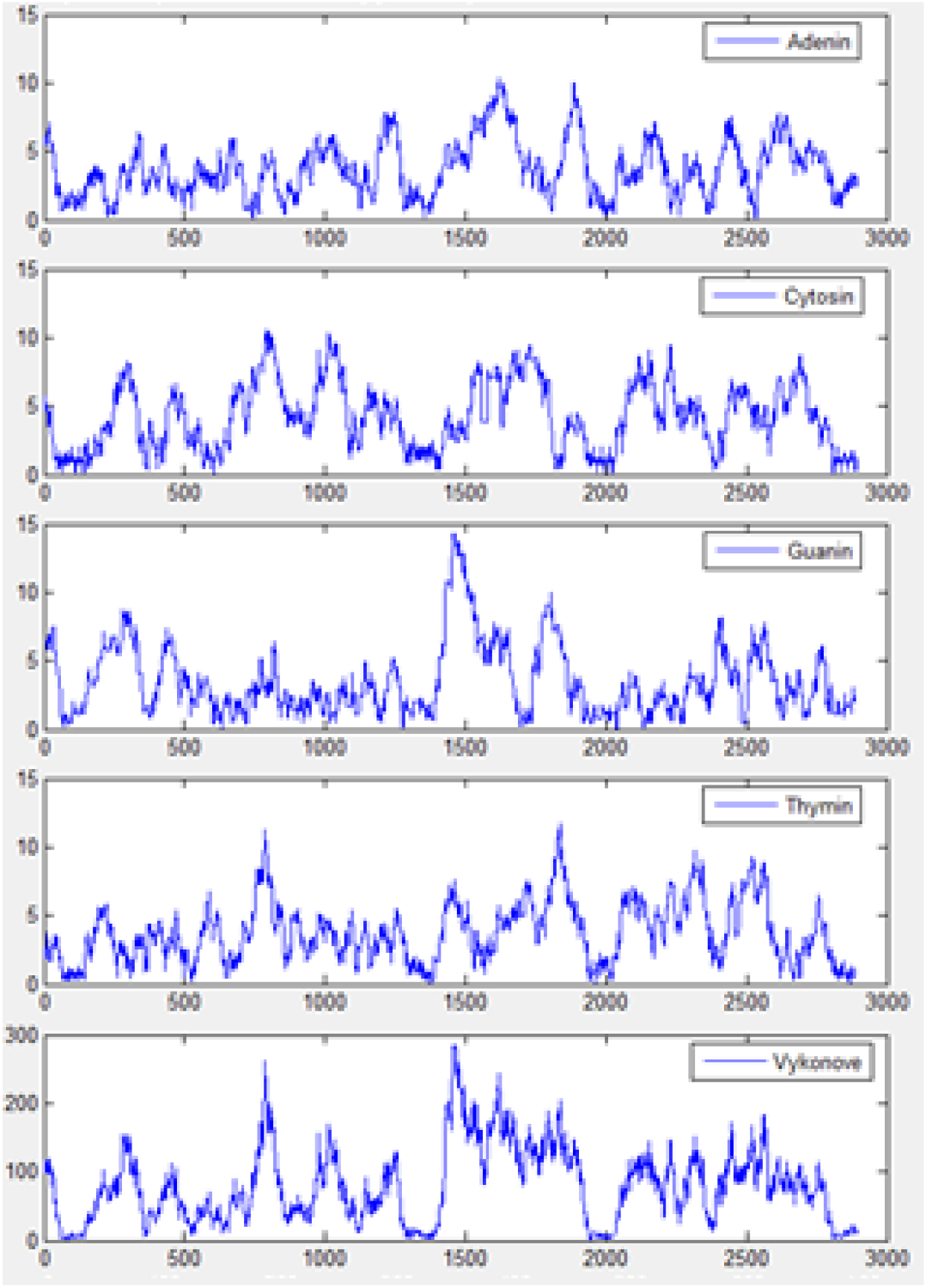
Obrázek 23: Poziční závislosti různých spektrálních koeficientů pro adenin

Příklad testovací sekvence s obsahem trinukleotidových TR – grafy nastavení vyhledávání postupně 2, 3, 4 nukleotidy dlouhé TR. Obdélníky ve druhém grafu značí výskyt trinukleotidových repetic. Tato obdélníková struktura je jasným indikátorem výskytu TR.

Je vidět, že vyhledávání krátkých TR touto metodou je snadné a účinné. Bohužel se zvětšující se délkou TR klesá schopnost detekovat přesně délku pomocí volby spektrálních koeficientů. Nicméně analýza denzitních vektorů zůstává stejná i při výskytu velmi dlouhých satelitů. Pokud je okno pro výpočet denzitních vektorů zvoleno o přesně stejné délce, jako je délka vyskytující se TR, budou oblasti s výskytem reprezentovány vodorovnými čarami, přerušovanými jen lokálními mutacemi. Příklad na obrázku níže. Vyskytující se repetice mají délku 27.



Obrázek 24: Denzity sekvence s výskytem TR o délce 27 počítané oknem stejné délky



Obrázek 25: Poziční závislost amplitudy spektrálních koeficientů výše zmíněné sekvence

Po vykreslení grafického zobrazení si uživatel určí meze s výskytem TR, jak je patrné z daných grafů a sekvenci ohraničenou mezemi si zobrazí tlačítkem „Zobrazit vybraný úsek sekvence“. Sekvence je následně zobrazena v zabudovaném prohlížecím

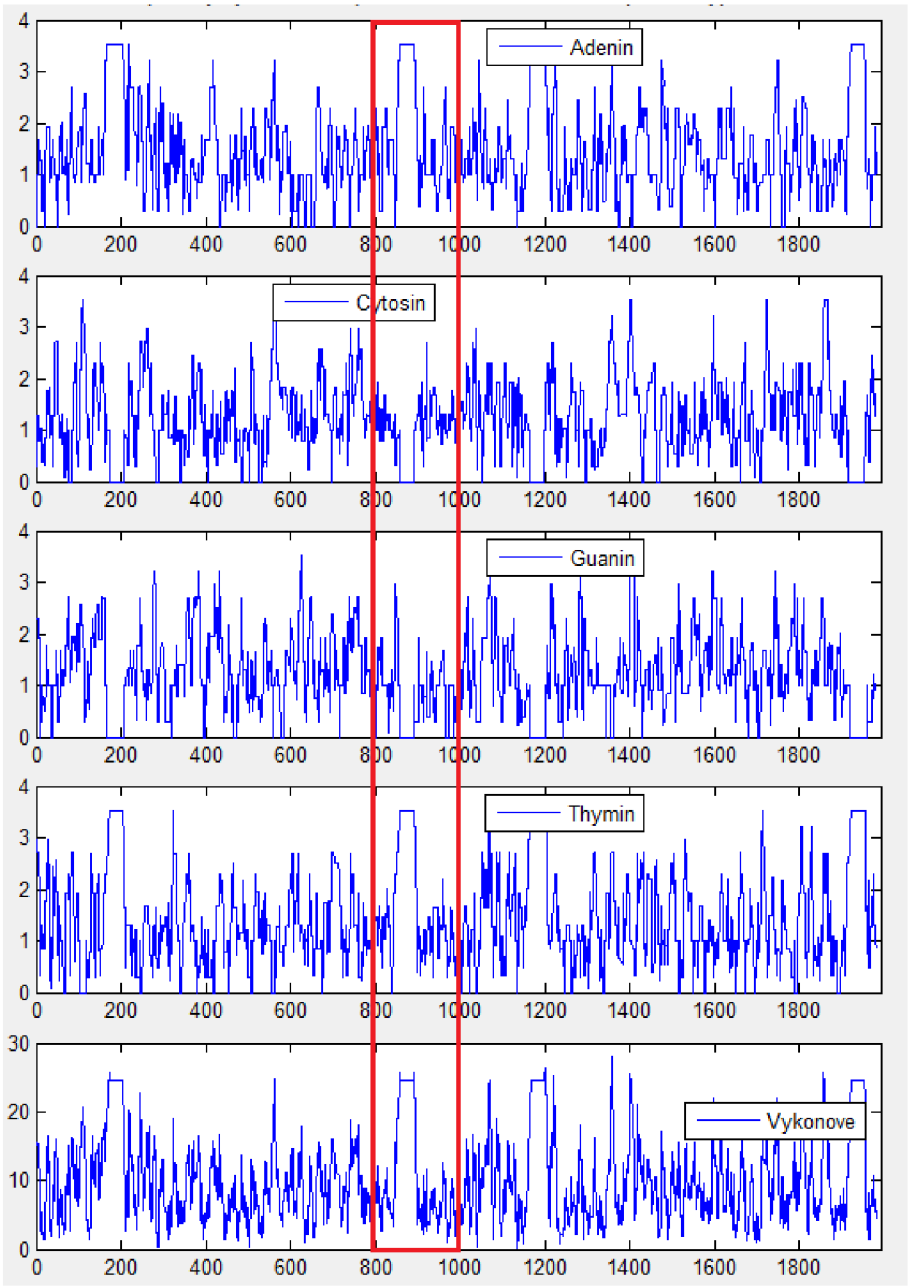
nástroji programového prostředí Matlab.

4.4 Srovnání s online vyhledávači

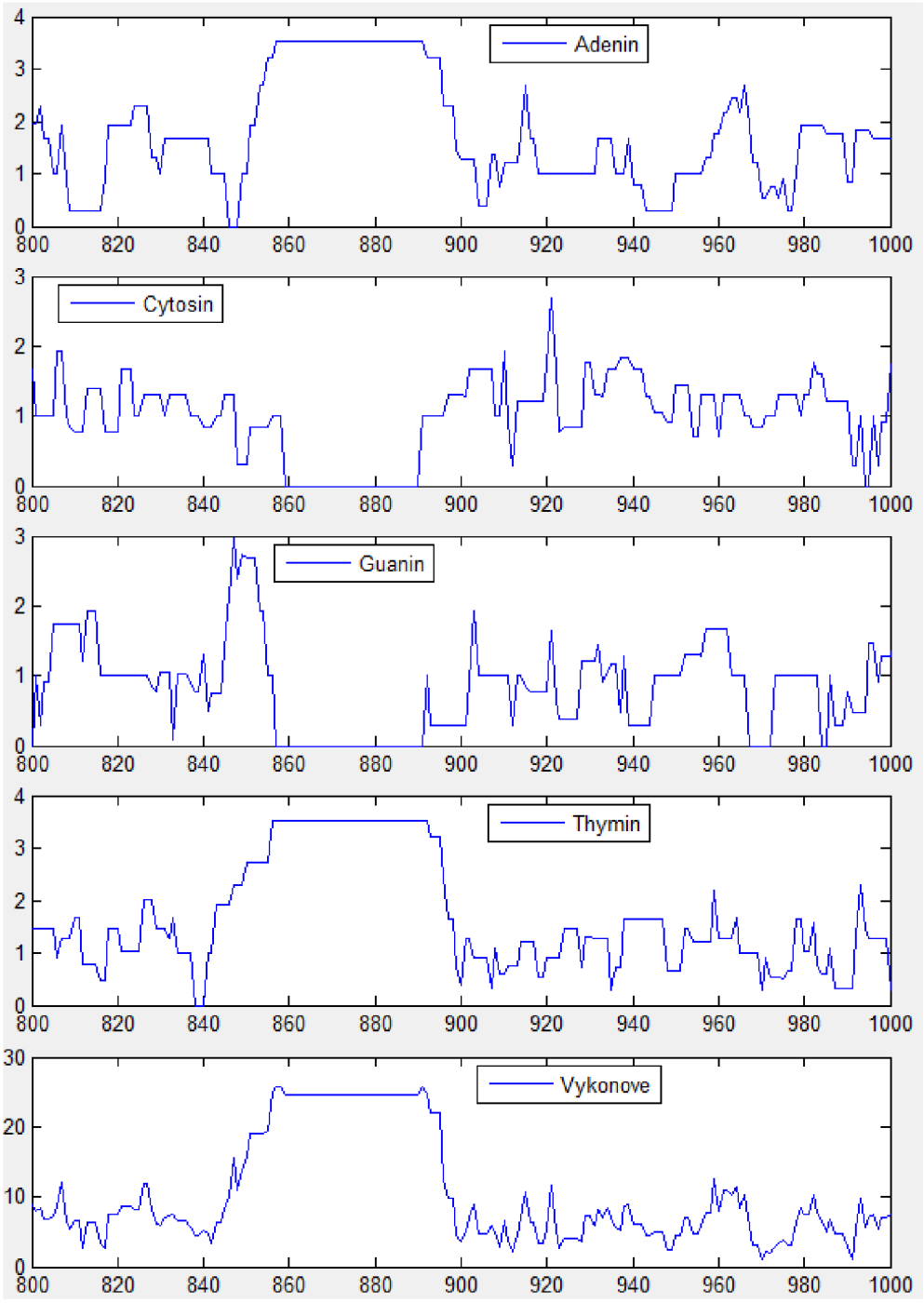
K testování naprogramovaného vyhledávače na principu analýzy denzitních vektorů byly použity stejné sekvence, kterými byly testovány vyhledávače viz kapitola 4.3.1.

Test_1.fasta – dinukleotidové repetice bez mutací

Podle očekávání byly lokace s výskytem bez problémů graficky lokalizovány.



Obrázek 26: Amplitudy spektrálních koeficientů 1/2

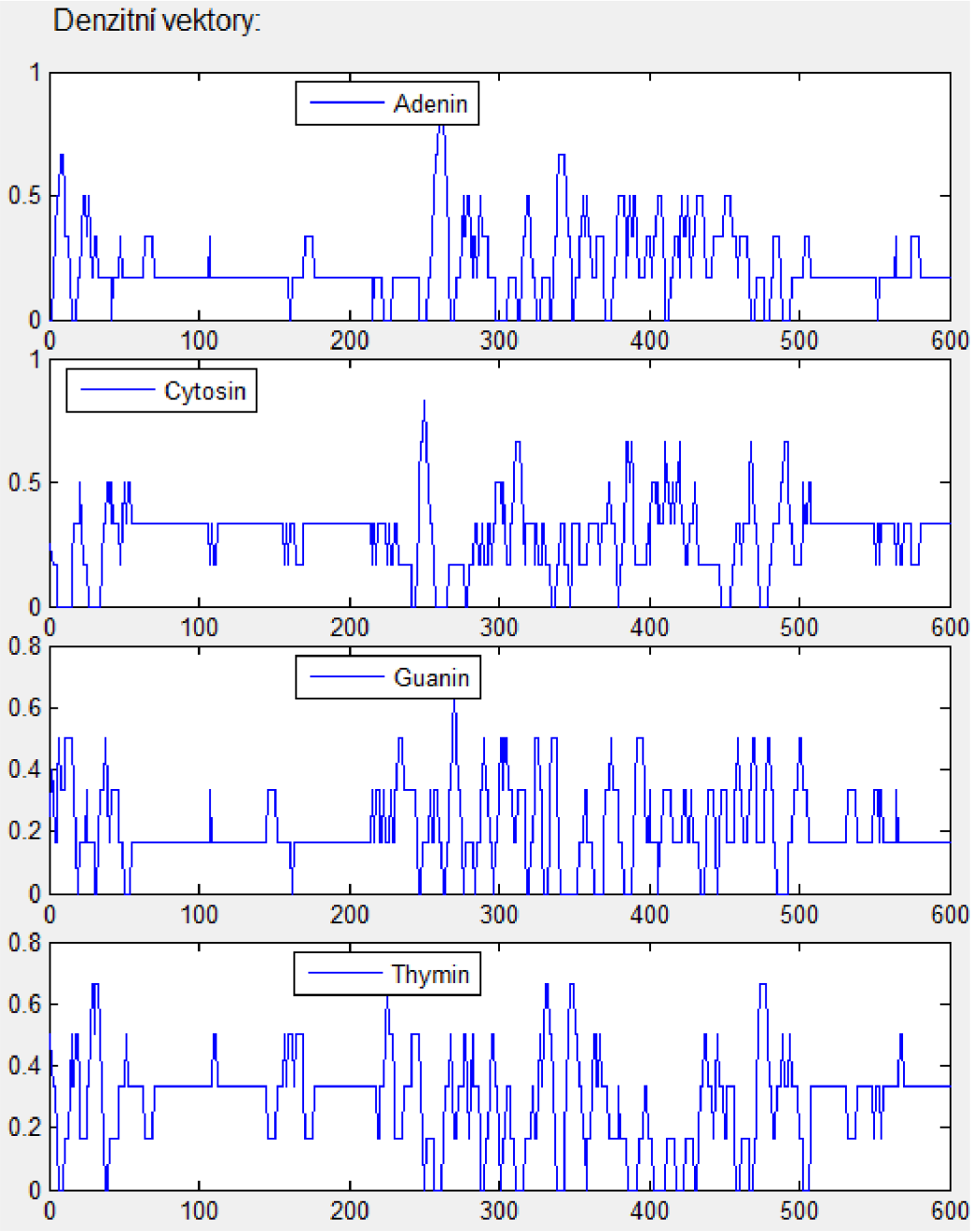


Obrázek 27: Detail na oblast vyznačenou v obrázku 27

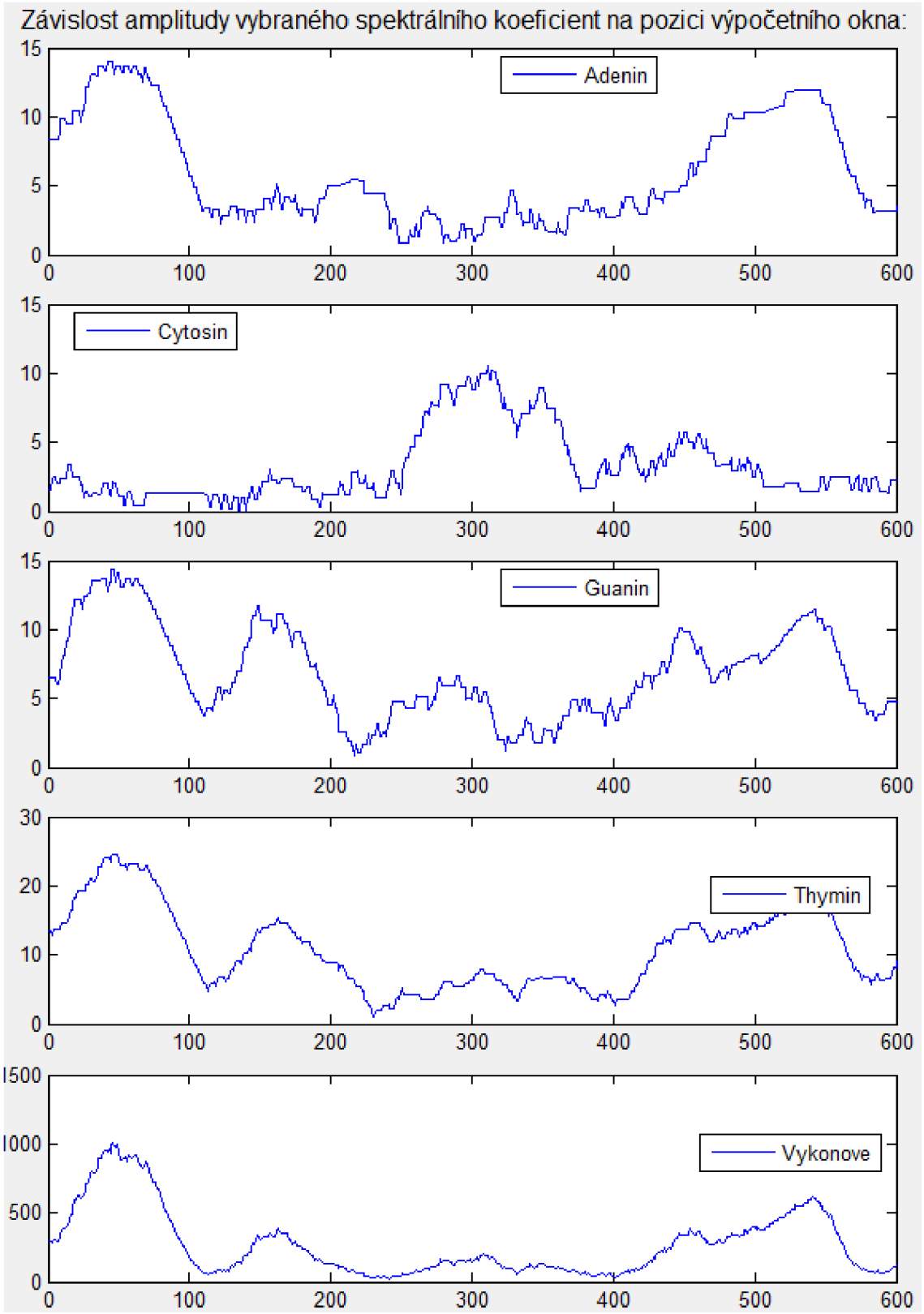
Test_2.fasta – běžná sekvence s výskytem mutací

V tomto případě mají grafy denzitních vektorů poměrně malou výpovědní hodnotu. Lze vyvodit závěr, že přesnost detekce analýzou denzitních vektorů je náchylná k mutacím. Volbou vhodných hodnot hledané repetice a velikosti okna lze na

grafu výkonového spektra rozeznat výskyt repetic ve formě píků. Přesná lokalizace je ovšem problematická.

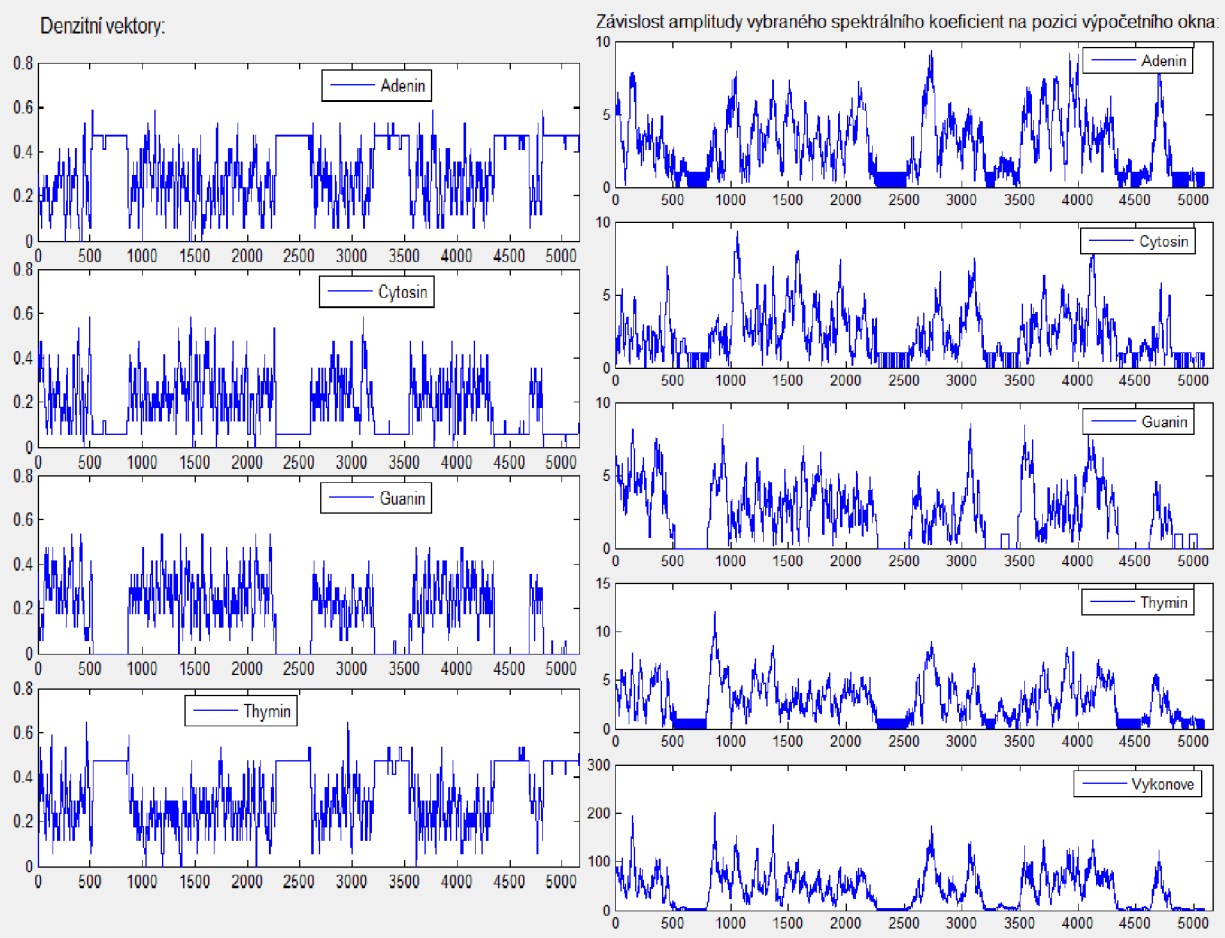


Obrázek 28: Detail denzitních vektorů



Obrázek 29: Detail amplitud spektrálních koeficientů

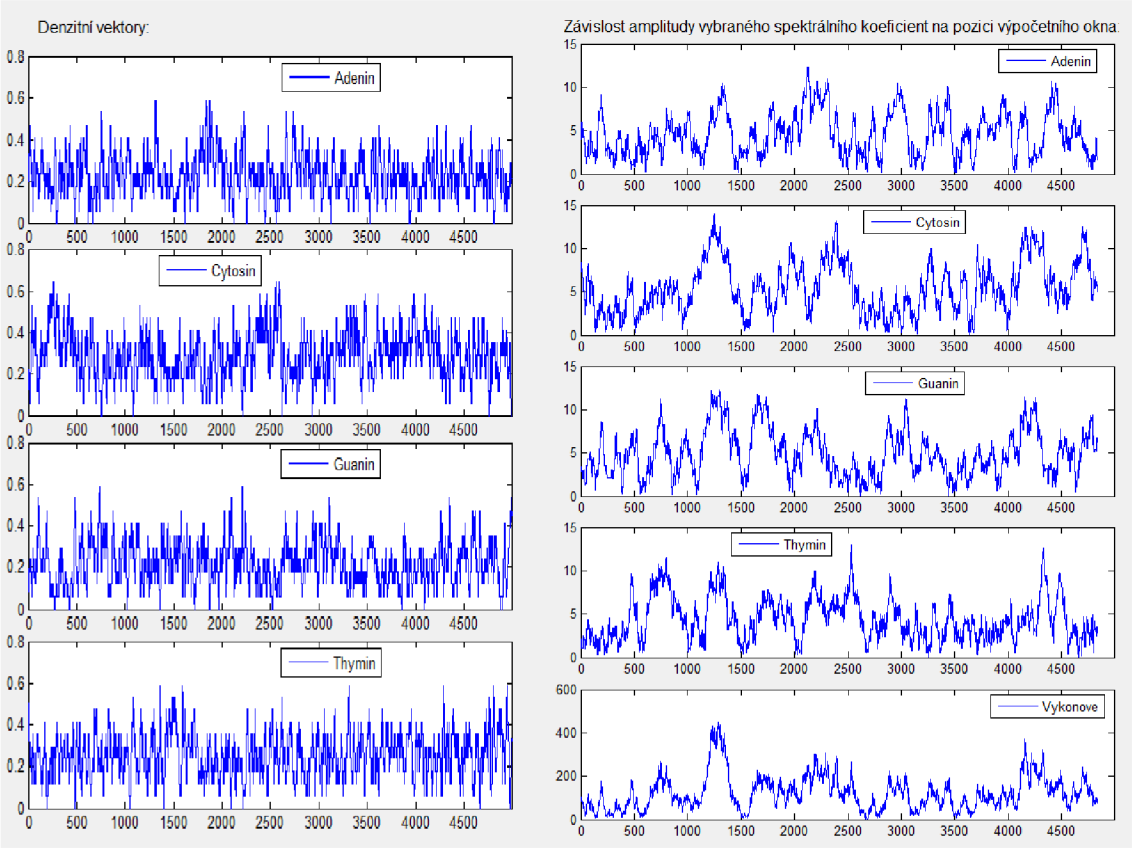
Test_3.fasta – detekce dlouhé, zdánlivé dinukleotidové repetice



Obrázek 30: Grafy denzitních vektorů a amplitud spektrálních koeficientů

Díky tomu, že je testovací sekvence málo zmutovaná, lze TR snadno lokalizovat z grafů denzitních vektorů. Jelikož se nejedná o automatizovanou detekci, nelze považovat detekci výskytu cytosinu za přednost.

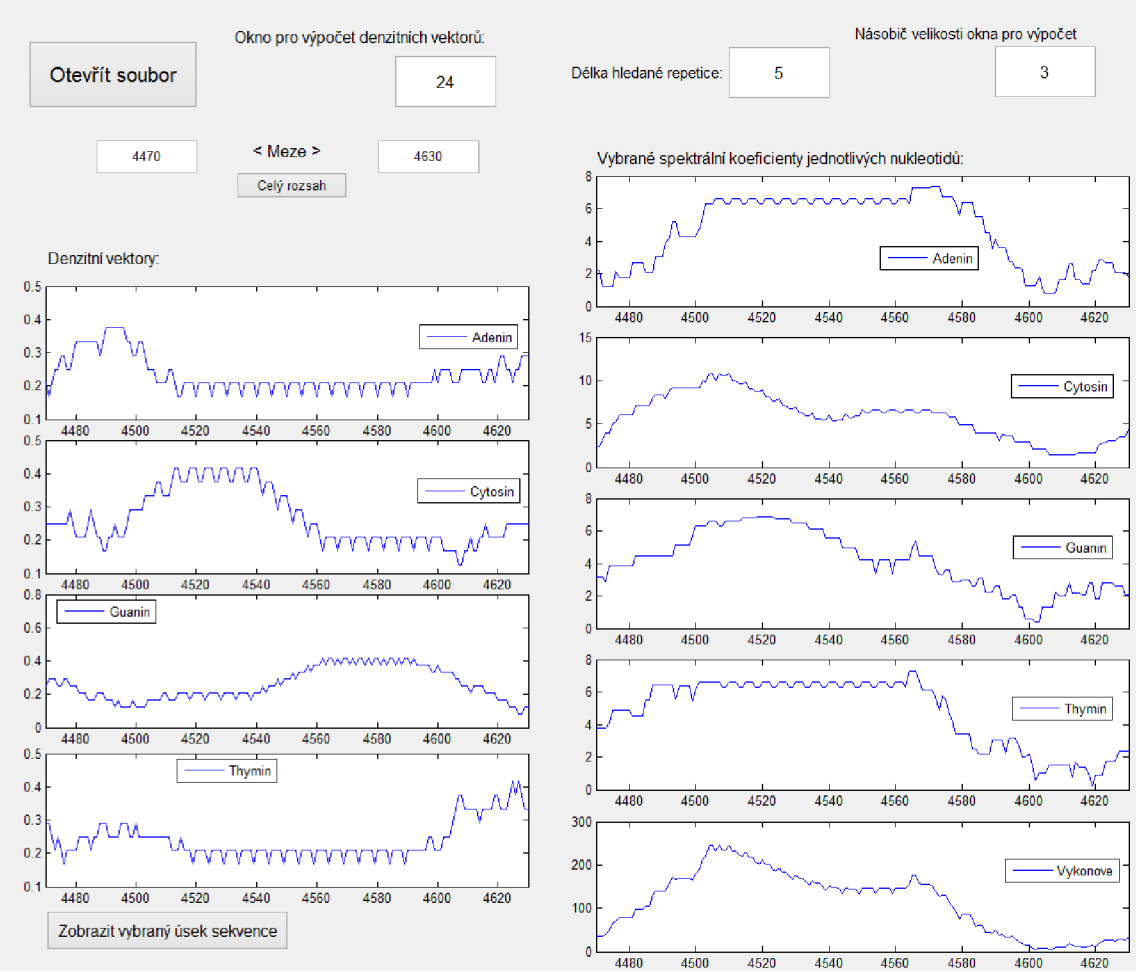
Test_4.fasta – silně zmutované repetice



Obrázek 31: Grafy denzitních vektorů a amplitud spektrálních koeficientů

V tomto případě program neposkytnul žádné vážně interpretovatelné informace. Vzhledem k míře mutace sekvence se jedná o očekávaný výsledek.

Test_5.fasta – evoluční mutace



Obrázek 32: Detail na evoluční mutaci TR

Program byl schopen nalézt vyskytnuvší se repetice, které jsou ale viditelné v detailu. Na obrázku si lze povšimnout, jak se měnily hodnoty denzit a spektrálních koeficientů při evoluční mutaci. Program se k evolučním mutacím chová jako k výskytu nového typu repetice, což ostatně evoluční mutace je, a detekuje ji obdobně. Menší problém nastává, jedná-li se o mutaci ve formě inzerce nebo delece. V takovém případě je nutno změnit hodnotu hledané repetice. Zda je to nutné, napoví mírná změna v hodnotách denzitních vektorů a zásadní změna v hodnotách korelačních koeficientů. Z toho vyplývá, že mutace zobrazená na obrázku je mutace bodová.

4.5 Shrnutí výsledků

Metoda vyhledávání tandemových repetic pomocí analýzy denzitních vektorů a vektorů vývoje spektrálních koeficientů se jeví jako metoda použitelná a efektivní především pro vyhledávání kratších tandemových repetic s nižší až střední mírou mutací. Programové a především konkrétní grafické zpracování limituje využití této metody pro vyhledávání delších repetic nebo v delších sekvencích. Náročnost výpočtu spektrálních koeficientů navíc roste exponenciálně s velikostí vyhledávané repetice.

V porovnání s ostatními programy má ovšem program značnou výhodu ve flexibilitě, především v možnosti okamžité úpravy parametrů a rozsahu zobrazení bez nutnosti opětovného nahrávání vstupních dat, jak je tomu u dvou ze tří ostatních vyhledávačů.

Nevýhodou je nutnost manuální alokace repetic při požadavku na přesné určení výskytu. Na druhou stranu je to nejpřesnější možný způsob detekce a program tuto alokaci usnadňuje.

ZÁVĚR

Cílem této práce bylo seznámení s pojmem tandemové repetice a s metodami jejich vyhledávání. Byly zmíněny pojmy DNA a genomika a jejich význam. Dále byly rozebírány definice a popis tandemových repetic, důvod jejich výskytu, jejich vznik a význam v DNA, resp. buněčném jádře, a jejich vliv na živý organismus. Krátce byly popsány možnosti jejich využití.

Další část práce se věnovala vyhledávání repetic. Byly zmíněny genomické internetové online nástroje k volnému využití, a to především vyhledávače tandemových repetic. Dále byly vybrány tři nástroje podle jejich dostupnosti a využitelnosti, které byly analyzovány tak, aby mohly být popsány algoritmy, podle kterých pracují, a protože jejich algoritmy jsou příliš složité, byl alespoň stručně popsán princip a vysvětlení parametrů. Dále byla zhodnocena jejich uživatelsky snadná ovladatelnost.

Pro další část byl v programovém prostředí Matlab naprogramován generátor sekvencí s obsahem tandemových repetic pro další testování vybraných internetových vyhledávačů. Generátor obsahuje velké množství vstupních parametrů pro vytvoření vhodných sekvencí k otestování možností a různých vlastností daných vyhledávačů. Poté bylo provedeno samotné testování. Testovalo se na pěti vygenerovaných sekvencích různé povahy. Dále byly prezentovány, srovnány a zhodnoceny dosažené výsledky jednotlivých vyhledávačů. Poté bylo provedeno zhodnocení samotných vyhledávačů.

V další části byla popsána nová metoda vyhledávání tandemových repetic, vytvořeny algoritmy a sestrojen program s uživatelským grafickým rozhraním v programovém prostředí Matlab. Induktivním způsobem bylo popsáno, jak program pracuje a jak se obsluhuje.

Tato metoda byla následně otestována na několika testovacích sekvencích, a to především na modelových případech, kterými byly testovány internetové vyhledávače. Výsledky testování programu byly srovnány s výsledky online vyhledávačů a následně zhodnoceny.

V poslední řadě následovalo zhodnocení této metody vyhledávání tandemových repetic.

LITERATURA

[1] Doc. RNDr. Jiří Dostál, CSc., RNDr. Hana Paulová, CSc, Mgr. Jiří Slanina, Ph.D., Prof. RNDr. Eva Táborská, CSc; Biochemie pro posluchače bakalářských oborů

[2] SNUSTAD, D. Peter a Michael J. SIMMONS. Genetika. 1. vydání. Brno : Nakladatelství Masarykovy univerzity, 2009. 894 s. ISBN 978-80-210-4852-2.

[3] SOUKUPOVÁ, Milena a František SOUKUP. *Kapitoly z lékařské biologie a genetiky II.* 1. vydání. Praha : Karolinum, 1998. 98 s.

[4] KAPRAS, Jan a Milada KOHOUTOVÁ. Kapitoly z lékařské biologie a genetiky III. 1. vydání. Praha : Karolinum, 1999. 101 s. ISBN 80-246-0001-3

[5] Fondon, J. W., & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences* **101**, 18058–18063 (2004)

[6] P. Z. Myers, Ph.D. (University of Minnesota, Morris) © 2007 Nature Education Citation: Myers, P. (2007) Tandem repeats and morphological variation. Nature

[7] Aktuální genetika - repetitivní DNA. ÚSTAV BIOLOGIE A LÉKAŘSKÉ GENETIKY 1.LF UK A VFN. *Multimediální učebnice lékařské biologie, genetiky a genomiky* [online]. 2005-2006 [cit. 2014-01-05]. Dostupné z: http://biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm

[8] KRISHNAN, A., TANG, F. Exhaustive whole-genome tandem repeats search. Bioinformatics, 2004, roč. 20, no. 16, s. 2702-2710.

[9] HAUTH, A. M., JOSEPH, D. A. Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinformatics, 2002, roč. 18, suppl. 1, s. S31-S37.

[10] Knight, Julian C. (2009). Human Genetic Diversity: Functional Consequences for Health and Disease. Oxford University Press. p. 167. ISBN 978-0-19-922769-3.

[11] Wyman AR, White R (November 1980). "A highly polymorphic locus in human DNA". Proc. Natl. Acad. Sci. U.S.A. 77 (11): 6754–8.

[12] D. B. Goldstein, A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman (1995). "An Evaluation of Genetic Distances for Use With Microsatellite Loci". Genetics 139 (1): 463–471.

[13] Q-Y. Huang, F-H. Xu, H. Shen, H-Y. Deng, Y-J. Liu, Y-Z. Liu, J-L. Li, R. R. Recker and H-W. Deng (2002). "Mutation Patterns at Dinucleotide Microsatellite Loci in Humans". *Am J Hum Genet.* **70** (3): 625–634.

[14] Kaukinen KH, Supernault KJ, and Miller KM (2004). "Enrichment of tetranucleotide microsatellite loci from invertebrate species". *Journal of Shellfish Research* **23** (2): 621.

[15] Variable Numbers of Tandem Repeats (VNTRs). ROYAL VETERINARY COLLEGE. *Royal Veterinary* College [online]. 2004 [cit. 2014-01-05]. Dostupné

z: http://www.rvc.ac.uk/review/DNA_1/4_VNTRs.cfm

[16] Torres-Cruz, Joshua and van der Woude, Marjan W. (2003). "Slipped-Strand Mispairing Can Function as a Phase Variation Mechanism in *Escherichia coli*". *Journal of Bacteriology* **185** (23): 6990–6994

[17] NCBI. *Molecular Biology of the Cell. 4th edition.* [online]. 2002 [cit. 2014-01-05]. Dostupné z: http://www.ncbi.nlm.nih.gov/books/NBK26837/

[18] What is FASTA format?. UNIVERSITY OF MICHIGAN. *Zhang Lab* [online]. 2007 [cit. 2014-01-05]. Dostupné z: http://zhanglab.ccmb.med.umich.edu/FASTA/

[19] TRedD software. BROOKLYN COLLEGE OF THE CITY UNIVERSITY OF N.Y. *Tandem Repeats Database* [online]. 2007 [cit. 2014-01-05]. Dostupné z: http://tandem.sci.brooklyn.cuny.edu/Algorithm.do

[20] How does Tandem Repeats Finder work?. *Laboratory for Biocomputing and Informatics* [online]. 2006 [cit. 2014-01-05]. Dostupné z: http://tandem.bu.edu/trf/trfdesc.html

[21] Mayer, Christoph, Phobos 3.3.11, 2006-2010, <http://www.rub.de/spezzoo/cm/cm_phobos.htm>.