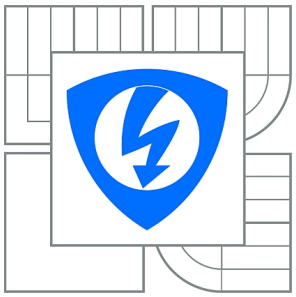
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY VYHLEDÁVÁNÍ TANDEMOVÝCH REPETIC V DNA SEKVENCÍCH

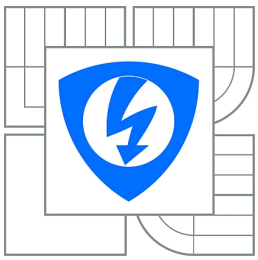
METHODS OF DNA TANDEM REPEATS ANALYSIS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — KRYŠTOF HAVLÍK
AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ
SUPERVISOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Student:	Kryštof Havlík	***ID:***	153523
Ročník:	3	***Akademický rok:***	2014/2015

NÁZEV TÉMATU:

Metody vyhledávání tandemových repetic v DNA sekvencích

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 reálných dat a uměle vytvořených sekvencích otestujte alespoň 3 volně dostupné vyhledávače. 3) V libovolném programovém prostředí naprogramujte převod sekvencí DNA do vhodně zvoleného numerického formátu. 4) Navrhněte metodu vyhledávání tandemových repetic pomocí analýzy genomických signálů, navrženou metodu implementujte v libovolném programovém prostředí a proveďte analýzu na souboru dat. 5) Výsledky porovnejte s volně dostupnými vyhledávači a diskutujte.

DOPORUČENÁ LITERATURA:

[1] KRISHNAN, A. a TANG, F. Exhaustive whole-genome tandem repeats search. Bioinformatics. 2004, 20(16), 2702-2710.
[2] HAUTH, A. M. a JOSEPH, D. A. Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinformatics. 2002, 18, suppl. 1, S31-S37.

Termín zadání:	9.2.2015	***Termín odevzdání:***	29.5.2015

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

ABSTRAKT

V této práci je objasněna základní problematika repetitivní DNA a jsou zde zhodnoceny vyhledávače tandemových repetic bezplatně dostupné široké veřejnosti. V praktické části práce je popsán algoritmus sloužící k vyhledávání tandemových repetic založený na převodu sekvence DNA do číselného formátu. Následuje zpracování vzniklého signálu pomocí krátkodobé Fourierovy transformace, vytvoření spektrogramu, jehož analýza slouží k nalezení pozice a obsahu repetitivních oblastí. Výsledkem práce je porovnání výstupů vybraných volně dostupných programů s vytvořeným programem a zhodnocení výhod a nevýhod.

KLÍČOVÁ SLOVA

Repetitivní DNA, tandemové repetice, Fourierova transformace, numericky reprezentovaná DNA.

ABSTRACT

This work clarifies basics of the repetitive DNA and evaluates thee tandem repeats finders, which are accessible to public free of payment. Second part describes an algorithm used for searching tandem repeats, based on converting DNA string into numerical signal. Then follows signal processing using short-time Fourier transform, formation of spectrogram and analysis for evaluating position and content of repetitive areas. Result of this work is comparison of outcomes provided by public accessible programs with results of created program and review of advantages and disadvantages.

KEYWORDS

Repetitive DNA, tandem repeats, Fourier transform, numerically represented DNA.

HAVLÍK, K. *Metody vyhledávání tandemových repetic v DNA sekvencích.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 70 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Metody vyhledávání tandemových repetic v DNA sekvencích 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 bakalářské práce ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne ....

(podpis autora)

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Všichni jedinci jednoho druhu mají stejnou sadu genů, a přesto se objevují výrazné fenotypové rozdíly mezi jedinci téhož druhu. Výzkum v poslední době naznačuje, že některé z těchto odlišností jsou způsobeny tandemovými repeticemi, jinak řečeno krátkými úseky DNA, které se opakují mnohokrát za sebou v určitém genu.

Nekódující části mohou u některých organismů dosahovat až 90 % celého genomu a přestože dříve panoval názor, že tato nadbytečná DNA nemá větší význam, v poslední době vědci nachází stále větší spojitosti s jejími opakujícími se částmi a projevem fenotypu u různých druhů. Setkáváme se rozptýlenými repeticemi, které jsou v podstatě roztroušené v celé DNA. Tyto repetice dosahují vysokého procenta obsahu v genomu, ale tím, že jsou roztroušené na dlouhém řetězci, není jejich význam tak velký, jako u tandemových repetic. Právě tandemové repetice a jejich vyhledávání je tématem této bakalářské práce.

V první části práce je popsáno, co to vlastně tandemové repetice jsou a jak vznikají. Tím, že se nachází ve větším počtu těsně za sebou, tak na rozdíl od rozptýlených repetic mají již větší dopad na organismus. Fenotypově od sebe oddělují jednotlivce jednoho druhu, ale jejich působení nemusí být jen takovéto. Mohou způsobovat velké množství nejrůznějších onemocnění a to nejen u člověka. K nejznámějším onemocněním způsobeným tandemovými repeticemi můžeme zařadit například Huntingtonovu chorobu, která je způsobena rozšířením několika opakování určité trojice nukleotidů za sebou, čímž vzniká nadbytek proteinu huntingtin. Této chorobě a dalším je také věnována jedna kapitola této práce. Další kapitola se věnuje numericky reprezentované DNA, což je vyjádření genetické informace v takové podobě, aby bylo možné při zpracování použít nejrůznější matematické operace, které nejsou aplikovatelné na řetězce znaků, což je vstupní informace pro jakýkoliv algoritmus zaměřený na mapování DNA. Mapování DNA v poslední době často souvisí s prokazováním identity, a to nejen u člověka. Tandemové repetice tedy mohou být důležité v tzv. DNA-fingerprintingu, což je metoda rozeznávání jednotlivců pomocí DNA. Tato metoda se nemusí vztahovat pouze na určování paternity, či využívání v kriminalistice, jak se domnívá široká veřejnost, ale dá se používat i při určování jednotlivých druhů bakterií.

Dále jsou popsány metody vyhledávání tandemových repetic a jsou vybrány tři volně přístupné vyhledávače: Tandem Repeats Finder, IMEx: Imperfect satellite extractor a Phobos. U těchto programů jsou zjednodušeně popsány jejich algoritmy výpočtu a vyhodnocování tandemových repetic, načež jsou otestovány na souboru různých dat. Část testovacích dat je uměle vytvořena v programovacím prostředí MATLAB, aby se dala otestovat přesnost vyhledávačů a jejich základní funkce.

V programové části je popsán algoritmus navržený pro vyhledávání tandemových repetic a všechny ostatní pomocné funkce. V programovém prostředí Matlab je vytvořen program srovnatelný s ostatními vyhledávači, které jsou veřejnosti volně přístupné.

Závěrečná část se věnuje testování vytvořeného programu a porovnává výsledky jím dosažené s výsledky ostatních testovaných vyhledávačů.

1. DNA

Jak důležitou roli má jádro buňky v dědičnosti začalo být zřejmé v 70. letech 19. století, při pozorování jader samčí a samičí pohlavní buňky při procesu oplození. Toto pozorování naznačovalo přítomnost něčeho uvnitř spermie a vajíčka, co způsobovalo dědění vlastností organismu. Dalším obrovským poznatkem bylo objevení chromozomů a pozorování jejich rozdělení při dělení buňky. Tato skutečnost zaručuje předání identické sady chromozomů, každé dceřiné buňce. Na přelomu 19. století bylo jasné, že počty chromozomů jsou vnitrodruhově identické, ale mezidruhově se liší. Pozdější výzkumy ukázaly, že chromozomy se vyskytují ve více stádiích a že tvoří samotnou DNA. Samotnou strukturu popsali a první správný trojrozměrný model pravotočivé dvoušroubovice představili v roce 1953 James Watson a Francis Crick na Cambridge University.

Základními stavebními jednotkami jsou nukleotidy[1], které obsahují jednu ze čtyř bází. Tyto čtyři báze jsou:

- Adenin (A)
- Guanin (G)
- Cytosin (C)
- Tymin (T)

V každém místě dvoušroubovice tvoří adenin pár s tyminem, podobně jako cytosin s guaninem. Tento jev se nazývá komplementarita bází a byl také popsán J. Watsonem a F. Crickem. Jedno vlákno DNA je spojeno s komplementárním protějškem vodíkovými můstky. Adenin s tyminem se pojí dvěma můstky, cytosin s guaninem třemi. Sousední nukleotidy ve vláknu jsou spojeny fosfodiesterovou vazbou, z čehož vyplývá, že na každém konci vlákna bude jiný konec. Konec, na kterém „vyčnívá“ cukernatá část označujeme jako 3' konec, opačný konec s fosfátovým zbytkem označujeme jako 5' konec. Vlákna se vždy spojují opačnými konci k sobě.

Přesný popis struktury vědcům naznačil, že sekvence DNA může být zkopírována použitím každého vlákna jako předlohy pro vznik nové komplementární sekvence bází. Dále ukázal, že genetická informace je zakódována ve sledu bází a že změny v genetické informaci vedou k mutacím. [1]

Jakožto komplexní struktura se dá DNA popsat v několika úrovních. Primární strukturou se rozumí posloupnost nukleotidů v jednom vlákně. Sekundární struktura popisuje DNA jako pravotočivou dvoušroubovici pospojovanou vodíkovými můstky. Terciární strukturou je dáno uspořádání celé DNA do chromatinu, a provázání s histony[2].

[1] Látky složené z dusíkaté báze (purinové, nebo pyrimidinové), deoxyribózy a zbytků kyseliny fosforečné

[2] Bílkoviny, které podporují stavbu chromatinu

2. Repetitivní DNA

Na konci šedesátých let dvacátého století začalo být vědcům jasné, že eukaryotický genom obsahuje oproti prokaryotickému vysoké procento nekódujících, často se opakujících částí DNA.

Při denaturaci DNA (rozpojování dvoušroubovice) teplem, dochází k rozpadu vazeb mezi nukleotidy dvou vláken. Díky těmto vazbám, tzv. vodíkovým můstkům vzniká pro DNA typická dvoušroubovice. První náznaky výskytu repetitivní DNA byly objeveny díky procesu renaturace. Tento proces se vyskytuje po denaturaci teplem a následným ochlazováním prostředí, ve kterém komplementární vlákna hledají možnost jak se spojit. S představou jednoho vlákna DNA o délce 10 000 nukleotidů, ve kterém se nenachází žádná opakující se sekvence, by renaturace trvala řádově týdny, nebo měsíce, než by došlo k opětovné asociaci komplementárního vlákna. Reálně by došlo k renaturaci takového vlákna v mnohem kratším čase.

Délka renaturace závisí na četnosti a délce repetitivních sekvencí. Množství opakujících se úseků může u určitých druhů přesahovat 90 % celkového objemu DNA. [22]

Množství DNA většiny organismů, se tedy může zdát jako nadbytečné. Na tento fakt poukazuje tzv. C-value paradox. Pro příklad haploidní buňka člověka obsahuje asi 3,3 miliard bp, oproti měňavce, jejíž DNA obsahuje dohromady asi 200 miliard bp. Průměrná kostnatá ryba má asi 300 miliard bp, ryby z řádu čtverzubcovitých mají DNA o délce asi 0,5 miliardy bp. Paradox tedy spočívá v tom, jak je možné, že je k „vytvoření“ průměrné kostnaté ryby potřebujeme 600x více DNA než k vytvoření jiné kostnaté ryby? Co dělá v genomu takové množství DNA, které se zdá nepotřebné? To je zásadní otázka, na kterou se hledá odpověď. [18]

V celém lidském genomu[3] tvoří okolo 1,2 % protein-kódující DNA, tato informace je základem pro expresi určitého genu. Značný podíl genetické informace tvoří nekódující sekvence, které byly dříve označovány za zbytečné, přebytkové nebo vyplňující. Tyto části nenesou přímou informaci o expresi genu a nejsou transkribovány. Do této skupiny jsou řazeny signální sekvence neboli regulační oblasti o expresi určitého genu, dále repetitivní sekvence, jejichž délka může být tak variabilní, že se dá použít k identifikaci jedince (tzv. DNA fingerprinting). Dále se sem řadí mobilní elementy a neklasifikovatelná DNA, kam patří tzv. mezerníky či spojky mezi transkripčními jednotkami. Vlastnosti této negenové DNA nejsou v současnosti přesně známy.

Dnes je repetitivní DNA již považována za nedílnou součást genetického kódu každé buňky. Proto je v poslední době důkladně prozkoumávána vědeckou obcí.

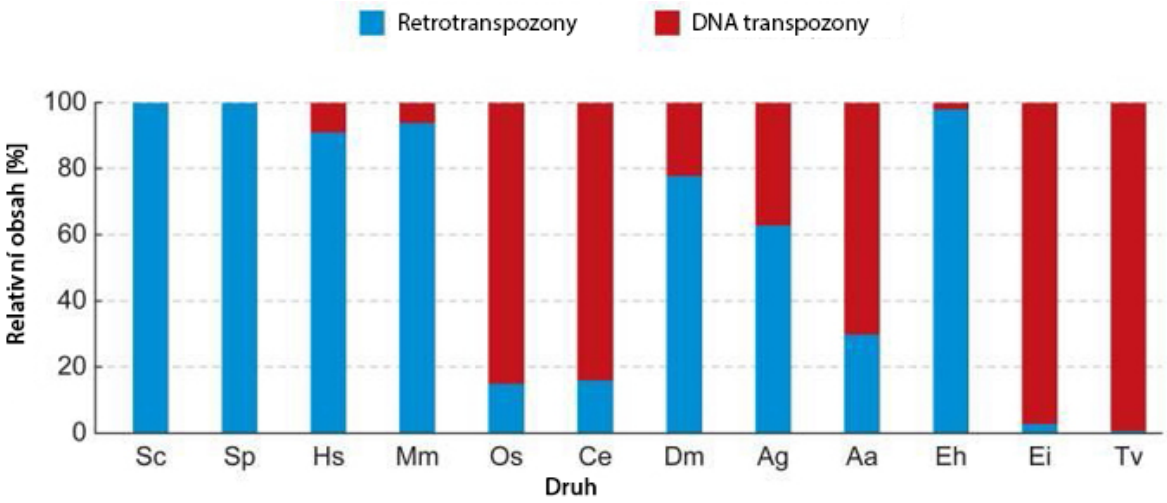
[3] Soubor všech genů jedince

Tyto opakující se sekvence mají různý charakter, délku i význam. Obecně se jedná o úseky DNA tvořené opakováním sekvenčního motivu neboli jednotkou repetice. Pokud jsou repetice uspořádány lineárně za sebou, jedná se o tzv. tandemové repetice. Druhou možností je náhodné opakování jednotek repetice v průběhu celého genomu, pak tyto repetice se nazývají rozptýlené nebo také transpozibilní elementy (TE). [22]

2.1 Transpozibilní elementy

Nejvíce se vyskytujícím druhem repetitivní DNA jsou právě rozptýlené sekvence, někdy nazývané „skákací geny“. Množství těchto opakování se odhaduje na 50 % celkového objemu savčí DNA a nachází se na mnoha různých místech v celém genomu. Individuální kopie identických, nebo velice podobných sekvencí, délky od 100 bp až do 10 kbp, se nachází v desetitisících až více než milionech opakování skrze celý genom. Tento rozptyl je dán opakovanou insercí tzv. transpozonů na stále nová místa v průběhu evoluce.

V dnešní době je známo mnoho druhů transpozibilních elementů, stejně jako mnoho kategorií do kterých se dají zařadit. Jedno z běžnějších dělení je na TE, které potřebují k přepisu reversní transkriptázu[4] a na ty, které ji nepotřebují. První zmíněný typ je znám pod názvem retrotranspozony, nebo také TE prvního řádu, zatímco druhý zmíněný typ se označuje jako DNA transpozony neboli TE druhého řádu. [9]



Obrázek 1: Poměr obsahu TE u různých eukaryotních druhů

[http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518 upraveno]

Výše uvedený obrázek popisuje relativní obsah retrotranspozony a DNA transpozonů u různých eukaryotních druhů. (Sc: *Saccharomyces cerevisiae*; Sp: *Schizosaccharomyces pombe*; Hs: *Homo sapiens*; Mm: *Mus musculus*; Os: *Oryza sativa*; Ce: *Caenorhabditis*

[4] Enzym umožňující přepis informace zpětně z RNA do DNA

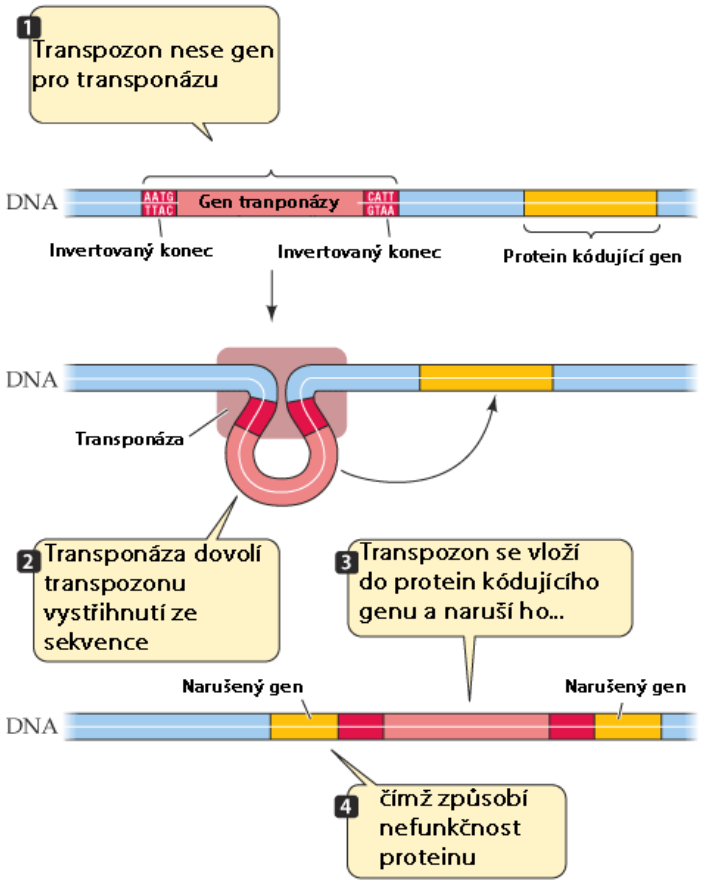
elegans; Dm: *Drosophila melanogaster*; Ag: *Anopheles gambiae*; Aa: *Aedes aegypti*; Eh: *Entamoeba histolytica*; Ei: *Entamoeba invadens*; Tv: *Trichomonas vaginalis*.) [9]

2.2 DNA transpozony

Tento druh TE je plně autonomní a kódují protein transponázu. Pomocí této bílkoviny se DNA transpozony oddělí ze svého původního místa v genu, a mohou se přesunout v podstatě na kterékoliv jiné místo v genomu. Tento pohyb po genomu se označuje jako „cut and paste“ (vyjmout a vložit). Bohužel některé transpozony nesou geny kódující enzymy inaktivující antibiotika například jako ampicilin nebo tetracyklin.

Pro TE druhého řádu je typické ohraničení krátkými repeticemi o 9 až 40 bázích, které jsou samy sobě komplementy. Například pokud se na jednom konci objeví sekvence ACGCTA, tak na druhém najdeme TGCGAT. Toto označení využívají TE k tomu, aby byly jednoduše rozpoznány transponázou.

Méně než 2 % lidského genomu je složeno z DNA transpozonů, což znamená, že významnější část genomu patří TE prvního řádu neboli retrotranspozonům. [9][19]



Obrázek 2: Schéma začlenění DNA transpozonu do genu

[http://www.hammiverse.com/lectures/19/2.html upraveno]

2.2.1 Retrotranspozony

Retrotranspozony se oproti TE prvního řádu pohybují za pomoci RNA zprostředkovatelů. Jinak řečeno nekódují enzym transponázu, ale produkují RNA transkripty, které jsou za pomoci enzymu reversní transkriptázy vloženy zpět do DNA na předem určené místo. To ovšem znamená, že při „skoku“ z jednoho místa v genomu na druhé, nedochází k přesunu originální sekvence, nýbrž k jejímu okopírování a vložení na nové místo a tím zdvojnásobení původního množství. Tento pohyb je tedy v literatuře popisován jako „copy and paste“ (kopírovat a vložit). Tento proces podléhá mnohým chybám a často vznikají porušené a neaktivní kopie retrotranspozonů postižené bodovými mutacemi, případně delecemi. Proto také obsahuje lidský genom až 45 procent takových sekvencí.

Retrotranspozony se mohou dělit na autonomní a neautonomní a to podle jejich schopnosti pohybu v genomu. Autonomní se mohou pohybovat samy, zatímco neautonomní vyžadují přítomnost jiného retrotranspozonu aby se mohly hýbat. To je dáno faktem, že tyto TE nenesou informaci pro reversní transkriptázu, která je pro transpozici potřebná. Proto si potřebují „půjčit“ enzym reversní transkriptázu od jiného elementu, aby mohly „přeskočit“. [9] [20]

LTR retrotranspozony

Na obou koncích těchto TE můžeme pozorovat dlouhé ukončovací repetice (anglicky: **L**ong **T**erminal **R**epeats, česky: dlouhé ukončující repetice). LTR retrotranspozony jsou strukturou a životním cyklem stejné jako retroviry, ale protein kódující geny jsou již neaktivní a zmutované, přesto je jejich životní cyklus velice podobný například retroviru HIV (Human Immunodeficiency Virus). V lidském genomu zabírají asi 8 %. Podobně jako DNA transpozony jsou díky své neaktivitě a neschopnosti transpozice považovány za neaktivní, tzv. fosilie.

Non-LTR retrotranspozony

Již podle názvu je jasné, že skupina non-LTR retrotranspozonů neobsahuje žádné ukončovací repetice.

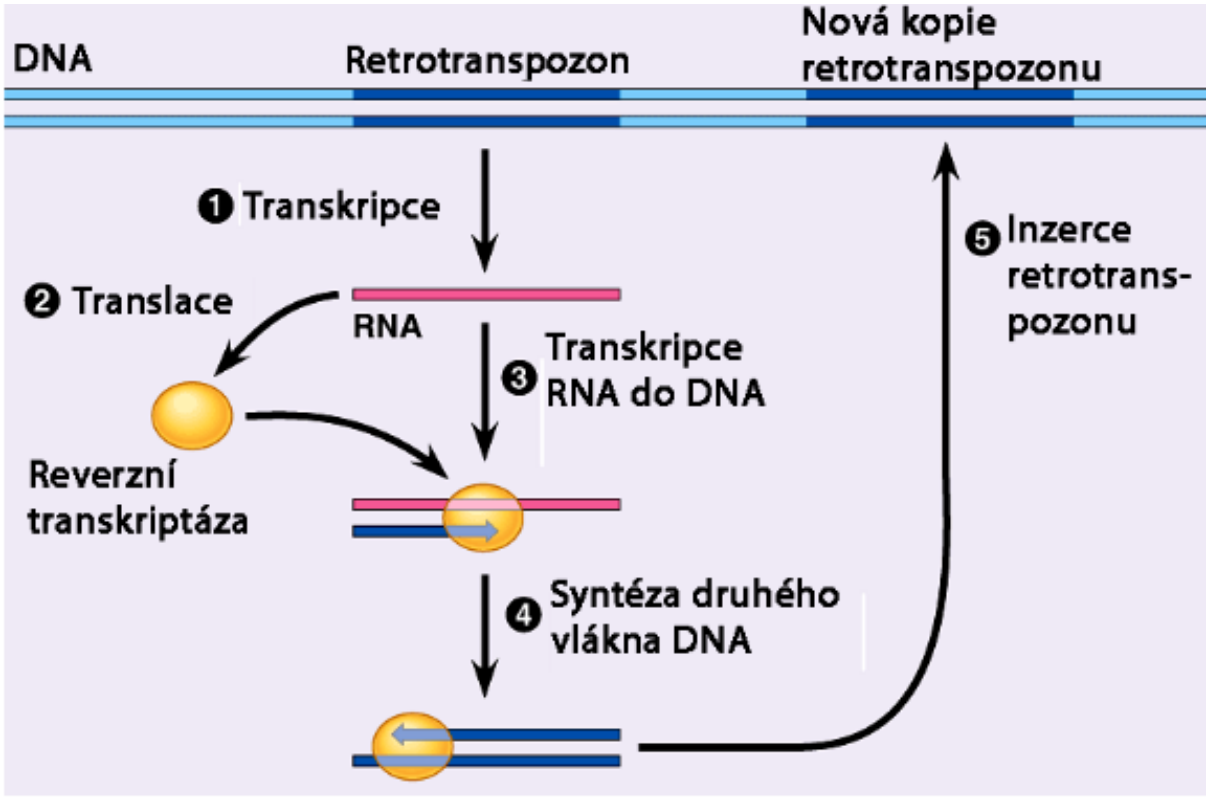
LINE (**L**ong **I**nterspersed **N**uclear **E**lements, česky: dlouhé rozptýlené jaderné elementy), jsou reversní transkriptázu kódující skupina. Savčí genom obsahuje zastaralou, již netransponující skupinu LINE2 a také mladší, stále aktivní skupinu LINE1, která se stále transponuje a její podíl narůstá (momentálně asi 21 %).

SINE (**S**hort **I**nterspersed **N**uclear **E**lements, česky: krátké rozptýlené jaderné elementy), jsou neautonomní TE, které se obecně pohybují v genomu díky reversní transkriptáze, produkovanou LINE. Nejznámější skupinou jsou Alu elementy, které se

vyskytují v CG bohatých oblastech, ovšem jejich funkce není jasná. Dále pak L1 elementy, které zabírají asi 17 % genomu.[18]

2.2.2 Vlastnosti transpozibilních elementů

Je faktem, že asi polovina lidského genomu je tvořena právě TE, s velkou mírou L1 a Alu retrotranspozonů. LINE a SINE jsou jediné stále aktivní TE v našem genomu, ostatní můžeme považovat za neaktivní fosilie. Jejich vlastnosti závisí hlavně na jejich cílovém místě, kam transponují. Pokud jsou přepsány do protein kódujícího genu, může dojít k mutaci. Tento fakt byl zjištěn, když vědci objevili vepsaný L1 do genu pro srážející faktor VII. Porucha struktury faktoru VII způsobuje nemoc zvanou hemofilie. O několik let později vědci objevili L1 vepsaný v buňkách rakoviny konečníku. Tento element nebyl objeven ve zdravých buňkách testovaného jedince. Je tedy zřejmé, že L1, transponovaný v somatických buňkách savců, mohou hrát velkou roli v rozvíjení nemocí.

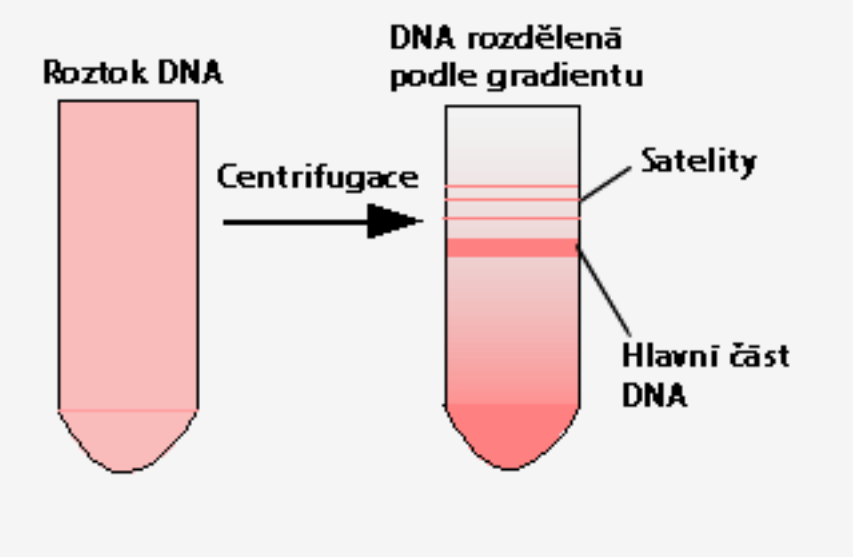


Obrázek 3: Proces transpozice retrotranspozonu v DNA [zdroj: http://www.hammiverse.com/lectures/19/images/2-8.png upraveno]

2.3 Tandemové repetice

Tandemové repetice jsou definovány jako dvě k sobě přiléhající kopie určité nukleotidové sekvence. Výskyt několika takových kopii za sebou se nazývá ATR (**A**pproximate **T**andem **R**epeat, česky: přibližná tandemová repetice). Byly objeveny při

centrifugaci DNA, díky své rozlišné vznášivé hustotě. Odtud pochází název satelitní DNA, z latinského satelles, což v češtině znamená průvodce. Tandemové repetice můžeme dělit podle počtu opakování jednotek repetice. Nejmenší jsou tzv. mikrosatelity a minisatelity, kde se nachází jednotky až desítky opakování. Větší jsou nazývány pouze satelity a dosahují délky v jednotkách Mbp. Tandemové repetice se často nachází v DNA, ale jejich vlastnosti nejsou úplně známy. Je známo, že tyto repetice mohou být původcem polymorfizmu, pozorovaného u jedinců téhož druhu. Vzhledem k tomuto zjištění jsou prozkoumávány potenciální vztahy mezi délkou tandemových repetic a genetickými poruchami. V popředí zájmu jsou například hyperkinetická porucha[5], roztroušená skleróza mozkomíšní, Alzheimerova choroba, nebo syndrom necitlivosti na androgeny.[8]



Obrázek 4: Dělení směsi DNA centrifugací (zdroj: http://cienciasdejoseleg.blogspot.cz/2014/06/repeticiones-en-satelite-del-adn.html upraveno)

2.3.1 Satelitní DNA

Satelitní DNA se skládá z nejdelších repetitivních motivů, které se délkou pohybují v rozmezí 5 – 300 bp v závislosti na druhu. Obvyklý počet opakování je 10^5 až 10^6. Vyskytují se hlavně v heterochromatinu a často tvoří centromery nebo telomery. Existuje nejméně 10 různých druhů satelitní DNA.

Lidská satelitní DNA má sklon k vysoké uspořádanosti. Satelitní α DNA, nacházející se v centromerách, je typicky 171 bp dlouhá, přítomná jako dimery (342 bp), ale může se vyskytovat až jako 16'mery (2763 bp), které slouží jako repetiční motiv. Obyčejně se neliší v délce tak zásadně jako mikrosatelity nebo minisatelity.

[5] Porucha pozornosti spojená s hyperaktivitou jedince

Lidská β satelitní DNA je přítomna jako 30 – 60 tisíc kopií 68 bp dlouhého monomeru (2 040 000 – 4 080 000 bp) na metacentrickém chromozomu 9 a na akrocentrických chromozomech 13, 14 15 21 a 22.

Jelikož genom obecně podléhá mutacím, pak můžeme i v satelitní DNA nalézat nepravidelnosti v opakováních, jako bodové mutace, inzerce a podobně. Kromě primárního opakovaného řetězce můžeme nacházet i sekundární motivy, které vznikají právě kvůli mutacím. Tyto mutace jsou dále kopírovány, až samy tvoří nějaký vzor.

Tabulka 1: Příklad satelitní DNA a jejího množství u *Drosophila virilus* [5]

Druh satelitu	**Primární sekvence**	**Celkový počet kopií**	**Obsah v genomu [%]**
I	ACAAACT	$1{,}1*10^7$	25
II	ATAAACT	$3{,}6*10^6$	8
III	ACAAATT	$1{,}6*10^6$	8
		Celkem v genomu	41

2.3.2 Minisatelity

Tyto tandemové repetice mají délku repetiční jednotky v rozmezí od 15 do 400 bp, s mediánovou hodnotou okolo 20. Běžně se motivy opakují 20 – 50x a tvoří tandemy dlouhé 1000 až 5000 bp, což je oproti satelitní DNA pouhý zlomek. Nejčastěji se vyskytují v telomerech a v subtelomerických oblastech chromozomů. Díky vysoké variabilitě, tzv. VNTR (**V**ariable **N**umber **T**andem **R**epeats, česky: tandemové repetice s proměnným počtem opakování) se stávají vhodným nástrojem pro srovnávání DNA vzorků, tzv. DNA fingerprinting. Bohužel kvůli délce VNTR není možná jednoduchá amplifikace[6] pomocí PCR a proto nejsou při srovnávání vzorků DNA příliš oblíbené.[1][20]

Poslední studie vedou k teorii, že VNTR podle svého počtu mohou ovlivňovat expresi genů, například produkci inzulinu. Různé typy diabetu jsou přirovnávány k různým počtům minisatelitní DNA před genem kódujícím produkci hormonu inzulin. Minisatelity mají v genomu pravděpodobně vždy jednu ze tří funkcí. Zaprvé jsou některé VNTR tvoří části ORF (**O**pen **R**eading **F**rame, česky: otevřený čtecí rámec), což je úsek DNA, vymezený iniciačním a terminačním kodonem kódující souvislý a dostatečně dlouhý polypeptidový řetězec. Proto se může přítomnost mikrosatelitu projevit jako polymorfismus mezi jedinci. Zadruhé je velice pravděpodobné, že některé VNTR se váží na proteiny, což má mnohé

[6] Zmnožení

následky. Například minisatelity na 5' konci genu se účastní regulace transkripce. Zatření mohou VNTR tvořit křehké, nestálé části chromozomů. [24]

2.3.3 **Mikrosatelity**

Jsou označovány jako SSTR (**S**imple **S**equence **T**andem **R**epeats, česky: tandemové repetice jednoduchých sekvencí). Repetiční jednotky jsou obecně dinukleotidy, trinukleotidy, tetranukleotidy, nebo pentanukleotidy. Setkáváme se tedy maximálně se sekvencemi o 500 bp. U člověka jsou některé genetické poruchy způsobeny expanzí trinukleotidů do kódujících oblastí. Mikrosatelity jsou vysoce polymorfní a proto se podobně jako minisatelity dají využít jako genetické markery. Díky jejich délce a jednoduchému rozpoznání pomocí PCR probíhá jejich amplifikace jednodušeji než u minisatelitů. Proto mohou být SSTR využívány k testování paternity, tvoření genetických map nebo mohou být užitečné ve forenzních vědách.[11]

Jak je zmíněno výše SSTR mohou někdy expandovat do kódujících částí DNA. Toto rozšíření je zodpovědné za více než dvacet závažných neuromuskulárních a neurodegenerativních onemocnění. Porozumění patogenním mechanismům TNR (**T**ri **N**ucleotide **R**epeats, česky: trinukleotidové repetice) v poslední době vysoce pokročilo, přesto mnoho aspektů jejich mutačních mechanismů zůstává nevysvětlených.

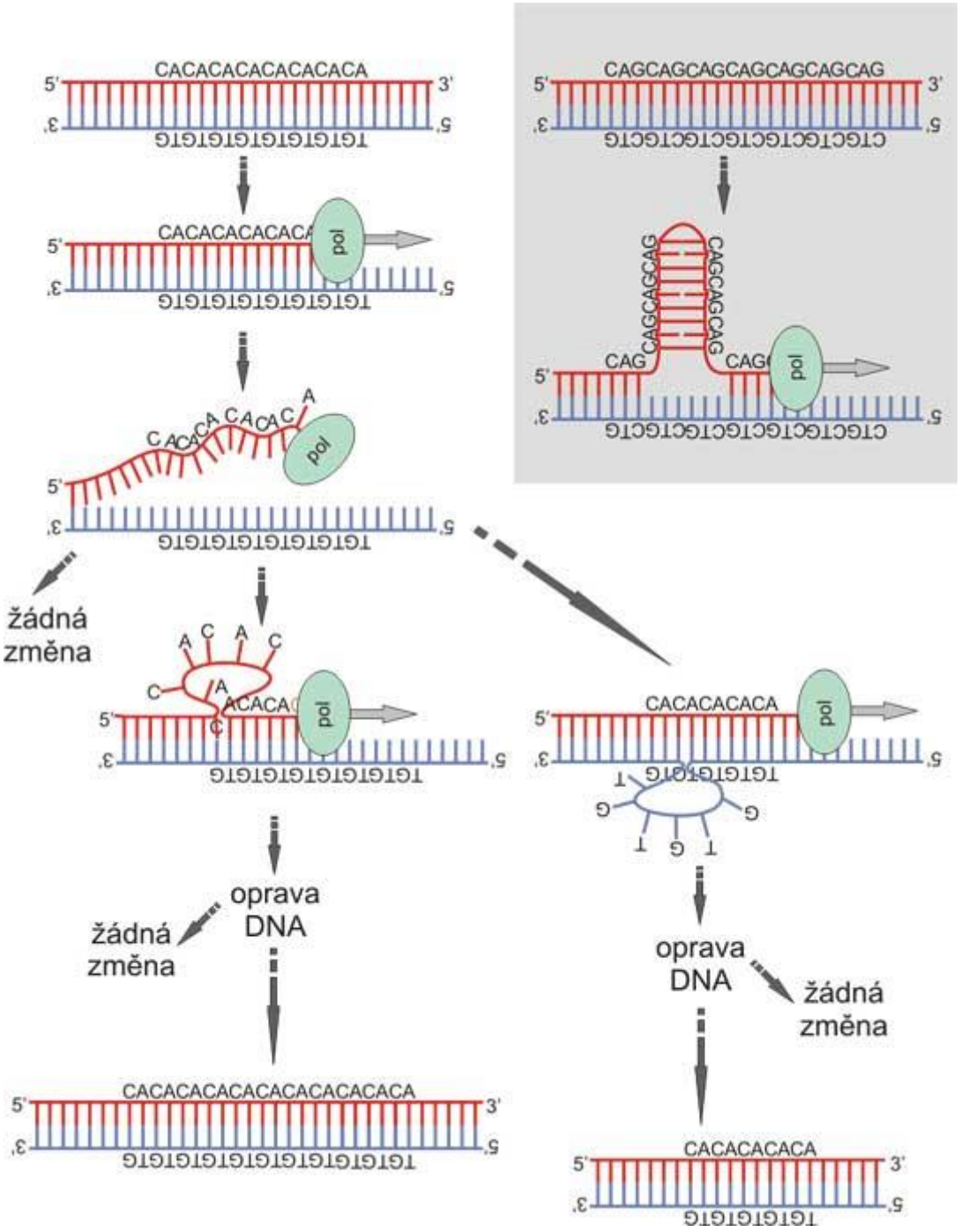
Savci si vytvořili systém, který zabraňuje rychlým změnám v DNA, které by mohly být kritické pro celý druh. Pokud tyto repetice překročí svou délkou určitý práh, pak restrikční mechanismy nefungují a dochází nejčastěji k přenosu TNR z rodiče na dítě a také k jejich expanzi při vývoji plodu. Změny v délce TNR mohou být velice podstatné. V kódujících sekvencích mohou být nestabilní již od cca 29 – 35 bp. Expanze TNR v kódujících oblastech je menší než 10 repetic za generaci. Při přenosu z rodiče na dítě roste TNR mimo kódující oblasti o 100 – 10 000 repetic za generaci. Jak pro kódující tak pro nekódující oblasti TNR rostou za hranici určitého prahu a zároveň se zvyšuje pravděpodobnost další nestabilní mutace. Pokud dojde k projevu nějaké poruchy, pak se postupně s generacemi stává více závažnou a začíná se projevovat v nižším věku. Tento fenomén nazýváme anticipace. Důvod proč některé repetice expandují rychleji a některé pomaleji není zatím přesně znám.[12]

2.3.4 **Způsob prodlužování/zkracování tandemových repetic**

Původně bylo prodlužování nukleotidů limitováno pouze na trinukleotidy, nyní je již jasné, že dochází k prodlužování úseků tetranukelotidů (CCTG), pentanukleotidů (AATCT) a dokonce dodekanukleotidů (C_4GC_4CGC), které mohou způsobovat různá onemocnění.[13]

Obecně nejvíce přijímaným mechanismem expanze in vivo je tzv. klouzání vlákna. Vlivem teplotního kolísání může dojít k disociaci 3‘-konce vlákna, který se při reasociaci

naváže na jiné místo repetice a tím pádem vznikne tzv. klička. Pokud tato klička vznikne v polymerázovém řetězci, dojde k prodloužení repetice, pokud vznikne na templátovém řetězci, tak dojde ke zkrácení repetice. [21][23]



Obrázek 5: Klouzání sekvence (zdroj: http://biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm)

Jiným mechanismem, který přispívá k prodlužování a zkracování většinou delších repetic je nerovnoměrný crossing-over. Při správném crossing-overu dochází nejdříve ke splynutí dvou homologních chromozomů, dále se rozpojí DNA dvoušroubovice a následuje vzájemná výměna jejich částí (dochází k překřížení), tím vzniká chromozom, jehož část je

totožná s chromozomem původní mateřské sady, zatímco druhá část je stejná, jako byla u otce. Při nerovnoměrném crossing-overu dochází mezi nehomologními chromozomy ke špatnému napojení a tím pádem podobně jako u klouzání vlákna k duplikaci, nebo deleci části sekvence DNA.[21]

2.3.5 Vliv repetic na morfologii

Jak již bylo zmíněno, mezi jedinci téhož druhu jsou menší, či větší rozlišnosti i přes to, že jejich genetická výbava je téměř totožná. Zde je nejpodstatnější slovo „téměř“, protože vliv nekódujících oblastí na expresi genu je zásadní a právě v těchto oblastech najdeme rozlišností nejvíce.[15]

V roce 2004 ve své studii, se vědci John Fondon a Harold Garner rozhodli prozkoumat populace, které vykazují zřetelné morfologické odlišnosti mezi jejich jedinci. Poté srovnali genomy těchto druhů a pokusili se najít korelaci mezi vnitrodruhovými rozdíly a počtem tandemových repetic. Tato dvojice vědců vybrala jako reprezentativní populaci psa domácího jednoduše proto, že chovatelé jsou známí svou vybíravostí a posedlostí čistokrevného mazlíčka, na němž jsou vzácné právě odlišnosti od ostatních psů stejného druhu.[1][15]

Většina genů se zdá tandemovými repeticemi nějak neovlivněná, ovšem pět genů, které vědci zkoumali, jsou opakovanými sekvencemi vysoce ovlivněny. Jde o geny: *Six-3* (Δ54 bp), *Hox-a7* (Δ33 bp), *Runx-2* (ins45 bp), *Hox-d8* (Δ30 bp) a *Alx-4* (Δ51 bp). Přestože jsou tyto alely velice netypické, tak jsou vhodnou příležitostí pro zkoumání viditelných efektů na určitých plemenech psů.

Například, pokud se alela $Alx\text{-}4^{\Delta 51}$ objeví u homozygotního jedince Pyrenejského horského psa, projeví se u něj bilaterální polydaktylie prvního prstu.

Naopak *Runx-2* je za normálních okolností hlavní gen regulující diferenciaci osteoblastů, což jsou buňky zodpovědné za formování kostí. Lidský *Runx-2* obsahuje repetitivní sekvenci kódující 23 glutaminů a za nimi 17 alaninů a při mutaci dochází k cleidocraniální dysplasii, což je syndrom charakterizovaný různými craniofaciálními[7] a jinými skeletálními malformacemi. *Runx-2* u psů obsahuje 18 až 20 glutaminů následovaných 12 až 17 alaniny a mezi různými plemeny je vysoce variabilní. Pearsonova korelační analýza odhalila jasnou spojitost mezi celkovou délkou alely, dorsoventrálním sklonem nosu a délkou střední části obličeje. Porovnávané hodnoty byly změřené z 3D vypočítaných modelů lebek dvaceti morfologicky odlišných čistokrevných psů a sedmi míšenců.[1][14][15]

[7] Oblast lebky a obličeje

Významné změny, které se objevily u domestikovaných psích druhů za posledních 150 let, díky vybíravosti chovatelů a postupné změně standardů na různá plemena, ukazují potenciál savčího genomu rychle odpovídat na silnou selekci.

Přes nedostatek genetické odlišnosti, která je způsobena strukturou psí populace a historií, jsou tato plemena schopná kontinuálně vytvářet nové morfologické rysy ve velice rychlém a stabilním tempu. [1][15]

Nemoci způsobené expanzí trinukleotidů

TRED (**T**riplet **R**epeat **E**xpansion **D**iseases, česky: Nemoci způsobené expanzí trinukleotidů), jsou charakterizovány amplifkací trinukleotidů ve specifických genech. Nejčastěji expandují CAG – GTC, CGG – GCC a GAA – TTC trinukleotidové páry.

První z prokázaných TRED je **Syndrom fragilního X** (také známý jako FRAXA, nebo FMR1), při kterém dochází ke zmnožení repetic CGG tripletu v oblasti kódující promotor genu FMR1 (**F**ragile X **M**ental **R**retardation). K rozvinutí onemocnění dochází při expanzi premutační alely nad 200 bp skoro výhradně při předání od matky, jejíž premutační alela je již delší než 90 bp. Hlavním projevem FRAXA je mentální retardace a typické protažení obličeje a velké ušní boltce.[14]

Podobně může docházet k jiným typům syndromu fragilního chromozomu X a to v oblasti kódující promotor genu FMR2 (vzniká tzv. FRAXE), nebo v oblasti kódující gen FAM11A (vzniká tzv. FRAXF).[20]

Bulbospinální svalová atrofie je vzácné degenerativní pomalu postupující onemocnění postihující převážně spinální a bulbární periferní motoneuron. Genetickým podkladem je expanze CAG tripletů na prvním exonu androgenového receptorového genu X chromosomu. Jedná se o onemocnění s gonozomálně recesivní dědičnosti, které se projevuje u mužů. V rozvinutém stádiu jsou typické svalové atrofie, slabost a fascikulace[8], zvláště v bulbární oblasti (jazyk a mimické svaly) a na končetinách. Dalšími příznaky je například tremor rukou, svalové bolesti a gynekomastie. Toto onemocnění se naplno projevuje až ve čtvrté až páté dekádě života, některé příznaky (gynekomastie, chronická únava) se projevují již od první až druhé dekády.[18]

Jedno z nejznámějších onemocnění způsobených expanzí trinukleotidů je **Huntingtonova choroba.** Jedná se autosomálně dědičné neurodegenerativní onemocnění mozku postihující jedince obojího pohlaví. Obvyklým klinickým obrazem jsou mimovolní pohyby, abnormální způsob chůze a porucha řeči. Dále se u nemocných projevuje úbytek rozumových schopností, poruchy nálady a chování (tzv. presenilní demence). Jedná se o poměrně vzácné onemocnění projevující se nejčastěji mezi 20 – 45 rokem věku.

[8] Samovolné záškuby svalových vláken

V molekulární oblasti se jedná o expanzi CAG od pozice 18. aminokyseliny v oblasti kódující gen huntingtin. Normálně fungující huntingtin má v řadě 6-34 CAG tripletů a se zvyšujícím se počtem přichází dřívější projevy nemoci.[12][14][22]

Další onemocnění způsobené jsou spinocerebrální ataxie, dentatorubral-pallidolusian atrofie (DRPLA), choroba Machado-Josepha nebo myotonická dystrofie.[14]

3. Numericky reprezentovaná DNA

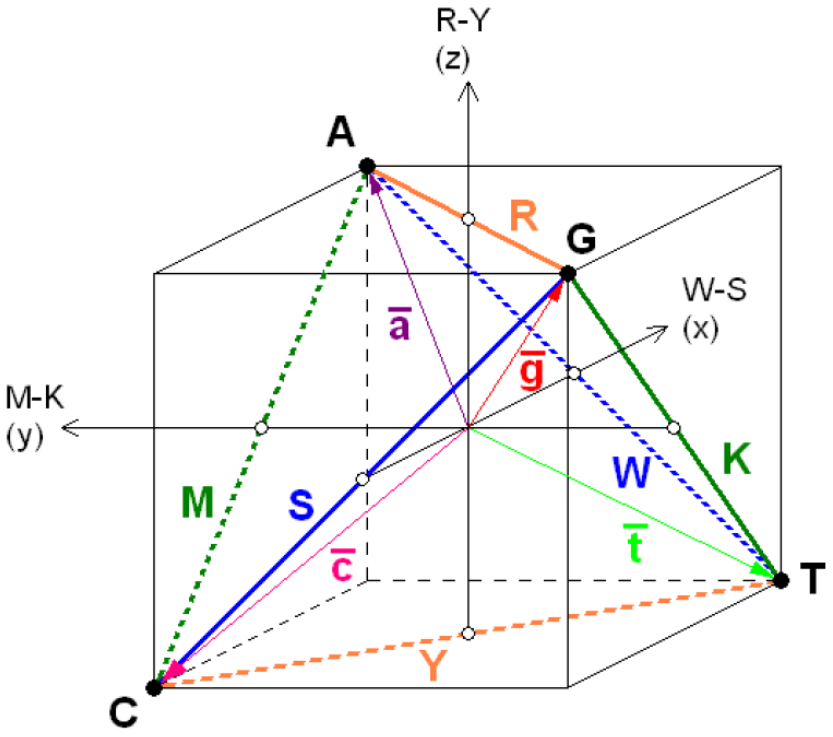
Numerická reprezentace DNA je klíčová pro použití širokého spektra analýz, zahrnujících zpracování signálu nejrůznějšími systémy (např. neuronové sítě). Díky určitým typům numerických reprezentací lze vyjádřit nejen, která báze se na jakém místě genomu nachází, ale vyjádřit částečně její biologické, chemické, nebo fyzikální vlastnosti. Samotný výběr reprezentace je klíčový pro výpovědní hodnotu náhradního číselného řetězce.[1][8]

Reprezentace čtyřstěnem (tetrahedron)

Nukleotidové báze můžeme dělit podle tří hlavních kritérií:

- Molekulární struktura – A a G jsou purinové báze (R), C a T jsou báze pyrimidinové (Y),
- Síla vazby – G a C jsou k sobě vázány třemi vodíkovými můstky, tedy silnou vazbou (S), A a T jsou vázány pouze dvěma vodíkovými můstky a tedy slabou vazbou (W),
- Přítomnost radikálu – A a C se řadí do amino (NH_3) skupiny (M), G a T obsahují zbytek keto (C = O) skupiny (K).

Tyto vlastnosti lze vyjádřit jako tetrahedron, jehož hrany reprezentují jednotlivé vlastnosti. Každá z os v trojrozměrném kartézském souřadném systému určuje jednu skupinu vlastností jako je síla vazby na ose x, přítomnost radikálu na ose y a molekulární strukturu na ose z.[2][25]



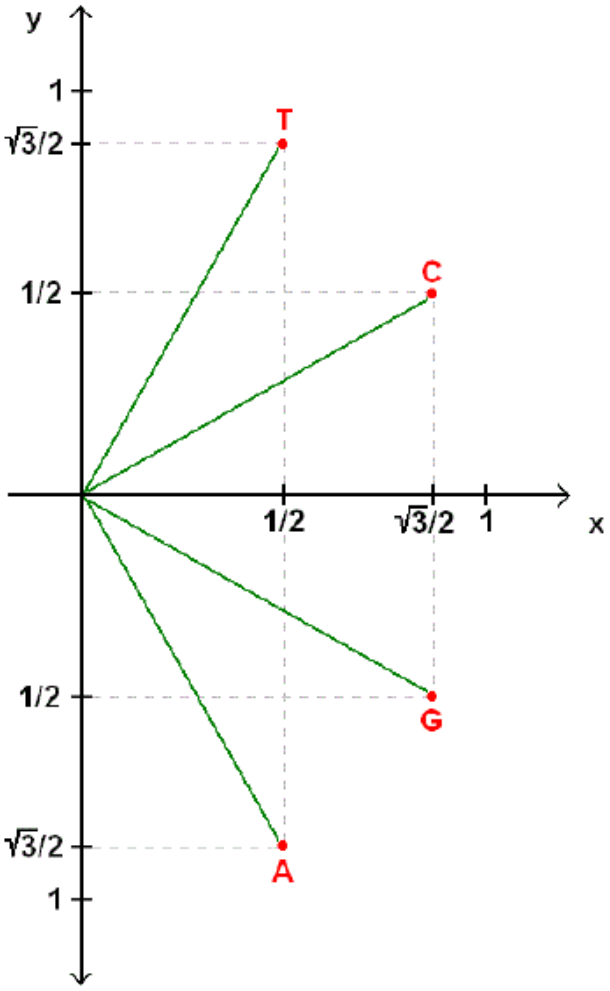
Obrázek 6: Nukleotidový čtyřstěn v pomocné krychli[25]

1D reprezentace reálnými čísly

Pravděpodobně nejjednodušší metoda numerické reprezentace DNA, je 1D reprezentace reálnými čísly, kdy je každému nukleotidu přiřazeno jedno číslo. Nevýhodou je ztráta informace o biochemických vlastnostech nukleotidů. Přidává jinou informaci. Pokud přiřadíme každé bázi jedno reálné číslo, např.: A = 1, C = 2, G = 3 a T = 4, pak můžeme říci, že A < C < G < T. Tato metoda je pro mapování genomu nevhodná, ovšem pro některé jednoduché operace a předzpracování je dostačující.[1][2]

2D reprezentace

Metoda 2D reprezentace zobrazuje reprezentovanou DNA v kartézském souřadnicovém systému. Čtyřem bázím A, C, G, T jsou přiřazeny čtyři směry, (- *x*), (+ *x*), (- *y*), a (+ *y*). Tato reprezentace je ovšem dosti nevyhovující kvůli možnosti zacyklení, křížení se nebo stejné reprezentace rozdílných sekvencí. Ke zdokonalení se používá reprezentace pouze v 1. a 4. kvadrantu kartézského souřadnicového systému, díky čemuž nemůže již dojít k zacyklení překrytí nebo jiné ztrátě informace.[1][2]



Obrázek 7: 2D reprezentace reálnými čísly v 1. a 4. kvadrantu [25]

4D binární reprezentace

Tato metoda vytváří čtyři indikační vektory $u_A(n)$, $u_T(n)$, $u_C(n)$ a $u_G(n)$, které vyjadřují nepřítomnost nebo přítomnost daného nukleotidu na pozici n v sekvenci. Je vhodná pro využití při zpracování reprezentované DNA Fourierovou transformací, a proto je velice často využívaná. Příkladem je sekvence AAGCGTCA, jejíž indikační vektory jsou u_A = 11000001, u_T = 00000100, u_C = 00010010 a u_T = 00000100.[2][25]

3D numerická reprezentace redukcí 4D binární reprezentace

Další možností reprezentace DNA dat je redukce 4D reprezentace do RGB spektra, čehož se využívá například při spektrální analýze.[25]

Každé bázi přiřazujeme bez jakékoliv ztráty informace vektor směřující ze středu jednoho ze čtyř vrcholů pravidelného čtyřstěnu takto:

$$A = (a_R, a_G, a_B) = (0,0,1)$$
$$C = (c_R, c_G, c_B) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$$
$$G = (g_R, g_G, g_B) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$$
$$T = (t_R, t_G, t_B) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right) \quad (1)$$

Tyto vektory jsou poté převedeny pouze do tří numerických sekvencí x_R, x_G, x_B, které odpovídají lineárnímu zápisu:[2][25]

$$x_R(n) = \frac{\sqrt{2}}{3}[2u_T(n) - u_C(n) - u_G(n)]$$
$$x_G(n) = \frac{\sqrt{6}}{3}[u_C(n) - u_G(n)]$$
$$x_B(n) = \frac{1}{3}[3u_A(n) - u_T(n) - u_C(n) - u_G(n)] \quad (2)$$

Existuje pochopitelně mnoho jiných reprezentací pro zjednodušení mapování genomu a vyhledávání tandemových repetic. Metody jsou stále pokročilejší a složitější a využívá se například reprezentace celých kodonů, což posouvá reprezentaci na jinou úroveň.[1][2][25]

4. Diskrétní Fourierova transformace (DFT)

V úsecích signálu, kde dochází k tandemovým repeticím, očekáváme pravidelné opakování určitého repetičního motivu. V tom případě tedy dochází i k pravidelnému opakování jednotlivých nukleotidů. Této periodicity je možno využít při číslicovém zpracování DNA sekvencí. Jako ideální se jeví zpracování pomocí tzv. Fourierovy transformace. Ta je ovšem definována pro spojité signály, což zkoumané nukleotidové sekvence nejsou. Proto je nutno využít jiného vyjádření pro získání spektra, a to diskrétní Fourierovy transformace

Diskrétní periodické signály lze popsat diskrétní Fourierovou řadou. Při popisu aperiodických signálů lze využít Fourierovu transformaci diskrétního signálu (DTFT). Výstupem DTFT je spojité spektrum, které je nevhodné pro počítačové zpracování, proto zavedeme pojem diskrétní Fourierova transformace. Tato matematická operace vzorkům časového průběhu, v našem případě jednotlivým nukleotidům v sekvenci, přiřadí posloupnost konečné délky (čárové frekvenční spektrum). Diskrétní Fourierova transformace je dána vztahem:

$$F(m) = \sum_{k=0}^{N-1} f(k) e^{-\frac{jm2\pi k}{N}} = \sum_{k=0}^{N-1} f(k) W^{km} \qquad (4)$$

, kde koeficient $m = 0, 1 \ldots N-1$ a určuje řád harmonické složky

$k = 0,1 \ldots N-1$ a určuje pořadí odebraného vzorku v časové oblasti

N značí počet odebraných vzorků.

Chceme-li výstup transformace interpretovat ve významu blízkém koncepci spektra spojitého signálu, tedy jako funkci frekvence, je vhodnější psát definiční vztah takto:

$$F(k\Omega) = \sum_{k=0}^{N-1} f(nT) e^{-jk\Omega nt} \qquad (5)$$

Protože je zde názorně patrno, že koeficienty diskrétního spektra přísluší jistým frekvencím $k\Omega$, zatímco vzorky signálu přísluší jistým časovým okamžikům nT.[7]

Výstupem DFT je frekvenční spektrum signálu, které ukazuje sílu signálu na jednotlivých frekvencích. Při provedení DFT na signál o délce např. 10kbp, kde je očekávaná repetice trinukleotidu o délce 60bp bude zvýšení amplitudy v tomto úseku naprosto nepatrné. Proto je tedy vhodné použít tzv. krátkodobou Fourierovu transformaci (STFT), což je operace totožná s DFT, ale provedená na vstupní sekvenci postupně, tzv. klouzavým oknem.

5. Metody vyhledávání tandemových repetic

Objev tandemových repetic, jejich biologický potenciál, jejich vliv na vznik a vývoj řady závažných onemocnění vyžaduje vyvinout dostatečně vhodný a sensitivní algoritmus, který bude schopný tyto repetice rychle a včas odhalit. Až poté může následovat zpracování a vyhodnocení výsledků, takže bez kvalitního algoritmu nemohou vznikat kvalitní závěry.[2][8]

Samozřejmě již existuje velké množství algoritmů, které tandemové repetice vyhledávají, ovšem často se u nich vyskytují jistá omezení. Navíc v celém genomu stále probíhají mutace, což v podstatě zaručuje, že sekvence, která je vyhledávána, bude částečně zmutovaná (nejčastěji bodové mutace). To je další fakt, který nepřispívá k jednoduchosti řešení vyhledávání tandemových repetic.[5][8]

Jednoduše řečeno, vyhledávač repetic je program skládající se ze tří hlavních částí: detekční část, filtrační část a výstup. Samotným „srdcem“ vyhledávače je první část: detekční jednotka. Ta udává, jak dlouho bude výpočet trvat, jaké na něj jsou kladeny nároky a celkovou vhodnost použití. Algoritmus založený na určitých kritériích (heuristický, statistický, nebo počítající pomocí skórovacích matic) detekuje vzory (motivy repetice, nebo repetice celé), tak jak je při vstupu specifikuje uživatel. Výstup tohoto procesu, sekvence, která je možná repetitivní musí dále podstoupit filtraci, aby byly odstraněny různé nadbytečnosti. V tomto procesu se přístupné vyhledávače nejvíce liší.[2][5]

Z pohledu uživatele jsou ještě před vyhledáváním důležité dvě otázky: zaprvé jestli hledání bude specifické, nebo nespecifické (to záleží na povaze hledané repetice), zadruhé jaké typy repetic budou vyhledávány (shodné, neshodné, přerušované, komplexní).[1][5][8]

Tabulka 2: Typy repetic u mikrosatelitů

Definice repetice	**Vlastnosti**	**Příklad**
Shodné	100% shodné kopie	$(A)_n$, $(ACT)_n$
Neshodné	Dochází k substitucím	$(AC)_nAT(AC)_m$
Přerušované	Dochází k substitucím, inzercím, delecím (přerušení)	$(AT)_nCGAG(AT)_m$
Komplexní	Objevuje se více motivů, periody opakování, substituce	$(ACG)_nT(TC)_m$, ATcgc | ATgcc | ATccc | ATcgg…

Nejjednodušší metodou vyhledávání tandemových repetic se může zdát přímé vyhledávání v sekvenci. Metody založené na vyhledávání z řetězce znaků jsou nejčastěji založeny na lokálním zarovnávání nějakých dříve vytipovaných sekvencí. Ty jsou vytipovány podle nějakého modelu, nejčastěji pravděpodobnostního. Pro využití matematických znalostí je spíše využíváno numericky reprezentované DNA.[1]

Na internetu je volně přístupné množství vyhledávačů tandemových repetic nebo obecněji programů na mapování nukleotidových sekvencí. Některé se zakládají na jednoduchém zarovnání sekvencí a nejsou příliš sofistikované a „odolné“ vůči bodovým mutacím v genomu, které ztěžují vyhledávání. Jiné jsou oproti tomu založené na velice sofistikovaných metodách vyhledávání, buď ze sekvence znaků, nebo jinak reprezentované DNA. Některé pracují přímo se sekvencí, jiné pracují s pravděpodobnostními modely výskytu různých repetic.

5.1 Tandem Repeats Finder (TRF)

Tandem Repats Finder je založený na pravděpodobnostním modelu tandemových repetic. Modeluje zarovnání a hodnocení dvou tandemových kopií vzoru délky *n* pomocí sekvence
n-nezávislých Bernoulliho pokusů (hodů mincí – „hlava nebo orel“). Pravděpodobnost úspěchu, p_m vyjadřuje průměrnou shodu mezi jednotlivými kopiemi. Každá *„hlava“* v Bernoulliho sekvenci značí shodu mezi nukleotidy, zatímco každý „orel“ značí neshodu, nebo mutaci. Pravděpodobnost p_i ukazuje střední hodnotu insercí a delecí mezi kopiemi.

Program se skládá ze dvou hlavních komponent, detekční a analyzační. Detekční komponenta se zakládá na pravděpodobnosti, že dvě přilehlé kopie budou obsahovat alespoň nějaké shodné znaky na odpovídajících místech. Tahle část se zakládá na pravděpodobnostech p_m *a* p_i.

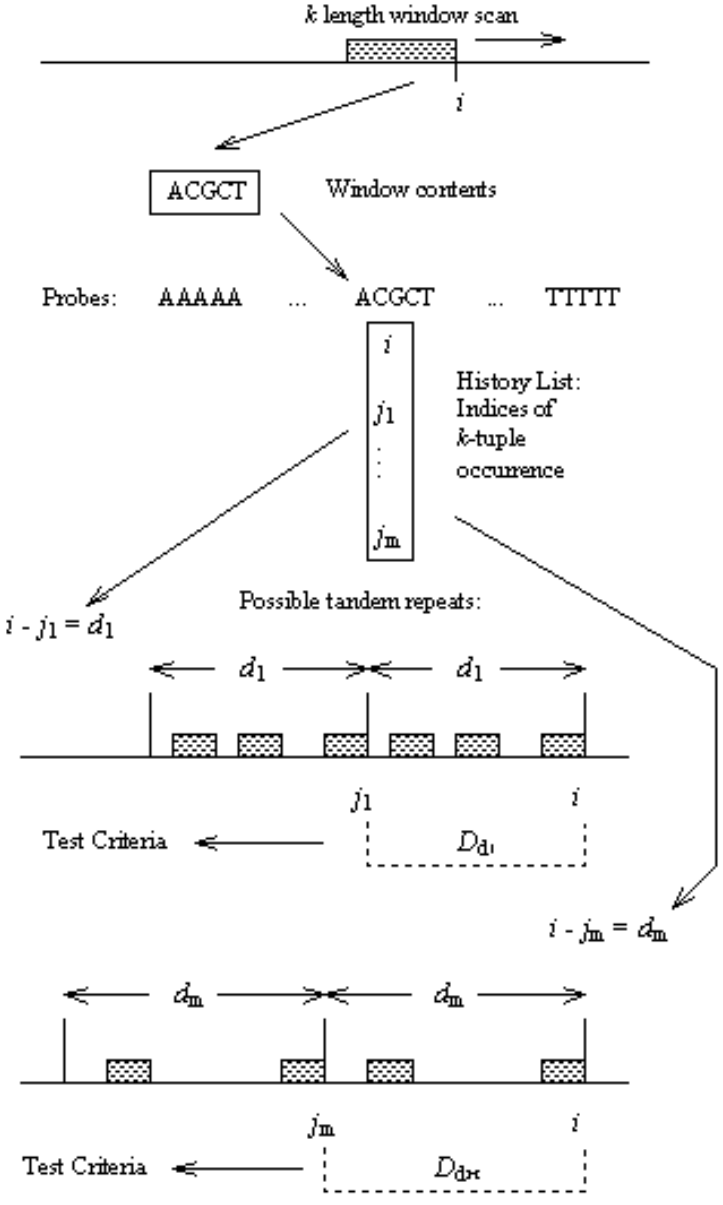
Algoritmus vyhledává shodné nukleotidy oddělené běžnou vzdáleností *d*, která není předem určena. Pro nejlepší výkon je vybrána sekvence, ve které se nachází *k* opakování (tzv. *k-tuple*, česky: k-tice). K-tice je okno o *k* za sebou jdoucích znaků z nukleotidové sekvence. Právě kvůli vybrané k-tici není tedy možné odhalit všechny shodné nukleotidy sekvence, ale pouze *k*.

Základní princip detekce je zobrazen na obrázku (Obrázek 8). Pokud *S* je nukleotidová sekvence, pak je vybráno malé celé číslo ($k = 5$ např.) a vytvoří se banka všech možných řetězců délky *k* (v genetickém kódu je 4^k možností sekvencí), které se nazývají vyhledávače. Projížděním okna po sekvenci je detekován vyhledávač na každé pozici *i* v sekvenci *S*. Pro každý jednotlivý vyhledávač *p* vzatý z banky je vytvořena historie H_p pozic, ve kterých se *p* nachází.

Vždy když je pozice i přidána do H_p, jsou v H_p vyhledávány všechny předchozí výskyty p. Pokud se jedna ze sekvencí p dříve vyskytla na místě j, pak vzdálenost $d = i - j$ je možná velikost vzoru repetice. Seznam distancí D_d obsahuje pozice i a zároveň jejich celkovou sumu.

Statistické vyhodnocení je založeno na Bernoulliho sekvencích, korespondujících se shodami detekovanými a uschovanými v seznamu distancí. Záleží na délce repetice, pravděpodobnosti shody p_m, pravděpodobnosti neshody p_i a k. Poté musí sekvence obstát v několika dalších testech kritérií, než je odeslána do analyzační komponenty.

Analyzační komponenta označí za tandemovou repetici takovou sekvenci, která je vytvořena z $j+1 \ldots i$ pozic nukleotidové sekvence a zarovnána ve svém okolí více než dvakrát za sebou.[2]



Obrázek 8: Schéma detekce tandemových repetic programem TRF[2]

5.2 IMEx: Imperfect Microsatellite Extractor

IMEx využívá k detekci tandemových repetic přístup s klouzavým oknem. Jedná se o algoritmus složený ze dvou oddělených částí, a to:

- Identifikace určité oblasti, kde se nachází perfektní tandemové repetice,
- Prodlužování těchto úseků na obě dvě strany, dokud jsou všechna kritéria pro pokračování tandemové repetice splněna.

Těmito kritérii je zaprvé určité číslo *k*, které určuje počet nedokonalostí v repetičním motivu a které nesmí přesáhnout určitý práh. Zadruhé se jedná o procentuální vyjádření celkových neshod, podle vzorce

$$procntuální\ odlišnost = \frac{počet\ mutací\ v\ pozorované\ sekvenci}{celkový\ počet\ bazí\ v\ dokonalé\ sekvenci} \qquad (3)$$

Hodnota *k* může být uživatelem nastavena na jakékoliv celé číslo, mezi 0 a *m*, kde *m* vyjadřuje délku repetičního motivu.

Blokové schéma (**Chyba! Nenalezen zdroj odkazů.**) ukazuje, jak program funguje. IMEx postupně projíždí řetězec znaků a vyhledává centra, kde se nachází ideální tandemové repetice. Začíná s nejdelší možnou délkou motivu, v tomto programu jsou to hexanukleotidy a pokračuje až k mononukleotidovým sekvencím. Pokud nenajde na žádné pozici *i* v sekvenci *S* repetice hexanukleotidů, vrátí se a hledá pentanukleotidy, jinak řečeno $m = m - 1$. Redundantní repetice automaticky vyřazuje. Tedy například sekvenci (ACCCGTACCCGT) označí jako repetici $(ACCCGT)_2$ a vnitřní, redundantní repetice C již ignoruje. Zároveň také dochází k uchování dat o bodových mutacích a dochází ke vzájemnému zarovnání vyznačené tandemové repetice a jejího ideálního protějšku. V poslední řadě dochází k uchování detailů jako repetiční jednotka, počet iterací, délka repetice, procenta nedokonalosti, skladba nukleotidů a zařazení do kódující nebo nekódující oblasti.[15]

5.3 Phobos

Phobos je program vhodný pro vyhledávání tandemových repetic v DNA sekvencích jakékoliv velikosti, včetně celých genomů. Dokáže detekovat jak přesné repetice, tak nedokonalé. Umí detekovat repetice o velikosti repetičního motivu od 1 bp až do několika tisíc párů bází. Měl by si umět poradit i s bodovými mutacemi jako indely a bodovými mutacemi. Dále je také schopný detekovat překrývající se repetice a sub-satelity (vnořené repetice).

Na každé pozici v sekvenci začne Phobos hledat vhodné místo pro výskyt tandemové repetice porovnáváním okolí. Díky tomu nemusí být přítomna žádná knihovna možných

repetic. Když je místo vybráno jako vhodné, pak se repetice začne rozšiřovat do obou směrů, a to tak daleko jak je to jen možné. Phobos také stále prohledává již nalezené repetice pro jinou délku repetičního motivu, čímž umožňuje vyhledat již zmíněné sub-satelity. Například při nalezení sekvence $(ATATAG)_{10}$ ji nejprve označí jako nedokonalou repetici motivu (AT), ovšem při dalším postupu ji označí také jako dokonalou repetici hexanukleotidu.

Pro najití vhodného motivu repetice dochází k zarovnávání krátkých úseků mezi sebou a hledání optimálního skóre. Každá shoda v určitém zarovnání má hodnotu 1, neshoda dostane hodnotu přidělenou uživatelem. Tyto hodnoty se ukládají a hledá se nevyšší pro začátek rozšiřování repetice. Při nalezení překrývajících se sekvencí dostane přednost zarovnání s vyšším skóre.

Pro vyhledávání dokonalých repetic nejsou dovoleny žádné mezery nebo neshody, na rozdíl od hledání nedokonalých repetic. Pro označení repetice za nalezenou musí práh skóre překročit předem určenou prahovou hodnotu.

Pro zjednodušení vyhledávání Phobos na rozdíl od dvou ostatních vyhledávačů dovoluje nastavení vysokého množství jak vstupních, tak výstupních dat. Může se volit mezi metodami vyhledávání pouze dokonalých nebo i nedokonalých repetic. Dá se nastavit přesné rozmezí vyhledávání v sekvenci a také rozsah hledaných repetičních jednotek. Samozřejmě dovoluje libovolné nastavení skórování a vlastnosti satelitu, aby byl správně označen za repetici (minimální délka satelitu, minimální skóre, minimální a maximální procento shody). Na rozdíl TRF a IMEx je Phobos pouze ke stažení a proto jsou jeho výstupy ukládány jako textový soubor a ne jako tabulka. [10]

5.4 Umělé testovací sekvence

Pro vytvoření umělých testovacích sekvencí bylo použito programovací prostředí MATLAB. Tvorba probíhá jednoduše díky bioinformatickému toolboxu. Nejprve je vybrán vhodný motiv podle různých typů repetic vyskytujících se u mikrosatelitů (viz. Tabulka 2). Ten je libovolným množstvím kopírován za sebe příkazem `repmat`, čímž vznikne vektor znaků, který vyjadřuje satelit. Pomocí funkce `randi`, která vybere určitý počet čísel v zadaném rozmezí, se vytvoří pozice inzercí, delecí a substitucí. Na těchto pozicích pak dojde v cyklu `for` k bodové mutaci. Vzniklý satelit je pak vložen mezi náhodné sekvence o libovolné délce, vzniklé příkazem `randseq`. Nakonec je vytvořen soubor s koncovkou FASTA, ve kterém se nachází typický první řádek začínající znakem „>“ a nadpisem. Od druhého řádku až do konce textového souboru je pak vložena samotná sekvence. Následující tabulka shrnuje vlastnosti uměle vytvořených sekvencí.

Tabulka 3: Vlastnosti uměle vytvořených sekvencí pro testování

Soubor [*.fasta]	Celková délka sekvence	Repetiční motiv	Pozice satelitů	Počet opakování	Počet indelů	Počet substitucí
TEST1	2734	ACGT	[101:629]	132	0	0
		ACGT	[1129:1657]	132	0	0
		ACGT	[2157:2685]	132	0	0
TEST2	2452	CGAGAT	[351:1068]	118	26	30
		CGGA	[1447:2302]	215	20	12
TEST3	4855	$(CT)_5A$	[1334:2386]	96	20	22
		$(CT)_5A$	[3075:4105]	96	85	46
TEST4	7574	CCTGA	[1251:2506]	256	110	75
		CCTGA	[3172:4428]	256	225	80
		CCTGA	[5276:6581]	256	325	150

5.4.1 Testování sekvence TEST1.fasta

V testování této základní sekvence obstávají všechny tři programy na 100 %. Všem se daří přesně určit repetiční jednotku, začátek repetice, počet opakování a repetici správně ukončit. Tato sekvence je naprosto základní a všechny programy by ji měli jednoduše a perfektně odhalit, pokud slibují nacházení tandemových repetic u reálných sekvencí.

5.4.2 Testování sekvence TEST2.fasta

V této sekvenci se vyskytují dva satelity. Jeden více mutovaný (CGAGAT), druhý méně (CGGA). Na těchto lehce zmutovaných sekvencích je vidět opravdová funkčnost programů.

TRF

V tomto případě si Tandem Repeats Finder výborně poradí s určením obou satelitů a jejich přesných poloh.

```
Sequence: TEST2_2 repetice mutované|TEST2
Parameters: 2 7 7 80 10 50 500
Length:  2452

Tables:   1

This is table  1  of  1  ( 2 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)
351--1068	6	118.0	6	88	5	1061	32	16	33	16	1.92
1447--2302	4	215.0	4	93	4	1493	24	25	49	0	1.55

Obrázek 9: Výstup vyhledávání pro soubor TEST2.fasta programu TRF

IMEx

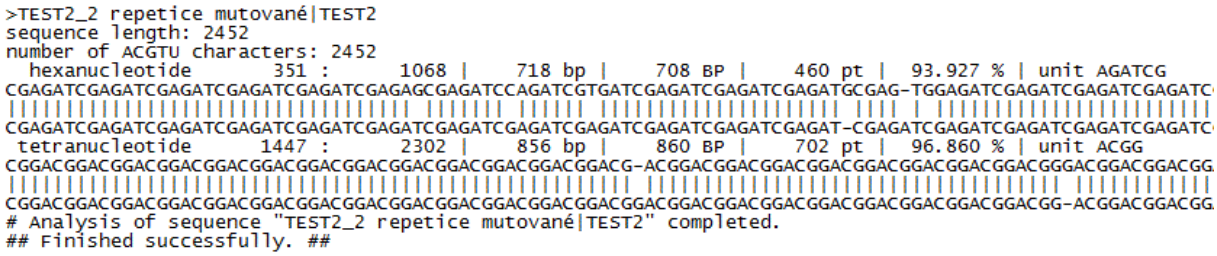
V programu IMEx se již bohužel nedaří nalézt první více mutovanou sekvenci. Při různém nastavení parametrů je schopný vyhledat pouze druhý satelit. První satelit rozděluje na několik samotných, alespoň se správným určením repetičního motivu (přestože je posunutý) a začátku první repetice.

S.No	Consensus	Rep. Size	Iterations	Tract-size	Start	End
1	GAGATC	6	27	163	430	592
2	GAGATC	6	26	155	626	780
3	GAGATC	6	23	138	932	1069
4	CGGA	4	220	859	1447	2305

Obrázek 10: Výstup vyhledávání pro soubor TEST2.fasta programu IMEx

Phobos

Phobos dokázal na rozdíl od programu IMEx rozeznat jak repetiční motiv, tak začátky a konce obou satelitů správně.



Obrázek 11: Výstup vyhledávání pro soubor TEST2.fasta programu IMEx

5.4.3 Testování sekvence TEST3.fasta

Záměrem této sekvence je otestovat přesnost programu při rozlišování delších repetičních motivů a bodových mutací. Jedná o zkoušku, zdali programy přiřadí adenin z repetičního motivu $(CT)_5A$ správně do motivu nebo jestli ho budou brát za bodovou mutaci. Pokud se to programu podaří, je v sekvenci skrytý ještě druhý stejný satelit ovšem již mnohem více zmutovaný. Jedná se o tzv. přerušovanou repetici.

TRF

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
1332--2384	2	486.0	2	78	18	1016	9	44	0	44	1.41
1334--2386	11	96.1	11	93	3	1815	9	44	0	44	1.41
3072--4104	2	494.5	2	70	22	557	10	44	1	43	1.47
3079--4105	11	95.6	11	81	12	1198	10	44	1	43	1.47
3079--4105	75	13.4	79	75	19	780	10	44	1	43	1.47

Obrázek 12: Výstup vyhledávání pro soubor TEST3.fasta programu TRF

Z obrázku výše lze po lepším prozkoumání hodnot zjistit fakt, že program TRF funguje uspokojivě i při těžším úkolu. Na první pohled sice označil mnohem více satelitů, než očekáváme, ovšem nalezl správně motivy $(CT)_5A$ u obou satelitů. Zároveň však za repetice označil i motiv $(CT)_5$ a všechny molekuly adeninu za bodové mutace. U obou satelitů správně nalezl začátky a konce repetic.

Jedinou zvláštností je poslední vyhledaný satelit na místě 3079 až 4105 o délce 75 bp a motivu (CTCTCTACTCTCTCTCTACTCTCTCTCTACTCCTCACTCTCTCTCTCTACTCTCTC TCCCTCTACTCTCTCTCTACTCT). Je zvláštní, že i tento motiv program označil vedle obou správně vyhledaných také.

IMEx

IMEx bohužel nepodporuje vyhledávání repetic delších, než jsou hexanukleotidy, proto by výsledky testování byly naprosto nevypovídající a nesmyslné. Proto byl program z testu souboru TEST3.fasta vynechán.

Phobos

Phobos nachází správně pouze dva vložené satelity s motivy $(CT)_5A$. Ani při nastavení hledání motivů až do délky 100 bp neoznačuje Phobos sekvenci, kterou TRF označuje za repetici.

```
>TEST3_ dlouhý motiv mutované|TEST3
sequence length: 4855
number of ACGTU characters: 4855
   11-nucleotide      1334 :      2386 |   1053 bp |   1057 BP |    844 pt |  96.689 % | unit ACTCTCTCTCT
CTCTCTCTCTAC-CTCTCTCTACTCTCTCTCTACTC-CTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCACTACTCTCTCTCTACTACTC
CTCTCTACTC-CTCTCTACTCTCTCTCTACTCTCTCTCTAC
|||||||||||| ||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||||| |||||||||||||||| |||
|||||||||| ||||||||||||||||||||||||||||||
CTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACT-CTC
CTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTAC
   11-nucleotide      3079 :      4105 |   1027 bp |   1052 BP |    429 pt |  89.924 % | unit ACTCTCTCTCT
CTCTCTAGTCTCTCTC-ACTCTCTCT-TACTCTCTCTCTGCTCTCTCTATACTCTCACCTCTACTCTCTCTCTAACTCTCTCTCCTACTCTCTCTCTACTCTCTC
CTCTCTATCTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTA
||||||| |||||||| ||||||||| |||||||||||| |||||||| |||||||  ||||||||||||||| ||||||||| ||||||||||||||||||
||||||| ||||||||||||||||||||||||||||||||||||||||||||
CTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTC-TCTCTACTCTCTCTCTA-CTCTCTCTC-TACTCTCTCTCTACTCTCTC
CTCTCTA-CTCTCTCTCTACTCTCTCTCTACTCTCTCTCTACTCTCTCTCTA
# Analysis of sequence "TEST3_ dlouhý motiv mutované|TEST3" completed.
## Finished successfully. ##
```

Obrázek 13: Výstup vyhledávání pro soubor TEST3.fasta programu Phobos

5.4.4 Testování sekvence TEST4.fasta

Sekvence je vytvořena pro zjištění maximální rozlišovací schopnosti vyhledávačů a nachází se v ní tři satelity. První satelit je zmutovány z 15 %, druhý z 24 % a třetí z 36 %.

TRF

První dvě méně mutované sekvence se ještě daří programu TRF nalézt naprosto precizně. Z nejvíce mutované sekvence nachází pouze satelit z místa 6294 do 6570. Dále podobně jako v testování předešlého souboru označuje dvě repetice, jednu s motivem 61 bp, druhou s 76 bp, které nejsou vědomě vloženy do umělé sekvence.

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
1251--2505	5	255.8	5	85	10	1681	20	39	19	20	1.93
3174--4426	5	251.0	5	70	20	696	20	36	21	20	1.95
3256--3902	61	10.1	64	70	22	209	21	37	21	19	1.94
3731--3902	76	2.3	76	81	9	189	22	37	22	18	1.94
6294--6570	5	54.6	5	79	9	276	21	37	19	20	1.94

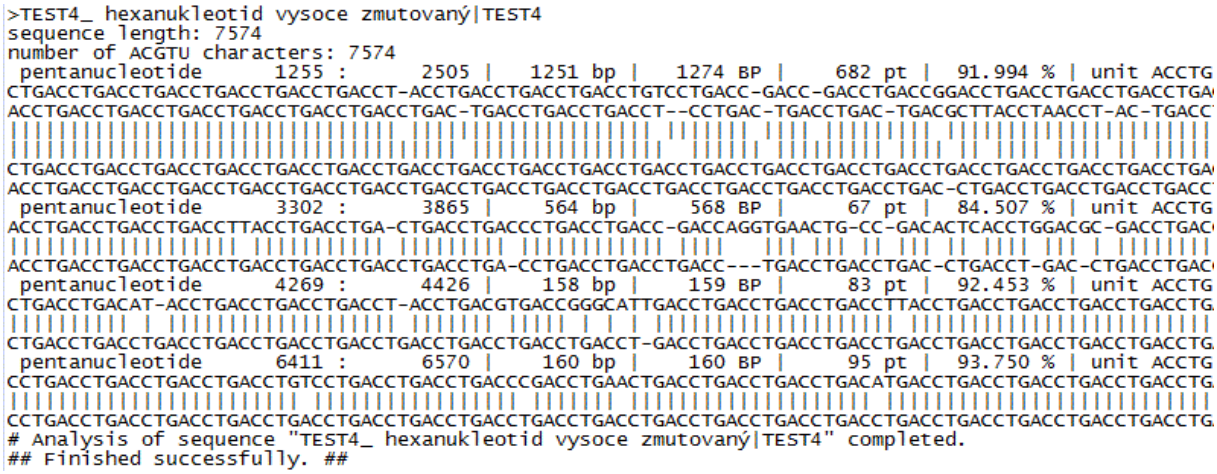
Obrázek 14: Výstup vyhledávání pro soubor TEST4.fasta programu TRF

IMEx

Tento program bohužel nezvládá ani v rámci možností vyhodnotit silně zmutované sekvence. Při změnách nastavení se dostává pouze k vyhledání 29 různých satelitů náhodně rozmístěných po celé testované sekvenci.

Phobos

Programu Phobos se daří nalézt první satelit bez jakýchkoliv nepřesností. Poté označuje dva satelity s pentanukleotidy (ACCTG). Jeden na místě 3302 až 3865, druhý na místě 4269 až 4426, čímž označuje méně než 60 % uměle vloženého satelitu. Ze třetího satelitu nachází pouze úsek 6411 až 6570, což není ani 20 % celkové délky.



Obrázek 15: Výstup vyhledávání pro soubor TEST4.fasta programu Phobos

Tabulka 4: Tabulka zhodnocení programů při testování umělých sekvencí

	TRF	**IMEx**	**Phobos**
TEST1	Správné, okamžité vyhledání.	Správné, okamžité vyhledání.	Správné, okamžité vyhledání.
TEST2	Správné, okamžité vyhledání.	Již tady špatně rozděluje první satelit na více samostatných.	Správné, okamžité vyhledání.
TEST3	Správné vyhledání umělých satelitů, nesprávné označení přerušovaných repetic za bodové mutace	Nereálné nesmyslné výsledky, protože nepodporuje vyhledávání motivů delších než 6 bp.	Správné, okamžité vyhledání.
TEST4	První dvě repetice nachází v pořádku. třetí již rozděluje a identifikuje špatně.	Vyhledá množství repetic, které nejsou vytvořeny, opravdu vytvořené nenachází	První satelit nachází správně, z druhého. asi 60 %, ze třetího pouze 20%

5.5 Reálná testovací data

Pro testování reálných byly vybrány tři sekvence. Sekvence byly staženy ve formátu .fasta přes webovou stránku http://www.ncbi.nlm.nih.gov/. Následující tabulka shrnuje jejich vlastnosti.

Tabulka 5: Vlastnosti reálných testovacích dat

Soubor [*.fasta]	**Popis**	**Původce sekvence**	**Délka testované sekvence**	**Očekávané výsledky**
vrrA_b_anthracis	Protein kódující gen.	*Bacillus cereus biovar anthracis*	747 bp	Jediný satelit, 12-nukleotid.
HCH12ATP	Gen pro atropin, nacházející se na chromosomu 12.	*Homo sapiens*	17 859 bp	Různé množství kopií trinukleotidu CAG/CAA.
PF3D7CH4	Celý chromosom 4.	*Plasmodium falciparum 3D7*	1 204 112 bp	Velké množství repetic, jedná se o zátěžový test.

5.5.1 Testování sekvence vrrA_b_anthracis.fasta

Sekvence vrrA (variable repeat region A, česky: úsek variabilního opakování) u *Bacillus anthracis* se vyznačuje satelitem s několikanásobným opakováním motivu (CAATATCAACAA). Počet opakování se mezi jedinci liší.

TRF

TRF v tomto případě správně našel a označil jediný satelit.

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
229--266	12	3.2	12	92	0	58	52	26	5	15	1.64

Obrázek 16: Výstup vyhledávání pro soubor vrrA_b_anthracis.fasta programu TRF

IMEx

U programu IMEx je největší nevýhodou omezení vyhledávání tandemových repetic, jejichž motiv je delší než šest nukleotidů. Proto v tomto případě program nenachází hledanou sekvenci, ale bohužel ani žádnou přerušovanou sekvenci, která by se nacházela poblíž hledané.

S.No	Consensus	Rep. Size	Iterations	Tract-size	Start	End
1	CAGCAA	6	2	12	208	219
2	AAACG	5	2	10	351	360
3	AATTT	5	2	10	418	427
4	AGAAAA	6	3	18	643	660
5	CAAAAC	6	2	12	725	736

Obrázek 17: Výstup vyhledávání pro soubor vrrA_b_anthracis.fasta programu IMEx

Phobos

Při nastavení minimálního skóre zarovnání 5, nachází Phobos hledanou tandemovou repetici s motivem 12-nukleotidu. Oproti TRF ovšem nachází i další mikrosatelit.

```
>gi|301051741:4013814-4014560 Bacillus cereus biovar anthracis str. CI chromosome, complete genome
sequence length: 747
number of ACGTU characters: 747
   12-nucleotide       229 :       266 |     38 bp |     38 BP |     16 pt |  94.737 % | unit AACAACAATATC
CAATATCGACAGCAATATCAACAACAATATCAACAACA
|||||||  ||| ||||||||||||||||||||||||||
CAATATCAACAACAATATCAACAACAATATCAACAACA
  hexanucleotide       641 :       661 |     21 bp |     22 BP |     11 pt |  95.455 % | unit AAAAAG
AAAA-AAAAAGAAAAAGAAAAA
||||  ||||||||||||||||
AAAAGAAAAAGAAAAAGAAAAA
# Analysis of sequence "gi|301051741:4013814-4014560 Bacillus cereus biovar anthracis str. CI chromosome, c
## Finished successfully. ##
```

Obrázek 18: Výstup vyhledávání pro soubor vrrA_b_anthracis.fasta programu IMEx

5.5.2 Testování sekvence HCH12ATP.fasta

Jedná se o část lidského chromozomu 12, ve které se může vyskytovat 7 až 13, nebo 49 až 75 kopií trinukleotidu CAG/CAA. Tato tandemová repetice je spojována s výskytem neurodegenerativního onemocnění, tzv. dentatorubropallidoluisické atrofie.

TRF

Díky jasnému a přehlednému uspořádání výsledků do tabulky, je po chvíli vidět, že očekávaný trinukleotid, v tomto případě CAG, se nachází v sekvencí na místě 12 255. Jeho motiv se opakuje se 19.7x. Poslední místo repetice by mělo být spíše označeno jako bodová mutace, a motiv by se tedy opakoval dvacetkrát.

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
639--680	2	21.0	2	90	0	66	0	4	50	45	1.23
2498--2536	20	2.0	20	84	0	51	15	48	7	28	1.72
3214--3243	15	2.1	15	93	6	53	0	63	13	23	1.29
3583--3634	18	2.9	18	91	2	79	0	32	67	0	0.91
6339--6373	4	8.5	4	93	6	61	77	0	0	22	0.78
7573--7599	1	27.0	1	100	0	54	100	0	0	0	0.00
6948--7597	320	2.1	320	85	7	900	34	21	27	17	1.96
12255--12313	3	19.7	3	92	0	100	37	33	28	0	1.58
12509--12537	15	1.9	15	100	0	58	10	48	0	41	1.37
13546--13589	18	2.4	18	88	0	61	31	20	47	0	1.50
13555--13591	18	2.1	18	94	0	65	29	24	45	0	1.53

Obrázek 19: Výstup vyhledávání pro soubor HCH12ATP.fasta programu TRF

IMEx

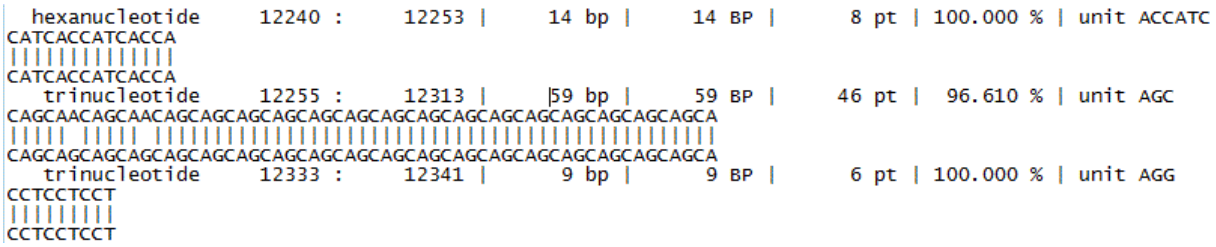
Program IMEx v tomto případě označuje 58 různých mikrosatelitů. Nachází i očekávanou repetici od místa 12 246 v počtu 24 iterací, tedy o pět iterací více, než nalezl TRF.

39	CAGCAA	6	4	24	12255	12278
40	CAG	3	24	72	12246	12317
41	TCTTCC	6	2	12	12510	12521

Obrázek 20: Výstup vyhledávání pro soubor HCH12ATP.fasta programu IMEx

Phobos

Tento program nachází totožný výsledek s TRF, tedy 19,7 násobek opakování trinukleotidu ACG od místa 12 255. Bohužel vyhodnocení této sekvence zabírá několik minut navíc oproti ostatním programům, které ji vyhodnotily za méně než jednu vteřinu. Dále na rozdíl od ostatních programů nachází několik dalších desítek sekvencí.



Obrázek 21: Výstup vyhledávání pro soubor HCH12ATP.fasta programu Phobos

5.5.3 Testování sekvence PF3D7CH4.fasta

Sekvence s několika stovkami tandemových repetic slouží jako zátěžový test. Celková délka 1,2 Mbp zkouší, zda jsou programy i při větším množství vstupních dat schopné dosahovat srozumitelných výsledků. Přesnější srovnání vyhodnocených dat všemi programy je v podstatě nemožné. Výstupem všech programů je nalezení několika tisíc tandemových repetic. U TRF jsou nalezeny asi 4000, u IMEx dokonce 7963. Dalším velkým rozdílem je čas vyhodnocení, který u TRF činí méně než minutu u IMEx výsledky připravuje asi 10 minut a Phobos se k výsledkům dostává po necelých 4 hodinách.

Tabulka 6: Tabulka zhodnocení programů při testování reálných dat

	TRF	**IMEx**	**Phobos**
vrrA_b_anthracis	Správné, okamžité vyhledání očekávané tandemové repetice. (>1 s)	Nenachází námi hledanou sekvenci. (>1 s)	Správné, okamžité vyhledání očekávané tandemové repetice. (>1 s)
HCH12ATP	Nachází hledanou repetici a několik dalších. (>1 s)	Nachází hledanou repetici ve větším rozsahu než ostatní vyhledávače. (2:36 min)	Nachází hledanou repetici a několik dalších. (>1 s)
PF3D7CH4	Za čas kratší než 1 minuta, nachází přibližně 4000 mikrosatelitů. (<1 min)	Během cca 10 minut nachází necelých 8000 repetic. (cca 10 min)	Po necelých 4 hodinách nachází více než 15 tisíc repetic. (cca 4 hodiny)

5.5.4 Zhodnocení testování veřejně dostupných vyhledávačů

Při postupném testování jak uměle vytvořených sekvencí, tak reálných dat se ukazují jednotlivé výhody a nevýhody programů.

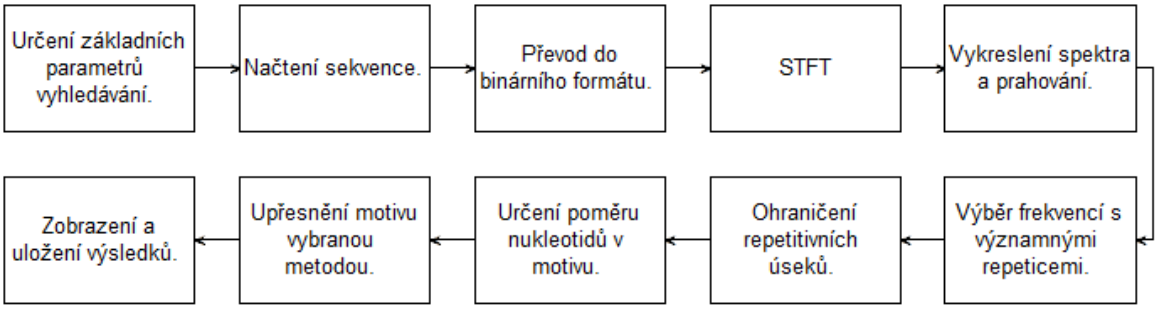
Program IMEx se po otestování na zadaných sekvencích jeví jako nejméně konkurence schopný. Jeho omezení spočívá hlavně v jeho algoritmu, který nebyl schopný vyhledat motivy repetic delší než šest nukleotidů. Při vyhledávání kratších motivů vyhledával dostatečně. Bohužel začal velice brzy rozdělovat repetice, u kterých se vyskytuje větší množství bodových mutací. Při hledání repetic v testovacích sekvencích ve většině případů bohužel selhal. Dostatečně si poradil se zátěžovým testem a v testování sekvence HCH12ATP.fasta označuje hledanou repetici jako delší, než ostatní dva programy. Program je tedy vhodný pro vyhledávání spíše nemutovaných repetic s motivem o rozsahu 1 až 6 nukleotidů.

Program TRF je svým pravděpodobnostním algoritmem velice složitý, ovšem v testování obstál výborně. V testu umělých sekvencí obstál naprosto dostatečně. Jeden z horších výsledků se objevil při testování sekvence TEST3.fasta, kdy program správně označil přerušovanou repetici $(CT)_5A$ jako samostatnou repetici, ale také ji označil za repetici s motivem CT, a každý adenin označil za bodovou mutaci. Ve výsledcích programu TRF se často objevují velice dlouhé motivy (více než 100 bp), s méně než dvěma opakováními, které ostatní programy neoznačují za repetice. V testování reálných sekvencí také obstál velice dobře, a to zejména při zátěžovém testu sekvence o délce 1,2 Mbp, kterou zvládl vyhodnotit za méně než jednu minutu. Celkově v testu obstál TRF výborně. Program je vhodný pro vyhledávání v neznámých sekvencích, díky přehlednosti výsledků se v nich dá dobře orientovat.

Program Phobos si vedl v testování umělých sekvencí pravděpodobně nejlépe, problém začaly dělat více zmutované sekvence, což je vidět na testu souboru TEST4.fasta, kde ze tří sekvencí nalezl správně pouze jednu. Z dalších dvou již nalezl bohužel jen části. U poslední vytvořené tandemové repetice označil pouze 20 % za repetici, její zbytek nenalezl. Při vyhledávání tandemových repetic v reálných sekvencích se osvědčil dobře. Na rozdíl od obou předchozích programů je jeho algoritmus mnohem více výpočetně náročný, a tak již při testování sekvence o 17 kbp trvalo vyhledávání více než 2 minuty. Při zátěžovém testu trvalo vyhledávání tandemových repetic necelé 4 hodiny. Program Phobos dopadl podobně jako TRF. Obstál dobře ve všech testech, jednou z nevýhod je jeho výpočetní náročnost. Další nevýhodou je výstup v podobě textu, který je mnohem méně přehledný, zvláště na první pohled, než výstupy programů IMEx a TRF.

6. Struktura vlastního programu

Program pro vyhledávání tandemových repetic je založen na převedení testované sekvence do vhodného numerického formátu, následné Fourierově transformaci získaného signálu a jeho spektrální analýze. Pro lepší přehled a jednodušší představu funkčnosti programu je dále uvedeno obecné blokové schéma programu.



Obrázek 22: Postupné blokové schéma programu

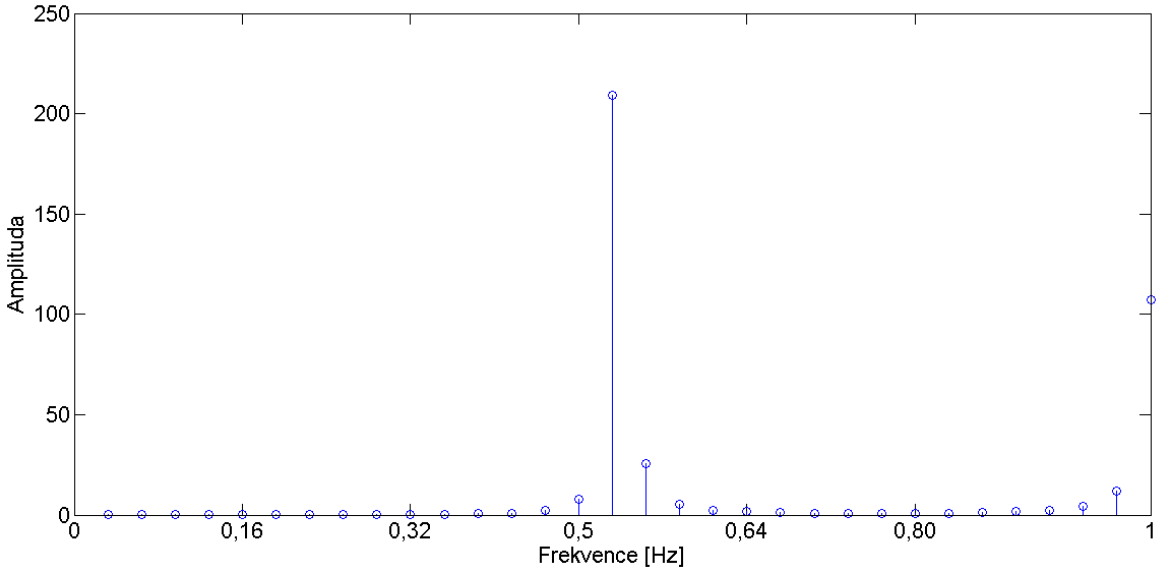
Po určení základních parametrů vyhledávání, které budou vysvětleny posléze a po načtení sekvence ze souboru ve formátu *FASTA* se program dostává k převodu sekvence do binárního formátu, který je podrobně popsaný v kapitole 3. Tímto převodem je vytvořena matice o 4 řádcích, kde jednotlivé řádky reprezentují jednotlivé nukleotidy, a N sloupcích, kde N značí celkovou, délku sekvence. V tuhle chvíli je již možné na numericky reprezentovanou sekvenci použít matematické operace, které nejsou na sekvenci znaků aplikovatelné jako je STFT.

Po provedení Fourierovy transformace s klouzavým oknem pro každou ze čtyř numerických sekvencí jsou výstupem čtyři spektra, každé pro jednotlivý nukleotid. Pro zisk celkového spektra a zvýšení jeho výpovědní hodnoty je vypočteno výkonové spektrum.

$$S(k) = |A(k)|^2 + |C(k)|^2 + |G(k)|^2 + |T(k)|^2 \qquad (6)$$

,kde A, C, G a T reprezentují vektory spektrálních koeficientů nukleotidy adenin, cytosin, guanin a tymin.

Následující obrázek (Obrázek 23) reprezentuje spektrum pro adenin v sekvenci TEST1.fasta v repetitivní oblasti, kde se opakuje motiv ACGT o délce 16 bp a adenin se tedy opakuje s frekvencí ¼. Na této frekvenci je také jasně zřetelný pík. Toto spektrum vznikne v každém cyklu programu a je vloženo do spektrogramu, který může být zobrazen.



Obrázek 23: Spektrum pro adenin

Po výpočtu výkonového spektra z výstupu STFT jsou tyto hodnoty vloženy do spektrogramu, ve kterém se zvýrazní úseky s vyšší maximální hodnotou amplitudy spektra, tedy úseky, v nichž je jistá periodicita vstupního signálu.

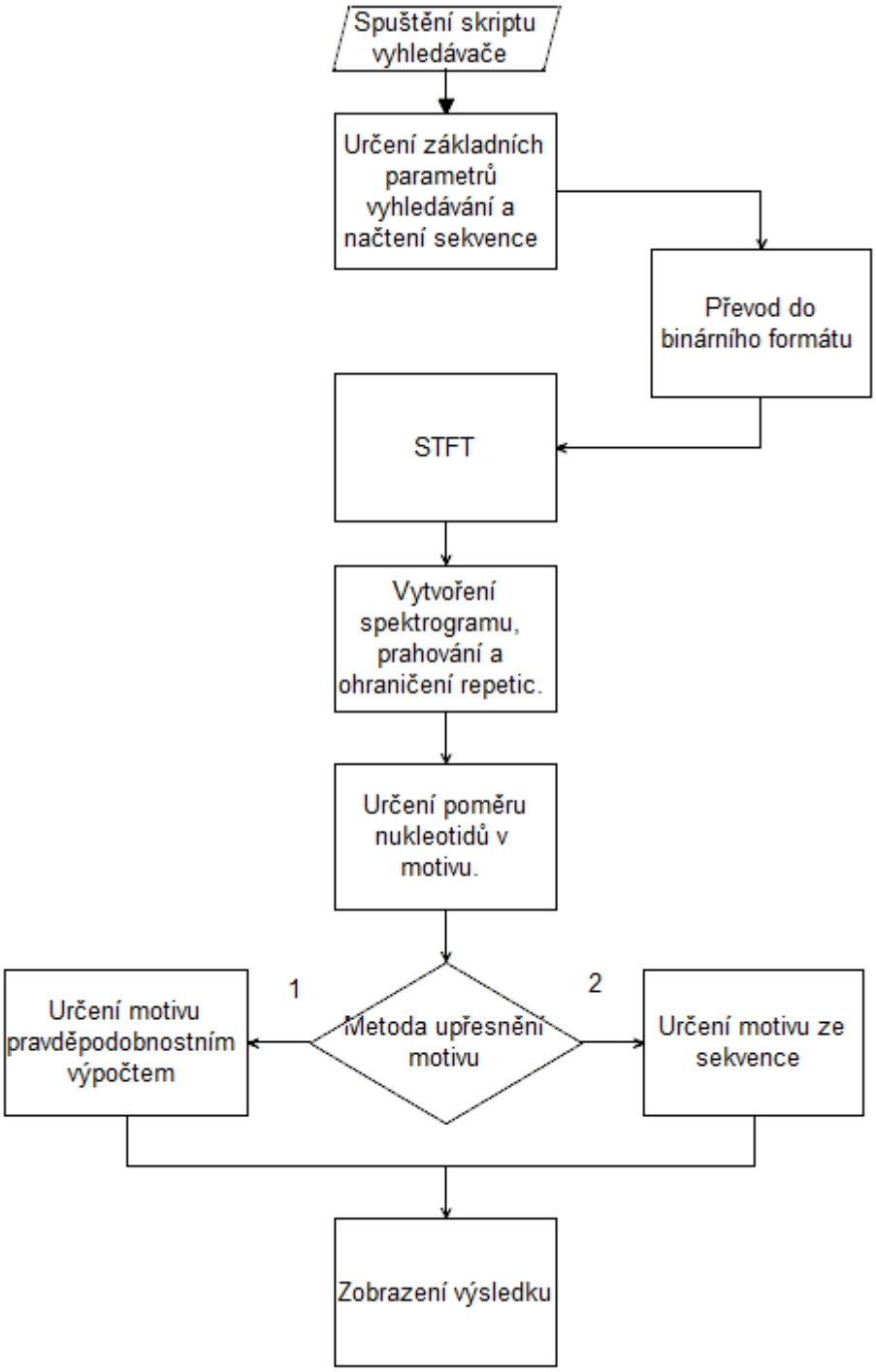
Dále již program pracuje se spektrogramem, přesněji s jeho jednotlivými řádky, ve kterých vybírá hodnoty, které jsou vyšší než určený práh. Pozice těchto hodnot uloží, a pokud je mezi nimi větší mezery, než stanovená hodnota `maximalni_mezera`, označí hodnotu před mezerou jako konec jedné repetice a pozici za mezerou jako začátek další repetitivní sekvence. První nadprahová hodnota je zároveň začátkem první tandemové repetice a poslední hodnota je koncem poslední tandemové repetice. Tímto postupem tedy program najde začátek a konec repetitivní oblasti, který uloží. Délka motivu repetice je také určena ze spektra a to z hodnot *y* osy, které odpovídají frekvenci opakování. Tedy pokud se motiv opakuje po každém třetím vzorku, pak je jeho odhadnutá délka 3.

Dále algoritmus najde střed této repetitivní oblasti, a pomocí DFT určí spektrum pro všechny nukleotidy a z velikosti amplitudy jednotlivých spekter určí poměr jednotlivých nukleotidů v motivu repetice. Např. pro motiv ATCT bude maximální amplituda spektra adeninu v ¼, cytosinu také v ¼, amplituda guaninu bude blízká nule na všech frekvencích a hodnota amplitudy spektra T bude nejvyšší v ½ a zároveň bude přibližně dvojnásobná oproti maximálním hodnotám spekter adeninu a cytosinu. Tyto dílčí informace dají dohromady vcelku přesný odhad poměru nukleotidů v motivu. Ovšem přesnou pozici jednotlivých nukleotidů v motivu není jistý, přesto pomocí tohoto poměru výrazně zúžíme možnosti pro odhalení přesného motivu. Pro motiv o délce tří nukleotidů existuje 4^3 možných zarovnání. Pokud je poměr nukleotidů v tomto motivu [A:C:G:T] [0:2:0:1], pak existují pouze 3 možnosti, jak motiv reálně vypadá. Poté je tedy mnohem jednodušší motiv přesně určit.

Výstupem celého programu je uspořádaná tabulka všech vyhledaných repetic, s označením jejich začátků a konců, určenou délkou repetitivní oblasti, určenou délkou motivu, počtem opakování motivu a v poslední řadě poměrem jednotlivých nukleotidů v motivu.

6.1 Realizace algoritmu

Metoda vyhledávání tandemových repetic v sekvencích DNA byla řešena v programovém prostředí Matlab verze R2014a s použitím Bioinformatics_Toobox a Signal_Processing_Toolbox.



Obrázek 24: Blokové schéma navrhnutého algoritmu

6.1.1 Funkce pro výpočet Fourierovy transformace

Skript ***ftrans.m*** obsahuje funkci pro výpočet Fourierovy transformace, což je základ celého navrhnutého algoritmu. Tato funkce je volána v několika sekcích a proto je vhodné ji popsat. Vstupní hodnotou je sekvence převedená do binárního formátu, výstupem je spektrum vstupní sekvence.

Funkce nejprve určí počet vzorků vstupní sekvence `N`. Dále se vytvoří sekvence nul `SEK` pro vložení vypočítaných hodnot spektra. V cyklu poté provádí samotnou FT podle vzorce 4. Počítá se `N/2` spektrálních koeficientů. Posledním krokem je výpočet absolutní hodnoty výsledku, protože výstupem FT jsou čísla s komplexním přírůstkem, ale informativní je pro nás pouze reálná část.

6.1.2 Funkce pro převod DNA do binárních vektorů

Přestože je převedení sekvence do binárního formátu naprosto triviální záležitostí, je převod sekvence do vhodného numerického formátu naprosto zásadním předpokladem pro použití Fourierovy transformace na analýzu sekvence.

Ve funkci ***ntTObin*** nejprve dochází k ujištění, že celá sekvence se skládá pouze z velkých písmen funkcí ***upper***. Výstupní proměnnou je vektor matice `bin` o čtyřech řádcích a stejné délce, jako je délka vstupní sekvence. Jednotlivé řádky reprezentují jednotlivé nukleotidy a plní se jedničkami a nulami podle toho jaké písmeno se nachází na vybrané pozici. Pokud se ve vstupním vektoru nachází jiný znak než A, C, G nebo T pak je na místě, na kterém se nachází ve všech vektorech uložena 0.

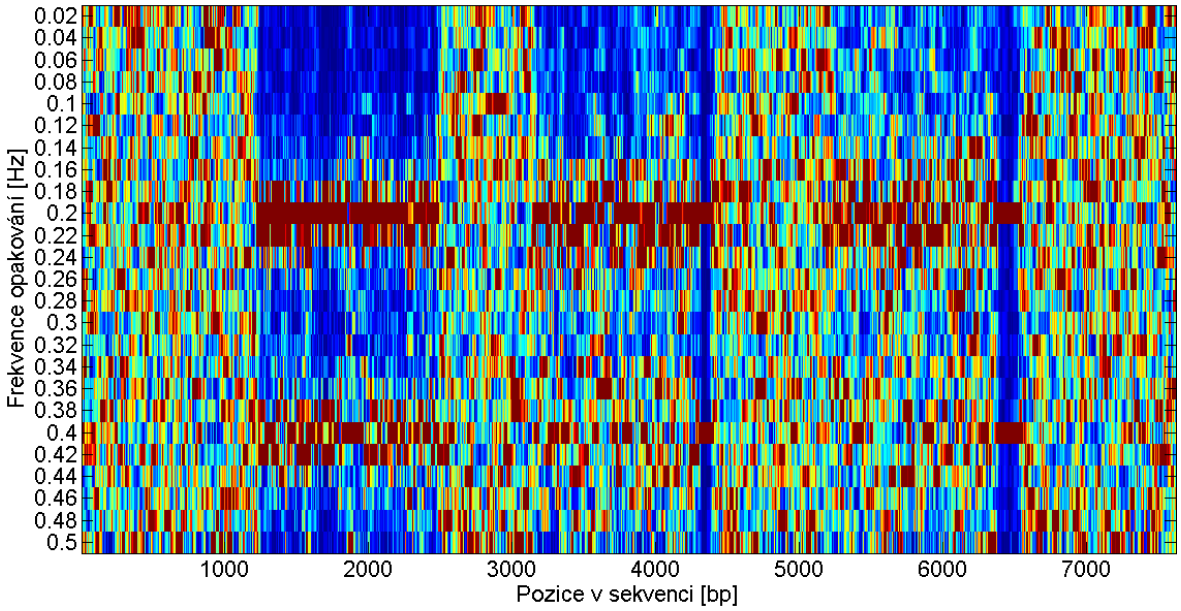
6.1.3 Funkce pro vytvoření spektrogramu a analýzu repetitivních oblastí

Dále popisovaná funkce ***Vyhledavac*** funguje samostatně, nebo ji lze spustit z jednoduchého uživatelského prostředí přes funkci ***GUI***. Pro samostatně funkční skript je ovšem nutné odkomentovat řádky 4 až 11 a 171 až 180.

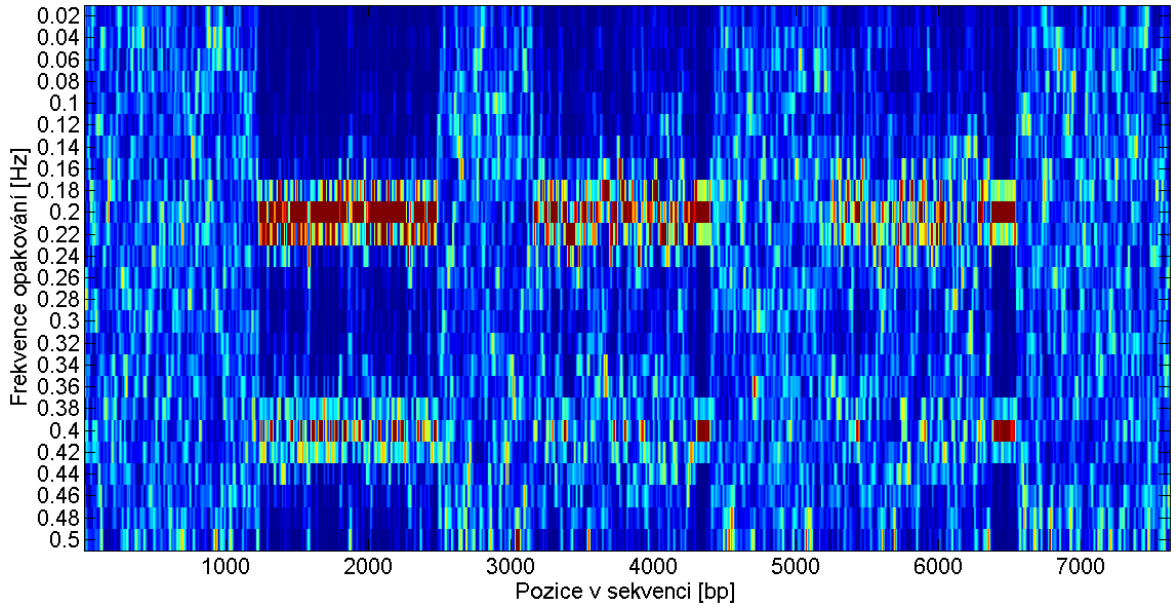
Hlavní skript pro spuštění programu je *Vyhledavac.m*. Algoritmus výpočtu začíná učením základních parametrů pro vyhledávání. Nejprve se určuje proměnná `delka_okna`, která určuje délku sekvence postupně podstupující FT. Čím delší je okno pro výpočet, tím delší je čas výpočtu výsledku, protože každý bod výstupního spektra je ovlivněn všemi vstupními body. Přesně je délka výpočtu FT závislá na čtverci délky vstupního řetězce. Dále potřebujeme určit krok, s jakým se bude klouzavé okno posouvat po sekvenci. Proměnná `krok` je pro přesné ohraničení repetitivní oblasti nastavena na 1. Proměnná `minrepl` určuje minimální délku repetice, `minrepc` pak minimální počet opakování motivu. Tyto hodnoty jsou v základu nastaveny na `delka_okna=50`, `minrepl=20` a `minrepc=10`.

Po spuštění skriptu je uživatel v command window vyzván k výběru metody vyhledání repetic, tou je buď výběr všech možných tandemových repetic, nebo je uživateli umožněno vybrat ze spektra určité frekvence, na kterých vyhledávat. Nastavení proměnné `maximalni_mezera` bude vysvětleno dále.

V následující sekci skriptu je nejprve načtená sekvence na konci doplněna o náhodnou sekvenci o délce dvou oken pro FT, čímž je zaručeno vyhledání repetic v celé sekvenci. Kdyby nebylo zavedeno toto opatření, pak Fourierova transformace neskončí na posledním místě sekvence, ale na místě, které je vzdáleno o délku okna od konce sekvence. Důležité je také určení konečné hodnoty znaků `N`. Následuje převedení řetězce znaků funkcí ***ntTObin*** do binárního formátu a uložení do proměnné `bin`. Pro tuto proměnnou proběhne STFT (funkce ***ftrans***) a postupně všechna vypočtená spektra vloží do spektrogramu, který může uživatel nechat zobrazit. Pro lepší vizualizaci je každý úsek vstupující do funkce ***ftrans*** násoben Hannovým oknem.



Obrázek 25: Spektogram vytvořený bez hodnot násobených Hannovým oknem



Obrázek 26: Spektrogram vytvořený po násobení hodnot Hannovým oknem

Pro ohraničení repetitivních oblastí ze spektra je zásadní výběr vhodných hodnot ze spektrogramu. Toho je dosáhnuto určením prahu, který byl experimentálně určen jako

$$práh\ spektra = \left(\frac{celková\ suma\ spektra}{délka\ okna * délka\ sekvence}\right) * 6 \qquad (7)$$

protože při této hodnotě dosahuje nejlepších výsledků. Tímto prahem projdou pouze hodnoty vyšší než 60 % maxima. Samozřejmě není obtížné ho změnit a tím citlivost vyhledávání zvýšit, nebo snížit. `Prahovane_spektrum` je proměnná která obsahuje pouze hodnoty přesahující určený práh násobené 100krát pro zlepšení jejich viditelnosti.

Program umí vyhodnotit buď tandemové repetice v celém spektrogramu, nebo pokud uživatel chce, může sám vybrat řádek (frekvenci) z prahovaného spektrogramu. Pokud si uživatel zvolí vlastní určení oblastí, otevře se okno s prahovaným spektrem, kde dvojklikem vybere řádek, pro který se tandemové repetice vyhledají a vyhodnotí.

Po výběru metody v algoritmu následuje buď pro jediný řádek, nebo postupně pro všechny řádky spektrogramu označení pozic nadprahových hodnot do vektoru `SEQ`. Pokud není první hodnota této pozice nevýznamná krátká repetice uloží se první hodnota vektoru jako začátek první repetice. Pokud se v sekvenci nachází mezera mezi dvěma repeticemi delší než je hodnota `maximalni_mezera`, pak se stávající hodnota `SEQ` uloží jako konec předešlé tandemové repetice a následující hodnota `SEQ` se uloží jako začátek následující repetice.

Při testování vyhledávače se objevila tendence posouvat celé repetice o několik hodnot zpět v sekvenci. Eliminace tohoto posunutí je jednoduše dosaženo přičtením poloviny délky okna k vyhledaným pozicím. Hodnota posunu byla stanovena experimentálně postupným zkoumáním a testováním a porovnáváním se známými pozicemi tandemových repetic v testovacích sekvencích.

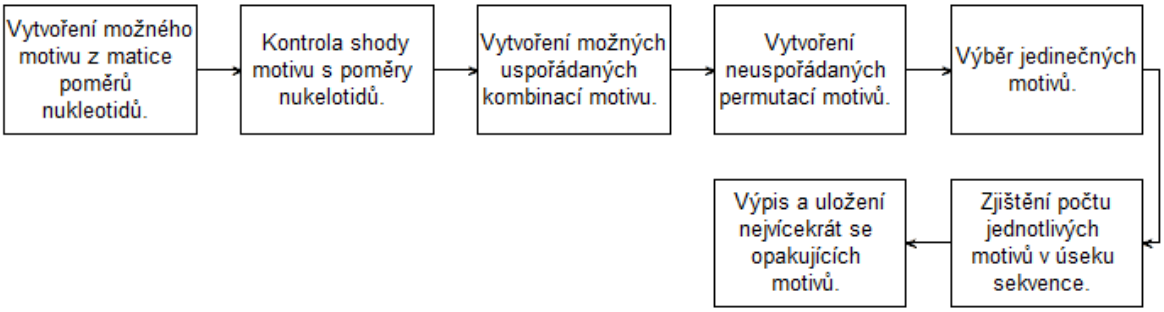
Určení poměru nukleotidů v motivu vyhledané repetice probíhá v další sekci skriptu. Délka motivu je určena pomocí frekvence opakování a je potřeba je určit pro každý prohledávaný řádek spektra. Tyto hodnoty jsou uloženy v proměnné `delka_motivu`. Podělením celkové délky tandemové repetice délkou motivu je určen počet opakování motivu.

Pro určení poměru nukleotidů v motivu repetice se vypočítá DFT pro jednotlivé ohraničené repetice. Určí se maximální hodnoty amplitudy jednotlivých spekter a je vypočítán jejich poměr, který samozřejmě nemusí přesně odpovídat délce motivu, což může být způsobeno krátkou repetitivní sekvencí, jejíž motiv je určován i z hodnot za její hranicí, nebo vysokou četností náhodných mutací ve vybraném úseku okolo středu repetice.

Předposlední sekce tohoto skriptu se snaží přesněji určit motivy jednotlivých nalezených repetic. Jednotlivé vytvořené funkce jsou popsány dále.

6.1.4 Funkce pro určení motivu repetice pravděpodobnostním postupem

Jedná se o funkci pro upřesnění možných motivů nalezených tandemových repetic. Vstupními hodnotami je jeden řádek matice poměru motivů `mot`, úsek s vyhledanou tandemovou repeticí a délka motivu této repetice.



Obrázek 27: Blokové schéma funkce pmotivu

Funkce prochází poměry nukleotidů a podle toho jaké číslo se nachází v právě vybraném sloupci, tolikrát se zopakuje a uloží jako možný motiv.

V matici poměrů může dojít k nepřesnosti a to, že poměry jednotlivých nukleotidů nebudou přesně souhlasit s délkou motivu. Pokud je součet jednotlivých čísel v poměru vyšší, než je určená délka motivu, není problém dále pokračovat, je-li součet jednotlivých čísel v poměru nižší než předpokládaná délka, může být posledním znakem jakýkoliv nukleotid. Například pokud je poměr nukleotidů určen jako [1:2:1:1] ale délka motivu je určena jako 6 nevíme, jaký nukleotid bude na šestém místě. V tu chvíli je potřebná podmínka, která na poslední místo motivu přidá postupně všechny čtyři nukleotidy.

Následuje vytvoření všech možných kombinací z poměrů nukleotidů. Protože v DNA záleží na přesném pořadí jednotlivých nukleotidů, nestačí pouze vytvořit uspořádané kombinace, ale v dalším cyklu jsou vypočítány i všechny možné permutace. Tím dostaneme všechny možné sekvence nukleotidů, které by mohli být hledaným motivem. Zde nastává nevýhoda pravděpodobnostního algoritmu a tou je velká výpočetní náročnost, neboť možných permutací může být mnoho.

Pokud proběhne výpočet permutací bez obtíží, poslední cyklus vypočítá jednotlivé délky opakování všech permutací v úseku a ty motivy, které se opakují nejvíce krát určí jako možné.

6.1.1 Funkce pro určení motivu repetice ze sekvence znaků

Na rozdíl od předchozí funkce, tato nepracuje s pravděpodobnostmi, ale přímo s vloženou vstupní sekvencí, která obsahuje tandemovou repetici.

Pomocí cyklu prochází celý úsek a ukládá obsah posuvného okna o délce motivu do proměnné `motivy`. Pokud je délka motivu 3, pak postupně klouzavé okno projede sekvence a uloží všechny jeho obsahy do matice, ze které následně vybere jen ty jedinečné, neboli vymaže všechny, které jsou v matici vícekrát.

V dalším cyklu porovná poměr nukleotidů všech nalezených krátkých úseků s poměrem motivu, který do této funkce vstupuje jako proměnná. Jakýkoliv řádek v matici `motivy`, který nesplňuje vypočítaný poměr je vymazán a tím je výběr zúžen. Zde nastává problém této metody. Tím je výše zmíněné vyloučení motivů, které nemají shodný poměr nukleotidů. Pokud je tedy poměr určen nepřesně, nelze určit přesný motiv touto metodou.

Při správném vyřazení nechtěných motivů, následuje podobně jako ve funkci ***pmotivu*** vyhledání počtu opakování jednotlivých motivů a uložení pouze těch, které se opakují nejčastěji.

6.1.2 Uživatelská aplikace

Jednoduché uživatelské prostředí zjednodušuje nastavení vstupních parametrů a zároveň zlepšuje přehled výsledků.

V Příloha 2, je vidět vytvořené uživatelské prostředí. Vlevo je několik editovacích polí, ve kterých se dají jednoduše změnit důležité vstupní hodnoty pro výpočet. Pod nimi je tlačítko *Nastavit výchozí,* pomocí kterého se vrátí do editovacích polí přednastavené hodnoty. Pod tímto tlačítkem je pole, ve kterém se volí jedna ze dvou možných metod vyhledávání (z celého spektra, nebo pouze na vybrané frekvenci). Vedle se nachází další pole, ve kterém se vybírá metoda určení motivu. Pod těmito poli se nachází čtverec, který, pokud je zaškrtnutý při výpočtu přidá k výsledku zobrazení spektra.

Výsledky vyhledávání se vloží do tabulky na středu uživatelského rozhraní. Po kliknutí na jakoukoliv hodnotu v tabulce, se vpravo vedle tabulky zobrazí v textovém poli všechny možné určené motivy, té sekvence, na jejímž řádku je označené pole a pod tabulkou se zobrazí část sekvence pro vizuální kontrolu motivu.

Po výpočtu se v dolní části zobrazí tlačítko pro uložení výsledku do formátu XLS a pod tlačítkem pro spuštění výpočtu se zobrazí celkový výpočetní čas v sekundách.

6.2 Vyhledávání tandemových repetic pomocí navrhnutého algoritmu

Testování algoritmu probíhá s výchozími hodnotami délka okna = 50, velikost kroku Fourierovy transformace = 1, maximální mezera = 80, minimální délka repetice = 20 a minimální počet opakování motivu =10. Pro vyhledávání na testovacích sekvencích je využito označení repetic pouze na vybraných frekvencích, pro reálná data je použito vyhledávání v celém spektrogramu.

Testování probíhá na osobním počítači s operačním systémem Windows 7 Professional N Service Pack 1 s procesorem Intel® Core™ i7-2670QM CPU @ 2.2 GHz a s 8 GB paměti RAM.

6.2.1 Sekvence TEST1.fasta

Základní test program plní bez problému a s odchylkou pouze dvou nukleotidů vyznačuje všechny tři uměle vytvořené repetitivní sekvence. Tabulka 7 zobrazuje výsledky poskytnuté programem. Tabulka 8 uvádí přesný popis testované sekvence pro jednoduché srovnání s dosaženými výsledky.

Tabulka 7: Výsledek vyhledávání v sekvenci TEST1.fasta

Číslo repetice	Začátek	Konec	Celková délka	Délka motivu	Počet opakování
1	105	627	522	4	131
2	1129	1655	526	4	132
3	2157	5683	526	4	132

Tabulka 8: Přesný popis sekvence TEST1.fasta

Soubor [*.fasta]	Celková délka sekvence	Repetiční motiv	Pozice satelitů	Počet opakování	Počet indelů	Počet substitucí
TEST1	2734	ACGT	[101:629]	132	0	0
		ACGT	[1129:1657]	132	0	0
		ACGT	[2157:2685]	132	0	0

6.2.2 Sekvence TEST2.fasta

Lehce mutované sekvence nejsou pro program problémem a vyhledá je správně s minimální odchylkou. Následující tabulky (Tabulka 9 a Tabulka 10) ukazují dosažené výsledky a přesný popis testované sekvence.

Tabulka 9: Výsledek vyhledávání v sekvenci TEST2.fasta

Číslo repetice	Začátek	Konec	Celková délka	Délka motivu	Počet opakování
1	361	1062	701	6	117
2	1451	2302	851	4	213

Tabulka 10: Přesný popis sekvence TEST2.fasta

Soubor [*.fasta]	Celková délka sekvence	Repetiční motiv	Pozice satelitů	Počet opakování	Počet indelů	Počet substitucí
TEST2	2452	CGAGAT CGGA	[351:1068] [1447:2302]	118 215	26 20	30 12

6.2.3 Sekvence TEST3.fasta

V této testovací sekvenci nastává první problém vyhledávání pomocí Fourierovy transformace. Opakování adeninu v motivu $(CT)_5A$ je pro spektrum nevýznamné a program nenachází repetici s motivem 11 ale pouze 2. Každý jedenáctý adenin je tedy brán jako bodová mutace. Začátek první repetice je označen až o 18 nukleotidů později, druhá repetice je označena naprosto přesně. Následující tabulky (Tabulka 11 a Tabulka 12Tabulka 9) ukazují dosažené výsledky a přesný popis testované sekvence.

Tabulka 11: Výsledek vyhledávání v sekvenci TEST3.fasta

Číslo repetice	Začátek	Konec	Celková délka	Délka motivu	Počet opakování
1	1352	2387	1035	2	518
2	3075	4105	1030	2	515

Tabulka 12: Přesný popis sekvence TEST3.fasta

Soubor [*.fasta]	Celková délka sekvence	Repetiční motiv	Pozice satelitů	Počet opakování	Počet indelů	Počet substitucí
TEST3	4855	$(CT)_5A$	[1334:2386]	96	20	22
		$(CT)_5A$	[3075:4105]	96	85	46

6.2.4 Sekvence TEST4.fasta

V této sekvenci je první lehce mutovaná sekvence označena pouze s odchylkou dvou nukleotidů na začátku. U druhé repetice je začátek zpožděn o 9 nukleotidů a konec označen předčasně o 7 nukleotidů, což je stále velice dobrý výsledek. Třetí vysoce zmutovaná sekvence již rozdělená na tři kratší sekvence, jejichž spojením by se program správnému výsledku velice dobře přiblížil. Následující tabulky (Tabulka 13 a Tabulka 14Tabulka 9) ukazují dosažené výsledky a přesný popis testované sekvence.

Tabulka 13: Výsledek vyhledávání v sekvenci TEST3.fasta

Číslo repetice	Začátek	Konec	Celková délka	Délka motivu	Počet opakování
1	1253	2506	1253	5	251
2	3181	4421	1240	5	248
3	5285	6070	785	5	157
4	6166	6215	49	5	10
5	6302	6567	265	5	53

Tabulka 14: Přesný popis sekvence TEST4.fasta

Soubor [*.fasta]	Celková délka sekvence	Repetiční motiv	Pozice satelitů	Počet opakování	Počet indelů	Počet substitucí
TEST4	7574	CCTGA	[1251:2506]	256	110	75
		CCTGA	[3172:4428]	256	225	80
		CCTGA	[5276:6581]	256	325	150

6.2.5 Sekvence vrrA_b_anthracis.fasta

Při tomto testu opět vlastní algoritmus bohužel selhává a žádný 12-nukleotid neoznačuje za tandemovou repetici. Pouze repetice číslo 3 velice okrajově odpovídá hledané repetici, ale nelze ji označit za správný výsledek.

Tabulka 15: Výsledek vyhledávání v sekvenci vrrA_b_anthracis.fasta

Číslo repetice	Začátek	Konec	Celková délka	Délka motivu	Počet opakování
1	624	670	46	50	1
2	195	216	21	3	7
3	207	286	79	3	26
4	378	402	24	3	8
5	199	382	183	3	61
6	673	761	88	2	44

6.2.6 Sekvence HCH12ATP.fasta

Při testování sekvence je výsledkem 69 různých repetitivních sekvencí. Hledané trinukleotidy CAG/CAA. Jako hledaný výsledek může být označena repetice 36, která i při náhledu na její sekvenci ve vytvořeném uživatelském prostředí odpovídá očekávanému výsledku. Stejně jako u ostatních programů musí uživatel projít všechny výsledky sám a není možné samostatné označení pouze očekávané repetice.

Číslo repetice	Začátek	Konec	Celková dél...	Délka motivu	Počet opak...	Poměr Ade...	Poměr Cyt...	Poměr Gua...
27	365	438	73	4	18	1	1	1
28	13080	13104	24	4	6	1	1	1
29	2343	2364	21	4	5	1	2	0
30	11887	11991	104	3	35	1	1	1
31	12248	12315	67	3	22	1	1	1
32	12927	13000	73	3	24	1	1	1
33	14632	14702	70	3	23	1	1	0
34	3783	3849	66	3	22	1	1	1
35	11893	11991	98	3	33	1	1	0
36	12243	12316	73	3	24	1	1	1
37	14655	14693	38	3	13	1	1	0
38	17013	17089	76	3	25	1	1	1
39	3799	3859	60	3	20	0	2	1
40	12240	12315	75	3	25	1	1	1
41	16027	16123	96	3	32	0	1	1
42	16666	16708	42	3	14	1	0	1
43	10967	10992	25	3	8	1	1	1
44	13662	13690	28	3	9	0	1	1

Obrázek 28: Výsledek vyhledávání v sekvenci HCH12ATP.fasta

6.2.7 Sekvence PF3D7CH4.fasta

Zátěžovým testem prošel program bez potíží, ale nachází 17562 repetitivních oblastí. Mnohem více než by měl být reálný počet výsledků. Vysoký počet výsledků může být způsoben opakovaným označením některých repetic, akorát s jinou délkou motivu. Program zvládá výpočet této sekvence dlouhé 1,2 Mbp za 28 minut.

Číslo repetice	Začátek	Konec	Celková dél...	Délka motivu	Počet opak...	Poměr Ade...	Poměr Cyt...	Poměr Gua...	Poměr Tym...
17543	1184820	1184984	164	2	82	1	0	0	1
17544	1185180	1185235	55	2	28	1	0	0	1
17545	1185373	1185413	40	2	20	1	0	0	1
17546	1186803	1186828	25	2	13	1	0	0	1
17547	1193099	1193136	37	2	19	1	0	1	1
17548	1193604	1193744	140	2	70	1	0	0	1
17549	1193947	1194213	266	2	133	1	0	0	1
17550	1194313	1194485	172	2	86	1	0	0	1
17551	1194574	1194617	43	2	22	0	0	0	1
17552	1194769	1194858	89	2	45	1	0	0	1
17553	1196158	1196238	80	2	40	1	0	0	1
17554	1196587	1196683	96	2	48	1	0	0	1
17555	1196902	1196980	78	2	39	1	1	0	0
17556	1197098	1197138	40	2	20	1	1	0	1
17557	1197693	1197725	32	2	16	1	0	0	1
17558	1199588	1199621	33	2	17	1	0	0	1
17559	1200004	1200064	60	2	30	1	1	0	0
17560	1201241	1201413	172	2	86	1	0	1	0
17561	1201788	1201824	36	2	18	1	0	1	1
17562	1202717	1203398	681	2	341	1	0	0	1

Obrázek 29: Výsledek vyhledávání v sekvenci PF3D7CH4.fasta

7. Zhodnocení výsledků testování

Při postupném testování různých sekvencí se ukázaly jednotlivé výhody a nevýhody navrženého algoritmu.

Na rozdíl od volně přístupných testovaných programů, nabízí `GUI.m` a `Vyhledavac.m` výhodu ve zobrazení spektrogramu signálu, což je dobrá pomůcka pro vizualizaci obsahu sekvence. Díky metodě výpočtu je navíc možné vyhledat repetice jen na vybraných frekvencích, tzn. o předem stanovených délkách motivu. Tyto možnosti žádný z ostatních programů neposkytuje. Při vyhledávání repetice, kterou uživatel předem zná, jen potřebuje zjistit její polohu, případně jiné vlastnosti je metoda časově úspornější a výsledky jsou mnohem přehlednější díky menšímu počtu výstupních dat.

I když se metoda vyhledávání z numericky reprezentované DNA zdá výpočetně a časově náročnější, než vyhledávání ze sekvence znaků, při zátěžovém testu si vlastní program vedl mnohem lépe než Phobos a výsledek poskytl již po 28 minutách oproti asi 4 hodinám u Phobosu.

Při testování uměle vytvořených sekvencí, jsou výsledky algoritmu velice dobré. První nedostatek se objevil při testování sekvence TEST3.fasta, kdy program nebyl schopen nalézt repetici s motivem $(CT)_5A$ a označil pouze repetici s motivem dinukleotidu CT. Druhý nedostatek se objevil při testování sekvence TEST4.fasta, kdy program správně určil první dvě mutované TR, třetí repetitivní sekvenci ale rozdělil do tří samostatných.

Výsledky testování jsou velice dobré a dají se přirovnat k testování programů TRF a Phobos. Následující tabulka shrnuje výhody a nevýhody vlastního algoritmu oproti ostatním testovaným.

Tabulka 16: Výhody a nevýhody vlastního programu

Srovnávaný program	**Vlastní algoritmus**
IMEx	Výhody: Vyhledává i delší repetice než s motivem hexanukleotidy. Nerozděluje repetitivní oblasti tak často jako IMEx. Přesnější výsledky Nevýhody: Časová a výpočetní náročnost.
TRF	Výhody: Žádné, výsledky jsou si velice podobné. Nevýhody: Časová a výpočetní náročnost.
Phobos	Výhody: Mnohem lepší výsledek při testování sekvence TEST4.fasta. Přehlednější výsledky. Časově úspornější. Nevýhody: Méně přesný.

Závěr

Cílem práce bylo vybrat tři volně dostupné vyhledávače tandemových repetic, každý pracující na jiném algoritmu a otestovat jejich funkčnost a vhodnost a dále pak navrhnout v libovolném programovacím jazyce algoritmus, který zvládne také vyhledat tandemové repetice, otestovat ho na stejném souboru dat jako volně přístupné programy a srovnat výsledky.

K testování byly vytvořeny čtyři sekvence s uměle vloženými tandemovými repeticemi na známých místech a se známými motivy. Tyto testovací sekvence se lišili délkou, počtem minisatelitů rozprostřených v sekvenci, délkou motivu a počtem opakování motivu. Pro skutečné zhodnocení výkonnosti jednotlivých programů byly sekvence podrobeny předem určenému počtu bodových mutací (inzercí, delecí a substitucí). Každá uměle vytvořená testovací sekvence se od sebe jistými vlastnostmi lišila tak, aby bylo otestováno co nejvíce vlastností programů. Z reálných dostupných dat byly vybrány tři sekvence. První sekvence vrrA je protein kódující gen druhu *Bacillus cereus biovar anthracis*. Tato sekvence byla vybrána, aby otestovala základní vlastnosti programu. Testovalo se, zda je schopný v krátké sekvenci najít několik jednoduchých opakování 12-nukleotidu. Sekvence v podstatě sloužila jako kontrola, jestli se nevyskytla chyba někde při tvorbě sekvencí umělých. Druhá sekvence, je část lidského chromozomu 12, která kóduje gen pro atropin a ve které se vyskytuje známá repetice. Jako třetí sekvence byl vybrán celý 4 chromozom druhu *Plasmodium falciparum 3D7* pro otestování rychlosti a efektivity algoritmů vybraných programů.

Tři vybrané programy: Tandem Repeats Finder, IMEx: Imperfect Microsatellite Extractor a Phobos se podrobily testování na vybraných sekvencích. Výsledky programu IMEx byly bohužel z velké části nedostatečné, ale v jednodušších testech si program vedl obstojně. Programy Phobos a Tandem Repeats Finder mají výpočetně náročnější algoritmy a jsou schopny vyhledávat mnohem delší motivy v sekvencích. Výsledky testování byly u těchto programů velice podobné. Největším rozdílem byla časová náročnost programu, kdy programu TRF stačilo pro analýzu sekvence o 1,2 Mbp několik desítek vteřin, zatímco Phobos prohledával data několik hodin a i tak se dostal pouze k motivům o maximální délce 250 bp, zatímco TRF testoval až do délky 500 bp. Pro zlepšení přehlednosti výsledků je v práci uvedeno několik tabulek a ukázky výstupů programů.

Vlastní program se zdá mnohem lepší než program IMEx. Výsledky testování jsou srovnatelné s výsledky programů TRF a Phobos. Každý algoritmus má své výhody a nevýhody, které jsou zhodnoceny v tabulce Tabulka 16. Navíc byl vlastní program doplněn o jednoduché uživatelské prostředí, ve kterém se dají snadno změnit parametry vyhledávání a které poskytuje mnohem přehlednější výsledky, které se dají navíc uložit do *XLS* souboru. GUI.m dělá navržený program více uživatelsky přátelský.

Seznam použitých zdrojů

[1] ARNIKER, S. B., Human Promoter Prediction Using DNA Numerical Representation. 2010. Electronic Theses and Disertations. University of Windsor

[2] BENSON, G., Tandem repeats finder: a program to analyze DNA sequences. Oxford University Press 1999, Vol. 27, No. 2, p. 573-580

[3] CRISTEA, P. D., Representation and analysis of DNA sequences. University Politechnica of Bucharest

[4] FONDON, J. W., GARNER H. R., Molecular origins of rapid and continuous morphological evolution. 2004 PNAS vol. 101 no. 52

[5] GOLDING B. Eukaryotic Genetics, Computational Biology at McMaster University, Hamilton, Ontario [cit. 29. září 2014]. Dostupné z URL: < http://helix.biology.mcmaster.ca/3I03.pdf>

[6] HARTL, D., JONES, E., Genetics: Principles and analysis 4th edition, Jones and Bartlett Publishers, Sudbury, Maschusetts, 1998, 1367s., ISBN 0-7637-0489-X

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

[8] KRISHNAN, A., TANG, F., Exhaustive Whole-Genome Tandem Repeats Search, Bionformatics Institute, Singapore

[9] LÓPEZ, M., PÉREZ, J. DNA Transposons: Nature and Applications in Genomics, Current Genomics, Bentham Science Publishers. Apr 2010;11(2): 115-128

[10] MAYER, C., Phobos: A tandem repeat search program, 2010. Dostupné z URL <http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm>

[11] McDONALD, D., Lecture 8. Population Genetics VI: Introduction to microsatellites: from theory to lab. practice, Dept. Zoology & Phsiology, University of Wyoming. [cit. 29. září 2014] Dosupné z URL < http://www.uwyo.edu/dbmcd/molmark/lect08/lect8.html>

[12] McMURRAY, C., Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet. Nov 2010; 11(11): 786-799

[13] MIRKIN, S., DNA structures, repeat expansions and human hereditary disorders. Current Opinion in Structural Biology 2006, 16: 1-8

[14] MITAS, M., Trinucleotide repeats associated with human disease, Department of Biochemistry and Molecular Biology, Oklahoma State University, Nucl. Acids. Res. (1997) 25 (12): 2245-2253

[15] MUDUNURI, S. B., NAGARAJARAM, H. A., IMEx: Imperfect Microsatellite Extractor. Bioinformatics Vol. 23 no. 10 2007, p. 1181-1187

[16] MYERS, P. Z., Ph.D., Tandem Repeats and Morphological Variation. 2007 Nature Education 1(1):1

[17] PRAY, L, Ph.D. Transposons: The jumping genes. Nature Education 1(1):204, 2008

[18] RIDZOŇ, P., MUDr., Bulbospinální svalová atrofie (Kennedyho nemoc), Neurologie pro praxi, 2006; 1:27-28

[19] SMIT, A. The origin of interspersed repeats in the human genome. Department of Molecular Biology, University of Washington, Seattle, USA. 1997

[20] SHAW, MA., CHIURAZZI, P., A novele gene FAM11A, associated with the FRAXF CpG islands is transcriptionally silent in FRAXF full mutation. Eur J Hum Genet, 2002 Nov; 10(11): 767-772

[21] ŠEDA, O., MUDr., PhD., LIŠKA, F., MUDr., PhD., ŠEDOVÁ, L., PharmDr., PhD., Aktuální genetika – multimediální učebnice lékařské biologie, genetiky a genomiky, Ústav biologie a lékařské genetiky 1. LF UK a VFN, Praha 2005 – 2006, [cit. 29. září 2014]. Dostupné z URL: < http://biol.lf1.cuni.cz/ucebnice/repetitivni_dna.htm>

[22] TAMARIN, R. Principles of Genetics 7th edition, Tata McGraw – Hill Education, 2004, 696s., ISBN 978-00-712-4320-9

[23] TUNTIWECHAPIKUL, W., SALAZAR, M., Mechanism of in Vitro Expansion of Long DNA Repeats. Biochemistry 2002, 41, 854-860

[24] VERGNAUD, G., DENOEUD, F., Minisatellites: Mutability and Genome Architecture, Institut de Génétique et Microbiologie, Université Paris Sud, Genome Research 2000. 10: 899-907

[25] VUT BRNO, FEKT, ÚBMI, studijní materiály k předmětu Bioinformatika (2011 – 2012), garant předmětu: PROVAZNÍK, I., prof. Ing., PhD.

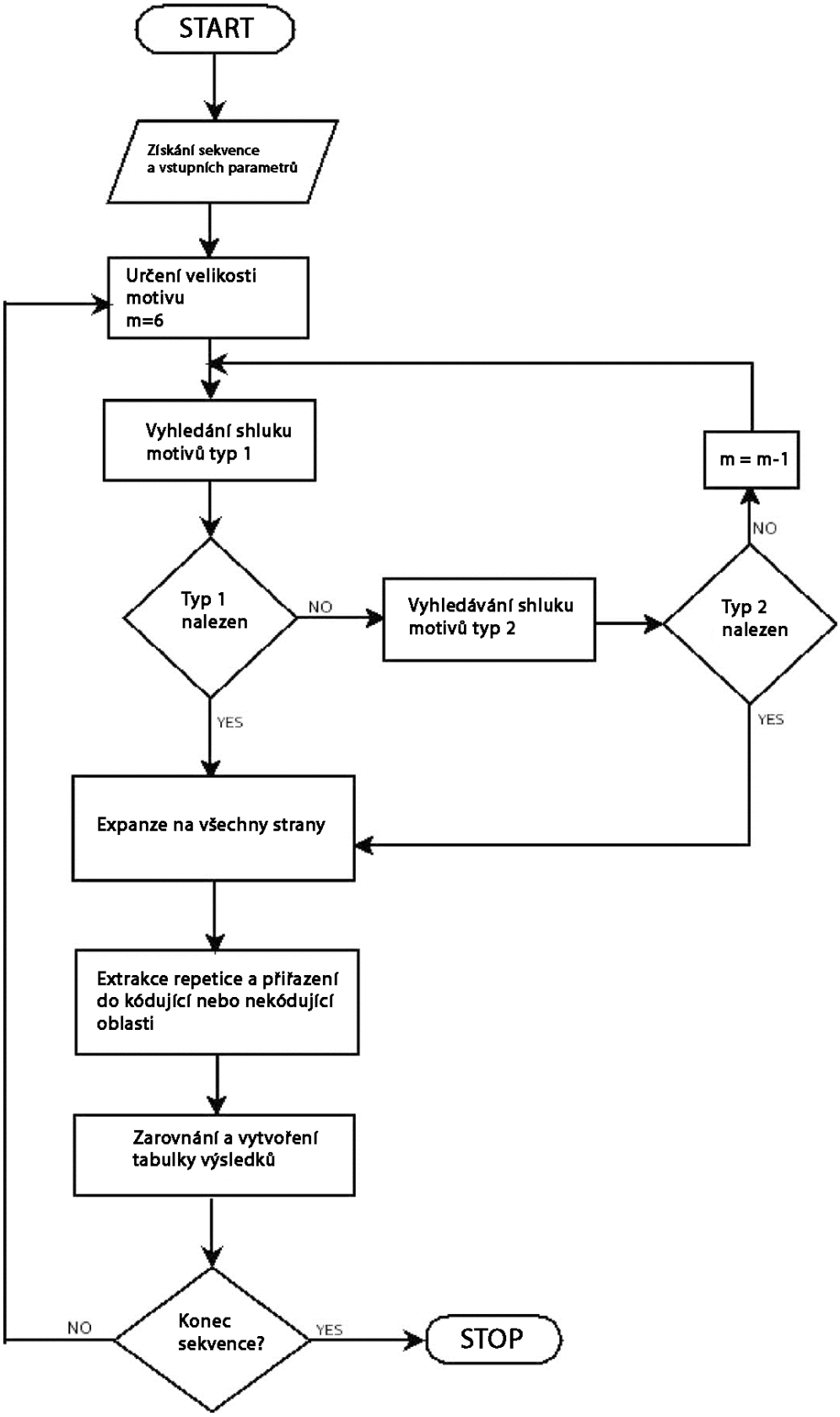
Seznam použitých zkratek a symbolů

A	adenin
bp	base pair, česky: pár bází; jednotka délky sekvence
ATR	**A**pproximate **T**andem **R**epeat, česky: přibližná tandemová repetice
C	cytosin
DNA	deoxyribonukleová kyselina
DFT	**D**iscrete **F**ourier **T**ransform
DFT	**D**iscrete **T**ime **F**ourier **T**ransform
G	guanin
HIV	Human Immunodeficiency Virus
LINE	**L**ong **I**nterspersed **N**uclear **E**lements, česky: dlouhé rozptýlené jaderné elementy
LTR	**L**ong **T**erminal **R**epeats, česky: dlouhé ukončující repetice
ORF	**O**pen **R**eading **F**rame, česky: otevřený čtecí rámec
PCR	**P**olymerase **C**hain **R**eaction – polymerázová řetězová reakce
SINE	**S**hort **I**nterspersed **N**uclear **E**lements, česky: krátké rozptýlené jaderné elementy
SSTR	**S**imple **S**equence **T**andem **R**epeats, česky: tandemové repetice jednoduchých sekvencí
STFT	**S**hort **T**ime **F**ourier **T**ransform
T	tymin
TE	transpozibilní elementy
TNR	**T**ri **N**ucleotide **R**epeats, česky: trinukleotidové repetice
TRF	**T**andem **R**epeats **F**inder
VNTR	**V**ariable **N**umber **T**andem **R**epeats, česky: tandemové repetice s proměnným počtem opakování

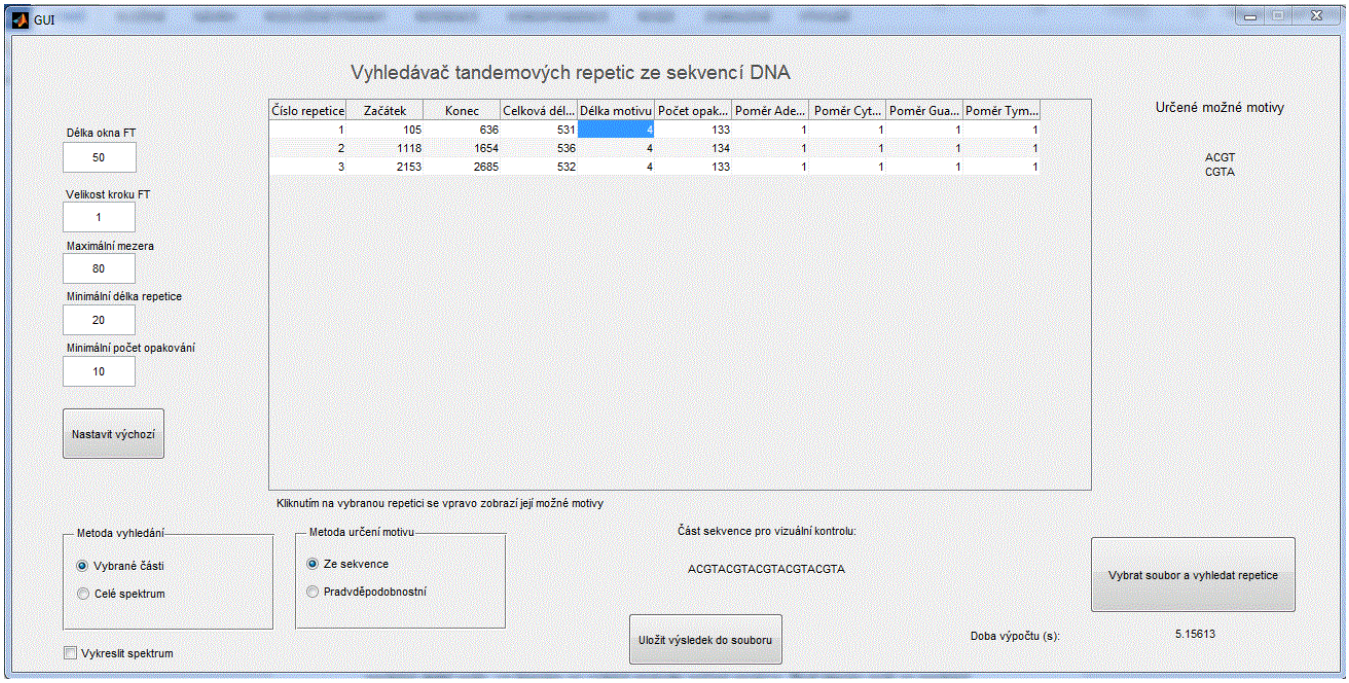
Seznam příloh

Přílohy

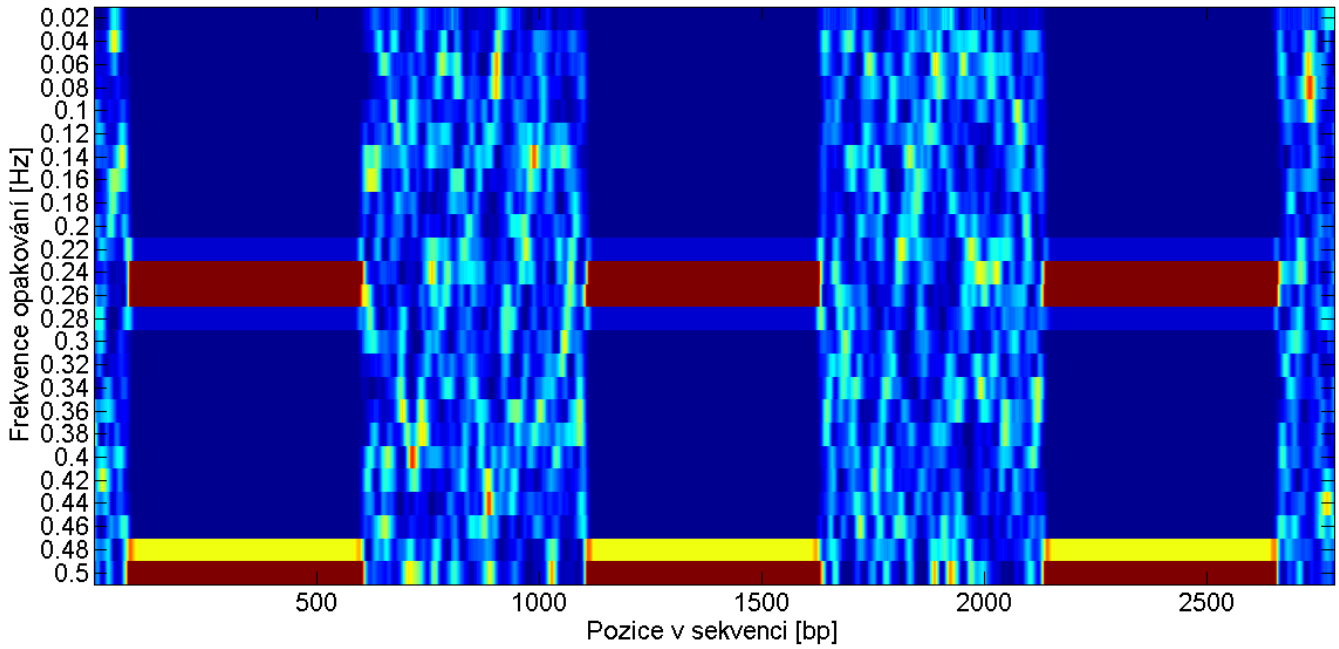
Příloha 1: Blokové schéma programu IMEx [15]



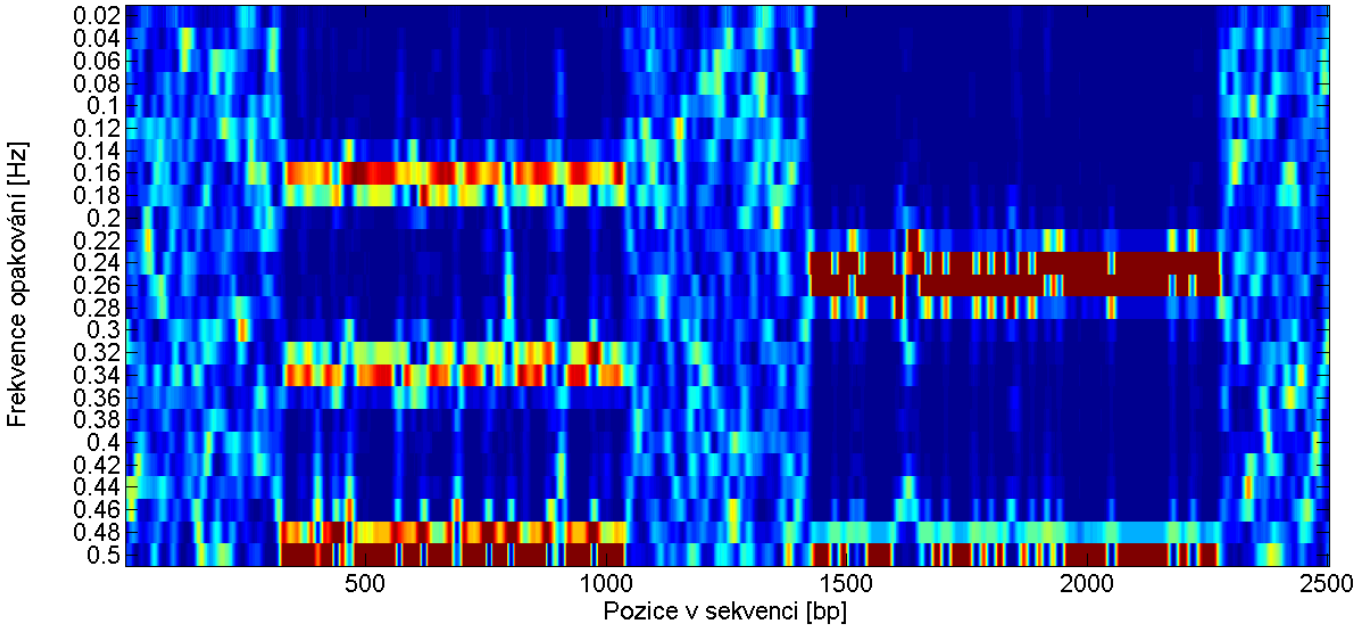
Příloha 2: Uživatelské prostředí



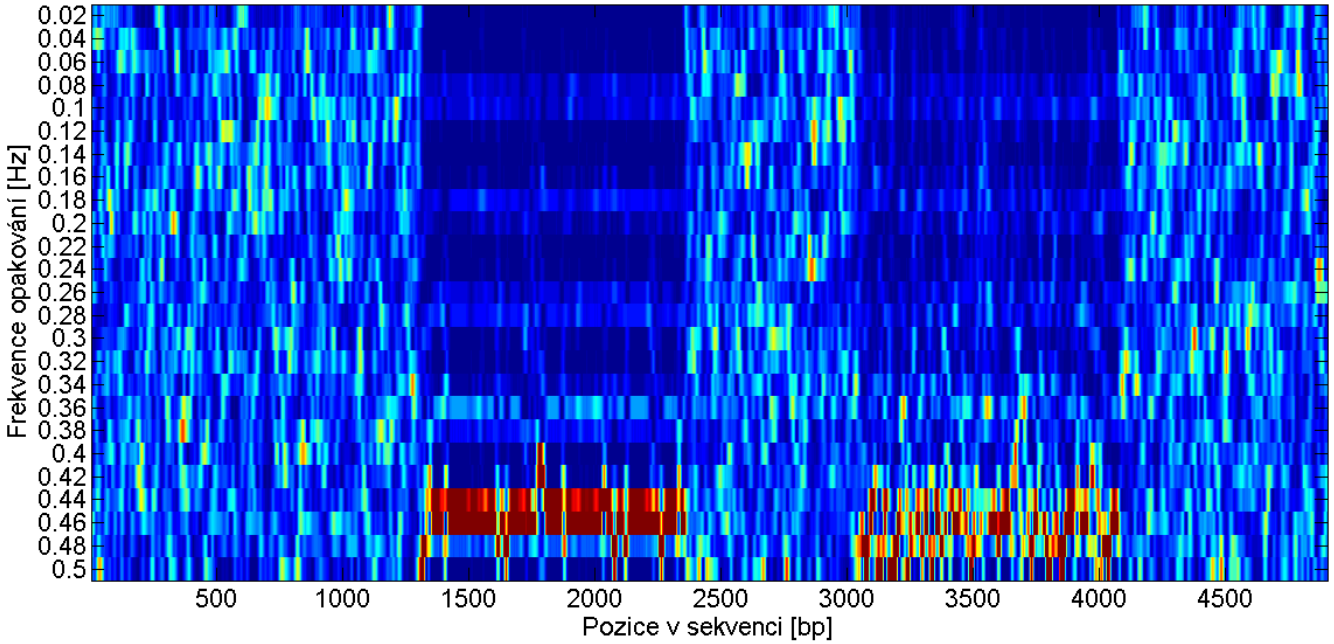
Příloha 3: Spektrogram sekvence TEST1



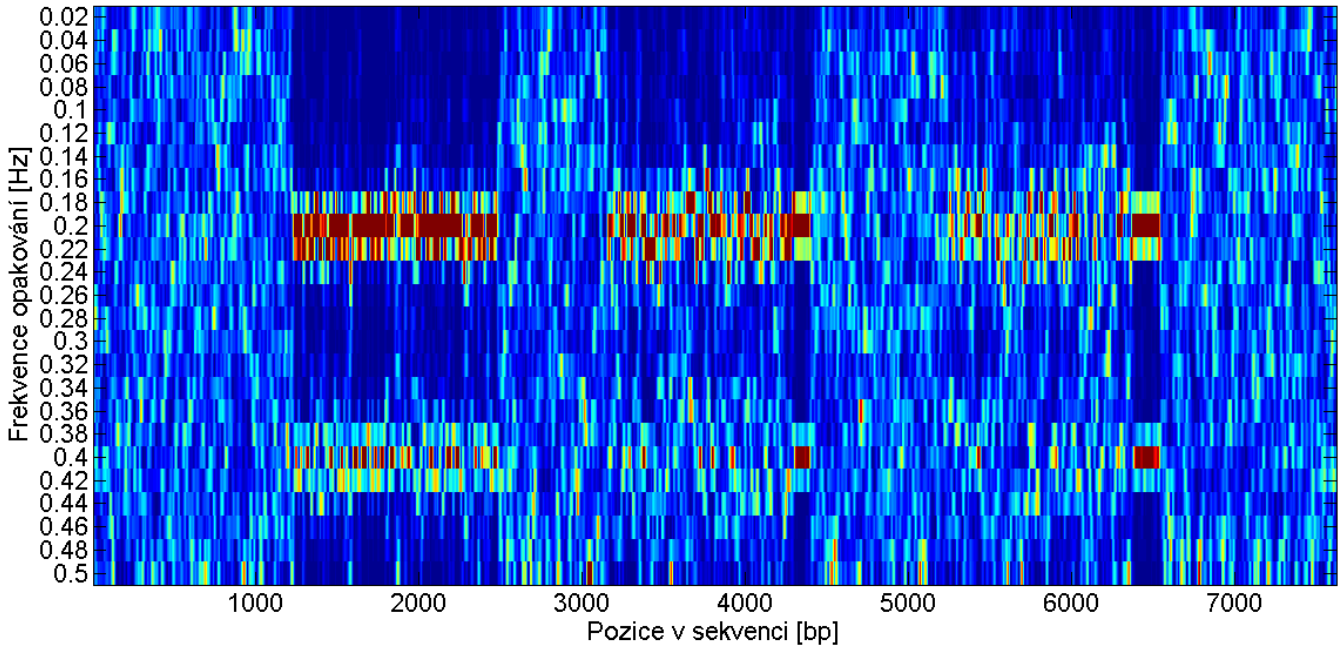
Příloha 4: Spektrogram sekvence TEST2



Příloha 5: Spektrogram sekvence TEST3



Příloha 6: Spektrogram sekvence TEST4



Obsah přiloženého CD

1. Textová část práce *Kryštof_Havlík_BP.pdf*
2. Programová část práce *Kryštof_Havlík_BP_přílohy.pdf*
    a. Kódy v Matlabu
    b. Testované DNA sekvence ve *FASTA* formátu