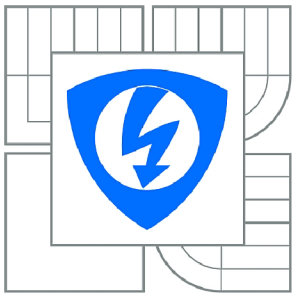
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

BIOINFORMATIKA LIDSKÉHO GENOMU

BIOINFORMATICS OF HUMAN GENOME

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — KAROLÍNA LANKAŠOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2011

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Student: Karolína Lankašová ***ID:*** 115103
Ročník: 3 ***Akademický rok:*** 2010/11

NÁZEV TÉMATU:

Bioinformatika lidského genomu

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se se stavbou DNA a projektem lidského genomu. Blíže prostudujte literaturu zaměřenou na oblasti bohaté na guanin a adenin a zjistěte, ve kterých oblastech genomu se nacházejí. Navrhněte metodu vyhledávání požadovaných sekvencí z genomu.
Prohledejte lidský genom a nalezněte úseky, ve kterých se vyskytuje pouze guanin a adenin. Porovnejte rozsah sekvencí v kódující a nekódující oblasti. Porovnejte výsledky z lidského genomu s evolučně blízkými živočichy a zvažte náhodnost, funkci a konzervovanost těchto úseků.

DOPORUČENÁ LITERATURA:

[1] HRABCOVA, I., KYPR, J., The longest (A+T) and (G+C) blocks in the human and other genomes, Journal of Biomolecular Structure & Dynamics, 2008, vol. 25, pp. 337-345.
[2] MOUNT, D.W. Bioinformatics. Cold Spring Harbour, New York, 2004.

Termín zadání: 7.2.2011 ***Termín odevzdání:*** 27.5.2011

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce: RNDr. Jaroslav Kypr, CSc.

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

UPOZORNĚNÍ:

ABSTRAKT

Bakalářská práce se zabývá úseky DNA obsahujícími pouze adenin a guanin. V teoretické části je popsána struktura a složení deoxyribonukleové kyseliny, chromosomů a genů. Jsou zde uvedeny základní informace o lidském a šimpanzím genomu a o konformacích řetězců obsahujících adenin s guaninem. Praktickou část tvoří program, který vyhledá požadované úseky v sekvencích, zobrazí je a uloží. Součástí práce je analýza genů společných pro člověka a šimpanze, které byly zkoumány pro stanovení náhodnosti, funkčnosti a konzervovanosti těchto úseků.

KLÍČOVÁ SLOVA

DNA, gen, genom, adenin, guanin, intron, exon, sekvence, FASTA, člověk, šimpanz.

ABSTRACT

Bachelor thesis deals with DNA segments containing only adenine and guanine. The theoretical part describes the structure and composition of deoxyribonucleic acid, chromosomes and genes. There are basic informations about chimpanzees and the human genome and conformations of the chains containing adenine with guanine. The practical part consists of a program that searches for the required sections in sequence, it displays and saves it. The thesis includes analysis of genes which are common for humans and chimpanzees, which were analyzed to determine randomness, functionality a preserving of these sections.

KEYWORDS

DNA, gene, genome, adenine, guanine, intron, exon, sequence, FASTA, human, chimpanzee.

LANKAŠOVÁ, Karolína *Bioinformatika lidského genomu*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2011. 54 s. Vedoucí práce byl Ing. Denisa Maděránková

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Bioinformatika lidského genomu“ jsem vypracovala samostatně pod vedením vedoucí 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 neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji své vedoucí bakalářské práce Ing. Denise Maděránkové za příkladné vedení, cenné rady a trpělivost a svému konzultantovi RNDr. Jaroslavu Kyprovi, CSc., za inspirující zadání a náměty, bez kterých by tato práce nemohla vzniknout. Práce byla provedena v rámci projektu IAA500040903 poskytnutému Grantovou agenturou Akademie věd ČR.

V Brně dne .....

.....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Od objevu šimpanzů a goril západní civilizací lidé hledají jejich podobnosti i odlišnosti s člověkem. Snaha určit rozdíly a jejich příčiny s rozvojem metod molekulární biologie přerostla v potřebu odlišnosti pojmenovat, zmapovat a kvantifikovat. Mnoho lidí je přesvědčeno, že stanovením rozdílů mezi genomy člověka a ostatních živočišných druhů objevíme základ lidství [1].

Je třeba mít neustále na paměti, že z podstaty své struktury se žádné dva genomy nemohou mezi sebou lišit o více než 75%. Sdílení 25% podobnosti genetické informace člověka s leknínem však nutně neznamená, že jsme z jedné čtvrtiny květiny. Člověka odlišuje od zbytku živé přírody rozdílnost ekologická, demogafická, anatomická a mentální [1].

Cílem předložené práce je nalézt v genomu člověka úseky obsahující pouze adenin a guanin a porovnat jejich rozsah v kódující a nekódující oblasti. Srovnáním s úseky v genomu šimpanze pak zvážit jejich náhodnost, funkčnost a konzervovanost.

První kapitola se věnuje teoretickému úvodu a principům nutným k pochopení bakalářské práce. Popisuje složení a strukturu DNA, její typy, možné konformace, kondenzaci do vyšších struktur. Zabývá se geny, jejich stavbou a umístěním. Dále obsahuje informace o lidském genomu, jeho stavbě a projektu, jehož cílem bylo genom člověka osekvenovat. Uvádí podobnost s šimpanzem a gorilou, popisuje šimpanzí genom. Další podkapitoly se zabývají konformacemi, které nabývá DNA obsahující oblasti bohaté na adenin a guanin, programovým prostředím MATLAB a formátem FASTA.

Druhá kapitola obsahuje praktickou část bakalářské práce. Je zde nastíněn algoritmus vytvořeného programu `Bakalarka.m`, který v sekvencích ve FASTA formátu nebo v náhodně vygenerovaných sekvencích vyhledává oblasti, které jsou tvořeny pouze adeninem s guaninem a jsou delší než 9 nt. Program může oblasti sečíst, rozdělit do skupin podle délky a vykreslit histogram. Úseky lze uložit do textového souboru pro další zpracování dat. V programu je možné si sekvence procházet po nukleotidech nebo nalezených úsecích. Druhou podkapitolu tvoří analýza genů, které byly vybrány z lidského a šimpanzího genomu. Jsou zde uvedeny informace o genech, jejich analýza i zhodnocení výsledků.

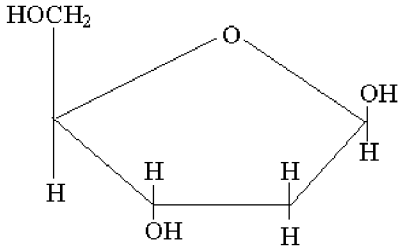
Práce je doplněna o přílohy obsahující uživatelský manuál k vytvořenému programu a přehledné výsledky analýzy genových dat.

1 TEORETICKÁ ČÁST

1.1 DNA

Kyselina deoxyribonukleová (DNA) hraje v každém živém organismu nezastupitelnou roli buněčné paměti. Uchovává, přijímá a uvolňuje veškeré informace o struktuře, funkci a reprodukci buněk. U mnohobuněčných organismů zapříčiňuje chování jedince jako celku. Jedná se o úlohu pasivní, přesto klíčovou. Tvoří totiž genetickou výbavu drtivé většiny živé přírody. Pokud nedojde k mutacím, struktura se v průběhu života nemění a je předávána dalším generacím [2], [3].

Jako nukleová kyselina je tvořena řetězcem nukleotidů (vláknem, pentlicí) složených ze tří částí – báze, cukru a fosfátu. Organické dusíkaté zásady (báze) dělíme na puriny, mezi které patří adenin (6-aminopurin) a guanin (2-amino-6-oxopurin) a pyrimidiny, zahrnující cytosin (6-amino-2-oxopyrimidin) s thyminem (2,6-dioxo-5-methylpyrimidin). Zbytek kyseliny fosforečné se estericky váže na pentózu 2-deoxy-D-ribosu (obr. 1.1) a způsobuje kyselý charakter celé molekuly. Pokud je z nukleotidu odštěpena, nazýváme jej nukleozid [2].

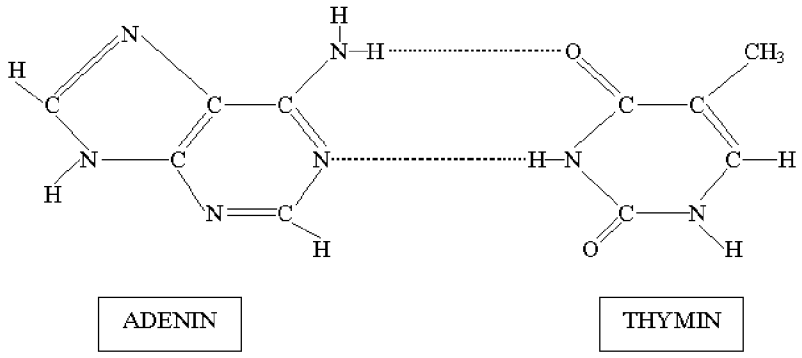


Obr. 1.1: 2-deoxy-D-ribosa [4].

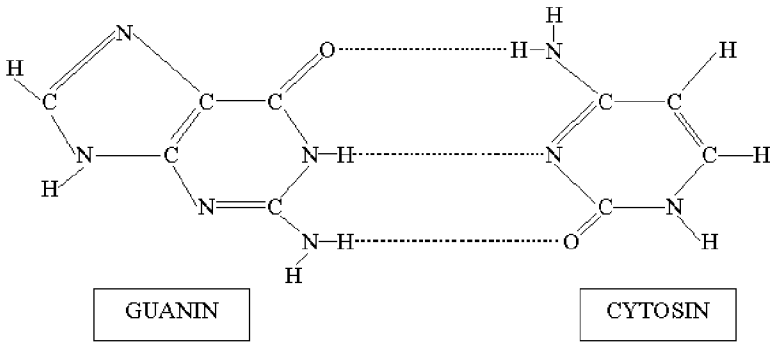
Primární struktura DNA vzniká kovalentními fosfodiesterovými vazbami mezi jednotlivými nukleotidy, konkrétně mezi hydroxyly v poloze 3' jedné a 5' druhé pentosy. Na této cukr-fosfátové kostře je β-N-glykosidovou vazbou poutána na pentosu báze. Řetězec je polarizován, má volnou OH-skupinu pentosy na svém 3' konci a fosfátový zbytek, nebo navíc také volnou OH-skupinu na 5' konci. Tato polarita je důležitá při prodlužování (elongaci) řetězce, které vždy probíhá ve směru od 5' ke 3' konci. V buňce, prostředí s pH neutrálním, se nukleotidový řetězec ionizuje na fosfátových zbytcích. Vzájemným odpuzováním záporných nábojů je molekula stabilizována [5]. Délku řetězce odvozujeme od počtu nukleotidů. Pro pojmenování používáme řecké číslovky – mononukleotid, dinukleotid atd. Jako oligonukleotidy

označujeme řetězce dlouhé do několika desítek nukleotidů, polynukleotidy používáme pro sto a více [3].

Sekundární strukturu DNA formují vodíkové můstky mezi bázemi, částečně van der Waalsovy interakce. Díky nim vzniká dvouvláknová DNA tvořená dvěma antiparalelními řetězci. Na obou koncích molekuly je 3' konec jednoho a 5' konec druhého řetězce. Adenin se páruje s thyminem dvěma vodíkovými můstky (obr. 1.2), guanin s cytosinem třemi (obr. 1.3), díky tomu jsou poutány pevněji [5].



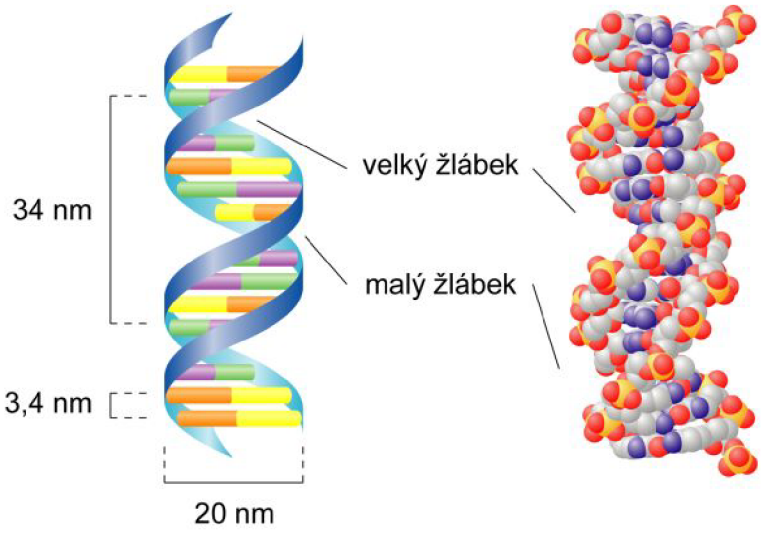
Obr. 1.2: Pár bazí adenin - thymin [4].



Obr. 1.3: Pár bazí guanin - cytosin [4].

Komplementarita, vazba pyrimidinové a purinové báze, je základním předpokladem pro expresi genetické informace – vodíkové můstky jsou dostatečně silné pro udržení řetězců, ale snadno oddělitelné při přenosu informace. Díky principu párování je v nukleové kyselině stejné zastoupení purinových a pyrimidinových bází [5]. Počet dvojic A-T a G-C je odlišný, poměr párů je proto konstantou typickou pro daný typ DNA [2].

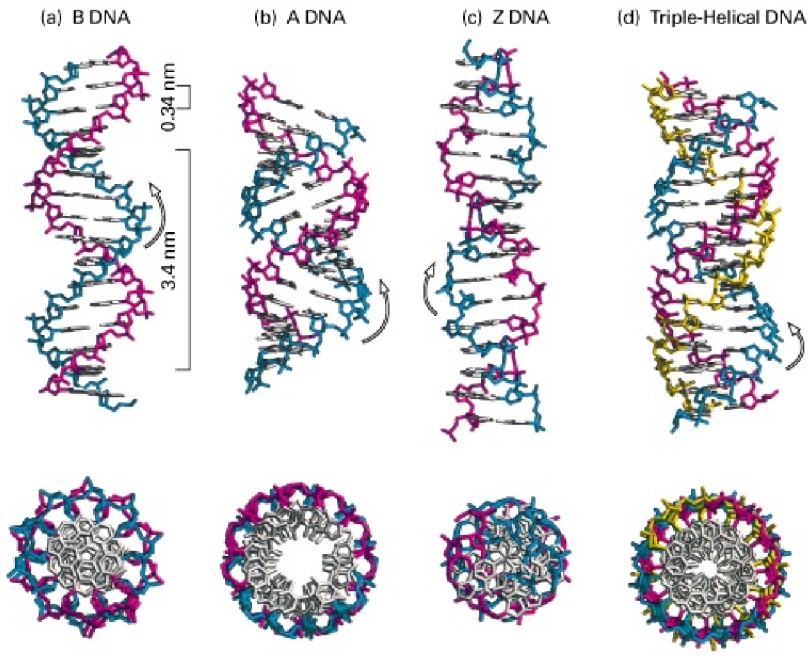
Rozeznáváme několik typů DNA – DNA jádra (DNA), DNA komplementární k RNA (cDNA), dvouřetězcovou DNA (double-stranded molecule, dsDNA, obr. 1.4), jednořetězcovou DNA (ssDNA) a mitochondriální (mtDNA). Jednotlivé typy jsou charakterizovány celkovým počtem, poměrem a sekvencí nukleotidů. Díky různé délce je možné je separovat na základě rozdílných sedimentačních rychlostí [5]. Počet nukleotidů (nt) či bází (b) nahrazujeme, pokud chceme zdůraznit dvouřetězcovou strukturu, počtem bázových párů (bp) [3].



Obr. 1.4: Struktura dsDNA [6].

Polynukleotidový řetězec nabývá různých konformací (tvarů, obr. 1.5). Základní struktury vznikají automaticky jako stav s minimální volnou energií. Jsou udržovány vnitromolekulárními silami, vodíkovými můstky a hydrofobními interakcemi [2]. Jaderná DNA se většinou vyskytuje v podobě dvoušroubovice. Za objev její struktury obdrželi James Watson a Francis Crick roku 1962 Nobelovu cenu za fyziologii a lékařství. Šroubovice v jádře eukaryot je pravotočivá, její cukrofosfátová kostra je hydrofilní, fosfátové zbytky se natáčejí ven do vodní fáze, báze se obracejí k její ose. Výše popsaná forma je jednou z šesti dosud známých forem dsDNA, konkrétně B-forma. Vyznačuje se antiparalelním vinutím řetězců s komplementárními páry umístěnými kolmo na osu šroubovice. Jednotlivé závity jsou od sebe vzdáleny 3,4 nm, na jedno otočení připadá 10 bp. Velký žlábek je přístupný pro navázání specifických látek. U pravotočivé A-formy připadá na otáčku 11 bp, závity jsou vzdáleny 2,6 nm. Páry nejsou kolmé na osu šroubovice, velký žlábek není přístupný bílkovinám a pentosa má jinou konfiguraci. Z-forma je levotočivá, navíjení komplementárních řetězců není plynulé, jednu otočku tvoří 12 bp a vzdálenost dvou závitů je 5,7 nm [5]. Vytváří se v úsecích, kde se pravidelně střídá purin s pyrimidinem (GCGCGCGC) [2]. Další

konformace nebyly in vivo prokázány. In vitro existují i C, D a E DNA, liší se počtem párů bazí na závit a schopností vytvářet třířetězcové úseky [3].



Obr. 1.5: B, A, Z a trojřetězcová DNA [7].

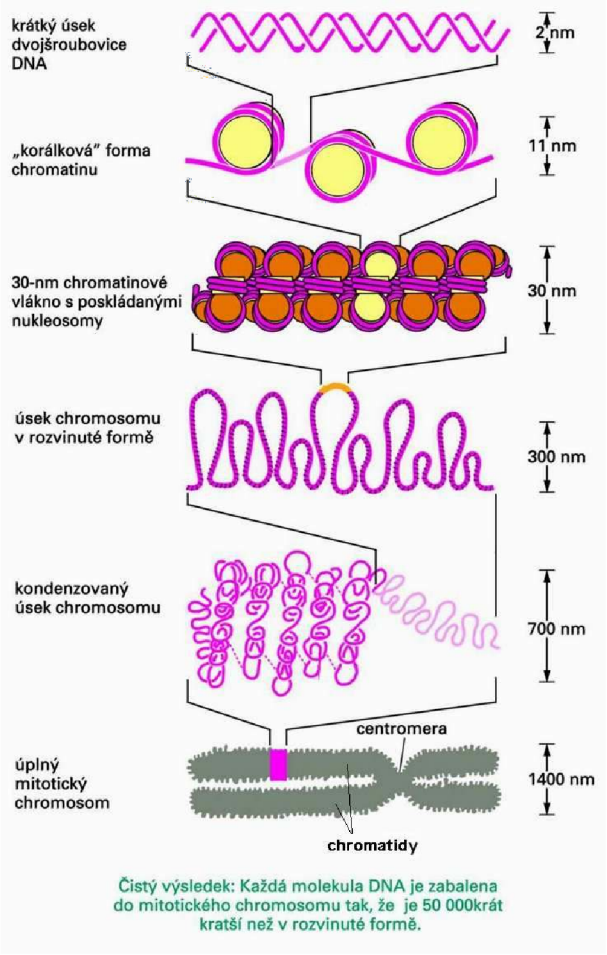
DNA má choulostivou stukturu. Vazby mezi bázemi naruší extrémní hodnoty pH, močovina a teplota nad 60 °C. Dvoušroubovice se rozdělí na řetězce – denaturuje. Rychlost reakce závisí na poměru jednotlivých nukleotidů (počtu vodíkových můstků mezi bázemi) a na iontové síle prostředí. Může se jednat o vratný děj, potom hovoříme o renaturaci. Při denaturaci stoupá absorpční maximum ultrafialového světla, při renaturaci naopak klesá zpět k 260 nm [5]. Přerušení obou vláken v jednom místě, depolymeraci (degradaci) způsobí větší příliv energie a působení enzymů, nukleáz, které narušují vazby mezi nukleotidy [2].

1.2 Chromosom

DNA je součástí jaderného chromatinu, který se dále skládá z bílkovin tvořících asi dvě třetiny objemu a nepatrného množství kyseliny ribonukleové (RNA) [5].

Genetický materiál by se díky svému rozsahu vzhledem k velikosti buněčného jádra do něj nevešel. Proto je DNA kondenzovaná, u eukaryotických buněk dokonce na několika stupních [5].

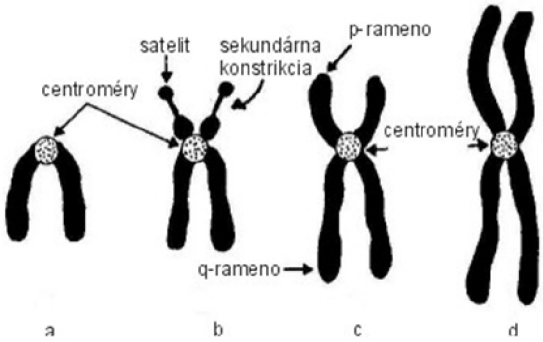
V chromosomu je DNA ve formě superhelixu (nadšroubovice) fixovaném koncem na proteinu. Druhý konec rotuje a tím šroubovici zhušťuje (mluvíme pak o tzv. pozitivním superhelixu) nebo rozvolňuje (negativní nadšroubovice) [5]. Délka lineární molekuly DNA v jednom lidském chromosomu dosahuje 1,7–8,5 cm, obsahuje 5–25 miliard bp [2].



Obr. 1.6: Kondenzace DNA [8].

Kondenzace (obr. 1.6) je způsobena ovinutím 1,7 závitu DNA (asi 146 nukleotidů) kolem osmi molekul bazických bílkovin s kladným nábojem umožňujícím pevnou vazbu s DNA, tzv. histonů (2xH2A, 2xH2B, 2xH3, 2xH4). Vznikají diskové útvary, tzv. nukleosomy [2], obsažené v haploidním genomu člověka v počtu sedmnácti milionů [5]. Mezi nukleosomy je vždy asi 50 nukleotidů spojníkové (volné) DNA. Histon H1 je umístěn vně. Spojuje histony tvořící nukleozomy s DNA v místě, kde přechází ve spojníkovou. Vzniká opakující se úsek dlouhý 200 nukleotidů. Množství histonů za sebou vytváří vlákna, která se dále spojují do tvaru solenoidu, na jehož jeden závit připadá 6 nukleozonů. Vzniká chromatinové vlákno (fibrila) s průměrem

30 nm. Vlákno obtáčí nosné bílkoviny a tvoří paprskovitě vybíhající kličky obsahující po 50–100 kbp. Takto vzniklý tyčinkovitý útvar nazýváme chromatidou. V mitóze dosahuje až 1,0 μm, protože vysoká kondenzovanost útvaru usnadňuje pohyb chromosomů v průběhu buněčného dělení. Hustota závitů v chromatidě není konstantní, liší se dle umístění nosných bílkovin. Řidší úseky nazýváme euchromatin, kondenzované heterochromatin. Euchromatin je schopný zahájit expresi genů a během mitózy je zhuštěný. Společně s fakultativním heterochromatinem obsahuje geny a je schopen je transkribovat. Konstitutivní heterochromatin je trvale kondenzovaný [3].



Obr. 1.7: Poloha centromery [9].

Chromosomy v jádře eukaryotických buněk mění svou morfologii v průběhu buněčného dělení díky kyselým bílkovinám. Jsou pozorovatelné během interfáze, kdy dochází k jejich dekondenzaci. Podle rozložení intenzity rozvolnění můžeme jednotlivé chromosomy rozpoznat. Liší se dále délkou obsažené DNA (sekundárně jde o velikost) a umístěním centroméry, úseku obsahujícím repetitivní alfa-satelitní sekvence (obr. 1.7). Centroméra může být umístěna po celé délce chromosomu, dělí jej na krátká a dlouhá raménka. Pro identifikaci chromosomů je označujeme čísly podle jejich relativní velikosti [10], raménka značíme q (dlouhá) a p (krátká). Úseky číslujeme od centromery směrem ke koncům ramen. Ramena jsou zakončena speciální sekvencí, tzv. telomérou [3].

Chromosomy obsahují skupiny genů, které jsou přenášeny společně. Umístění genů se může lišit v závislosti na typu živočicha. Geny, co do počtu, jsou v chromosomech rozděleny nerovnoměrně. Nejvíce genů se nachází v chromosomech 1, 11, 17, 19 a 22. Stejně tak jsou některé geny transkribovány mnohem častěji než jiné. Například informační obsah chromosomu X je jeden z nejmenších, ale bylo v něm zatím identifikováno nejvíce genů prokazatelně spojených s určitými chorobami [3].

U člověka nalezneme DNA mimo jádro ještě v mitochondriích ve formě dvoušroubovicového prstence. Je základem mitochondriální dědičnosti.

1.3 Gen

Základní jednotkou dědičnosti, schopnosti předávat genetickou informaci z mateřské generace na dceřinou, je gen. Jedná se o úsek DNA kódující enzym. Nachází se na specifické části chromosomu zvaný lokus. Uspořádání lokusů je specifické pro jednotlivé druhy. Výjimku tvoří skákavé geny bez stálého umístění - transposomy [5].

Dle velikosti se geny třídí na malé, středně velké, tvořící nejpočetnější skupinu, veliké a obrovské. Délka známých genů leží v intervalu ohraničeném známým nejkratším a nejdelším genem: (0,8–2300) kb [3].

Pro eukaryotické organismy jsou typické složené strukturní geny, skládající se z exonů, oblastí kódujících protein a intronů. Exony jsou velmi podobné i u vzdálených druhů, rozdíly intronových úseků jsou mnohem pestřejší. Jeden gen nemusí být nutně tvořen pouze jedním exonem. Před použitím jsou introny, které tvoří podstatně dlouhý úsek především velkých genů, odstraněny. Pokud úsek DNA nese informaci pro více proteinů, jedná se o překrývající se strukturní geny, jejichž část leží v oblasti sousedního genu [2].

Z celé savčí DNA podléhá transkripci asi desetina. Zbytek, tzv. negenová DNA, obsahuje úseky s regulační funkcí, kterou plní navázáním specifických proteinů a oblastí, o nichž doposud příliš nevíme. Mají podobu opakujících se (repetitivních) sekvencí, variabilních natolik, že mohou sloužit k odlišení jedinců [2]. Regulační oblasti mají různou délku a umístění vůči genu, na něhož mají pozitivní (tzv. enhancemery) nebo negativní (tzv. supresory) a odstupňovaný účinek. Některé jsou naprosto nezbytné, jedná se především o specifické sekvence umožňující začít a ukončit transkripci, jiné, jsou-li odstraněny, nemusí funkci genů ovlivnit [3].

Díky existenci mnoha kombinací genových produktů vzniklých umístěním více promotorů (počátků transkripce) v rámci jednoho exonu je složité určit počátek a konec jednotlivých genů. Slouží nám k tomu tzv. genetický kód, charakteristický způsob zápisu, který slouží k vyhledávání proteinotvorných struktur v DNA. Sekvencím, které použití metody umožňují, říkáme otevřený čtecí rámec (ORF), označující přítomnost genu. Zbývá nám jen najít začátek a dostatečně vzdálený konec. Na gen můžou dále ukazovat specifické sekvence, např. CpG ostrovy na počátku genu [3].

Genom eukaryot se skládá z unikátních a repetitivních sekvencí, tvořících 30% lidské DNA. Rozeznáváme středně a silně repetitivní sekvence, které mohou obsahovat až statisíce nukleotidů. DNA obsahuje speciální sekvence, které lze číst stejně z obou konců, palindromy. Pokud se vyskytnou na protilehlých vláknech, vzniká vlásenková struktura. Její funkce, kromě signálního účelu, se pouze odhaduje [5].

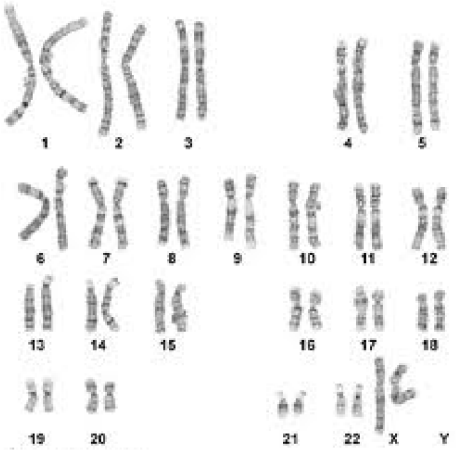
Některé úseky nemají stálou polohu, nazýváme je pohyblivé (mobilní) elementy. Patří mezi ně např. retropozony, neaktivní geny připomínající cDNA. Mohou se za

určitých podmínek přesunout do jiné části genomu, chovají se jako mutageny, mohou totiž měnit strukturu i funkci genů. Celou skupinu označujeme pseudogeny [3].

1.4 Genom

Pojem genom označuje soubor genů v buňce nebo zahrnuje všechny molekuly DNA. V našem případě je myšlen soubor veškeré genové i negenové DNA [2].

Lidské buňky, s výjimkou gamet a buněk některých orgánů, obsahují dvě kompletní sady genomu. Genom člověka obsahuje 3,0–3,5 miliard bp. Přes všeobecný názor nepatří lidský genom k největším, mnohem rozsáhlejší je např. u obojživelníků [3].



Obr. 1.8: Karyotyp ženy [4].

Struktury obsahující geny (genofory) eukaryotické buňky jsou jaderné chromosomy, mimojaderné (mitochondriální) chromosomy a plazmidy. Jaderné chromosomy tvoří jaderný genom, nejaderné struktury plazmon. Soubor chromosomů v jádře buňky se nazývá karyotyp (obr. 1.8). U člověka ho tvoří 23 párů chromosomů, 44 autosomů a 2 gonosomy (pohlavní chromosomy) [2].

Důležitými parametry genomu jsou jeho struktura a obsah. S rostoucí délkou genomu začínají geny tvořit čím dál menší část, přibývá úseků DNA s dosud neznámým účelem. Zároveň stoupá složitost genomu. Koeficient komplexnosti Cot vyjadřuje počet a délku různých sekvencí v DNA. Stanovuje se na základě rychlosti obnovy vodíkových můstků po denaturaci nukleové kyseliny. Krátké repetitivní sekvence reasociují nejrychleji, nejpomaleji dlouhé jedinečné sekvence, geny [3].

S rostoucí velikostí genomu stoupá počet genů v něm obsažených. Lidský genom obsahuje 30-40 tisíc genů. S pokrokem v sekvenční analýze se bude číslo zpřesňovat.

Dnes je popsáno kompletně jen několik chromosomů. Z prozatím dosažených výsledků je usuzováno, že rozdělení genů mezi chromosomy není rovnoměrné [3].

Některé geny, tzv. údržbáři (housekeeping genes), se exprimují nepřetržitě, jiné občas nebo jen v určitém typu buněk. V lidském těle se nejvíce genů transkribuje v mozku. Existují geny nezastupitelné, u jiných si může buňka vybírat. U diploidních buněk jsou někdy nezbytné obě alely (konkrétní formy genu), u některých je potřebná jen jedna [3].

Vědci se snaží zjistit, zda jsou všechny geny nezbytné a jestli jsou všechny používány. Z výsledků metody zablokování funkce genů (knock out) na pokusných organismech (drozofila, kvasinka, myš) se ukazuje, že některé geny lze vyřadit bez viditelných následků. Podíl těchto genů se liší dle druhu živočicha [3].

Ačkoliv každá somatická buňka obsahuje celý genom, je v kterémkoliv okamžiku aktivně využívána pouze jeho část. O expresi genů v jednotlivé buňce rozhodují určité faktory, signální molekuly, mezi které patří steroidní hormony a další látky, proteiny a methylovaná DNA [2].

1.4.1 Projekt lidský genom

Projekt lidský genom začal v říjnu roku 1990 v USA. Byl naplánován na 15 let, ale díky postupným pokrokům v technologii sekvenace byl dokončen již roku 2003. Oslavil tak padesátileté výročí objasnění struktury DNA. Účastnily se ho laboratoře osmnácti různých zemí, další země pomohly formou studia metod molekulární biologie pro průzkum genomu [11].

Cílem projektu bylo rozpoznat všechny báze v lidském genomu, najít geny, uchovat získané informace v databázích pro pozdější využití, zlepšit metody pro analýzu dat, převést technologie spojené s projektem do soukromého sektoru zavedením patentů a grantů a vyřešit vzniklé etické otázky. Pro dosažení cílů byly kromě člověka osekvenovány další organismy (Escherichia coli, Octomilka, Laboratorní myš) [11].

Projekt byl sponzorován Národním institutem pro výzkum lidského genomu Národního institutu zdraví a Programem lidského genomu Ministerstva energetiky USA. Zájem ministerstva vznikl po vynálezu atomové bomby. Průzkum genomu, rozpoznání a možná náprava mutací jsou klíčové pro pochopení účinku radioaktivního záření na lidský organismus. Dále se angažovala v etických, sociálních a právních otázkách týkajících se výzkumu genomu [11].

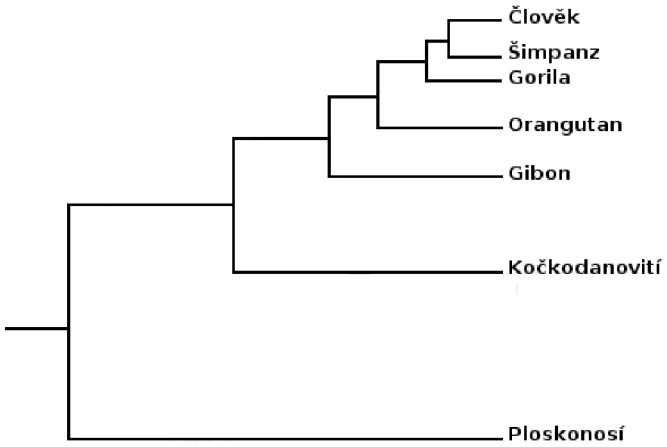
Navazující práce se zabývají sekvenací genomů dalších organismů pro podrobnější pochopení evoluce. V sekvencích se hledají geny, zkoumá se jejich poloha, funkce a regulace. Popisují se chromosomy, hledá se spojení s fenotypem. Genomy stejných organismů se začínají srovnávat mezi sebou.

Projekt lidský genom odstartoval lavinu výzkumů, vznikly nové obory, medicína dokáže diagnostikovat a léčit některé nemoci dříve, než se projeví první příznaky. Nesporně se jedná o projekt, který výrazně ovlivnil vědecký pokrok 21. století.

1.4.2 Genom šimpanze

Šimpanzi, konkrétně šimpanz učenlivý (*Pan troglodytes*) a šimpanz bonobo (*Pan paniscus*) a gorila nížinná (*Gorilla gorilla*) jsou v evoluci naši nejbližší příbuzní, jak je patrné z obr. 1.9. Podle odhadů jsme byli před sedmi miliony let součástí jednoho druhu. Naše druhy se vyznačují rozením živých mláďat, mezi savci chápavou rukou, nehty a pouze jedním párem bradavek a mezi primáty velmi pohyblivým ramenem a absencí ocasu [1].

Šimpanzí genom je lidskému podobný natolik (uvádí se 99,44% [1]), že je obtížné rozhodnout, zda sekvence v lidském genomu, která u šimpanze není, byla přidána během lidské evoluce nebo vymizela v rodokmenu šimpanzů. Navíc není každý nalezený rozdíl mezi sekvencemi nutně podstatný. Může se jednat o přirozenou variantu v rámci druhu. Buňka šimpanze obsahuje o 10% více DNA než buňka lidská. Lidské genomy se mezi sebou liší asi v jedné z tisíce bází a šimpanzí dokonce v jedné z dvěstěpadesáti bází. Abychom mohli posoudit rozdíly mezi genomy, musí se nejprve osekvenovat několik rodů šimpanzů, získat více genomů od společných předků a příbuzných lidí a šimpanzů a také nalézt sekvence více jedinců téhož druhu [12].



Obr. 1.9: Fylogenetický strom primátů.

Díky rozsahu znalostí, které máme o šimpanzi učenlivém, se logicky stal prvním osekvenovaným primátem pro pochopení vývoje člověka. Může se stát významným

ukazatelem vzniku geneticky podmíněných chorob. Existuje množství studií srovnávajících závažnost a frekvenci onemocnění společných pro člověka a lidoopy. Porovnání genomů může v budoucnu odpovědět, proč se některé nemoci vyskytují pouze u člověka, některé pouze u šimpanze a další se liší svými projevy [10].

Karyotyp šimpanze tvoří 24 párů chromosomů, 46 autosomů a 2 gonosomy. Značení je stejné jako člověka. Jediný rozdíl je u chromosomu 2. Šimpanz má na rozdíl od člověka čtyři. Značíme je 2A a 2B, protože genové složení je stejné jako u chromosomu 2 člověka. Porovnáním karyotypů se vzdálenějšími příbuznými vědci zjistili, že ve vývoji lidské linie došlo ke splynutí dvou chromosomů v jeden [1].

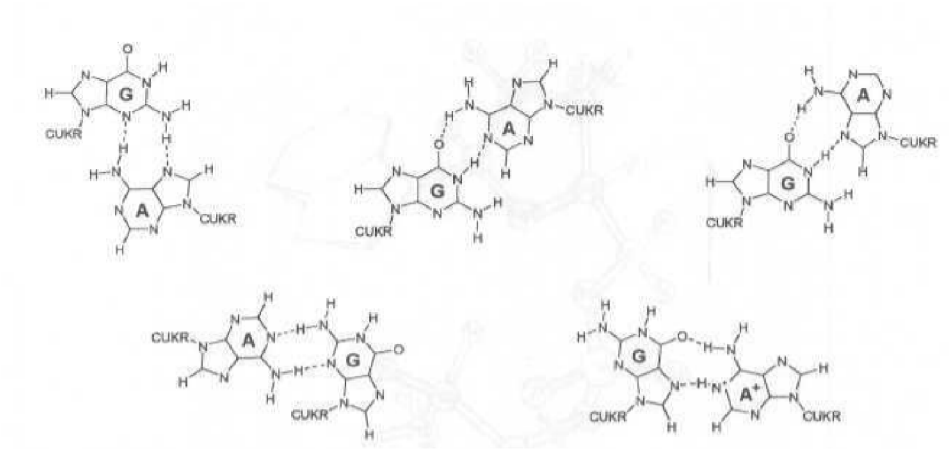
1.5 Sekvence tvořené adeninem a guaninem

Dlouhé úseky složené pouze z purinových bází s pyrimidinovými bázemi na komplementárním vláknu se nacházejí v lidském genomu zejména před geny. Zakódovaná informace je většinou v řetězci s převahou adeninu a guaninu [13].

Oblasti bohaté na purinové báze mohou vytvářet různé struktury. Pokud se adenin pravidelně střídá s guaninem, získá DNA zajímavé vlastnosti. Purinové báze a jejich páry stabilizují dvoušroubovici. Oligonukleotidy obsahující pouze pravidelně se střídající adenin s guaninem mohou tvořit duplex s paralelní i antiparalelní orientací. V paralelní dvoušroubovici vznikají nekanonické páry. Existuje několik konformací purinových párů lišících se v umístění vodíkových vazeb a orientaci bází vůči deoxyribose (obr. 1.10). Párování je závislé na pH a stavbě okolní DNA. Jako celek vytváří dvoušroubovice konformaci B s enormním zavinutím. Orientace řetězců může být antiparalelní i paralelní s vysokou stabilitou v širokém rozmezí pH. Dále může vzniknout levotočivá dvoušroubovice Z-DNA. V přítomnosti nízkých koncentrací zinečnatých iontů tvoří antiparalelní duplex, který je pravděpodobně součástí triplexu DNA společně s $(TC)_n$ řetězcem [13].

V rozmezí kyselého pH vzniká paralelní šroubovice s protonizovanými adeniny. S klesajícím pH tvoří oligonukleotid jednořetězcovou pravidelnou strukturu připomínající α-šroubovici, která je pravděpodobně levotočivá. Paralelní orientaci mají i dvoušroubovice tvořené oligonukleotidy CGATCG, TCGATCGA a CGATCGATCG. Jsou však stabilní jen při nízkém pH (3,8–4,4) [13].

Oktamer GCGAAAGC a heptamer GCGAAGC vytváří stabilní monomolekulární vlásenky. Nachází se v replikačních počátcích a promotorech transkripce. Termostabilita vlásenek souvisí s interakcí guaninu s adeninem. Vzácně se v DNA objevují vlásenky se smyčkami tvořenými jediným nukleotidem, tzv. vystrčenou bází. Tvar i stabilita vlásenek závisí na obsažených nukleotidech a jejich pořadí [13].



Obr. 1.10: Možnosti G-A párů a G-A^+ párů [13].

DNA obsahující tandemová opakování tripletů CNG zmnožená nad limit bývá příčinou neurodegenerativních onemocnění. Konformační labilitu molekuly DNA způsobují úseky $(GAA)_n$. Souvisí s výskytem Friedreichovy ataxie - kritická oblast se nachází v prvním intronu genu kódujícího protein frataxin na devátém chromosomu. Neobvyklé konformace DNA mohou být i fyziologické v jevech, kde je nestabilita deoxyribonukleové kyseliny žádoucí [13].

Pokud je jednořetězcový polynukleotid složený pouze z purinových bází tvořen alespoň z třetiny guaninem, vzniká čtyřřetězcová struktura. Dalším zajímavým motivem je tzv. element GAGA. Jedná se o úseky DNA tvořené především purinovými bázemi, které slouží v promotorech k rozpoznání transkripčními faktory. Stávají se tak nezbytnými pro expresi genů. Při mutaci popsaného úseku dojde k vypnutí genu [13].

1.6 MATLAB

MATLAB je programové prostředí a skriptovací programovací jazyk, který vychází z jazyka Fortran. První verze byla vydána již v roce 1985. Název vznikl ze spojení MATrix LABoratory. Jak název napovídá, hlavní datovou strukturou jsou matice, které nemusí nutně obsahovat pouze čísla. Programovací jazyk je určen především pro vědeckotechnické výpočty, simulace apod. [14].

Pro ulehčení výpočtů jsou vytvořeny různé nástroje, tzv. toolboxy. Jedná se o skupinu často používaných algoritmů zaměřených na různé oblasti. Pro bioinformatiku existuje Bioinformatic Toolbox, obsahuje např. algoritmy pro zarovnání sekvencí a práci s genetickými daty.

GUIDE (Graphical User Interface Development Environment) je prostředí MATLABu, ve kterém je možné vytvářet aplikace s grafickým rozhraním. Hlavní výhodou grafického prostředí je jednodušší ovládání uživatelem, proto byl použit i v předložené práci.

1.7 Sekvenční data

Sekvenční data jsou zápisy posloupnosti monomerů, v případě deoxyribonukleové kyseliny se jedná o pořadí nukleotidů od 5' ke 3' konci molekuly. Nukleotidy značíme písmeny stanovenými Mezinárodní unií pro čistou a aplikovanou chemii (IUPAC) [15] a jejich přehled je uveden v příloze.

Sekvence zapsané jako řetězce IUPAC znaků jsou surová data. Původní soubory obsahovaly pouze surová data. Veškeré informace o sekvenci uložené v souboru byly zapsány do jeho názvu. Vzhledem k omezení počtu znaků názvu je nahradily speciální formáty schopné uchovávat sekvenci i s její specifikací [15].

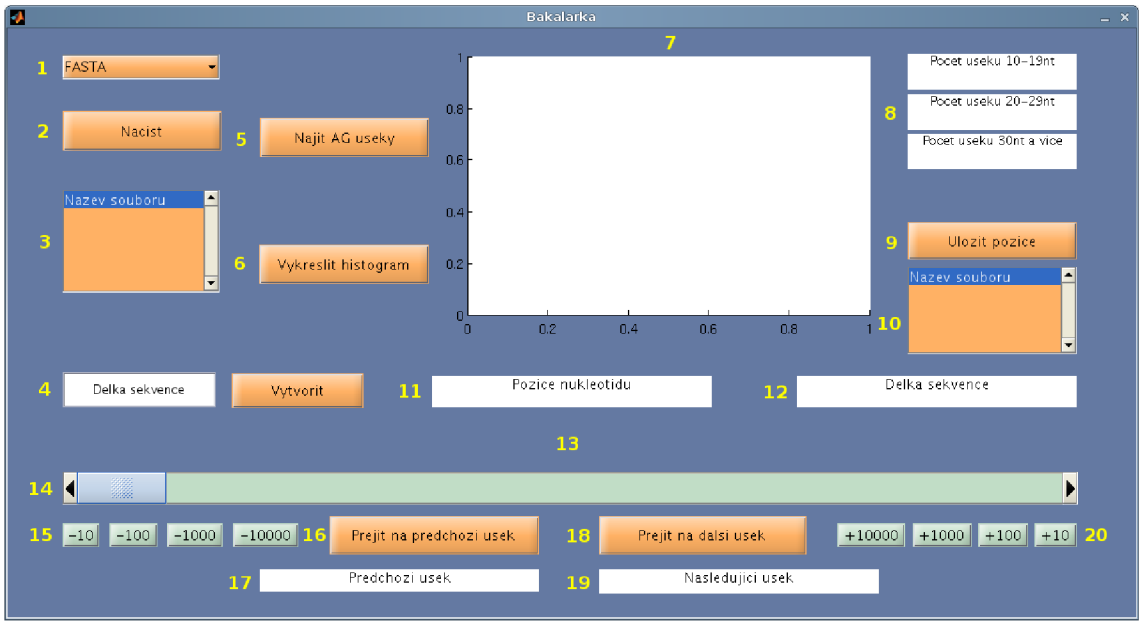
Pro svou jednoduchost se nejčastěji používá formát FASTA. Používá se pro sekvence nukleotidů i aminokyselin. FASTA je textový soubor. Jeho první řádek nazývaný hlavička (head) začíná znakem > (větší než). Obsahuje údaje o sekvenci včetně jedinečného identifikátoru sekvence. Na druhém řádku začíná surová sekvence, která tvoří zbytek souboru [15]. Sekvence by neměla obsahovat kromě písmen žádné další alfanumerické znaky. Při tvorbě programu, který bude s FASTA pracovat, je však nutné počítat i s možností dalších znaků, nejčastěji mezer. V případě více sekvencí v jednom souboru se toleruje prázdný řádek pro jejich oddělení.

V dnešní době existuje množství veřejných databází, na kterých jsou publikovaná primární sekvenční data. Data použitá v bakalářské práci pochází z databáze GenBank. GenBank obsahuje všechny veřejně dostupné DNA sekvence. Veškerá data jsou denně aktualizována a neustále zálohována. Databáze je přístupná prostřednictvím internetu. Díky velkému množství zveřejněných dat nemohou být všechna ověřována a není tak zaručena jejich správnost [15].

2 PRAKTICKÁ ČÁST

2.1 Program

Program `Bakalarka.m` byl vytvořen v prostředí MATLAB 7.9.0.529 R2009b. Je realizován v grafickém prostředí GUIDE. Slouží k načtení a zpracování sekvence DNA. Může pracovat s FASTA souborem zvoleným uživatelem nebo s náhodnou sekvencí o zvolené délce. Program v upravené sekvenci nalezne úseky obsahující pouze adenin a guanin delší než 9 nt. Pro účely dalšího zpracování dat, které uživatel získá, program vypíše délku sekvence, počty úseků o délkách nastavených v programu a vykreslí histogram pro představu o četnosti úseků. Veškeré nalezené úseky lze uložit do textového souboru, který si uživatel sám zvolí. Poslední částí programu je zobrazení sekvence, kterou si může uživatel procházet oknem o délce 50 nt. Nalezené úseky jsou barevně odlišeny. Speciálními tlačítky se uživatel může přesouvat o větší počet nukleotidů nebo přemisťovat jen po nalezených úsecích. Podrobný popis programu a uživatelského prostředí jsou uvedeny v následujících kapitolách. Manuál k programu je zařazen v příloze.



Obr. 2.1: Uživatelské prostředí programu `Bakalarka.m`.

2.1.1 Načtení a úprava sekvence

Program pracuje se sekvencemi ve FASTA formátu nebo náhodně generovanými sekvencemi. Pro další práci s daty je sekvenci v obou případech nutné upravit.

Funkce `pbNacti_Callback` se spustí po zmáčknutí tlačítka `pbNacti` (obr. B.1-2) uživatelem. V případě volby FASTA formátu v `pmVyber` (obr. B.1-1) se zobrazí vyskakovací okno, ve kterém si uživatel vybere soubor obsahující požadovanou sekvenci ve formátu .txt. Funkce si uloží název souboru a cestu k němu a poté jej otevře. Název souboru se zapíše do listboxu `lbSoubor` (obr. B.1-3), který slouží jako kontrola pro uživatele, že načtení proběhlo v pořádku. Při problému s načtením souboru je uživatel informován dialogovým oknem.

Soubor se začne číst po jednotlivých řádcích. Z důvodu alokace paměti pro sekvenci, která výrazně omezí dobu načítání, se zjistí velikost souboru. Vektor, který bude obsahovat sekvenci, naplníme nulami. Jejich počet určíme z rozdílu velikosti souboru a délky prvního řádku. První řádek se ukládá jako hlavička souboru a sekvenci neobsahuje. Proto by bylo zbytečné pro něj vyhrazovat místo v paměti. Všechny ostatní řádky se postupně uloží jako sekvence, pro správné umístění do vektoru je nutné ověřovat si délku každého řádku. Po skončení čtení se soubor uzavře.

Dalším krokem je smazání zbylých nul z vektoru společně s bílými znaky a nepotřebnými písmeny, které byly uloženy v souboru. Do konce programu pracujeme pouze se sekvencí obsahující čtyři písmena značící nukleotidy. Délka nově vzniklého vektoru odpovídá délce sekvence, s kterou dále pracujeme. Uživateli se pro informaci zobrazí v textovém poli `tDelka` (obr. B.1-12).

Pro usnadnění práce se sekvencí v programu, se převede posloupnost nukleotidů na čísla. Adenin je nahrazen číslem jedna, cytosin číslem dva, guanin číslem tři a thymin číslem čtyři. Zápis je využíván všude tam, kde urychlí práci s daty nebo usnadní zápis kódu.

Pro kontrolu uživatele se během načítání sekvence, její úpravě a převodu na čísla zobrazuje postup programu v procentech. Ukazatel byl do programu přidán z důvodu časově náročného zpracování dlouhých sekvencí.

Pokud zvolíme náhodně generovanou sekvenci, funkce vypíše do listboxu `lbSoubor` (obr. B.1-3) slovo Náhoda. Je nutné, aby uživatel zadal požadovanou délku vzniklé sekvence do editovatelného textového pole `eDelka` (obr. B.1-4) a tu následně potvrdil tlačítkem `pbVytvor` vedle. Pokud zapomene délku zadat, je upozorněn dialogovým oknem. Zavolá se funkce `pbVytvor_Callback`. Délka se načte a zobrazí se uživateli v poli `tDelka` (obr. B.1-12). Funkce vytvoří vektor náhodných čísel od nuly do jedné o zadané délce. Podmínkou, která rozděluje interval (0,1) na čtyři části podle procentuálního zastoupení bází ve skutečném lidském genomu (30% připadá na adenin a thymin, po 20% na cytosin a guanin), změní vektor na posloupnost

čísel jedna až čtyři nahrazující nám báze stejně jako v dříve popsaném převodu nukleotidů na čísla. Pro další práci s daty se sekvence převede i na nukleotidy.

Průběh vytváření sekvence v procentech je uživateli opět zobrazován. Když je sekvence hotová, objeví se vyskakovací okno, do kterého uživatel zadá název právě vzniklé sekvence a uloží si ji. Po úspěšném uložení se název zobrazí v listboxu `lbUlozene` (obr. B.1-10). Sekvence se uloží jako FASTA formát, v jeho hlavičce je uveden název, který uživatel zadal. Díky tomu, lze sekvenci opakovaně zpracovávat programem.

2.1.2 Vyhledání úseků tvořených adeninem a guaninem

Zmáčknutím tlačítka `pbNalez` (obr. B.1-5) se zavolá funkce `pbNalez_Callback`. Funkce postupně prochází sekvenci. Pokud nalezne adenin nebo guanin, uloží jeho pozici a postupuje, dokud nenalezne první cytosin nebo thymin. Poté uloží pozici posledního nalezeného hledaného nukleotidu. Zjistí vzdálenost mezi pozicemi. Pokud je delší než udává podmínka, uloží začátek, konec i délku úseků do vektorů. Podmínka je zavedena proto, aby se zkrátila další práce s daty. Pokud pracujeme s geny, je vhodné hledat úseky delší než devět nukleotidů. Při práci s celými chromosomy můžeme hranici zvýšit na devadesátdevět nukleotidů.

Pokud jsou nalezeny jakékoliv úseky vyhovující podmínce, funkce zobrazí první pozice ve vektorech začátků a konců jako polohu nejbližšího nalezeného úseku do textového pole `tNasledujici` (obr. B.1-19). Podrobněji se jím budeme zabývat v další kapitole.

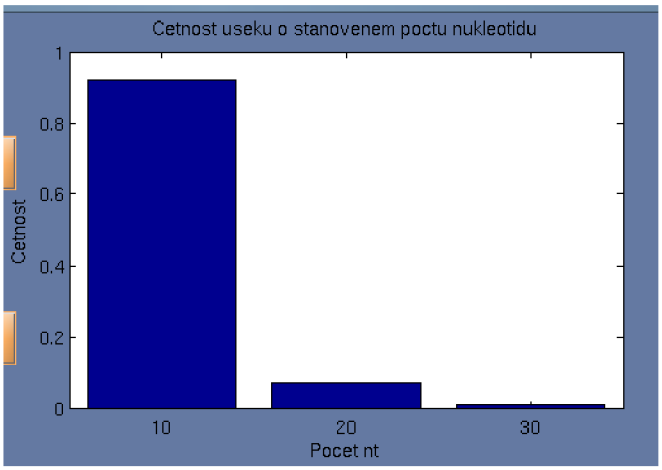
Pokud není nalezen žádný úsek, objeví se dialogové okno, které uživatele na tento fakt upozorní.

2.1.3 Zobrazení a práce s výsledky

Po nalezení úseků obsahujících pouze adenin a guanin si uživatel může výsledky zobrazit v histogramu, uložit je a procházet si sekvenci. V případě, že žádné úseky nebyly nalezeny, je uživatel upozorněn dialogovým oknem a některé funkce se neprovádí jako zbytečné.

Histogram (obr. 2.2) zobrazí uživatel zmáčknutím tlačítka `pbHist` (obr. B.1-6). Zavolá se funkce `pbHist_Callback`, která rozdělí nalezené úseky podle délky do tří skupin. Při práci s geny je vhodné dělit úseky na intervaly 10–19 nt, 20–29 nt a 30 nt a delší. Pokud pracujeme s celým chromosomem je lépe nastavit intervaly 100–149 nt, 150–199 nt a 200 nt a delší. Program sečte počty úseků v jednotlivých intervalech a poté je vydělí celkovým počtem úseků. Procentuální zastoupení úseků pak vykreslí jako sloupcový graf do pole grafu (obr. B.1-7).

Současně se do textových polí `t10`, `t20` a `t30` (obr. B.1-8) vypíše počet úseků v prvním, druhém a třetím intervalu.



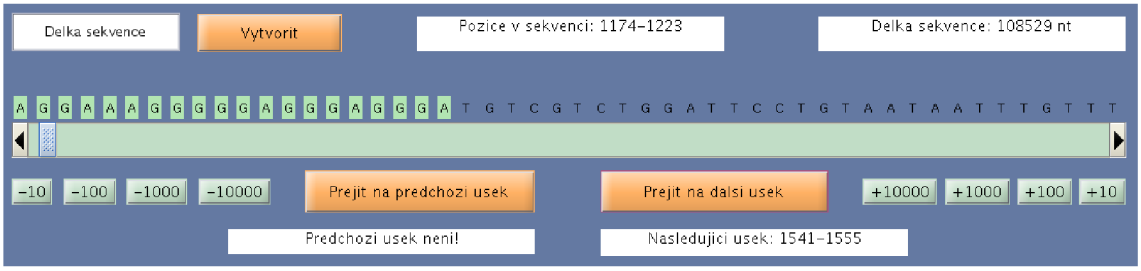
Obr. 2.2: Ukázka histogramu programu `Bakalarka.m`.

Pokud chceme s nalezenými úseky dále pracovat, je možné si je uložit do souboru .txt. Stisk tlačítka `pbUloz` (obr. B.1-9) zavolá funkci `pbUloz_Callback`. Funkce zobrazí vyskakovací okno, ve kterém uživatel zadá název a umístění textového souboru. Funkce uloží jeho název a zobrazí ho v listboxu `lbUlozene` (obr. B.1-10). Dále vytvoří buňkové pole, do kterého po řádcích uloží začátky, konce i délky nalezených úseků. Poté soubor otevře, data do něj zapíše a zase jej uzavře.

Poslední možností je procházení sekvence po nukleotidech. Zmáčknutím tlačítek s šipkami posuvníku `sPosun` (obr. B.1-14) ho aktivujeme. Díky funkci `sPosun_Callback` můžeme sekvenci procházet oknem délky 50 nt. Funkce nejprve zjistí aktuální pozici posuvníku a zaokrouhlí ji na celé číslo, protože pracujeme s nukleotidy. Rozsah posuvníku se nastaví na interval od jedné do konce sekvence bez délky posledního okna. Krok posuvníku se nastaví na jeden nukleotid. Poté se vypíše aktuální pozice okna v sekvenci do textového pole `tPozice` (obr. B.1-11) jako interval od aktuální pozice do délky okna. Nukleotidy v sekvenci odpovídající pozici okna se zobrazí jako graf v poli `axes2` (obr. B.1-13).

Pokud byly nalezeny úseky tvořené adeninem a guaninem, je dalším krokem je barevně odlišit v grafu, jak můžete vidět na obr. 2.3. Funkce nalezne všechny počátky úseků, které se nacházejí v okně. Pokud začátek existuje, začne od něj až po konec úseku nebo konec okna, vykreslovat barevně odlišené nukleotidy. Podobně hledá všechny konce úseků a od nich vykresluje nukleotidy pozpátku až na začátek úseku nebo okna. Při práci s celými chromosomy se hledají úseky delší než je rozsah okna. Proto funkce zjišťuje, zda existuje úsek, jehož začátek leží před aktuálním oknem a

konec za ním. Pokud je podmínka splněna, zabírá úsek celé okno, jsou vykresleny barevně všechny nukleotidy.



Obr. 2.3: Ukázka zobrazení sekvence programem `Bakalarka.m`.

Funkce dále hledá úsek, který se nachází nejblíže před úseky v aktuálním okně, respektive jeho konec. Pozici úseku, pokud existuje, vypíše do textového pole `tPredchozi` (obr. B.1-17). Zmáčknutím tlačítka `pbPredesly` (obr. B.1-16) si uživatel může úsek zobrazit. Funkce `pbPredesly_Callback` předá pozici předcházejícího úseku funkci `sPosun_Callback`. Začátek okna se nastaví na počátek úseku.

Obdobně funkce hledá úsek, který je první za úseky zobrazenými v okně, přesněji jeho začátek. Interval, ve kterém je úsek, zapíše do textového pole `tNasledujici` (obr. B.1-19). Zmáčknutím tlačítka `pbPrejdi` (obr. B.1-18) uživatelem se zavolá funkce `pbPrejdi_Callback`, která předá pozici úseku funkci posuvníku po ověření, že začátek úseku v sekvenci není dále než pozice posledního okna.

Další možností uživatele je přemisťování v sekvenci tlačítky `pb10` (obr. B.1-20), `pb100`, `pb1000` a `pb10000` nebo `pbm10` (obr. B.1-15), `pbm100`, `pbm1000` a `pbm10000`. Jimi zavolané funkce programu `pb10_Callback`, `pb100_Callback`, `pb1000_Callback` a `pb10000_Callback` nebo `pbm10_Callback`, `pbm100_Callback`, `pbm1000_Callback` a `pbm10000_Callback` zvětší nebo zmenší pozici posuvníku o 10, 100, 1000 a 10000 nukleotidů. Tlačítka slouží pro snadnější posun okna uživatelem. Při práci s celými chromosomy je vhodné posuny zvětšit.

2.2 Analýza dat

Zpracování celého genomu se v průběhu řešení práce ukázalo nesplnitelné. Postupné úpravy kódu umožnily zpracovat celý chromosom, jak se ukázalo při testování sekvence dlouhé 150 Mb (největší lidský chromosom má asi 250 Mb). Sekvence byla vytvořena uměle jako opakující se úsek chromosomu, který pochází z databáze GenBank. Problém je však získat sekvenci celého chromosomu. V databázi se nachází

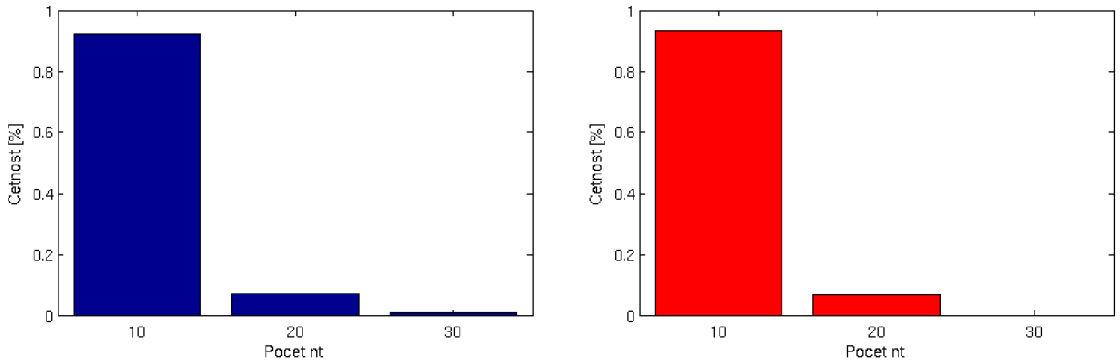
pouze popsané úseky. Mezi nimi se nachází mezery, které obsahují i statisíce nukleotidů. Velké mezery chybí zmapovat převážně u genomů primátů. Dalším úskalím je rozpoznání intronů a exonů. Algoritmy, které dokáží odhadnout pozici genů, jsou velmi komplikované a jejich implementace je nad rámec bakalářské práce.

Z uvedených důvodů jsme se rozhodli zpracovávat konkrétní geny z databáze GenBank, které jsou společné pro člověka i šimpanze. Veškeré geny člověka pochází z verze Build 37.2 a veškeré geny šimpanze z verze Build 2.1 Záměrně byly voleny geny obsahující introny i exony, aby bylo možné porovnávat četnost a podobnost polohy úseků tvořených adeninem a guaninem v rámci druhů. Geny mají různou délku i počet intronů. Všechny geny se nachází na chromosomu 21 šimpanze i člověka. Gen člověka bude srovnán i s náhodnou sekvencí stejné délky pro porovnání náhodnosti úseků.

2.2.1 COL18A1

Gen kóduje alfa řetězec kolagenu XVIII. Jedná se o protein mezibuněčné hmoty, který proteolýzou tvoří endostatin, silný antiangiogenní protein. Mutace genu je spojena s Knoblochovým syndromem, který způsobuje abnormality ve struktuře ledvin [16].

Gen COL18A1 člověka je úsek dlouhý 108529 nt. Skládá se z 42 exonů a 41 intronů. Program v genu nalezl celkem 114 úseků obsahujících adenin a guanin delších než 9 nt (obr. 2.4a). COL18A1 obsahuje 105 úseků dlouhých 10–19 nt, z toho 7 úseků se nachází v exonech, 8 úseků dlouhých 20–29 nt, všechny jsou v intronech a jeden úsek dlouhý 30 nt, umístěný v intronu.



(a) Člověk (b) Šimpanz

Obr. 2.4: Histogram genu COL18A1.

Gen COL18A1 šimpanze obsahuje 63798 nt. Skládá se z 42 exonů a 41 intronů. Program v genu nalezl celkem 61 úseků obsahujících adenin a guanin delších než 9 nt

(obr. 2.4b). Gen obsahuje 57 úseků dlouhých 10–19 nt, z toho 7 úseků se nachází v exonech a 4 úseků dlouhých 20–29 nt, všechny jsou v intronech.

Náhodná sekvence vytvořená programem o délce 108529 nt obsahovala 16 úseků dlouhých 10–19 nt.

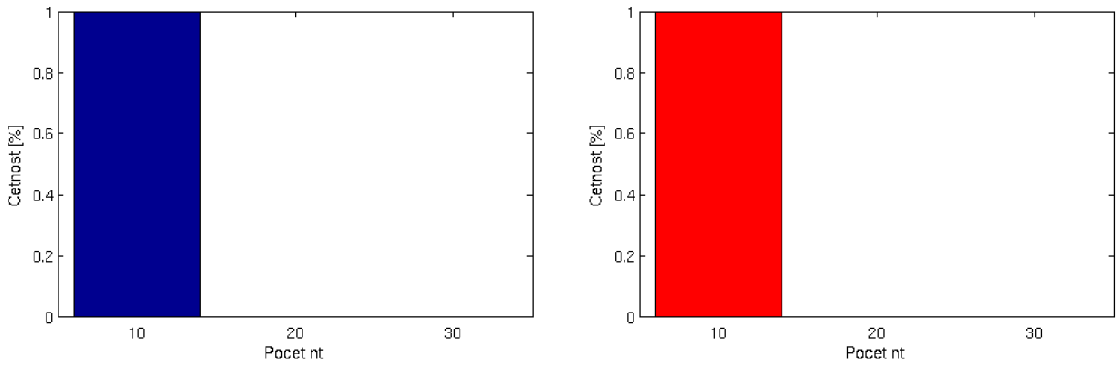
2.2.2 CSTB

Cystatin B (stefin B) je intracelulární inhibitor cysteinové proteázy. Je považován za součást ochrany proti prosakování proteázy z lysozomů. Nalezením mutací genu je diagnostikována progresivní myoklonická epilepsie [16].

Gen CSTB člověka je úsek dlouhý 2429 nt. Skládá se z 3 exonů a 2 intronů. Program v genu nalezl 1 úsek obsahující adenin a guanin dlouhý 14 nt (obr. 2.5a).

Gen CSTB šimpanze obsahuje 2099 nt. Skládá se z 3 exonů a 2 intronů. Program v genu nalezl 1 úsek obsahující adenin a guanin dlouhý 14 nt (obr. 2.5b).

Náhodná sekvence dlouhá 2429 nt neobsahovala žádný úsek delší než 9 nt.



(a) Člověk (b) Šimpanz

Obr. 2.5: Histogram genu CSTB.

2.2.3 IL10RB

Interleukin 10 receptor beta patří mezi cytokininové receptory. Jedná se o postranní řetězec, který je nezbytný pro receptorový komplex interleukin 10 [16].

Gen IL10RB člověka je úsek dlouhý 30849 nt. Skládá se z 7 exonů a 6 intronů. Program v genu nalezl celkem 37 úseků obsahujících pouze adenin a guanin delších než 9 nt (obr. 2.6a). IL10RB obsahuje 32 úseků dlouhých 10–19 nt, z toho 1 úsek se nachází v exonu, 5 úseků dlouhých 20–29 nt, jeden z nich v exonu (tab. 2.1).

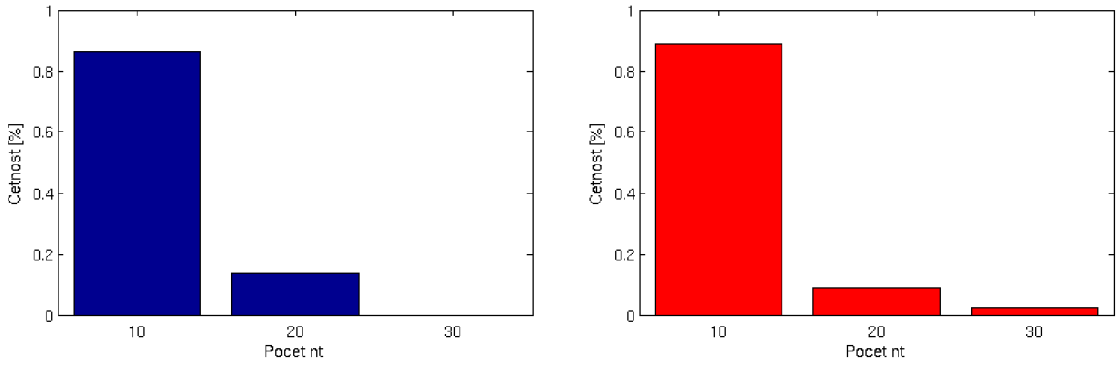
Gen IL10RB šimpanze obsahuje 30840 nt. Skládá se z 7 exonů a 6 intronů. Program v genu nalezl celkem 45 úseků obsahujících pouze adenin a guanin delších

Člověk	Šimpanz
AAAAAAAAAAAAAAAAAAAA GGGAGAAGGAAAGAGAGAAGG AAAAAAAAAAAAAAAAGAAAAGAAAA GAAAAAAAAAAAAAAAAAAAAG AAAAAAAAAAAAAAAAAA	AAAAAAAAAAAAAAAAAAAA GAAAAAAAAAAAAAAAAAAG AAAAAAAAAAAAAAAAGAAAAGAAAA AAAAAAAAAAAAAAAAAAAAAAAA

Tab. 2.1: Nalezené úseky v genu IL10RB o délce 20–29 nt.

než 9 nt (obr. 2.6b). Gen obsahuje 40 úseků dlouhých 10–19 nt, z toho 2 úseky se nachází v exonech, 4 úseky dlouhé 20–29 nt, všechny v intronech (tab. 2.1) a jeden úsek dlouhý 33 nt, nachází se v exonu.

Náhodná sekvence dlouhá 30849 nt obsahovala 10 úseků dlouhých 10–19 nt.



(a) Člověk (b) Šimpanz

Obr. 2.6: Histogram genu IL10RB.

2.2.4 PRSS7

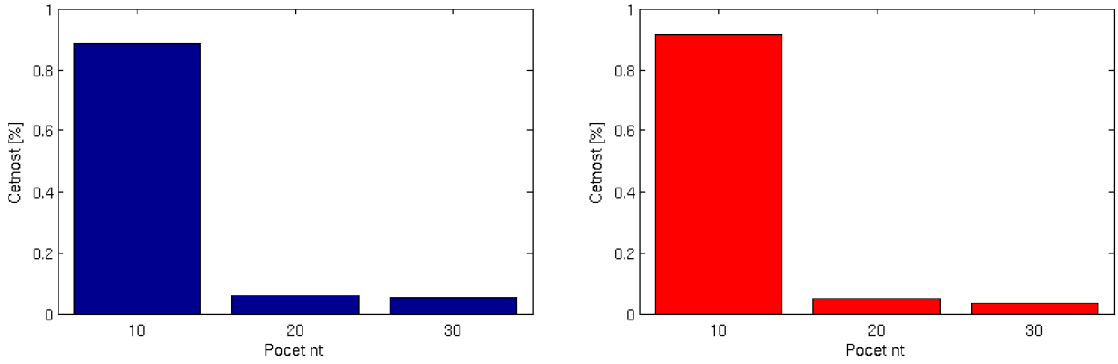
Transmembránová proteáza serin 15 patří mezi trypsinové peptidázy. Kóduje enzym, který mění pankreatický proenzym trypsinogen na trypsin. Mutace genu způsobuje nedostatek enterokinázy a poruchu malabsorpce [16].

Gen PRSS7 člověka je úsek dlouhý 134538 nt. Skládá se z 25 exonů a 24 intronů. Program v genu nalezl celkem 165 úseků obsahujících pouze adenin a guanin delších než 9 nt (obr. 2.7a). PRSS7 obsahuje 146 úseků dlouhých 10–19 nt, z toho 4 úseky se nachází v exonech, 10 úseků dlouhých 20–29 nt, všechny v intronech a 9 úseků delších než 29 nt, nachází se v intronech.

Gen PRSS7 šimpanze obsahuje 134272 nt. Skládá se z 25 exonů a 24 intronů. Program v genu nalezl celkem 167 úseků obsahujících pouze adenin a guanin delších než 9 nt (obr. 2.7b). Gen obsahuje 153 úseků dlouhých 10–19 nt, z toho 4 úseky se

nachází v exonech, 8 úseků dlouhých 20–29 nt, všechny v intronech a 6 úseků delších než 29 nt, nachází se v intronech.

Náhodná sekvence dlouhá 134538 nt obsahovala 36 úseků dlouhých 10–19 nt.



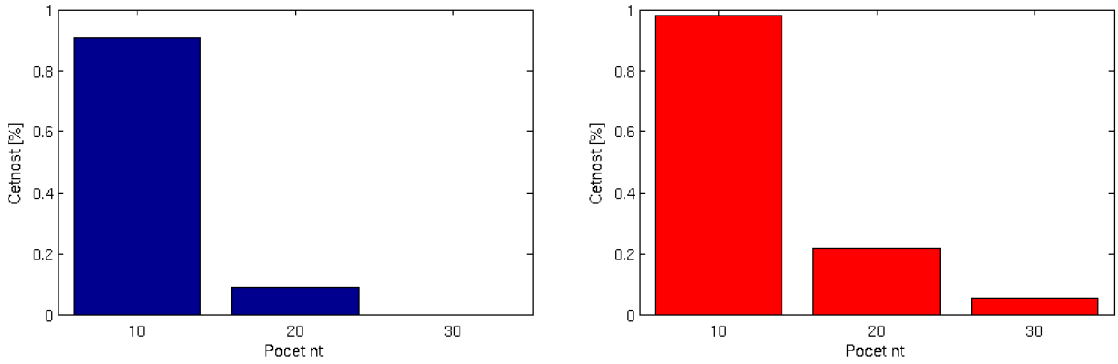
(a) Člověk (b) Šimpanz

Obr. 2.7: Histogram genu PRSS7.

2.2.5 TMEM50B

Jedná se o transmembránový protein 50B [16].

Gen TMEM50B člověka je úsek dlouhý 30834 nt. Skládá se z 7 exonů a 6 intronů. Program v genu nalezl celkem 11 úseků obsahujících adenin a guanin delších než 9 nt (obr. 2.8a). TMEM50B obsahuje 10 úseků dlouhých 10–19 nt, z toho 3 úseky se nachází v exonech a 1 úsek dlouhý 22 nt, nachází se v intronu.



(a) Člověk (b) Šimpanz

Obr. 2.8: Histogram genu TMEM50B.

Gen TMEM50B šimpanze obsahuje 52496 nt. Skládá se z 9 exonů a 8 intronů. Program v genu nalezl celkem 23 úseků obsahujících pouze adenin a guanin delších

než 9 nt (obr. 2.8b). Gen obsahuje 18 úseků dlouhých 10–19 nt, jeden z nich se nachází v exonu, 4 úseky dlouhé 20–29 nt, všechny v intronech a 1 úsek dlouhý 45 nt v intronu.

Náhodná sekvence dlouhá 30834 nt obsahovala 7 úseků dlouhých 10–19 nt.

2.2.6 Shrnutí výsledků

Z výsledků analýzy dat lze usuzovat, že s rostoucí délkou sekvence stoupá počet nalezených úseků tvořených adeninem a guaninem. Pokud porovnáme počty úseků v genomu člověka a šimpanze u srovnatelně dlouhých sekvencí, je jich podobné množství. Počet úseků v exonech je u všech genů podstatně menší, jsou převážně kratších délek. V intronech je počet úseků větší a navíc dosahují větších délek. Z histogramů jednotlivých genů lze usuzovat, že celkový poměr úseků všech tří délek je u člověka srovnatelný se šimpanzem. Podívejme se na gen IL10RB, který má v intronech stejný počet úseků o délce 20–29 nt (2.1). Úseky jsou si, pokud zanedbáme jejich pořadí, velmi podobné. Na první pohled poznáme jediný úsek, který se nachází v exonu. Z podobnosti úseků je patrné, že úseky mají v genu podstatnou funkci a jsou v evoluci konzervovány. Porovnáním s náhodnými sekvencemi, ve kterých se objevoval vždy menší počet úseků a žádný z nich nepřevyšoval délku 20 nt, je patrné, že námi nalezené úseky delší než 19 nt, které se nacházejí téměř výlučně v intronech genů nevznikly náhodou, což potvrzuje jejich důležitou úlohu.

3 ZÁVĚR

Cílem bakalářské práce bylo vyhledat úseky, ve kterých se vyskytuje pouze adenin s guanininem a porovnat jejich rozsah v kódující a nekódující oblasti. Srovnáním s evolučně blízkými živočichy měla být zvážena náhodnost, funkce a konzervovanost úseků.

Teoretická část se věnuje složení, struktuře, typům a konformacím deoxyribonukleové kyseliny. Je popsána kondenzace DNA do formy chromosomů, jejich složení a značení. Dále se zabývá definicí genu, jeho typům a účelům. Podrobněji je rozebrán genom a projekt, jehož cílem bylo získat kompletní genom člověka. Je uvedena podobnost člověka, šimpanze a gorily a rozdíl genomu šimpanze od lidského. Je známo, že úseky DNA tvořené pouze adeninem a guaninem nabývají zajímavých konformací, jejich popisu je věnována další podkapitola. V závěrečné části je popsáno programové prostředí MATLAB a hojně užívaný formát sekvenčních dat - FASTA.

V praktické části bakalářské práce je uveden popis programu `Bakalarka.m` vytvořený v grafickém prostředí MATLABu. Program dokáže načíst sekvenci ve FASTA formátu, zpracovat ji a vyhledat v ní úseky obsahující pouze adenin a guanin delší než 9 nt. Je možné pomocí programu vygenerovat náhodnou sekvenci, uložit ji jako FASTA a pracovat s ní obdobně. Program umožňuje nalezené úseky rozdělit do tří skupin podle délky a vykreslit z jejich poměru histogram. Úseky si lze uložit do textového souboru, do kterého jsou ukládány záměrně tak, aby byly snadno zpracovatelné tabulkovými editory pro další analýzu. Další možností je si sekvenci prohlížet v grafu, posouvat se po jednotlivých nukleotidech, jejich skupinách nebo přeskakovat mezi nalezenými úseky. Pro přehlednost jsou úseky tvořené pouze adeninem a guaninem delší než 9 nt barevně odlišeny.

Další část tvoří analýza dat. V průběhu tvorby bakalářké práce se ukázala obtížnost nalezení kompletních sekvencí chromosomů. V současné době jsou zpracovány především zajímavé oblasti, v ostatních částech chromosomů, zejména v nekódujících oblastech mezi geny, chybí zmapovat velké množství úseků, jsou zde pouze zaznačeny tzv. mezery. U lidského genomu je to asi desetina celkového rozsahu [1], v genomu primátů podstatně více. Rozlišování intronů a exonů v sekvencích patří mezi pokročilé algoritmy svou náročností převyšující rámec bakalářské práce. Proto bylo po dohodě s vedoucím práce i souhlasem konzultanta rozhodnuto, že se program ověří na malém množství genových dat z databáze GenBank.

Analýza byla provedena na pěti genech společných pro člověka i šimpanze. Záměrně byly voleny geny obsahující introny i exony o různých délkách, aby bylo možné porovnat rozsah úseků v kódující a nekódující oblasti. Pro každý lidský gen byla programem vygenerována náhodná sekvence o stejné délce pro posouzení náhodnosti úseků.

Z výsledku rozboru genů lze usuzovat, že s rostoucí délkou sekvence stoupá počet úseků složených pouze z adeninu s guaninem v ní obsažených. V případě náhodných sekvencí narůstá pouze počet úseků, u genů se zvětšuje navíc i jejich délka. Můžeme tedy rozhodnout, že úseky delší než 19 nt v sekvencích nevznikají náhodně.

Pokud srovnáme počty nalezených úseků mezi introny a exony, je patrné, že jich mnohem více obsahují nekódující oblasti, u úseků delších než 19 nt se vyskytují téměř výhradně v intronech. Z histogramů můžeme vyčíst, že poměry počtů úseků ve třech námi zavedených skupinách podle délky jsou u člověka a šimpanze srovnatelné.

Srovnáním sekvencí úseků o délce 20–29 nt nalezených v nekódujících oblastech genu IL10RB u člověka a šimpanze je naprosto zřetelné, že úseky jsou vysoce konzervované. Jejich rozdíly jsou nepatrné, liší se pouze v délce. Vzhledem k tomu, že úseky jsou si natolik podobné a zachovaly se v evoluci, lze se domnívat, že mají v organismu nezastupitelnou funkci.

Přehled výsledků zpracování genů programem je společně s uživatelským návodem pro použití programu uveden v příloze.

LITERATURA

[1] MARKS, Jonathan. *Jsme téměř 100% šimpanzi?: Lidoopi, lidé a geny*. Vydání první. Praha: Academia, 2006. 349 s. ISBN 80-200-1321-0.

[2] NEČAS, Oldřich, et al. *Obecná biologie pro lékařské fakulty*. 3. přepracované vydání. Praha: Nakladatelství H&H, 2000. 555 s. ISBN 80-86022-46-3.

[3] BRDIČKA, Radim. *Lidský genom: Na rozhraní tisíciletí*. Vydání první. Praha: Grada Publishing, 2001. 249 s. ISBN 80-2470-118-9.

[4] ŠÍPEK, A. jr. *Genetika - Váš zdroj informací o genetice*. [online], 25. 11. 2003, poslední aktualizace 2. 10. 2008 [cit. 29. 12. 2010]. Dostupné z URL: <http://genetika.wz.cz/>.

[5] LEDVINA, Miroslav; STOKLASOVÁ, Alena; CERMAN, Jaroslav. *Biochemie pro studující medicíny: I. díl*. Vydání první. Praha: Karolinum, 2004. 274 s. ISBN 80-246-0849-9.

[6] RAMBOUSEK, Lukáš. *Předmětová komise biologie a chemie*. [online], 2007, poslední aktualizace 25. 4. 2011 [cit. 12. 5. 2011]. Dostupné z URL: <http://giobio.ic.cz/obrazky/obecka.htm>.

[7] RINEHART, Claire A. *Introduction to Molecular and Cell Biology & Cell, Biol. 220*. [online], 2004, poslední aktualizace 2006 [cit. 29. 12. 2010]. Dostupné z URL: <http://bioweb.wku.edu/courses/biol22000/>.

[8] SOCHOR, Michal *Botanika - teorie a praxe*. [online], poslední aktualizace 24. 12. 2010 [cit. 29. 12. 2010]. Dostupné z URL: <http://botanika.borec.cz/index.php>.

[9] KOUKALOVÁ, Světlana *Zápisky zdravotní sestřičky*. [online], 29. 3. 2006, poslední aktualizace 16. 12. 2010 [cit. 29. 12. 2010]. Dostupné z URL: <http://vnl.xf.cz/index.php>.

[10] VARKI, Ajit. A Chimpanzee Genome Project Is a Biomedical Imperative. *Genome Research*. 2000, 10, s. 1065-1070.

[11] SHERYL, A. Martin. *Human Genome Project Information*. [online], poslední aktualizace 4. 11. 2010 [cit. 15. 12. 2010]. Dostupné z URL: <http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml>.

[12] DENNIS, Carina. Chimp genome: Branching out. *Nature*. 2005, 437, s. 17-19.

[13] KYPR, Jaroslav; VORLÍČKOVÁ, Michaela. Konformační vlastnosti řetězců DNA obsahujících guanin a adenin. *Chemické listy*. 1998, 92, s. 530-537.

[14] *The MathWorks, Inc.* [online], poslední aktualizace 22. 05. 2011 [cit. 17. 05. 2011]. Dostupné z URL: <http://www.mathworks.com/>.

[15] CVRČKOVÁ, Fatima. *Úvod do praktické bioinformatiky*. Vydání první. Praha: Academia, 2006. 148 s. ISBN 80-200-1360-1.

[16] *National Center for Biotechnology Information*. [online], poslední aktualizace 29. 11. 2010 [cit. 28. 12. 2010]. Dostupné z URL: <http://www.ncbi.nlm.nih.gov/guide/>.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

A,C,G,T	Adenin, Cytosin, Guanin, Thymin
b	Báze
bp	Pár bazí – Base Pair
Cot	Koeficient komplexnosti
DNA	Kyselina deoxyribonukleová – Deoxyribonucleic Acid
cDNA	DNA komplementární k RNA – Complementary DNA
dsDNA	Dvouřetězcová DNA – Double-Stranded DNA
ssDNA	Jednořetězcová DNA – Single-Stranded DNA
mtDNA	Mitochondriální DNA – Mitochondrial DNA
H	Histon
nt	Nukleotid
ORF	Otevřený čtecí rámec – Open Reading Frame
p,q	Krátká, dlouhá raménka
pH	Vodíkový exponent

SEZNAM PŘÍLOH

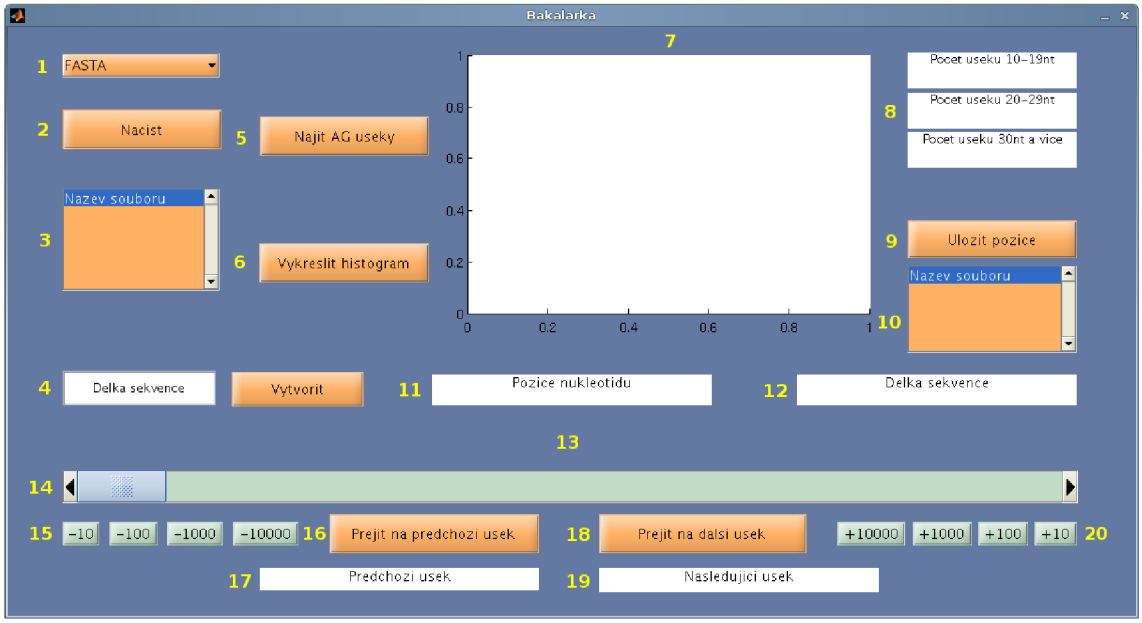
A IUPAC KÓDOVÁNÍ NUKLEOVÝCH KYSELIN

Kód	Báze
A	Adenin
C	Cytosin
G	Guanin
T	Thymin
R	Purin (A, G)
Y	Pyrimidin (C,T)
S	G, C (Strong)
W	A, T (Weak)
K	G, T (Keto)
M	A, C (Amino)
B	C, G, T
D	A, G, T
H	A, C, T
V	A, C, G
N	Cokoliv

Tab. A.1: Kódování DNA dle IUPAC

B UŽIVATELSKÝ MANUÁL

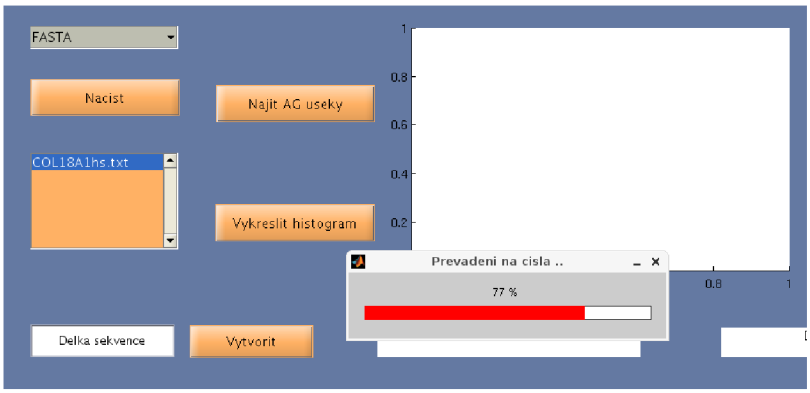
Program `Bakalarka.m` slouží pro načtení a zpracování sekvence ve FASTA formátu nebo vytvoření náhodné sekvence. Program nalezne úseky obsahující pouze adenin a guanin, které si lze zobrazit, prohlížet a uložit.



Obr. B.1: Uživatelské prostředí programu `Bakalarka.m`.

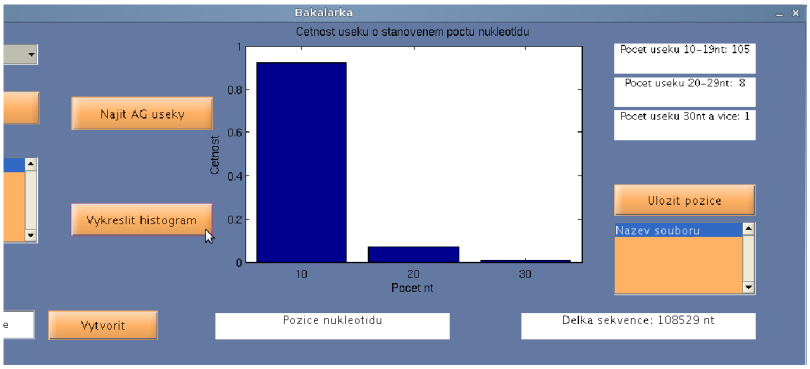
1. Program spusťte zmáčknutím F5 v souboru `Bakalarka.m` v MATLABu.
2. Nejprve si pomocí pop-up menu `FASTA / Nahodna sekvence` (obr. B.1-1) zvolte, zda chcete pracovat se sekvencí ve formátu FASTA nebo vygenerovat náhodnou sekvenci.
3. Svou volbu potvrďte tlačítkem `Nacist` (obr. B.1-2).
   (a) V případě volby `FASTA` se objeví vyskakovací okno, ve kterém si vyberte požadovaný soubor. V listboxu `Nazev souboru` (obr. B.1-3) se objeví jeho název a okamžitě se otevře dialogové okno, ve kterém můžete sledovat postup programu (obr. B.2), který upravuje sekvenci. Pokud se dialog neobjeví, zkontrolujte vybraný soubor. Pokud obsahuje na konci prázdné řádky, smažte je. Pokud vše proběhne bez chyby, dialogové okno se samo uzavře a v textovém poli `Delka sekvence` (obr. B.1-12) se objeví počet nukleotidů v načtené sekvenci.
   (b) V případě volby `Nahodna sekvence` se objeví v listboxu `Nazev souboru` (obr. B.1-3) slovo `Nahoda`. Do editovatelného pole `Delka sekvence` (obr.

B.1-4) zadejte počet nukleotidů sekvence, kterou chcete nechat vytvořit. Číslo potvrďte tlačítkem `Vytvorit`, které je umístěno vedle pole. Objeví se dialogové okno, ve kterém můžete kontrolovat pokrok programu. Po vytvoření sekvence se její délka objeví v textovém poli `Delka sekvence` (obr. B.1-12). Okamžitě se otevře vyskakovací okno, sekvenci si můžete uložit pro další práci ve formátu FASTA. Po uložení se název objeví v listboxu `Nazev souboru` (obr. B.1-10).



Obr. B.2: Pokrok programu.

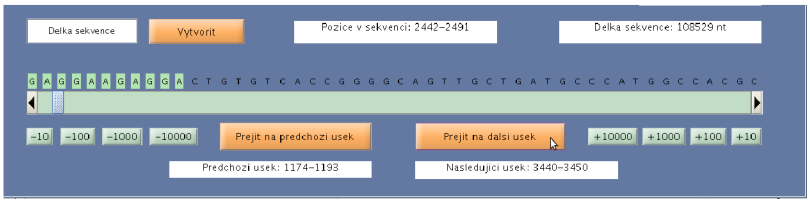
4. Pro nalezení úseků obsahujících adenin a guanin, které jsou delší než 9 nt, zmáčkněte tlačítko `Najit AG useky` (obr. B.1-5). Po nalezení všech úseků se v textovém poli `Nasledujici usek` (obr. B.1-19) objeví pozice prvního úseku v sekvenci. Pokud sekvence žádný úsek neobsahuje, objeví se dialogové okno.
5. Pro vykreslení histogramu zmáčkněte tlačítko `Vykreslit histogram` (obr. B.1-6). V poli grafu (obr. B.1-7) se objeví histogram. Do textových polí vedle grafu (obr. B.1-8) se vypíše počet úseků v délce 10–19 nt, 20–29 nt a delších než 29 nt (obr. B.3).
6. Pokud si chcete pozice nalezených úseků uložit do textového souboru, zmáčkněte tlačítko `Ulozit pozice` (obr. B.1-9). Objeví se vyskakovací okno, do něj zadejte název, pod kterým chcete úseky uložit. Program vytvoří textový soubor s třemi sloupci, které obsahují začátky, konce a délky úseků. Hodnoty lze zpracovávat např. pomocí OpenOffice Tabulky. Po úspěšném uložení souboru se jeho název objeví v listboxu `Nazev souboru` (obr. B.1-10).
7. Zmáčknutím šipky na posuvníku (obr. B.1-14) nebo jeho posunutím myší se v okně grafu (obr. B.1-13) objeví 50 nukleotidů. Nad grafem v textovém poli `Pozice nukleotidu` (obr. B.1-11) se objeví pozice nukleotidů v načtené sekvenci. Opakovaným zmáčknutím šipky posuvníku se posouváte o jeden



Obr. B.3: Histogram.

nukleotid vpřed nebo zpět v sekvenci. Nalezené úseky jsou v grafu barevně odlišeny.

8. Pokud se chcete pohybovat po větším počtu nukleotidů použijte tlačítka `-10, -100, -1000` a `-10000` (obr. B.1-15), které posunou posuvník o 10, 100, 1000 a 10000 nukleotidů doleva nebo tlačítka `10, 100, 1000` a `10000` (obr. B.1-20), které posunou posuvník o 10, 100, 1000 a 10000 nukleotidů doprava.



Obr. B.4: Zobrazení nalezeného úseku.

9. Pokud se chcete přemisťovat po nalezených úsecích, použijte tlačítko **`Prejit na predchozi usek`** (obr. B.1-16), které přesune okno posuvníku na začátek úseku před aktuálně zobrazeným oknem. Pozice je uvedena v textovém poli **`Predchozi usek`**. Použitím tlačítka **`Prejit na dalsi usek`** (obr. B.1-18) se posuvník přesune na úsek za aktuálním oknem, jehož pozice je zapsaná v textovém poli **`Nasledujici usek`** (obr. B.1-19).
10. Program ukončíte zmáčknutím křížku v pravém horním rohu okna.

C PŘEHLED VÝSLEDKŮ

	Homo sapiens	Pan troglodytes	Náhodná sekvence
Identifikace genu	NG_011903.1	NC_006488.2	-
Délka genu	108529 nt	63798 nt	108529 nt
Počet úseků 10–19 nt	105	57	16
- exony	7	7	-
- introny	98	50	-
Počet úseků 20–29 nt	8	4	0
- exony	0	0	-
- introny	8	4	-
Pozice úseků > 29 nt	1	0	0
- exony	0	0	-
- introny	1	0	-

Tab. C.1: Přehled výsledků pro gen COL18A1.

	Homo sapiens	Pan troglodytes	Náhodná sekvence
Identifikace genu	NG_011545.1	NC_006488.2	-
Délka genu	2429 nt	2099 nt	2429 nt
Počet úseků 10–19 nt	1	1	0
- exony	1	1	-
- introny	0	0	-
Počet úseků 20–29 nt	0	0	0
- exony	0	0	-
- introny	0	0	-
Pozice úseků > 29 nt	0	0	0
- exony	0	0	-
- introny	0	0	-

Tab. C.2: Přehled výsledků pro gen CSTB.

	Homo sapiens	Pan troglodytes	Náhodná sekvence
Identifikace genu	NG_012089.1	NC_006488.2	-
Délka genu	30849 nt	30840 nt	30849 nt
Počet úseků 10–19 nt	32	40	10
- exony	1	2	-
- introny	31	38	-
Počet úseků 20–29 nt	5	4	0
- exony	1	0	-
- introny	4	4	-
Pozice úseků > 29 nt	0	1	0
- exony	0	1	-
- introny	0	0	-

Tab. C.3: Přehled výsledků pro gen IL10RB.

	Homo sapiens	Pan troglodytes	Náhodná sekvence
Identifikace genu	NG_012207.1	NC_006488.2	-
Délka genu	134538 nt	134272 nt	134538 nt
Počet úseků 10–19 nt	146	153	36
- exony	4	4	-
- introny	142	149	-
Počet úseků 20–29 nt	10	8	0
- exony	0	0	-
- introny	10	8	-
Pozice úseků > 29 nt	9	6	0
- exony	0	0	-
- introny	9	6	-

Tab. C.4: Přehled výsledků pro gen PRSS7.

	Homo sapiens	Pan troglodytes	Náhodná sekvence
Identifikace genu	NC_000021.8	NC_006488.2	-
Délka genu	30834 nt	52496 nt	30834 nt
Počet úseků 10–19 nt	10	18	7
- exony	3	1	-
- introny	7	17	-
Počet úseků 20–29 nt	1	4	0
- exony	0	0	-
- introny	1	4	-
Pozice úseků > 29 nt	0	1	0
- exony	0	0	-
- introny	0	1	-

Tab. C.5: Přehled výsledků pro gen TMEM50B.

D ZDROJOVÉ KÓDY PROGRAMU BAKALARKA.M

```
function pbNacti_Callback(~, ~, handles)
% inicializace promennych

switch get(handles.pmVyber,'Value');
case 1;
   % otevreni souboru zvoleneho uzivatelem

   if ischar(celacesta) && exist(celacesta,'file') == 2;
      informace_o_souboru = dir(celacesta);
      velikost_souboru= informace_o_souboru.bytes;
      radek = fgetl(fid);
      pocet_nukleotidu = velikost_souboru -length(radek);
      poc = 1;
      sek = char(zeros(1,pocet_nukleotidu));
      rozmer = length(radek);
      % tvorba waitbar

      while ischar(radek);
         if poc == 1;
            Head = radek;
            pocind = 1;
         else
            sek(1,pocind:rozmer+pocind-1) = radek;
            pocind = pocind + rozmer;
         end
         radek = fgetl(fid);
         rozmer = length(radek);
         poc = poc + 1;
         % tvorba waitbar
      end

      fclose(fid);

   else
      warndlg('Nebyl vybran soubor!');
   end
```

```
    delka = sum(sek=='A'|sek=='C'|sek=='G'|sek =='T');
    Wsek = char(zeros(1,delka));
    prom = 1;
    % tvorba waitbar

    for n = 1:delka;
        if sek(n)=='A'||sek(n)=='C'||sek(n)=='G'||sek(n)=='T';
            Wsek(prom) = sek(n);
            prom = prom +1;
            % tvorba waitbar
        end
    end

    Seq = Wsek;
    Seqcisla = zeros(1,delka);
   % tvorba waitbar

    for i = 1 : delka;
       % prevedeni pismen oznacujich nukleotidy na cisla
       % tvorba waitbar
    end

    % tvorba waitbar
    set(handles.tDelka,'String',
        sprintf('Delka sekvence: %d nt',delka));

case 2
    Head = 'Nahoda';
    set(handles.lbSoubor, 'String', Head);

end
```

```
function pbVytvor_Callback(~, ~, handles)
pocetnt = get(handles.eDelka,'String');
% inicializace promennych

if isempty(delka) == 0
   % tvorba waitbar
   nahoda = rand(1,delka);
   Seqcisla = zeros(1,delka);

   for ind = 1:delka
      if nahoda(ind) < 0.3
         Seqcisla(ind) = 1;
      elseif nahoda(ind) < 0.6
         Seqcisla(ind) = 4;
      elseif nahoda(ind) < 0.8
         Seqcisla(ind) = 2;
      else
         Seqcisla(ind) = 3;
      end
      % tvorba waitbar
   end

   Seq=char(zeros(1,delka));

   for i = 1 : delka;
      % prevedeni pismen oznacujich nukleotidy na cisla
      % tvorba waitbar
   end

  % tvorba waitbar
   set(handles.tDelka,'String',
       sprintf('Delka sekvence: %d nt',delka));
  % vytvoreni a ulozeni promenne do formatu FASTA

else
  warndlg('Zadejte delku sekvence!');
end
```

```
function pbNalez_Callback(~, ~, handles)
% inicializace promennych

for i = 1:delka;
   if (zacatek==0 && (Seqcisla(i)==1||Seqcisla(i)==3));
      zacatek = i;
   elseif (zacatek~=0 && (Seqcisla(i)==2||Seqcisla(i)==4));
      konec = i - 1;
      pomoc = konec - zacatek + 1;
      if pomoc > 10;
         vektorzac = [vektorzac zacatek];
         vektorkon = [vektorkon konec];
         vektordel = [vektordel pomoc];
      end
      zacatek = 0;
   end
end

if isempty(vektorzac) == 0
   usekpo = [];
   usekpo(1) = vektorzac(1);
   usekpo(2) = vektorkon(1);
   set(handles.tNasledujici,'String',sprintf(% vypis));
else
   warndlg('Nebyly nalezeny zadne useky!');
   zadny_usek = 1;
end


function pbHist_Callback(~, ~, handles)
% inicializace promennych

if zadny_usek == 0
   axes(handles.axes1);
   for i1 = 1:celkem;
      if vektordel(i1) > 9 && vektordel(i1) < 20;
         d_deset = d_deset + 1;
      elseif vektordel(i1) > 19 && vektordel(i1) < 30;
         d_dvacet = d_dvacet + 1;
```

```
        elseif vektordel(i1) > 29;
            d_tricet = d_tricet + 1;
        end
    end
    clear i1;
    x = 10:10:30;
    y = [d_deset/celkem d_dvacet/celkem d_tricet/celkem];
    bar(x,y);
    % popis grafu
    set(handles.t10,'String',sprintf(% vypis));
    set(handles.t20,'String',sprintf(% vypis));
    set(handles.t30,'String',sprintf(% vypis));
else
    warndlg('Nebyly nalezeny zadne useky!');
end


function pbUloz_Callback(~, ~, handles)
global vektordel vektorzac vektorkon celkem Nazev
[nazevsouboru, cesta] = uiputfile('*.txt', 'Ulozte soubor');
Nazev = nazevsouboru;
set(handles.lbUlozene, 'String', Nazev);
data = {'Zacatek_useku' 'Konec_useku' 'Delka_useku'};

for i2 = 1:celkem;
    data{i2+1,1} = vektorzac(i2);
    data{i2+1,2} = vektorkon(i2);
    data{i2+1,3} = vektordel(i2);
end

[radku,~]= size(data);
fid = fopen(nazevsouboru, 'w');
fprintf(fid, '%s %s %s \n', data{1,:});

for radek = 2:radku;
    fprintf(fid, '%d %d %d \n', data{radek,:});
end
fclose(fid);
```

```
function sPosun_Callback(hObject, ~, handles)
% inicializace promennych
for i=1:50;
   text((i-1)/50+0.25/50,0,Seq(pozice+i-1));
end

if zadny_usek == 0

for k = 1:celkem;
  if vektorzac(k)>=pozice && vektorzac(k)<=(pozice+49);
    vektorzac1 = [vektorzac1 vektorzac(k)];
    vektorpoz1 = [vektorpoz1 k];
  end
end

for i = 1:(length(vektorzac1));
  for l = vektorzac1(i):vektorzac1(i)
          +vektordel(vektorpoz1(i))-1;
    if l-pozice+1 <= 50;
      text((l-pozice)/50+0.45/50,0,Seq(l),% parametry grafu);
    end
  end
end

for k = 1:celkem
  if vektorkon(k) >= pozice && vektorkon(k) <= (pozice+49);
    vektorkon1 = [vektorkon1 vektorkon(k)];
    vektorpoz2 = [vektorpoz2 k];
  end
end

for i = 1:(length(vektorkon1));
  for l = vektorkon1(i):-1:vektorkon1(i)-
          vektordel(vektorpoz2(i))+1;
    if l-pozice +1 > 0;
      text((l-pozice)/50+0.45/50,0,Seq(l),% parametry grafu);
    end
  end
end
```

```
for l = 1:celkem
  if vektorzac(l) < pozice && vektorkon(l) > pozice + 49
    for i = 1:50
      text((i-1)/50+0.45/50,0,Seq(pozice+i-1),
          % parametry grafu);
    end
    break
  end
end

n = 1;
while n < celkem +1 && vektorzac(n) < pozice + 50
  n = n + 1;
end

if n == celkem+1
  set(handles.tNasledujici,'String',sprintf(% vypis));
else
  usekpo(1) = vektorzac(n);
  usekpo(2) = vektorkon(n);
  set(handles.tNasledujici,'String',sprintf(% vypis));
end

n = celkem;
while n > 1 && vektorkon(n) > pozice - 1
  n = n - 1;
end

if n == 1
  set(handles.tPredchozi,'String',sprintf(% vypis));
else
  usekpred(1) = vektorzac(n-1);
  usekpred(2) = vektorkon(n-1);
  set(handles.tPredchozi,'String',sprintf(% vypis));
end

end
```

```
function pbPrejdi_Callback(~, eventdata, handles)
global delka usekpo
pozice = usekpo(1);
if pozice > delka - 49;
    pozice = delka - 49;
end
set(handles.sPosun,'Value',pozice);
sPosun_Callback(handles.sPosun,eventdata,handles);


function pbPredesly_Callback(~, eventdata, handles)
global usekpred
pozice = usekpred(1);
if pozice < 1;
    pozice = 1;
end
set(handles.sPosun,'Value',pozice);
sPosun_Callback(handles.sPosun,eventdata,handles);


function pbm10_Callback(~, eventdata, handles)
pozice = get(handles.sPosun,'Value');
pozice = pozice - 10;
if pozice < 1;
   pozice = 1;
end
set(handles.sPosun,'Value',pozice);
sPosun_Callback(handles.sPosun,eventdata,handles);


function pb10_Callback(~, eventdata, handles)
global delka
pozice = get(handles.sPosun,'Value');
pozice = pozice + 10;
if pozice > delka - 49;
   pozice = delka - 49;
end
set(handles.sPosun,'Value',pozice);
sPosun_Callback(handles.sPosun,eventdata,handles);
```

E OBSAH PŘILOŽENÉHO CD

latex/ Adresář se zdrojovými kódy bakalářské práce v přeložitelné podobě.

- kody/ Adresář se zdrojovými kódy uvedenými v textu práce.
- logo/ Adresář s logem fakulty.
- obrazky/ Adresář s obrázky použitými v práci.
- pdf/ Adresář se soubory .pdf vloženými do práce.
- text/ Adresář s texty práce.

matlab/ Adresář se zdrojovými kódy programu `Bakalarka.m`.

- sekvence/ Adresář s analyzovanými geny.
- useky/ Adresář s nalezenými úseky obsahujícími adenin a guanin.