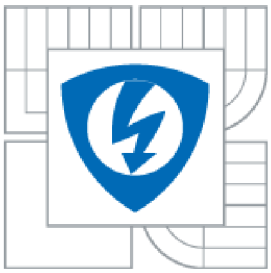
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGENIRING

VYHLEDÁVÁNÍ OBLASTÍ BOHATÝCH NA ADENIN A GUANIN

SEARCHING ADENINE AND GUANINE RICH REGIONS

BAKALÁŘSKÁ PRÁCE

BACHELOR´S THESIS

AUTOR PRÁCE — ALŽBĚTA VLACHYNSKÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor

Biomedicínská technika a bioinformatika

Studentka: Alžběta Vlachynská ***ID:*** 133991

Ročník: 3 ***Akademický rok:*** 2012/2013

NÁZEV TÉMATU:

Vyhledávání oblastí bohatých na adenin a guanin

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se se stavbou DNA a projektem lidského genomu. Blíže prostudujte literaturu zaměřenou na oblasti bohaté na guanin a adenin. 2) Navrhněte metodu pro vyhledávání adenin-guanin bohatých oblastí. 3) Pro navrženou metodu vytvořte pseudokód a vývojový diagram. 4) Navrženou metodu implementujte v libovolném programovém prostředí. 5) Pomocí programu proveďte analýzu části lidského, šimpanzího a myšího genomu na výskyt adenin-guanin bohatých úseku. Výsledky porovnejte a zhodnoťte. Zvažte náhodnost, funkci a konzervovanost nalezených úseku.

DOPORUCENÁ LITERATURA:

[1] HRABCOVA, I. a KYPR, J., The longest (A+T) and (G+C) blocks in the human and other genomes, Journal of Biomolecular Structure & Dynamics, 2008, roc. 25, s. 337-345.

[2] MOUNT, D.W. Bioinformatics. Cold Spring Harbour, New York, 2004.

Termín zadání: 11. 2. 2013 ***Termín odevzdání:*** 31. 5. 2013

Vedoucí práce: Ing. Denisa Maděránková

Konzultanti bakalářské práce:

prof. Ing. Ivo Provazník, Ph.D.

Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Práce se zabývá vyhledáváním oblastí s vysokým obsahem adeninu a guaninu v molekule DNA. Jejich nalezení a stanovení náhodnosti, funkčnosti a konzervovanosti by snad mohlo vést k lepšímu pochopení uložení genetické informace v sekvenci DNA. První část práce se věnuje složení a struktuře DNA, její replikaci, transkripci a translaci v buňce. Je v ní přiblížen pojem genom. Následuje kapitola, která se věnuje numerickému zpracování sekvenčních dat DNA, nezbytných k dalšímu zpracování za pomoci výpočetní techniky. Je uvedena metoda denzitních vektorů, vhodná pro vyhledání požadovaných úseků sekvence DNA. V následující části práce je navržena a pomocí programovacího prostředí MATLAB realizována aplikace pro vyhledávání oblastí bohatých na adenin a guanin. Poslední část práce tvoří analýza genů člověka, šimpanze a myši.

KLÍČOVÁ SLOVA

DNA, adenin, guanin, genové databáze, denzitní vektory

ABSTRACT

This work deals with searching adenine and guanine rich regions. Their finding which determines the randomness and functionality may lead to better understanding genetic data storage in DNA sequences. The first part of the work describes the chemical composition and structure of DNA, its replication, transcription and translation. It contains the basic information about the genome. The next chapter deals with numerical representation of DNA sequences, necessary for a computer processing. Nucleotide density is a method suitable for searching adenine and guanine rich regions. The practical part of the work is an application designed and implmented in MATLAB environment. The last part is the analysis of human, chimpanzee and mice genes.

KEYWORDS

DNA, adenine, guanine, gene database, nucleotide density

BIBLIOGRAFICKÁ CITACE

VLACHYNSKÁ, A. *Vyhledávání oblastí bohatých na adenin a guanin*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 52 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Vyhledávání oblastí bohatých na adenin a guanin jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne 27. května 2013

.....

podpis autora

PODĚKOVÁNÍ

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc, konzultace a další cenné rady při zpracování mé bakalářské práce. Dále bych ráda poděkovala svým rodičům za finanční i morální podporu při mých studiích.

V Brně dne 27. května 2013

.....

podpis autora

OBSAH

Úvod **8**
1. Teoretický náhled tématiky **9**
1.1. DNA 9
1.1.1. Historie výzkumu DNA 9
1.1.2. Struktura DNA 9
1.1.3. Replikace 11
1.1.4. Transkripce a translace 11
1.1.5. Oblasti bohaté na adenin a guanin 14
1.2. Genom (člověk, šimpanz, myš) 14
1.2.1. Mapování genomů, sekvenace 14
1.2.2. Projekt lidského genomu 16
1.2.3. Projekt myšího genomu 17
1.2.4. Projekt genomu šimpanze 17
1.3. Veřejné genové databáze 18
1.4. FASTA formát 19
2. Návrh metody vyhledávání A-G bohatých oblastí **20**
2.1. Numerická reprezentace 20
2.1.1. Denzitní vektory 21
2.2. Vývojový diagram 24
2.3. Pseudokód 25
3. Realizace algoritmu vyhledávání **27**
3.1. Rozbor programu 27
3.1.1. Funkce denzitni_vektory.m 27
3.1.2. Funkce urceni_pozic.m 29
3.1.3. Popis uživatelského prostředí GUI 31
4. Analýza genů **34**
4.1. Analýza vybraných genů 34
4.1.1. Gen ALB1 34
4.1.2. Gen HBB 37
4.1.3. Gen P53 39
4.1.4. Gen CDK2 41
4.1.5. Gen SHBG 43
4.2. Testování obsahu A-G bohatých oblastí v náhodné sekvenci 45
4.3. Vyhodnocení 46
Závěr **48**
Literatura **49**
Použité zkratky **51**
Obsah elektronické dokumentace **52**

SEZNAM OBRÁZKŮ

Obr. 1 Struktura čtyř běžných deoxyribonukleotidů přítomných v DNA [3] 10
Obr. 2 Znázornění dvoušroubovicové struktury DNA [4] 10
Obr. 3 Replikace DNA [5] 11
Obr. 4 Genetický kód [6] 12
Obr. 5 Transkripce na ribosomu [6] 13
Obr. 6 Sekvenování DNA metodou koncových terminátorů [2] 15
Obr. 7 Počítačový výstup nukleotidové sekvence [8] 16
Obr. 8 Loga největších veřejných genových databází 18
Obr. 9 Grafické znázornění sumy denzitních vektorů sekvence ACTCTTAGGC, W=5 22
Obr. 10 Sumy denzitních vektorů mitochondriálního genu COX1 dravce *Accipiter gentilis* dle biochemických vlastností nukleotidů, W = 25. [20] 23
Obr. 11 Vývojový diagram výpočtu denzitních vektorů ze sekvence DNA 24
Obr. 12 Vývojový digram zobrazující zpracování denzitních vektorů a algoritmu pro určení pozic v sekvenci s obsahem AG vyšším než práh 25
Obr. 13 Výstup funkce ***denzitni_vektory.m***, gen IRS1 – Homo sapiens, velikost okna W = 21 28
Obr. 14 Výstup funkce ***urceni_pozic.m***, gen IRS1 – Homo sapiens, velikost okna W = 21, prah = 0.9 29
Obr. 15 Uživatelské prostředí GUIDE 31
Obr. 16 GUI - nahrávání sekvence (sekvence IRS1 - Homo sapiens) 32
Obr. 17 GUI - výpočet denzitních vektorů a prahování (sekvence IRS1 -Homo sapiens) 33

SEZNAM TABULEK

Tab. 1 Příklad hlavičky formátu FASTA 19
Tab. 2 Podporované kódy nukleových kyselin 19
Tab. 3 Ukázka algoritmu arrayMax(A,n) v pseudokódu a v Javě [22] 25
Tab. 4 Procentuální obsah A-G bohatých oblastí v sekvenci genu ALB1 – bez ohledu na délku nalezeného úseku 35
Tab. 5 Procentuální obsah A-G bohatých oblastí v sekvenci genu ALB1 – nalezené úseky delší než W/2 36
Tab. 6 Procentuální obsah A-G bohatých oblastí v sekvenci genu ALB1 – nalezené úseky delší než W 36
Tab. 7 Procentuální obsah A-G bohatých oblastí v sekvenci genu HBB – bez ohledu na délku nalezeného úseku 37
Tab. 8 Procentuální obsah A-G bohatých oblastí v sekvenci genu HBB – nalezené úseky delší než W/2 38
Tab. 9 Procentuální obsah A-G bohatých oblastí v sekvenci genu HBB – nalezené úseky delší než W 39
Tab. 10 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 – bez ohledu na délku nalezeného úseku 40
Tab. 11 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 – nalezené úseky delší než W/2 40
Tab. 12 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53– nalezené úseky delší než W 41
Tab. 13 Procentuální obsah A-G bohatých oblastí v sekvenci genu CDK2 – bez ohledu na délku nalezeného úseku 42
Tab. 14 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 – nalezené úseky delší než W/2 42
Tab. 15 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 – nalezené úseky delší než W 43
Tab. 16 Procentuální obsah A-G bohatých oblastí v sekvenci genu SHBG – bez ohledu na délku nalezeného úseku 44
Tab. 17 Procentuální obsah A-G bohatých oblastí v sekvenci genu SHBG– nalezené úseky delší než W/2 45
Tab. 18 Procentuální obsah A-G bohatých oblastí v sekvenci genu SHBG – nalezené úseky delší než W 45
Tab. 19 Procentuální obsah A-G bohatých oblastí v náhodné sekvenci 46

Úvod

Vývoj genetiky zažívá v posledních letech rozmach. Objev DNA a rozvoj technologií umožňujících sekvenování dat v ní obsažených otevírají nové možnosti poznání. Vědci věří, že pokud se nám podaří plně pochopit způsob uložení informace v DNA, dostaneme odpověď na mnoho otázek týkajících se jak původu, tak i existence člověka. Již nyní jsou známy konkrétní mutace způsobující často velmi závažné choroby a probíhá aktivní vývoj genové terapie, kladoucí si za cíl najít způsob, jak na základě znalosti genů nemoci léčit – ať už tak, že daný gen „opravíme“ nebo ho „jen“ deaktivujeme a zabráníme tak nežádoucím projevům jeho transkripce. Mnohé nemoci jsou však multifaktoriální (ovlivňuje je více genů) a zatím není možné přesně určit mutaci, která je způsobuje. Navíc nám doposud není známo, jakým způsobem je regulována exprese genů. Ve výzkumu může být hodně nápomocná právě bioinformatika. Pomocí informačních technologií jsme schopni zpracovávat i poměrně velké množství informací a vyvozovat z nich závěry.

Informace je v DNA zakódována pomocí čtyř bází – adeninu (A), cytosinu (C), guaninu (G) a tyminu (T). Jak se však ukázalo, rozmístění jednotlivých bází v řetězci rozhodně není náhodné a často dochází k shlukování různých typů bází. Je možné, že právě takové úseky DNA by mohly indikovat přítomnost/nepřítomnost genů, či mít vliv na regulaci jejich exprese. Tato práce se zabývá vyhledáváním oblastí sekvence s vysokým obsahem A a G, s cílem zhodnotit náhodnost, funkci a konzervovanost nalezených úseků.

Součástí práce je program sloužící k vyhledávání A-G bohatých oblastí na základě výpočtu denzitních vektorů. Program je napsán v programovacím prostředí Matlab R2010a a obsahuje i uživatelské rozhraní. Proto uživatel prakticky vůbec nemusí přijít do styku s kódem programu. Pomocí tohoto programu je provedena analýza pěti genů člověka, šimpanze a myši na obsah A-G bohatých oblastí.

1. TEORETICKÝ NÁHLED TÉMATIKY

1.1. DNA

DNA neboli deoxyribonukleová kyselina je nositelkou genetické informace u naprosté většiny známých organizmů. Ve své struktuře kóduje informace o stavbě, funkci a reprodukci buněk, čímž předurčuje vývoj a vlastnosti celého organismu. Je také prostředkem, kterým jsou tyto informace předávány na další generace. [1]

1.1.1. Historie výzkumu DNA

Deoxyribonukleová kyselina byla popsána roku 1869 švýcarským lékařem Friedrichem Miescherem. O funkci DNA toho dlouho nebylo moc známo. Až v roce 1943 provedli Oswald Avery společně s Colinem MacLeodem a Maclynem McCaetym sérii pokusů s transformací pneumokoků (tz. Averyho-MacLeodův-McCartyho experiment) a zjistili, že DNA je genetickým materiálem buněk.

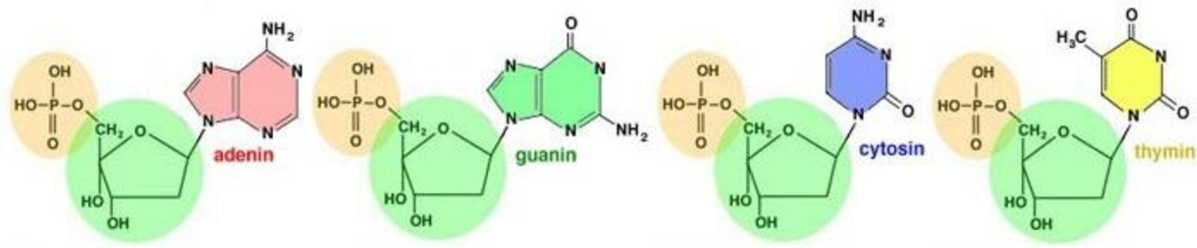
Patrně nejslavnějším milníkem ve výzkumu DNA bylo odhalení její trojrozměrné struktury. V roce 1953 byl v časopise Nature představen správný dvoušroubovicový model. Za tento objev byli autoři James D. Watson a Francis Crick v roce 1962 oceněni Nobelovou cenou za fyziologii a medicínu. Následně v roce 1957 předložil již slavný Crick sérii pravidel popisující vztahy mezi DNA, RNA a proteiny, dnes označovanou za centrální dogma molekulární biologie (viz kapitola 1.1.4).

O rok později se podařilo prozkoumat způsob replikace DNA v buňkách a na počátku 60. let byl rozluštěn genetický kód vědci H. Gobindem, R. W. Holleym a M. W. Nirenbergem. V letech 1972 až 1975 F. Stanger a W. Gilbert vyvinuli techniku sekvenování DNA, čím spustili celou novou éru výzkumu genetické informace. [1]

1.1.2. Struktura DNA

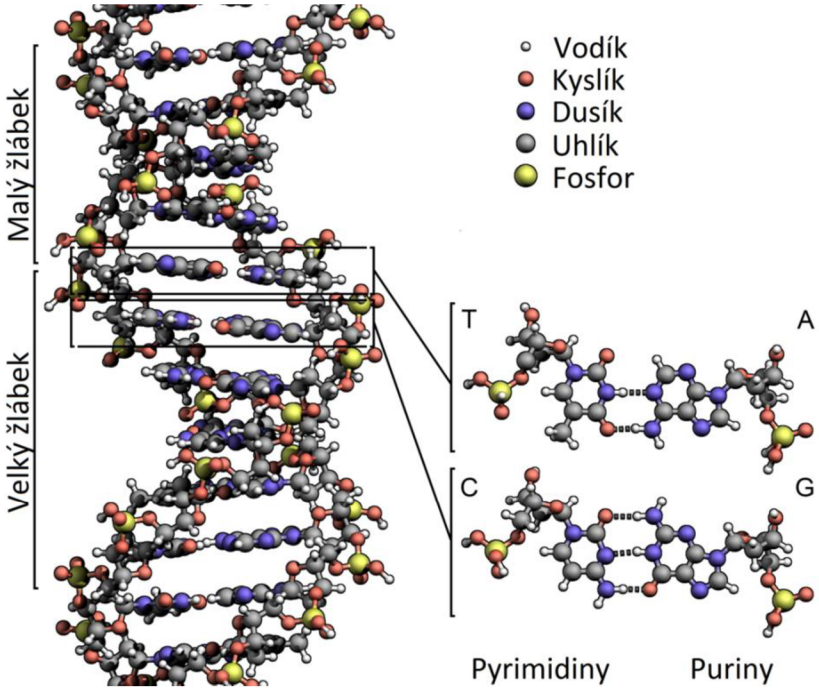
Stavbu DNA je možné zkoumat na několika úrovních. Pořadí nukleotidů v lineárním dvojvlákně nazýváme primární strukturou. Stáčení vlákna do typické, zpravidla pravotočivé, dvoušroubovice je záležitostí sekundární struktury. Pod terciální strukturou se pak obvykle rozumí nadšroubovicové vinutí, které usnadňuje kondenzaci DNA až do nadmolekulárních struktur, jakými jsou chromozomy. [2]

Každá nukleová kyselina je tvořena řetězcem nukleotidů, složených ze tří částí – fosfátovou skupinou, pětiuhlíkatým cukrem a cyklickou dusíkatou složkou zvanou báze. V DNA je cukrem 2-deoxyribóza a běžně se v ní vyskytují čtyři báze: adenin (A), cytosin (C), guanin (G) a tymin (T). Adenin a guanin jsou bicyklické báze zvané puriny, cytosin a tymin jsou monocyklické báze nazývané pyrimidiny. Jejich chemická struktura je vidět na Obr. 1. [1]



Obr. 1 Struktura čtyř běžných deoxyribonukleotidů přítomných v DNA [3]

Polynukleotidové řetězce jsou složeny ze sekvence nukleotidů spojených navzájem fosfodiesterovými vazbami, které propojují sousední deoxyribózové skupiny. Dvě polynukleotidová vlákna se otáčí kolem sebe ve spirále, přičemž jsou držena pohromadě ve šroubovicové konfiguraci vodíkovými vazbami mezi bázemi opačných vláken; výsledné páry bází jsou uspořádány nad sebou mezi dvěma řetězci kolmo k ose molekuly jako schody točitého schodiště. Párování je specifické a vyplývá ze schopnosti tvorby vodíkových vazeb bázemi v jejich normálních konfiguracích, guanin se vždy páruje s cytosinem třemi vodíkovými můstky a adenin tvoří s tyminem dvě vodíkové vazby. Všechny páry bází se tak skládají z jednoho purinu a jednoho pyrimidinu. [1]



Obr. 2 Znázornění dvoušroubovicové struktury DNA [4]

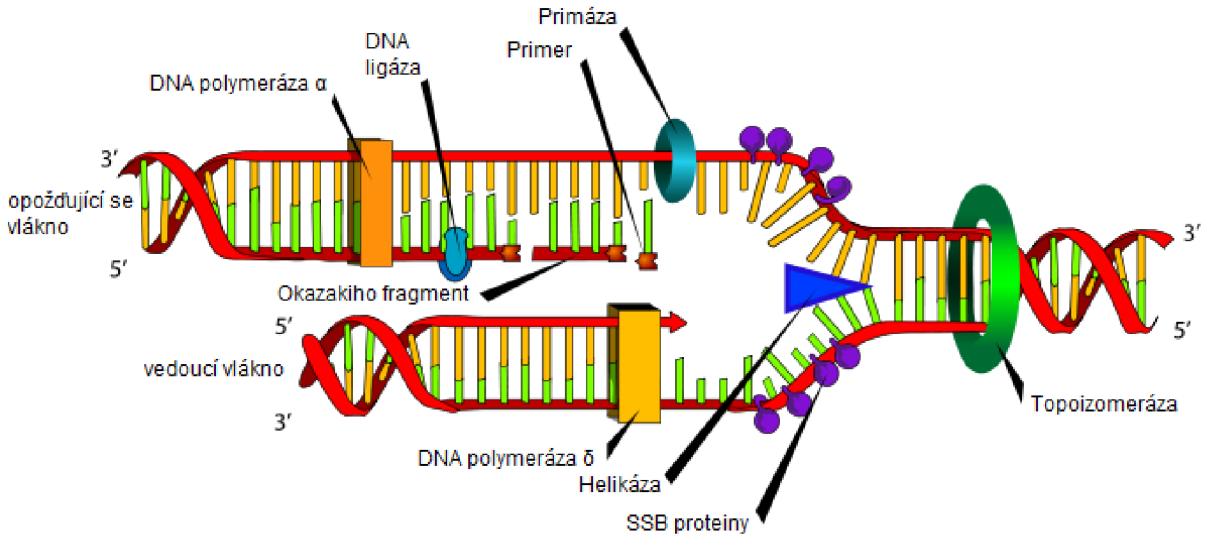
Jakmile je známa sekvence bází v jednom vlákně je možné na základě specifického párování bází určit i sekvenci druhého vlákna. Tato vlastnost DNA se nazývá komplementarita a poskytuje jedinečný nástroj pro uložení a přenos genetické informace z generace na generaci. (kapitola 1.1.3). [1]

Všechny funkční molekuly DNA, obsažené v živých buňkách jsou ještě nadšroubovicově vinuty (superspiralizovány). Během metafáze miózy a mitózy je DNA za pomoci proteinů ještě více sbalena a kondenzována do kompaktních chromozomů. [1]

1.1.3. Replikace

Replikace je proces, při kterém dochází k zmnožení molekul DNA, přičemž z jedné mateřské molekuly vzniknou dvě naprosto stejné dceřiné, každá s jedním vláknem původní DNA a jedním vláknem nově syntetizovaným, proto se tento proces nazývá semikonzervativní replikace. Vlákna mateřské DNA slouží jako templát a na základě párování bází jsou k nim postupně přidávány komplementární báze do vlákna dceřiného, tj. pořadí bází v každém dceřiném vlákně je podmíněno schopností bází mateřského vlákna tvořit vodíkové vazby. [1] [2]

Replikace se uplatňuje při rozmnožování, kdy zajišťuje identitu obou dceřiných buněk. U člověka probíhá syntéza nového řetězce DNA rychlostí přibližně 3000 nukleotidů za minutu. Aby byl proces efektivnější, neprobíhá replikace postupně po celém vlákně, ale hned na několika místech najednou. Jelikož ale hlavní enzym replikace DNA-polymeráza může katalyzovat syntézu pouze ve směru 5´→3´ konci je jedno vlákno tvořeno průběžně a další přerušovaně. Opožďující se vlákno roste syntézou krátkých fragmentů zvaných Okazakiho fragmenty (viz Obr. 3). [1] [2]

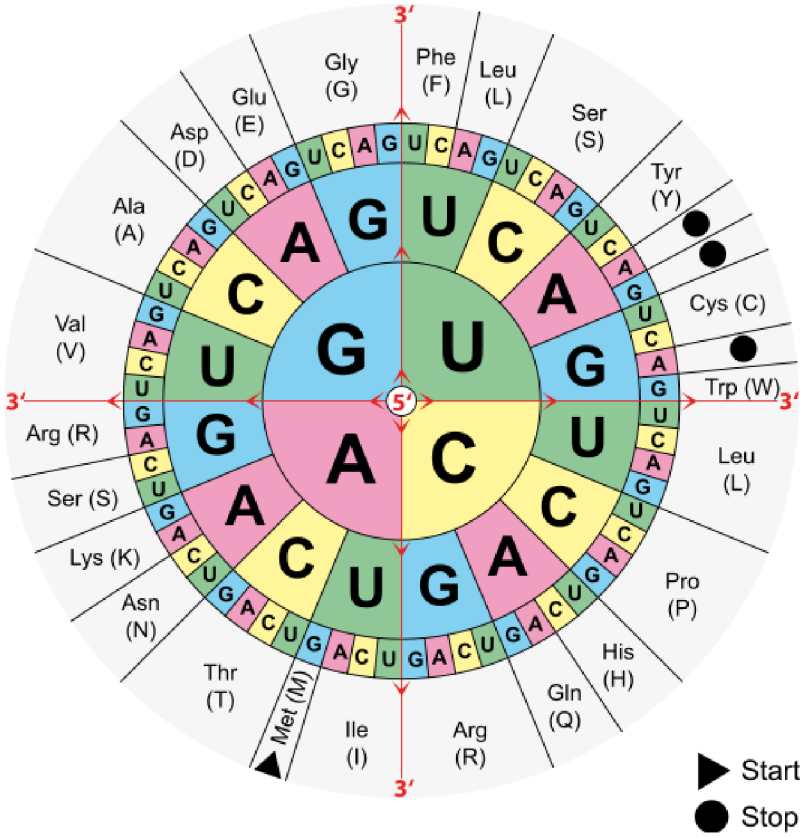


Obr. 3 Replikace DNA [5]

Celý proces je mnohem složitější, ale není objektem této práce, proto se jím nebudu zabývat. Mohu však ještě poznamenat, že přes složitost a současně velkou rychlost replikace je její přesnost opravdu ohromující – průměrně jedna chyba na miliardu včleněných nukleotidů. [1] [2]

1.1.4. Transkripce a translace

Centrální dogmata molekulární biologie popisuje cestu přenosu informace mezi biopolymery. Říká, že informace uchovaná v DNA se transkripcí přenáší do molekul RNA a translací do proteinů. Exprese genetické informace tedy probíhá ve dvou krocích. Při transkripci se jedno vlákno DNA použije jako templát pro syntézu vlákna RNA, označovaného jako transkript. Při translaci je následně nukleotidová sekvence RNA převedena do sekvence aminokyselin polypeptidového produktu genu. Tento převod je řízen genetickým kódem, který určuje aminokyseliny podle tripletů v transkriptu, zvaných kodony (Obr. 4). [1] [2]



Obr. 4 Genetický kód [6]

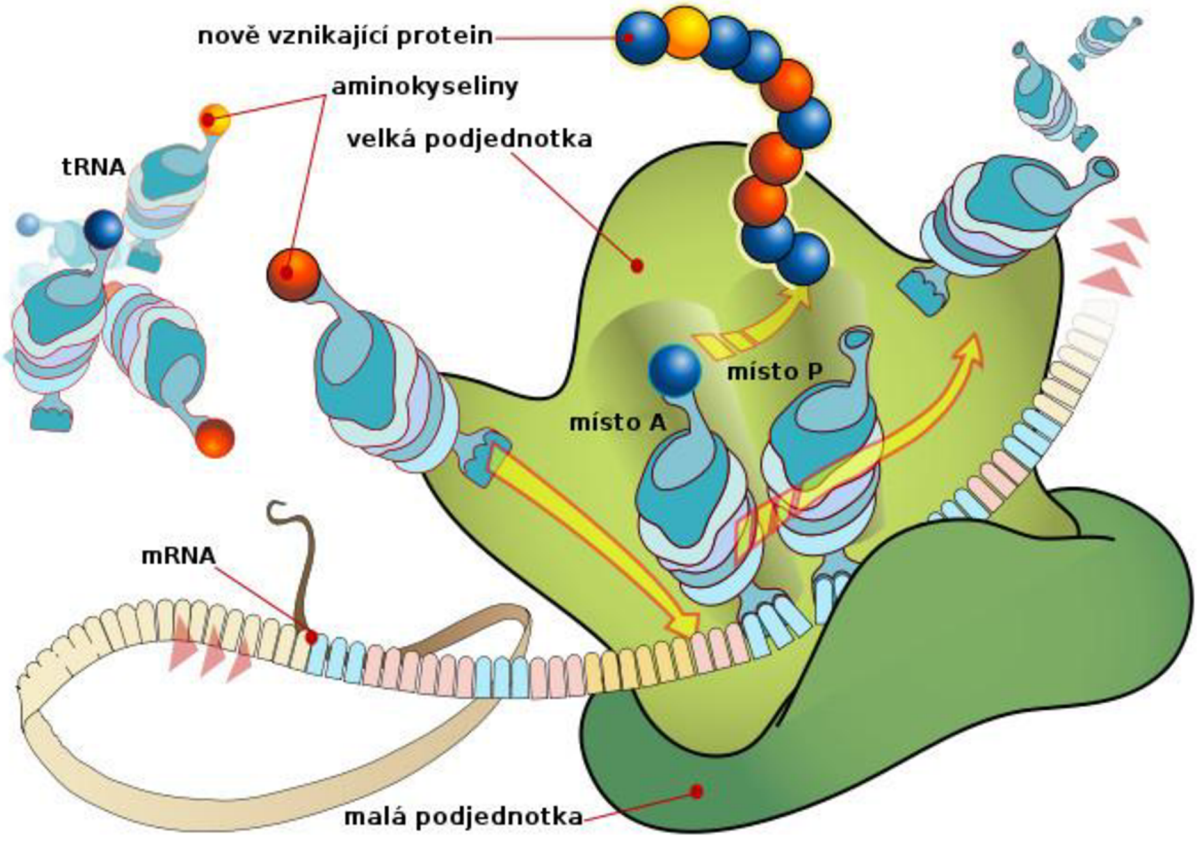
Translace probíhá v ribozomech, složitých makromolekulárních ústrojích složených ze tří až pěti molekul RNA a 50-90 různých proteinů. Molekuly RNA překládané v ribozomech nazýváme mediátorová RNA (mRNA). U prokaryot se mRNA shoduje s primárním transkriptem. U eukaryot však většina jaderných genů obsahuje nekódující sekvence zvané introny, které oddělují kódující sekvence zvané exony. Proto bývají primární transkripty (prekurzory mRNA neboli pre-mRNA) ještě v jádře zpracovány - introny jsou odstraněny sestřihovými reakcemi v makromolekulárních strukturách zvaných spliceozomy. Teprve takto upravená mRNA vystupuje z jádra ven do cytoplazmy, kde na ribozomech probíhá translace. [1] [2]

Translaci nukleotidové sekvence v molekule mRNA do aminokyselinové sekvence polypeptidového produktu lze rozdělit do tří fází: iniciace, elongace a terminace polypeptidového řetězce. [1] [2]

Syntézu polypeptidů zahajuje speciální transferová RNA (tRNA) jako odpověď na iniciační translační kodon (obvykle AUG a někdy GUG). Proto syntéza všech polypeptidů začíná methiolinem, který se ale z polypeptidů často později odštěpuje, takže funkční proteiny nemusí mít na aminokonci methionin. Spolu s tRNA se na mRNA naváží iniciační faktory a malá podjednotka ribozomu. U eukaryot je následná iniciace translace složitější než u prokaryot. Iniciační komplex prohlíží mRNA od 5´-konce a hledá kodon AUG pro iniciaci translace, přičemž účinnost, s jakou se daný kodon AUG pro iniciaci využívá, závisí na přiléhající nukleotidové sekvenci. Optimální iniciační sekvencí je 5´-GCC(A nebo G)CCAUGG-3´. Purin (A nebo G) umístěný tři báze proti směru translace od iniciačního kodonu a G, který za ním

bezprostředně následuje, jsou nejdůležitější – ovlivňují účinnost translace více než desetinásobně. Jakmile je triplet nalezen, iniciační faktory komplex opouští a místo nich se naváže velká podjednotka, čímž vzniká úplný ribozom. Komplex ribozom/mRNA/tRNA je připraven zahájit druhou fázi translace – elongaci. [1] [2]

Přidání každé aminokyseliny do rostoucího polypeptidu vyžaduje tři kroky. Nejdříve navázání aminoacyl-tRNA (tRNA s aminokyselinou odpovídající kodonu čtenému na mRNA → tRNA obsahuje tzv. antikodon) do místa A ribosomu. Druhým krokem elongace je tvorba peptidové vazby mezi aminoacyl-tRNA v místě A a karboxylovým koncem rostoucího polypeptidového řetězce připojeného k tRNA v místě P. Během třetí fáze dochází v důsledku posunu ribozomu o tři nukleotidy směrem k 3´-konci molekuly mRNA k translokaci peptid-tRNA z místa A do místa P a tRNA z místa P do místa E, kde se odštěpí od aminokyseliny již navázané na polypeptidový řetězec. Posunem se navíc uvolnilo místo A a ribozom je připraven zahájit další cyklus elongace řetězce (Obr. 5). [1] [2]



Obr. 5 Transkripce na ribosomu [6]

Elongace přechází do třetí fáze terminace, jakmile do místa A vstoupí kterýkoliv ze stop kodonů (UAA, UAG nebo UGA). Tyto tři kodony jsou rozeznány proteiny zvanými uvolňovací faktory, které způsobí, že ke karboxylovému konci vznikajícího polypeptidového řetězce je navázána molekula vody. Tím se ukončí syntéza polypeptidu, který je následně spolu s mRNA uvolněn z ribozomu. Terminace je dokončena rozpadem ribozomu na podjednotky. [1] [2]

1.1.5. Oblasti bohaté na adenin a guanin

Molekula DNA obsahuje, jak již bylo zmíněno výše, čtyři báze, z toho dvě purinové a dvě pyrimidinové, přičemž se páruje vždy báze purinová s bází pyrimidinovou. Čímž je dáno, že jejich obsah v molekule DNA musí být stejný. Rozložení purinových a pyrimidinových bází mezi jednotlivé řetězce DNA však už stejnoměrné být nemusí a také často není. V genomech různých organismů i člověka se nacházejí překvapivě dlouhé úseky, kde se na jednom řetězci shlukují báze A a G (purinové) a na druhém komplementární báze C a T (pyrimidinové). Takovéto úseky se často vyskytují před geny. I samotné geny bývají podobně asymetrické, přičemž kódující informace je zpravidla uložena na řetězci bohatém na A a G. [7]

Shluky purinových bází mohou utvářet několik sekvenčních motivů. Nejznámější jsou shluky guaninů tvořící tetraplexovou DNA. Vysoký obsah purinů v řetězci zpravidla přispívá ke konformační stabilitě DNA. V dinukleotidu ApG se sousední páry mezi řetězci překrývají a v daném místě je tak DNA extrémně zavinutá. [7]

Předpokládá se, že různý obsah purinových a pyrimidinových bází na jednotlivých řetězcích DNA má význam v regulaci molekulárně biologických procesů (transkripce genu, replikace DNA apod.). Podobným mechanismem by možná mohly být tyto procesy ovlivňovány i uměle.

1.2. Genom (člověk, šimpanz, myš)

Genom je veškerá genetická informace uložená v DNA konkrétního organismu. Úžeji jej můžeme chápat jako kompletní sekvenci jaderné DNA – tedy jedné sady chromozomů – avšak může do něj být zahrnuta navíc i kompletní DNA obsažená v buněčných organelách (u živočichů se jedná o mitochondriální DNA). [1] [2]

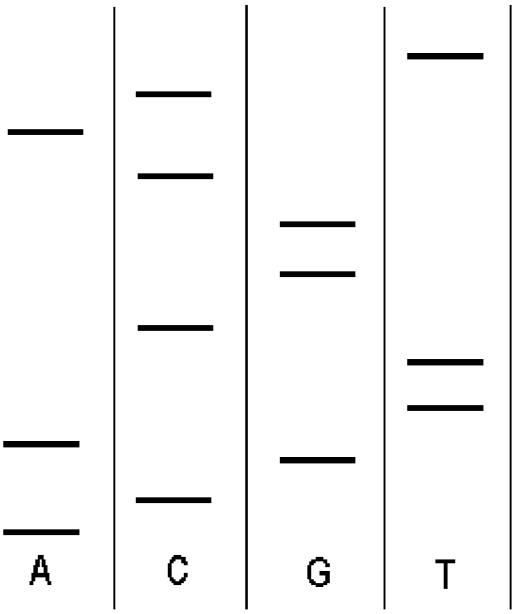
1.2.1. Mapování genomů, sekvenace

Při zkoumání genomu je potřeba zjistit následující fakta. V prvé řadě kolik má organismus chromozomů. Za druhé, na kterém chromozomu a kde na něm se který gen nachází. A konečně za třetí, v současné době je již celkem běžná i sekvenace genomu, při které zjistíme kompletní sekvenci molekul DNA. [2]

Naše dnešní schopnost sekvenovat v podstatě jakoukoliv molekulu DNA je výsledkem čtyř hlavních objevů. Nejdůležitějším průlomem byl objev restrikčních enzymů a jejich schopnost vyštěpit z chromozomu specifický úsek. Další velký pokrok představovalo zdokonalení gelové elektroforézy až tak, že mohou být rozděleny fragmenty DNA lišící se jen o délku jednoho nukleotidu. Stejně tak důležité bylo zdokonalení klonování genů, které umožní přípravu velkého množství konkrétních molekul. K tomuto účelu se využívá PCR (z anglického polymerase chain reaction = polymerázová řetězová reakce). A konečně vědci vynalezli postup, kterým mohou být stanoveny nukleotidové sekvence DNA. Nejčastěji se využívá postup vyvinutý v sedmdesátých letech minulého století Frederickem Sangerem a jeho kolegy, označovaný jako Sangerova metoda. [1] [2]

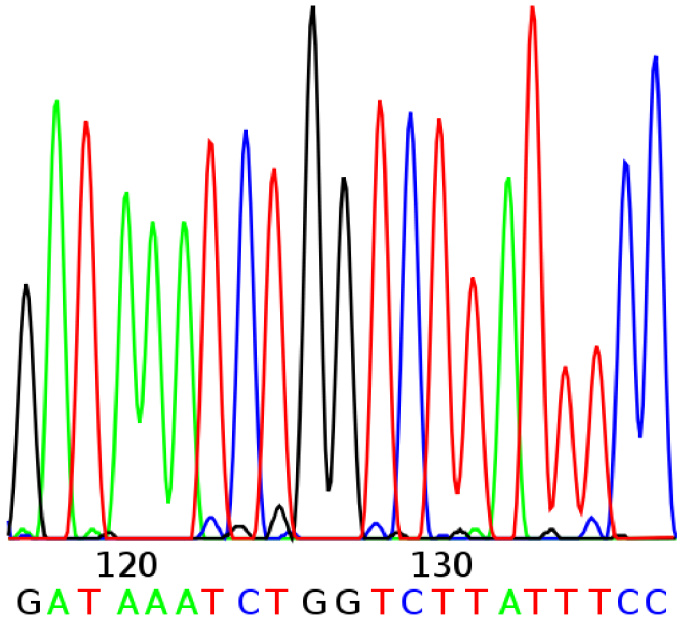
Sangerova metoda využívá in vitro syntézy DNA v přítomnosti upravených 2´,3´-dideoxyribonukleozidtrifosfátů (ddATP, ddCTP, ddGCP a ddTTP), které nemají na 3´uhlíku OH skupinu. Jestliže jsou zařazeny do prodlužujícího se řetězce syntetizované DNA, syntéza nemůže dále pokračovat a je ukončena. Pokud použijeme ve čtyřech oddělených reakcích pokaždé jiný dideoxyribonukletid jako terminátor (v jedné směsi ddATP, v další ddCTP atd.) budou vytvořeny čtyři soubory fragmentů, kde každý bude obsahovat řetězce, které budou vždy ukončeny stejnou bází (A, C, G nebo T). [1] [2]

V dané reakci je udržován poměr dXTP : ddXTP (kde X je jedna ze čtyř bází) přibližně 1:100, takže pravděpodobnost ukončení řetězce je 1/100. Vznikne tak soubor fragmentů končících ve všech možných místech sekvenovaného řetězce DNA. Poté jsou řetězce separovány polyakrylamidovou gelovou elektroforézou a jejich pozice v gelu se vizualizují autoradiografií. Nejkratší fragment urazí největší vzdálenost a dá vzniknout pruhu nejblíže k pozitivní elektrodě, každý následující pruh obsahuje řetězec o jeden nukleotid delší než předchozí. Toto je provedeno pro všechny čtyři soubory fragmentů. Postupným čtením pruhů na gelu, kde pozice pruhu určuje pozici v řetězci a to z jaké sady fragment pochází, udává druh báze na tomto místě, zjistíme pořadí bází ve studované sekvenci. (viz Obr. 6) [1] [2]



Obr. 6 Sekvenování DNA metodou koncových terminátorů – sekvence TCACGGCTTAGCA [2]

V současnosti jsou sekvenace DNA zpravidla prováděny na automatických sekvenačních přístrojích, které využívají metodu zmíněnou výše, jen lehce modifikovanou. Dideoxyribonukleotidy jsou značeny fluorescenčním barvivem díky čemuž je možná separace všech produktů v jedné dráze gelu či kapiláře, druh báze je určován na základě barevného značení. To může být pomocí fotobuňky snímáno a přenášeno do počítače, který analyzuje, zaznamenává a tiskne výsledky. (ukázka výstupu počítače viz Obr. 7) [1] [2]



Obr. 7 Počítačový výstup nukleotidové sekvence [8]

Značná automatizace, zrychlení a v neposlední řadě i zlevnění sekvenace zapříčinily obrovský nárůst množství sekvenovaných dat. V dnešní době se jich denně hromadí velké množství na veřejných internetových stránkách, kde jsou přístupné komukoli, kdo je chce použít.

1.2.2. Projekt lidského genomu

Projekt lidského genomu byl zahájen v roce 1990 s cílem sekvenovat celý lidský genom – asi tři miliardy nukleotidových párů – do roku 2005. Pod vedením Francise Collinse spolupracovalo na mezinárodním projektu osm výzkumných týmů ve Spojených státech amerických a jeden tým ve Velké Británii. V květnu 1998 však J. Craig Venter omráčil celý svět oznámením, že se spojil s firmou PerkinElmer Corporation a vytvořili novou společnost Celera Genomics, která chce skenovat lidský genom za pouhé tři roky. Závod o to, kdo identifikuje, sekvenuje a patentuje důležité geny, začal. [1]

Venterovo prohlášení bylo založeno na dvou klíčových vymoženostech. Prvním byla celogenomová „shotgun“ sekvenace, která je založena na rozložení celého genomu na menší části, skenování pouze konců těchto fragmentů a nakonec užití supervýkonného počítače pro sestavení kompletní sekvence, tak že najdou překrývající se části jednotlivých fragmentů. Výpočet, který provedl 500 miliónů biliónů porovnání dvojic bází, vyžadoval 64 gigabytů sdílené paměti superpočítače společnosti Celera Genomics a zabral více než 20 000 hodin procesoru. Vědci věří, že šlo o dosud vůbec největší "biologický" výpočet v historii lidstva. [1] [9]

Trocha konkurence a technologický pokrok projekt urychlily. 15. února 2001 (tedy o čtyři koky dříve než bylo v plánu) publikovalo The International Human Genome Sequencing Consortium podporované z veřejných prostředků svou první verzi sekvence lidského genomu v časopise Nature. Den na to 16. února 2001, soukromě financovaná skupina Celera Genomics publikovala svou verzi v časopise Science. Konsorcium pokračovalo ve své práci a v říjnu 2004

vydalo téměř kompletní sekvenci genomu pokrývající přibližně 99% euchromatické DNA. Překvapením bylo zjištění, že odhadovaný počet genů byl menší, než se odhadovalo (cca 30 000 proti předpokládaným 50-100 tisíci). [1]

Projekt lidského genomu byl jedinečný nejen vysokou úrovní mezinárodní spolupráce, ale také v tom, že se jako první velký vědecký projekt začal vážně zabývat nejen výzkumem jako takovým, ale také etickými, právními a sociálními otázkami spojenými se sekvenací genomu. Mezi nejčastěji zmiňované otázky patří, jaký přístup budou mít ke genetickým informacím pojišťovny, zaměstnavatelé, soudy, školy i jiné instituce a jakým způsobem bude tento přístup využíván. Dalším důležitým tématem je otázka důvěrnosti a ochrany genetických informací. Kdo ji zajistí? Neopominutelný je také psychologický dopad nově přístupných informací na jednotlivce i celou společnost. Tyto a podobná témata budou v budoucích letech nejspíše čím dál více aktuální a bude třeba se s nimi vyrovnat. [9] [10]

1.2.3. Projekt myšího genomu

Kromě lidského genomu byly sekvenovány také genomy mnohých dalších organismů. Jako první eukaryotický organismus byly sekvenovány pekařské kvasinky Saccharomyces cerevisiae v roce 1996, následovaly je další modelové systémy jako hlístice (Caenorhabditis elegans) v roce 1998 a v roce 2000 byly zveřejněny genomy octomilky (Drosophila melanogaster) a modelové rostliny Arabidopsis thaliana. Poté následovalo mapování genomů savců, např. psa, potkana a také myši a šimpanze. [1]

Myš (mus musculus) je nejběžnějším modelovým organismem pro biologii savců, lidské nemoci a evoluci genomů. Její genom byl sekvenován a výsledky byly zveřejněny v časopise Nature už v roce 2002. Sekvence genomu myši umožňuje lepší pochopení lidského genomu a důležitý nástroj pro biomedicínský výzkum. V současnosti jsou myši využívány k modelaci lidských dědičných chorob. Jako první je pro tento účel použil Oliver Smithies. Takovýmto modelem byla myš nesoucí stejnou mutaci, jaká u člověka vyvolává cystickou fibrózu. V dnešní době přesáhl počet myších modelů lidských dědičných chorob pět stovek. [11] [12]

1.2.4. Projekt genomu šimpanze

Pan troglogites neboli šimpanz je spolu s dalšími lidoopy nejbližší příbuzný člověka, proto je důležitým modelem pro studium biologie, nemocí a evoluce. Výzkum šimpanze poskytl nový vhled do poznání některých nemocí, např. AIDS, cukrovky či žloutenky. Rozdíly mezi šimpanzím a lidským genomem mohou definovat genetický základ vlastností, které odlišují člověka od ostatních primátů. [13]

V září roku 2005 byl v časopise Nature publikován článek o hrubém přečtení genomu šimpanze. Na tomto úkolu pracovalo 67 vědců z celkem 23 ústavů sdružených v Chimpanzee Sequencing and Analysis Consortium. Výsledky práce potvrzují jen minimální rozdíly mezi tímto lidoopem a člověkem. V genech (části genomu obsahující exony) činí rozdíl pouhého 1,2%. Rozdíl naroste na 4%, pokud vezmeme v úvahu i introny. Zajímavostí je, že nejvíce odlišností mezi oběma tvory se objevuje na samčím chromozomu Y, nejméně jich vědci našli na samičím

chromozomu X. Ačkoli se rozdíl mezi genomy zdá maličký, představuje ve více než 3 miliardách písmen milióny míst, kde se oba tvorové ve své DNA odlišili. Zatím však není jasné, které odchylky mají jaký význam. [14]

Je známo více teorií o genetické podstatě vzniku člověka. První počítá s tím, že se mutacemi změnily geny a s nimi i bílkoviny, které umožnili člověku evoluční úspěch, případně že člověku pomohla na evoluční výsluní ztráta některých genů, které bránily v evoluci našim zvířecím předkům. Jiná teorie tvrdí, že se geny příliš nezměnily, ale zato prošla významnou proměnou jejich regulace. Řečeno nadneseně, lidský genom pak podle této teorie hraje se stejným ansámblem genů jinou písničku než šimpanzí. Právě tato teorie by podle většiny odborníků mohla vysvětlovat velkou část genetických příčin vzestupu člověka. [14]

1.3. Veřejné genové databáze

Základy bioinformatiky byly položeny v 80. letech minulého století, kdy bylo vědecké úsilí vnímáno jako záležitost především veřejná (a také bylo zpravidla financováno z veřejných rozpočtů či alespoň z neziskových zdrojů) a měřítkem kvality výzkumu byly publikace, nikoliv patenty. Zřejmě z toho důvodu je vůbec možné, že se volné sdílení sekvenčních dat stalo standardní praxí, od níž už (snad) nelze upustit. Významné časopisy dnes po autorech publikujících sekvenční data vyžadují, aby je zpřístupnili odborné veřejnosti. [15]

V případě zveřejňování sekvencí DNA je standardem trojice primárních databází mezinárodního konsorcia (Internacional Nucleotide Sequnce Database Collaboration). Jedná se o americkou databázi GenBank provozovanou NCBI (the National Center for Biotechnology Information), evropskou EMBL (European Molecular Biology Laboratory Data Library) a japonskou DDBJ (DNA Data Bank od Japan). Jejich obsah je téměř totožný, jelikož si denně vyměňují veškeré změny v datech. V primárních databázích bylo v dubnu 2011 uloženo zhruba $191{,}4 \times 10^9$ párů bází v 135×10^6 sekvencích pocházejících z více než 380 000 organizmů. Vzhledem k tomu, že produkce sekvečních dat neustále narůstá, je údržba primárních databází úctyhodným úkolem. [1] [15]

Obr. 8 Loga největších veřejných genových databází

Aby databáze plnily svou funkci, musí být možné data v nich uložené aktivně třídit a vyhledávat v nich informace podle předem zvolených kritérií. K vyhledávání a stahování záznamů z databází slouží webové uživatelské rozhraní. Umožňuje vyhledávat záznamy na základě přístupových kódů, klíčových slov v názvech a anotacích, jmen autorů a literárních citací. Je

možné také putovat mezi záznamy, případně mezi databázemi, pomocí křížových odkazů. Dále jsou také k dispozici pokročilejší nástroje, např. prohledávání na základě podobnosti sekvencí samotných (funkce BLAST a pod.). [15]

1.4. FASTA formát

FASTA je softwarový balík sloužící při zarovnávání nukleotidových a aminokyselinových sekvencí. Poprvé byl popsán jako FASTP D. J.Lipmanem a W. R. Pearsonem v roce 1985, původně jako algoritmus pro vyhledávání podobností v proteinových sekvencích. [15] [16]

Sekvence se formátu FASTA, též nazývaném Pearsonův formát, je běžný textový soubor. Začíná jednořádkovým popisem a následují řádky sekvenčních dat. Popisný řádek je odlišen úvodním znakem > a slouží jako jedinečný identifikátor sekvence. Přesný formát hlavičky sekvence není dán. Obvykle obsahuje zkratku databáze, ze které sekvence pochází, identifikační kód sekvence v databázi, název sekvence, umístění a lokus genu. Ukázky možných formátů hlavičky jsou vidět v Tab. 1. [15] [16]

Tab. 1 Příklad hlavičky formátu FASTA

Název databáze	**Hlavička**
GenBank	gi|gi-number|gb|accession|locus
EMBL Data Library gi|gi	gi|gi-number|emb|accession|locus
DDBJ, DNA Database of Japan	gi|gi-number|dbj|accession|locus
GenBank	gi|gi-number|gb|accession|locus
Protein Research Foundation	prf||name
SWISS-PROT	sp|accession|name

Po hlavičce následují sekvenční data reprezentovány ve standardních kódech aminokyselin a nukleotidů dle IUB / IUPAC (viz http://www.iubmb.org/ a http://www.iupac.org/). Podporované kódy nukleových kyselin jsou uvedeny v Tab. 2 [16] [17]

Tab. 2 Podporované kódy nukleových kyselin

A	→	adenin	M	→	A C (amino skupina)
C	→	cytosin	S	→	G C (silná interakce)
G	→	guanin	W	→	A T (slabá interakce)
T	→	tymin	B	→	G T C
U	→	uracil	D	→	G A T
R	→	G A (purin)	H	→	A C T
Y	→	T C (pyrimidin)	V	→	G C A
K	→	G T (keto skupina)	N	→	A G C T (kterékoliv)

Ukázka části zápisu sekvence ve formátu FASTA:

```
>gi|291061370:c6360199-6181755 Pan troglodytes chromosome 6 - 2.1.4
AAGGGCGCACCTGCCGGGATAACAGGCCAGATGAAGTAAATAGAAAATCATCTGAGCTCCCCTACTGGCT
CCAGCTGTGGAGAAGGGGGAGGAGAAAACCCTGTGGGACAGGGGAGGAGGGTGAGGGCTCCTCTTAGGAA
GTTATTTAAGAGCCAACTGTCTTGTCTTTCCCGAGTCCGTTTGAGGAAGTCCCCGAGGCGCACAGAGCAA
GCCCACGCGAGGGCACCTCTGGAGGGGTGCGCCTGCAGGTAAGCCACCGCCCCTCCACCCTATGAGCCAG
GGCCCGCTGCGTCCACCTTCTGCACCTCGGTTTCCTGGTTGAACCAGCAAGCGGCTTGCTCTGGGCCCTG
TGGCGCCGGTCACAGGCAGCTCCACTTGCCCAATCCTGGCTTCCCGCCCCCAACTCCGCACCTGCCCAGC
TTATGGAGCTGAAACACAAAAGGTGGAGGGAGTGGGGCCAAGTGAGAACGCAGGAGGGGGCTGGAACCCC
GCCGGCTTGCTCCCCACTCCGCGCGAATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

2. NÁVRH METODY VYHLEDÁVÁNÍ A-G BOHATÝCH OBLASTÍ

2.1. Numerická reprezentace

Cílem této práce je návrh metody pro vyhledávání adenin-guanin bohatých oblastí. Pro počítačové zpracování genomické sekvence je zpravidla nutné ji nejdříve převést ze symbolického zápisu do numerického. Metody, jakými je sekvence DNA převedena z písmen do číselné abecedy, jsou různé. Nejjednodušším řešením, které se nabízí je písmena A, C, G, T nahradit reálnými čísly 1, 2, 3, 4. Toto řešení však není nejvhodnější, pro zavádějící číselné hodnoty jednotlivých bází, které nejsou nijak biologicky podložené. Pro jednoduché operace se však dá tato metoda použít. Jiný poměrně jednoduchý způsob spočívá v nahrazení písmen komplexními čísly, kdy A=1+j, C=-1-j, T=1-j, G=-1+j. Výhodou této metody je, že komplexní čísla v sobě mohou skrývat informaci o pozici v prostoru, což je možné využít pro grafické zobrazení sekvence. [18] [19]

Dalším způsobem zpracování genomických dat je binární reprezentace. Při této metodě se z jedné posloupnosti bází vytvoří čtyři vektory osahující nuly a jedničky. Každý řetězec odpovídá jednomu písmenu (bázi), přičemž jednička v daném řetězci odpovídá přítomnosti odpovídajícího písmene (báze) na dané pozici a nula znamená, že se dané písmeno (báze) na této pozici v řetězci nevyskytuje. Takto vzniklým řetězcům se říká indikační vektory a zpravidla se značí $uA[n]$, $uC[n]$, $uG[n]$, a $uT[n]$. Tuto reprezentaci jsem si vybrala pro mé zpracování, proto pro názornost uvádím příklad výpočtu indikačních vektorů k sekvenci ACTCTTAGGC: [18] [19]

uA=1000001000

uC=0101000001

uG=0000000110

uT=0010110000

Pro binární reprezentaci se čtyřmi vektory se používá označení 4D. Existují však i 3D binární reprezentace, kdy je sekvence interpretována pouze třemi vektory aniž by docházelo ke ztrátě informace. 3D reprezentace se využívají pro grafické znázornění v RGB spektru. Já ve své práci použiji 4D binární reprezentaci. [18] [19]

2.1.1. Denzitní vektory

Mým úkolem je nalézt úseky sekvence DNA s vysokým obsahem A a G. Možností, jak je identifikovat, je sledovat lokální zastoupení jednotlivých nukleotidů – tzv. nukleotidovou denzitu. Ta vyjadřuje průměrné zastoupení jednotlivých nukleotidů v dané části sekvence. Denzita nukleotidů je jednoduchá a efektivní metoda numerické reprezentace symbolické sekvence DNA. Při jejím výpočtu jsou nejdříve vytvořeny indikační vektory (viz kapitola 2.1) a následně spočteny denzitní vektory podle následujícího vzorce: [20]

$$d_X[n] = \frac{\sum_{i=n-\frac{W}{2}}^{n+\frac{W}{2}} u_X[i]}{W} \qquad pro\ n = 1 \dots N, \tag{1}$$

kde N je délka sekvence, W je velikost okna, X je typ nukleotidu. Velikost posuvného okna W musí být liché číslo, protože n je pozice centrálního prvku okna. Posuvné okno se pohybuje po celé délce indikačního vektoru a na pozici n vrátí průměr hodnot obsažných v okně. Pro eliminaci vlivu začátku a konce indikačních vektorů se na jejich začátek a konec přidává $W/2$ nul (podíl zaokrouhlujeme vždy dolů). Tímto způsobem získáme sadu čtyř denzitních vektorů. Pro názornost uvádím příklad výpočtu denzitních vektorů pro sekvenci uvedenou výše (v kapitole 2.1), hodnotu okna zvolíme W=5, délka sekvence N=10. V prvé řadě doplníme indikační vektory o $W/2$ nul, na začátku i konci tak přidáme dvě nuly: [20]

uA=**00**100000100**00**

uC=**00**010100000**100**

uG=**00**000000011**00**

uT=**00**001011000**00**

Nyní budeme vektory postupně procházet oknem o velikosti 5 a počítat průměr hodnot v něm obsažených. První hodnotu denzitního vektoru pro adenin spočteme tak, že vezmeme prvních pět hodnot rozšířeného vektoru uA a podělíme je délkou okna:

$d_A[1]=(0+0+1+0+0)/5=1/5$

Následně se okno posune o jeden prvek doprava a můžeme tak spočíst hodnotu denzitního vektoru na pozici dvě:

$d_A[2]=(0+1+0+0+0)/5=1/5$

Obdobně postupujeme i dále a získáme sadu čtyř denzitních vektorů:

$d_A=\frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ 0\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ 0$

$d_C=\frac{1}{5}\ \frac{2}{5}\ \frac{2}{5}\ \frac{2}{5}\ \frac{1}{5}\ \frac{1}{5}\ 0\ \frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}$

$d_G=0\ 0\ 0\ 0\ 0\ \frac{1}{5}\ \frac{2}{5}\ \frac{2}{5}\ \frac{2}{5}\ \frac{2}{5}$

$d_T=\frac{1}{5}\ \frac{1}{5}\ \frac{2}{5}\ \frac{3}{5}\ \frac{3}{5}\ \frac{2}{5}\ \frac{2}{5}\ \frac{1}{5}\ 0\ 0$

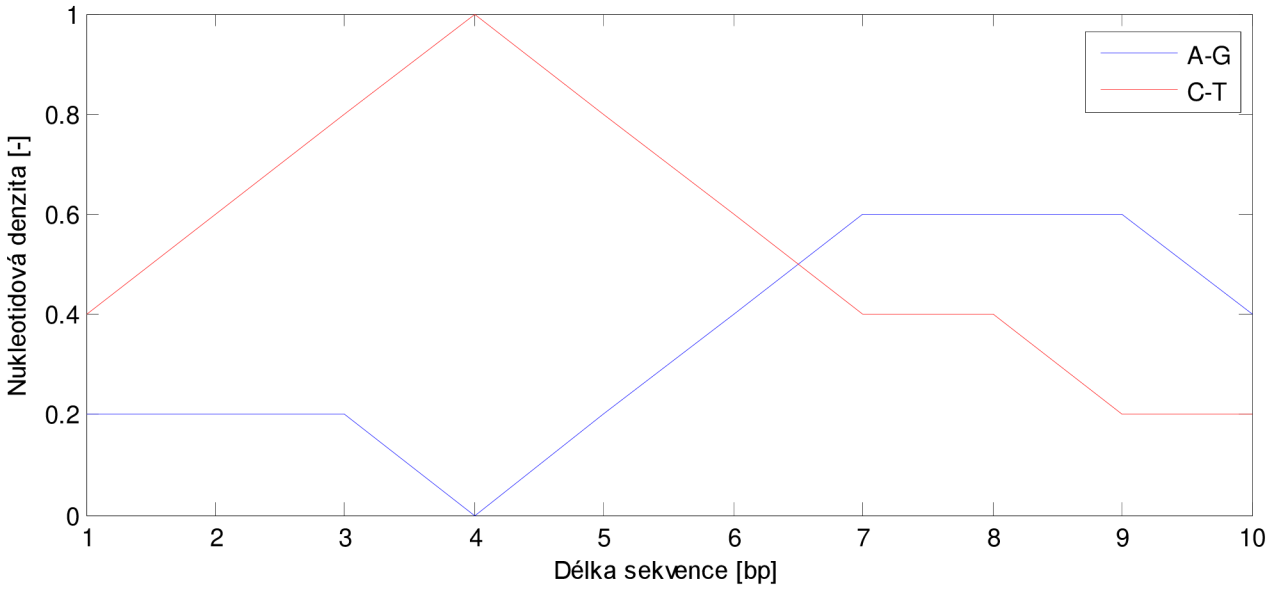
Délku okna je potřeba zvolit dle délky sekvence a požadované rozlišovací schopnosti. Jako minimální smysluplná velikost bývá uváděna délka okna 5. Při využití nukleotidové denzity pro další analýzy je vždy nutné délku okna optimalizovat.

Výhodou zpracování sekvencí pomocí nukleotidových denzit je možnost výsledky reprezentovat graficky. Nejjednodušším způsobem je prosté vykreslení samotných denzitních vektorů. Další možností je vizualizace sumací vektorů nukleotidů, které mají nějakou společnou biochemickou vlastnost, například dle dělení na purinvé/pyrimidinové nukleotidy, nukleotidy obsahující keto/amino skupinu a podobně. (viz kapitola 1.1.2) Tato možnost je vhodná právě v mém případě, kdy můžu zobrazit obsahy A a G, proti obsahu C a T. Potřebné vektory dostaneme jednoduše součtem odpovídajících hodnot denzitních vektrorů A a G a poté C a T. Budeme-li pokračovat v našem příkladu, dostaneme následující: [20]

$$d_{AG}=\frac{1}{5}\ \frac{1}{5}\ \frac{1}{5}\ 0\ \frac{1}{5}\ \frac{2}{5}\ \frac{3}{5}\ \frac{3}{5}\ \frac{3}{5}\ \frac{2}{5}$$

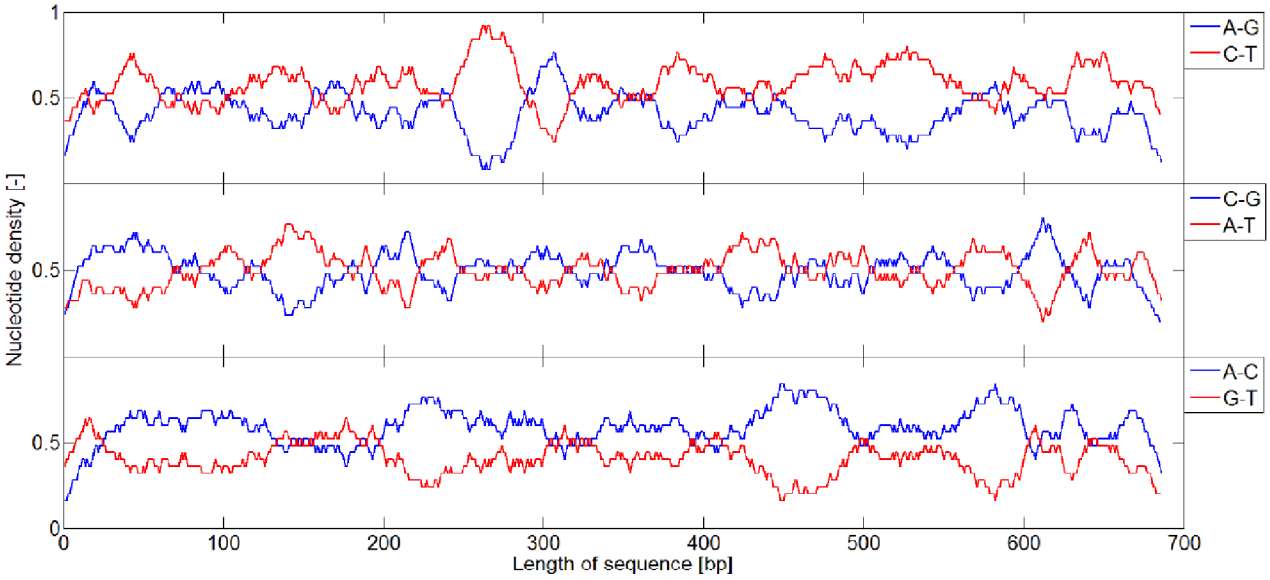
$$d_{CT}=\frac{2}{5}\ \frac{3}{5}\ \frac{4}{5}\ 1\ \frac{4}{5}\ \frac{3}{5}\ \frac{2}{5}\ \frac{2}{5}\ \frac{1}{5}\ \frac{1}{5}$$

Vizualizovaný výsledek naší vzorové sekvence je viditelný na Obr. 9. Pouhým pohledem na obrázek jsme schopni zjistit, že v první části sekvence převládaly báze C a T, zatímco v druhé části byla větší denzita bází A a G.



Obr. 9 Grafické znázornění sumy denzitních vektorů sekvence ACTCTTAGGC, W=5

Náš příklad byl samozřejmě velmi jednoduchý a především sekvence byla příliš krátká. Obvykle se zpracovávají sekvence mnohem delší – stovky až tisíce párů bází. Délce sekvence je pak uzpůsobena i velikost okna. Na Obr. 10 je vidět příklad zpracování denzitních vektorů pro mitochondriální gen COX1 dravce *Accipiter gentilis*.

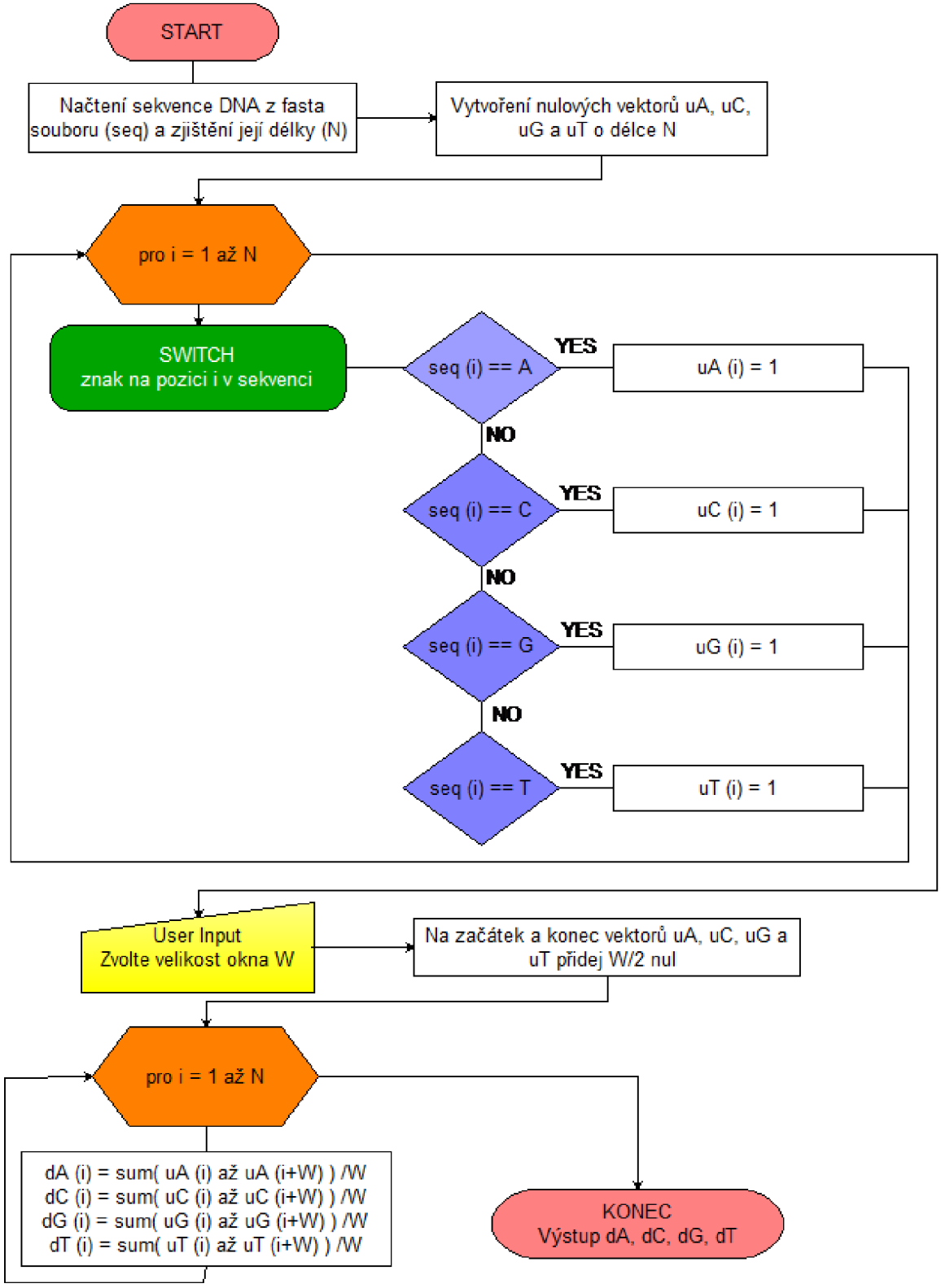


Obr. 10 Sumy denzitních vektorů mitochondriálního genu COX1 dravce *Accipiter gentilis* dle biochemických vlastností nukleotidů, W = 25. [20]

Tvorba denzitních vektorů je jednoduchá a efektivní metoda numerického zpracování, umožňující zobrazit průměrné zastoupení jednotlivých nukleotidů v sekvenci. Žádné dvě rozdílné sekvence nemají stejné nukleotidové denzity. Avšak při zpětné rekonstrukci symbolické sekvence v některých případech není možné určit přesnou polohu nukleotidu, především na začátku a konci sekvence – ztrácíme informaci o přesné pozici některých nukleotidů. Zpětná rekonstrukce proto není jednoznačná. [20]

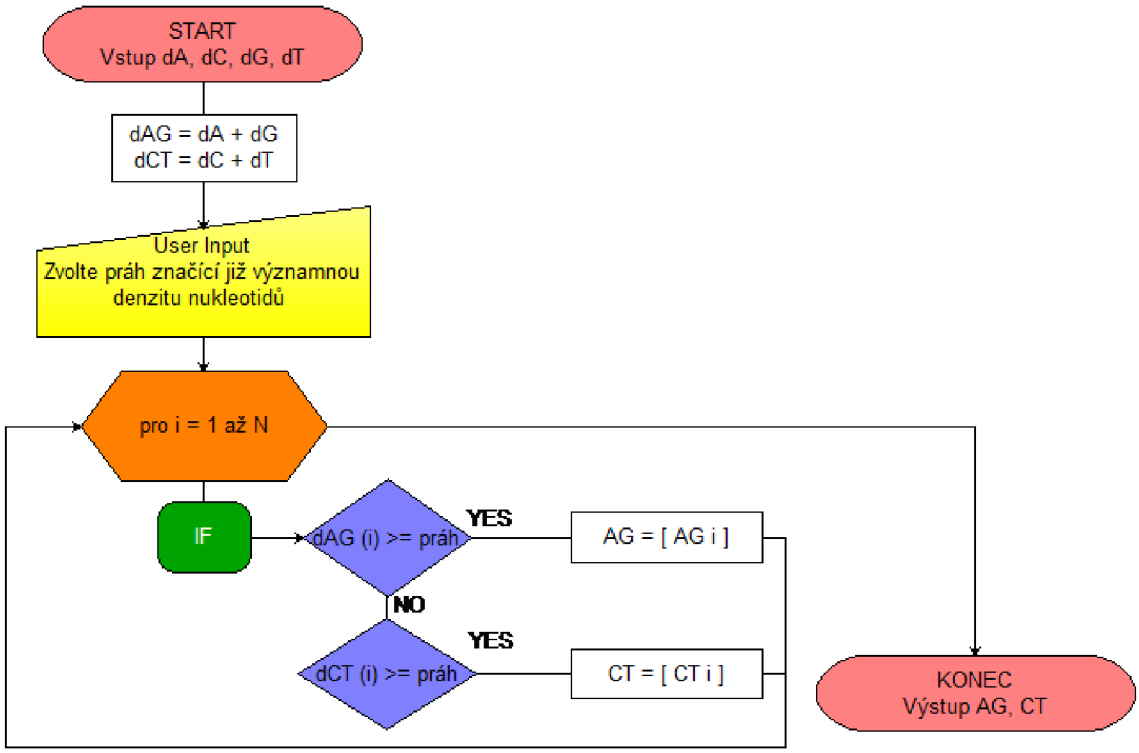
2.2. Vývojový diagram

Vývojový diagram je grafické znázornění algoritmu. Algoritmus je přesný postup, který vede k vyřešení problému. Jednotlivé operace prováděné v průběhu algoritmu jsou v diagramu znázorněny specifickými symboly spojenými šipkami. [21]



Obr. 11 Vývojový diagram výpočtu denzitních vektorů ze sekvence DNA

Na Obr. 11 Vývojový diagram výpočtu denzitních vektorů ze sekvence DNA je vidět vývojový diagram znázorňující algoritmus výpočtu denzitních vektorů ze symbolické sekvence DNA. Na Obr. 12 je vidět další postup při zpracování denzitních vektorů a realizace vyhledání a zapsání pozic míst v sekvenci s vysokým obsahem adeninu a guaninu.



Obr. 12 Vývojový digram zobrazující zpracování denzitních vektorů a algoritmu pro určení pozic v sekvenci s obsahem AG vyšším než práh

2.3. Pseudokód

Pseudokód (pseudo-code) je jazyk pro popis algoritmů nezávislý na programovacím jazyce. Je určen pro člověka nikoli pro počítač. Základní řídící konstrukce vychází z konstrukce programovacích jazyků. Zápis však bývá kratší a jednodušší, z důvodu, že se pracuje na vyšší abstraktní úrovni. Nemusí se dodržovat přesná syntaxe (středníky, závorky a podobně). V pseudokódu je vhodné využívat komentářů; vysvětlit význam parametrů a specifikovat vstupy a výstupy. [22]

Tab. 3 Ukázka algoritmu arrayMax(A,n) v pseudokódu a v Javě [22]

Algorithm arrayMax(A,n):
Input:Pole A obsahující n celých čísel.
Output:Maximální prvek v poli A.
currentMax ← A[0]
for i ← 1 to n-1 do
if currentMax < A[i] then
currentMax ← A[i]
return currentMax
	public class ArrayMaxProgram{
	static int arrayMax(int []A, int n){
	// vrací maximální prvek z *n* čísel pole *A*
	int currentMax=A[0];
	for(int i=1;i<n;i++)
	if(currentMax<A[i])
	currentMax=A[i];
	return currentMax;
	}
	}

Algoritmus arrayMax zapsaný pomocí **pseudokódu**	Algoritmus arrayMax v programovacím **jazyce Java**

Pseudokód algoritmu pro vyhledávání oblastí s vysokým obsahem A a G v sekvenci DNA:

searchAG(seq, W, práh)

```
% vstupem je sekvence obsahující A, C, T, G (seq), velikost posuvného okna (W) a prahová
% velikost denzity vektoru, považovaná již za významnou (práh)
% výstupem je vektor obsahující pozice v sekvenci s vysokou denzitou A a G
N ← délka zadané sekvence
uA ← nulový vektor o délce N          % tvorba nulových vektorů pro každou bázi
uC ← nulový vektor o délce N
uG ← nulový vektor o délce N
uT ← nulový vektor o délce N
for     i ← 1 to N do                 % označení místa výskytu dané báze přepsáním 0 na 1
        if      seq(i) == A
                uA(i) ← 1
        elseif  seq(i) == C
                uC(i) ← 1
        elseif  seq(i) == G
                uG(i) ← 1
        else
                uT(i) ← 1
uA ← přidání W/2 nul na začátek i konec vektoru uA
uC ← přidání W/2 nul na začátek i konec vektoru uC
uG ← přidání W/2 nul na začátek i konec vektoru uG
uT ← přidání W/2 nul na začátek i konec vektoru uT
for     i ← 1 to N do                 % procházení vektoru oknem W a průměrování hodnot
        dA(i) ← sum( uA(i) až uA(i+W) ) /W
        dC(i) ← sum( uC(i) až uC(i+W) ) /W
        dG(i) ← sum( uG(i) až uG(i+W) ) /W
        dT(i) ← sum( uT(i) až uT(i+W) ) /W
dAG ← dA + dG
dCT ← dC + dT
AG ← prázdný vektor                   % inicializace proměnné AG
CT ← prázdný vektor                   % inicializace proměnné CT
for     i ← 1 to N do                 % hledání a zapsání pozic s hodnotou dAG větší než práh
        if      dAG(i) ≥ práh
                AG ← [ AG i ]         % připsání pozice do AG
        elseif  dCT(i) ≥ práh
                CT ← [ CT i ]         % připsání pozice do CT
return AG, CT
```

3. REALIZACE ALGORITMU VYHLEDÁVÁNÍ

Mým cílem je vytvořit program, který dokáže vyhledat v sekvenci oblasti s vysokým obsahem adeninu a guaninu. Vyhledávání bude založeno na výpočtu denzitních vektorů (viz kapitola 2.1.1). Vstupem je sekvence ve formátu FASTA, stažená z některé z veřejných genových databází (viz kapitola 1.3). Uživatel si bude moci zvolit velikost okna W, určující rozlišovací schopnost výpočtu. Také si zvolí velikost prahu, odpovídající vysokému obsahu adeninu a guaninu. Výstupem pak bude vektor pozic v sekvenci, ve kterých denzita přesahuje prahovou hodnotu.

Rozhodla jsem se pro realizaci algoritmu pomocí programovacího prostředí MATLAB, který vyvinula firma MATHWORKS. Jeho název je odvozen z anglického MATrix LABoratory. Jedná se o velice výkonný nástroj pro vědecké a technické výpočty, zejména v maticových aplikacích. MATLAB obsahuje velké množství knihoven používaných pro velikou škálu lidských činností. Mimo jiné obsahuje také bioinformatický toolbox, obsahující funkce pro práci s biomedicínskými daty. Velkou výhodou je také možnost vytvořit grafické uživatelské prostředí - GUIDE, které zjednodušuje použití aplikace uživatelem, který tak vůbec nemusí přijít do kontaktu se samotným kódem programu. Veškeré funkce byly vytvořeny pomocí programovacího prostředí MATLAB verze R2010a (s knihovnou Bioinformatic Toolbox).

3.1. Rozbor programu

Základem programu je funkce tvořící uživatelské prostředí (jeho popis je uveden v kapitole 3.1.3) jejíž pomocí uživatel celý program ovládá. V prvé řadě obsahuje část pro načtení sekvence, dále uživatel zadává další parametry potřebné k analýze a pomocí uživatelského prostředí volá funkce tvořící funkční část programu – ***denzitni_vektory.m*** a ***urceni_pozic.m***. Tyto funkce jsou podrobně popsány dále.

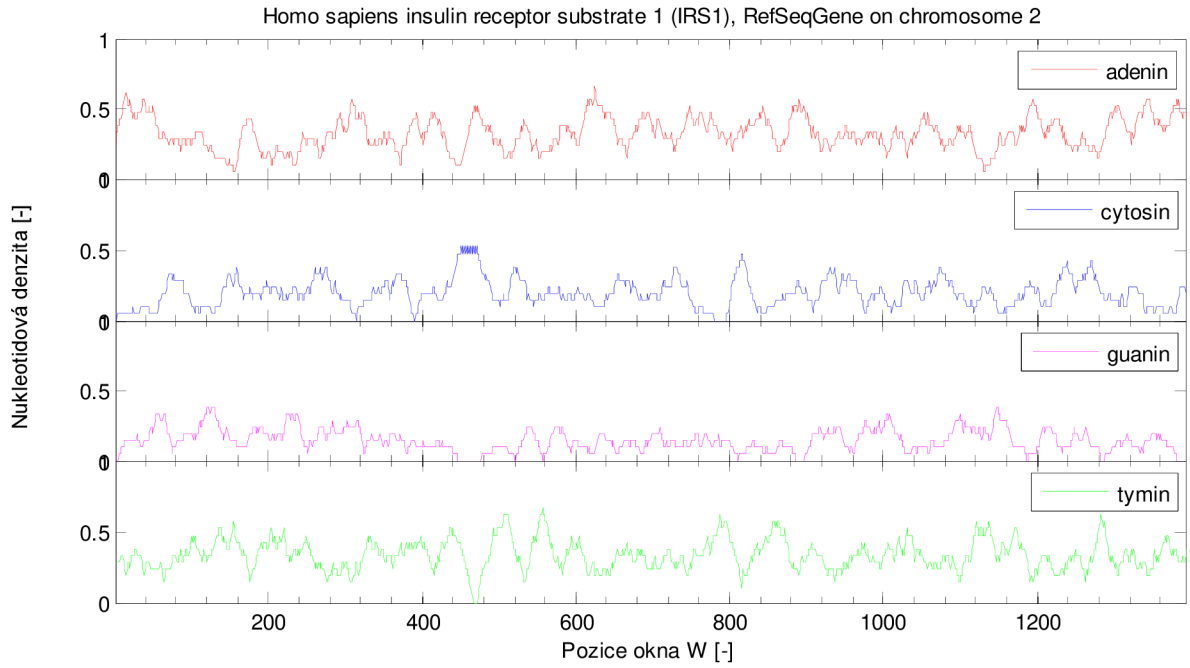
3.1.1. Funkce denzitni_vektory.m

První funkce realizuje algoritmus výpočtu denzitních vektorů. Její základní princip je viditelný na vývojovém diagramu na Obr. 11. Funkce se nazývá ***denzitni_vektory.m*** a jejími vstupy jsou hlavička sekvence a samotná sekvence ve formátu řetězce znaků. Posledním vstupem funkce pro výpočet denzitních vektorů je velikost okna W určující také do jisté míry rozlišovací schopnost algoritmu.

V první části funkce ***denzitni_vektory.m*** je potřeba ze sekvence vytvořit indikační vektory (viz. kapitola 2.1). Protože však sekvence může obsahovat i jiné znaky než jen A, G, C a T značící základní aminokyseliny DNA (viz kapitola 1.4), je potřeba i těmto znakům přiřadit odpovídající pravděpodobnosti výskytu jednotlivých bází na dané pozici. Například znak R značí výskyt purinu, tj. adeninu nebo guaninu, proto na danou pozici přiřadíme do indikačního vektoru adeninu i guaninu hodnotu 0,5 – což odpovídá 50% pravděpodobnosti výskytu každé z těchto bází. MATLAB disponuje funkcí tzv. podmínkového indexování, díky čemuž se můžeme vyhnout dlouhému cyklu s mnoha vnitřními podmínkami. Níže je zobrazena část programu tvořící indikační vektor pro adenin.

```
17    %vytvoreni indikacnich vektoru
18    L=length(sekvence);
19    uA = zeros(1, L);
20    uA(sekvence == 'A') = 1;
21    uA(sekvence == 'R') = 0.5;
22    uA(sekvence == 'M') = 0.5;
23    uA(sekvence == 'W') = 0.5;
24    uA(sekvence == 'D') = 1/3;
25    uA(sekvence == 'H') = 1/3;
26    uA(sekvence == 'V') = 1/3;
27    uA(sekvence == 'N') = 1/4;
```

V další části programu probíhá samotný výpočet denzitních vektorů podle vzorce (1) uvedeného v kapitole 2.1.1. Nakonec jsou jednotlivé denzitní vektory vykresleny pomocí funkcí subplot a plot a pomocí funkce sameaxis jsou upraveny společné osy. Ukázka výstupu programu je na Obr. 13.



Obr. 13 Výstup funkce ***denzitni_vektory.m***, gen IRS1 – Homo sapiens, velikost okna W = 21

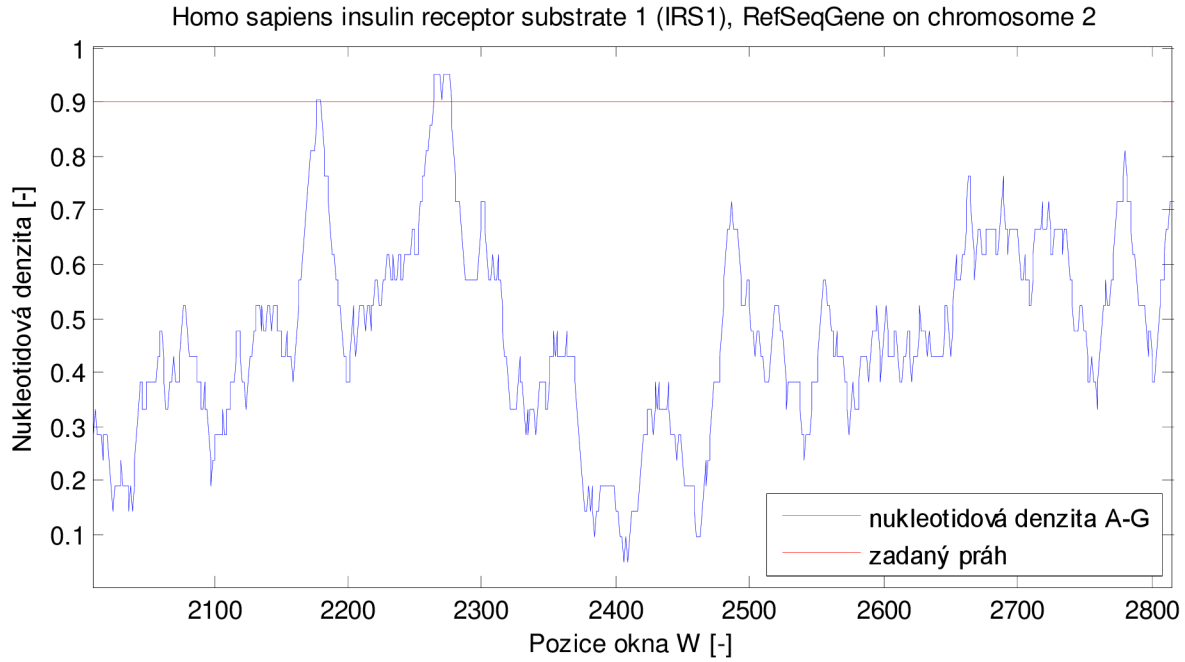
Funkce obsahuje také jednu podmínku hlídající správné použití funkce. Jak je uvedeno v kapitole 2.1.1, velikost okna W musí být vždy liché číslo a to z toho důvodu, že během výpočtu přiřazujeme průměrnou hodnotu denzity na pozici ležící uprostřed okna. Funkce proto obsahuje podmínku, a pokud uživatel zadá jako velikost okna W sudé číslo, vypíše chybu. Zda je číslo sudé, testuji pomocí funkce mod, která vrací zbytek po dělení – je-li tedy zbytek po dělení dvěma roven 0, běh skriptu se zastaví a vypíše se chybová hláška.

3.1.2. Funkce urceni_pozic.m

Druhou funkcí programu je ***urceni_pozic.m***, která realizuje samotné vyhledávání oblastí v sekvenci s vysokým obsahem zvolených bází (základ algoritmu je vidět na vývojovém diagramu na Obr. 12 Vývojový digram zobrazující zpracování denzitních vektorů a algoritmu pro určení pozic v sekvenci s obsahem AG vyšším než práh). Ačkoliv se tato práce zabývá pouze vyhledáváním adeninu a guaninu, funkce je napsána tak, že je možné si vybrat libovolnou dvojici nukleotidů. Na vstupu funkce je proto nutné zadat poměrně velké množství parametrů. V prvé řadě vypočtené denzitní vektory bází, jejichž denzita nás zajímá, dále je pro popis grafu na výstupu funkce potřeba zadat druh bází, jejichž denzitní vektory zadáváme a hlavičku zpracovávané sekvence. Dále je samozřejmě nutné zadat prahovou velikost denzity, která značí významný obsah požadovaného páru bází a jejíž pozici v sekvenci chceme získat. Posledním vstupním parametrem funkce je velikost okna W, se kterým byly počítány denzitní vektory, tato hodnota je použita pro zmenšení zkreslení výsledků, které vzniklo průměrováním hodnot v okně při výpočtu denzitních vektorů.

Takové velké množství vstupních parametrů by bylo pro uživatele velice nepřehledné, v tomto případě však nevadí, neboť funkce není určena k přímému použití, ale pracuje jako externí funkce uživatelského prostředí, pomocí kterého jsou všechny parametry získány.

Celá funkce je rozdělena na dvě části, pokud je zadána velikost prahu, vypočte pozice v sekvenci a vykreslí sumu denzitních vektorů i práh (viz Obr. 14 Výstup funkce ***urceni_pozic.m***, gen IRS1 – Homo sapiens, velikost okna W = 21, prah = 0.9). Pokud však velikost prahu zadána není, vykreslí pouze sumu denzitních vektorů požadovaného páru nukleotidů, aby uživatel před volbou velikosti prahu mohl prohlédnout výstupní data.



Obr. 14 Výstup funkce ***urceni_pozic.m***, gen IRS1 – Homo sapiens, velikost okna W = 21, prah = 0.9

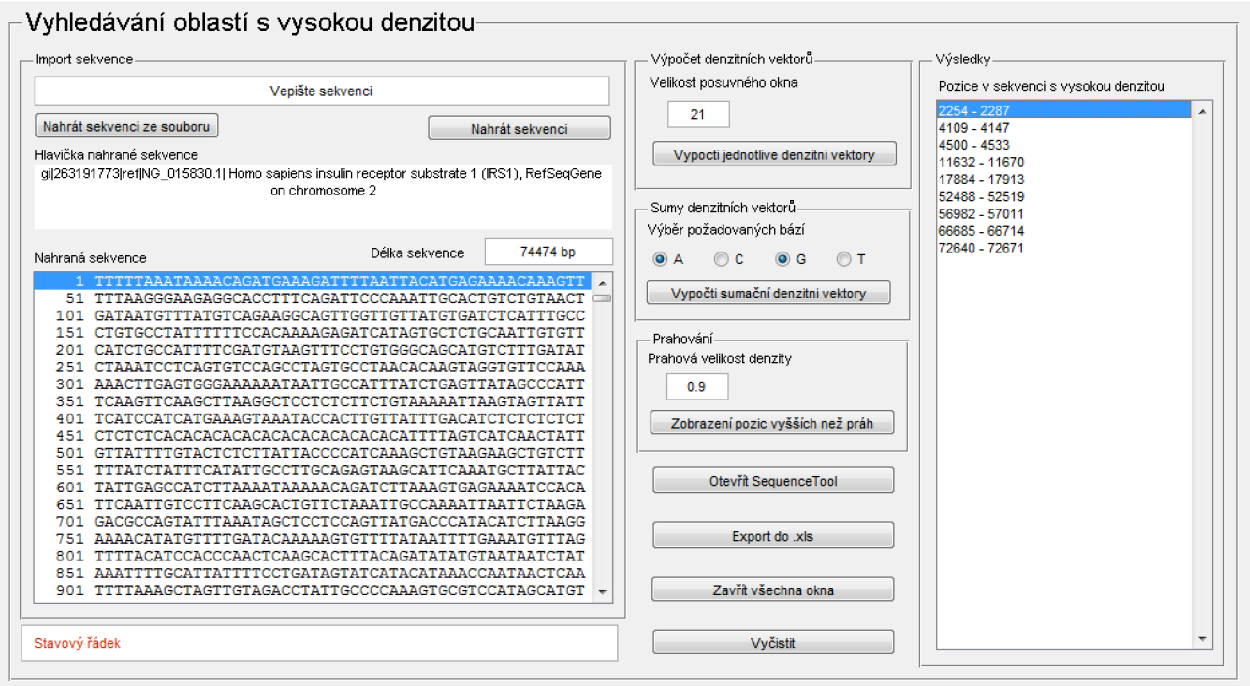
Věnujme se nyní části programu, která je aktivní v případě, že je zadána i velikost prahu. V prvé řadě je vypočten denzitní vektor odpovídající obsahu požadovaného páru bází, ten vznikne prostou sumací denzitních vektorů zvolených bází (viz kapitola 2.1.1). Následně takto vypočtený vektor pomocí for-cyklu procházím a porovnávám s velikostí prahu, pokud ji jeho hodnota přesáhne, zapíši si do pomocné proměnné pozici této hodnoty ve vektoru – přibližně odpovídá pozici v sekvenci. Tímto způsobem však dostaneme mnoho jednotlivých pozic, mezi nimi i takové, kdy jen místně skokově vzroste hodnota denzity nad zvolený práh, ale následně hned klesá. Abychom rozlišili významné (delší) úseky od těch malých a nevýznamných, je dále přidána podmínka, že hodnoty denzity musí přesáhnout práh alespoň na úseku stejně dlouhém, jako je polovina velikosti okna W použitá při výpočtu denzitních vektorů. Pokud tato podmínka není splněna, jsou nalezené pozice ignorovány. Naopak pokud splněna je, je k úseku ještě přidáno na každou stranu půl velikosti okna, čímž se snažím kompenzovat zkreslení, vzniklé průměrováním hodnot v okně při výpočtu denzitních vektorů. Teprve takto zpracované pozice jsou výstupem funkce.

Jelikož potřebuji dva různé formáty výstupu, jeden pro zobrazení v uživatelském prostředí GUIDE a druhý pro export do MS Excelu, jsou na výstupu funkce ***urceni_pozic.m*** dvě hodnoty, jedna obsahuje pozice ve formátu řetězce s pomlčkou mezi vypočtenými hodnotami (ve funkci se nazývá E) a druhá obsahuje pouze pozice počátků a konců úseků, uložené v numerickém formátu (ve funkci se nazývá B).

Nakonec následuje již jen prosté vykreslení sumačního denzitního vektoru a prahu pomocí funkce plot. Funkce také obsahuje jednoduchou podmínku, která hlídá uživatele při použití. Velikost prahu je nutno zvolit z intervalu <0;1>, jelikož i hodnoty denzity za žádných okolností nebudou nabývat jiných hodnot. Toto je hlídáno obdobnou podmínkou, jako je u funkce ***denzitni_vektory.m*** hlídána lichá velikost okna W.

3.1.3. Popis uživatelského prostředí GUI

Páteřní částí programu pro vyhledávání oblastí s vysokou denzitou je grafické uživatelské prostředí, které využívá výše popsané naprogramované funkce. Uživatelské prostředí je navrženo tak, aby bylo co nejsnáze pochopitelné a aby každé tlačítko mělo jasně viditelnou funkci. Z toho důvodu je opticky rozděleno na několik funkčních oddílů – panelů (viz Obr. 15).



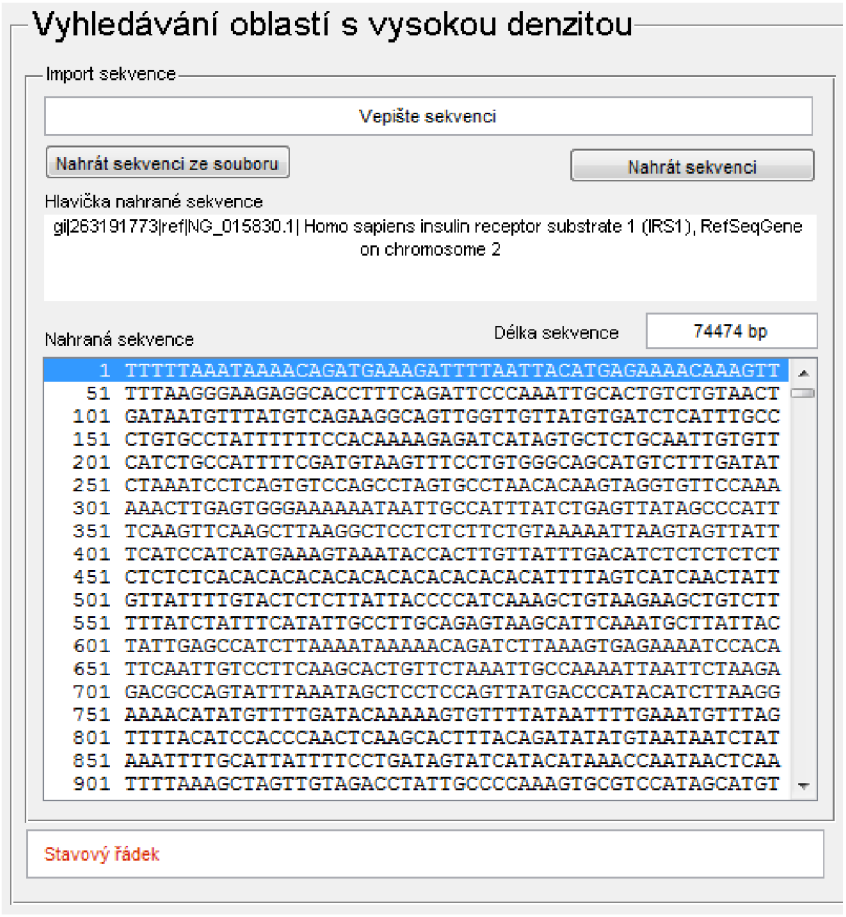
Obr. 15 Uživatelské prostředí GUIDE

První oddíl obsahuje část pro načtení sekvence. Uživatel si může vybrat ze dvou variant, buď může vepsat/vložit sekvenci do určeného okna umístěného zcela nahoře vlevo, a načíst ji pomocí tlačítka nahrát sekvenci. Nebo ji může pomocí stisku tlačítka nahrát ze souboru formátu FASTA (viz kapitola 1.4), tlačítko je popsané jako Nahrát sekvenci ze souboru. Pro nahrání sekvence ze souboru je využita funkce ***fastaread.m***, která je součástí biomedicínského toolboxu MATLABU. Po nahrání sekvence se sekvence přehledně zobrazí v okně s názvem Nahraná sekvence, dále se zobrazí její hlavička (v případě nahrávání sekvence ze souboru) a také její délka v bp. Tato část GUI je lépe viditelná na Obr. 16.

Dále uživatel přistoupí k výpočtu denzitních vektorů, k čemuž slouží další panel se stejným názvem. Zde si uživatel zvolí velikost okna W (je přednastaveno na hodnotu 21) a stiskem tlačítka Vypočti jednotlivé denzitní vektory se spustí funkce ***denzitni_vektory.m***, která provede výpočet a vykreslení denzitních vektorů, jak je popsáno výše v kapitole 3.1.1.

Následuje panel s názvem sumy denzitních vektorů. Zde si uživatel vybere dvojici bází, pro kterou chce provádět analýzu výskytu oblastí v sekvenci s vysokou denzitou. Báze si zvolí jednoduchým zakliknutím radiobuttonu. Vzhledem k zadání této práce je defaultně zvolena kombinace adenin – guanin. Podmínkami je ošetřeno, že není možné zakliknout více než dvě tlačítka radiobuton. Po stisku tlačítka vypočti sumační denzitní vektory se spustí první část funkce

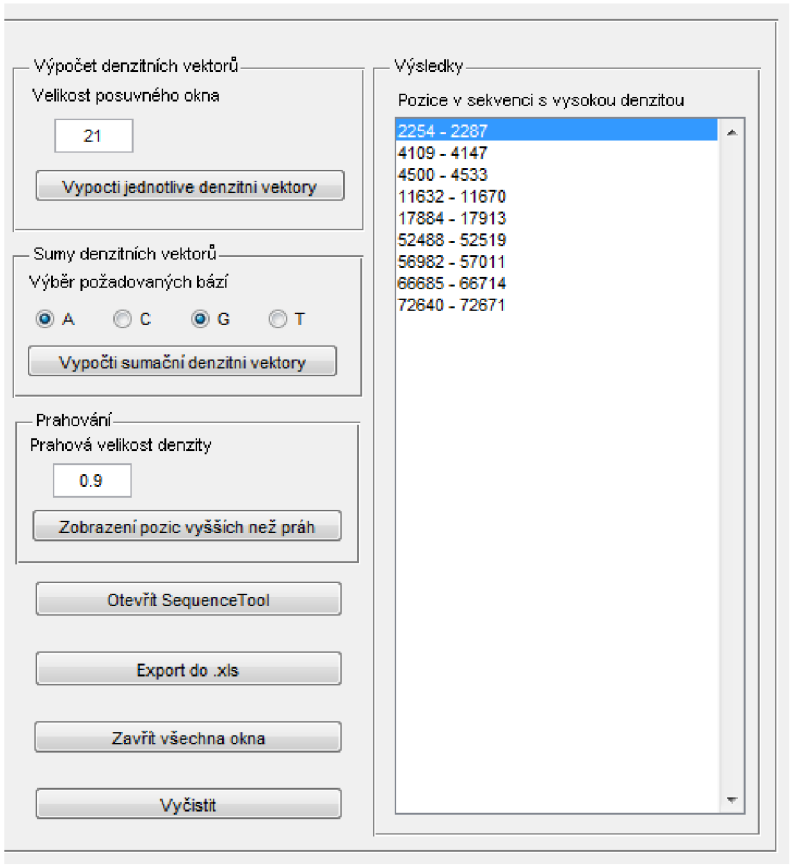
urceni_pozic.m, která vykreslí sumační denzitní vektor zvoleného páru nukleotidů. Její funkce je podrobně popsána v kapitole 3.1.2. Výše popsaná část uživatelského prostředí je lépe viditelná na Obr. 17.



Obr. 16 GUI - nahrávání sekvence (sekvence IRS1 - Homo sapiens)

Posledním funkčním oddílem je panel s názvem prahování. Zde uživatel zvolí velikost prahu, která značí významnou velikost denzity zvoleného páru bází. Fukce ***urceni_pozic.m*** pak vykreslí do grafu práh a sumační denzitní vektor požadovaného páru bází. Vypočtené pozice s hodnotou denzity vyšší než zadaný práh, se zobrazí v okně s názvem Pozice v sekvenci s vysokou denzitou. Tento oddíl je lépe viditelný na Obr. 17.

Uživatelské prostředí dále obsahuje několik tlačítek sloužících uživateli pro usnadnění práce s programem. V prvé řadě je to tlačítko s popisem Otevřít SequenceTool, po jeho stisknutí se spustí funkce ***sequencetool.m*** z biomedicínského toolboxu MATLABU. Zobrazí okno pro práci se sekvencí, umožňující zobrazit translaci i transkripci a obsahuje i další nástroje pro práci se sekvencí, například čítač nukleotidů. Dalším tlačítkem v řadě je Export do .xls, po jeho stisku se otevře dialogové okno, s jehož pomocí uložíme soubor ve formátu .xls, do nějž se zapíší vyhledané pozice do dvou sloupců, do prvního se vepíší pozice počátků úseků a do druhého pozice konců nalezených úseků.



Obr. 17 GUI - výpočet denzitních vektorů a prahování (sekvence IRS1 -Homo sapiens)

Následují poslední dvě tlačítka, jedno pro zavření všech oken s grafy a další slouží k vymazání všech dosud nahraných dat. Úplně dole vlevo je pak ještě stavový řádek. Zde se zobrazují chybové hlášky, např. pokud uživatel zadá špatnou hodnotu okna či chce pokračovat v analýze prahováním, aniž by nejdříve nechal vypočítat denzitní vektory. Ať chce či nechce, je tak donucen dodržovat při analýze předepsaný postup. Vzhled výše popisované části GUI je lépe viditelný na Obr. 16 a Obr. 17.

4. ANALÝZA GENŮ

Posledním úkolem této práce je provést analýzu části lidského, šimpanzího a myšího genomu na výskyt adenin-guanin bohatých úseků za pomoci vytvořeného programu. K tomuto účelu jsem si vybrala pět různých genů z databáze NCBI. Záměrně jsem zvolila geny obsahující introny i exony, aby bylo možné porovnávat četnost úseků tvořených adeninem a guaninem v kódujících i nekódujících oblastech sekvencí. Pro větší vypovídající hodnotu analýzy jsem se nažila vybrat geny ověřené, nikoli pouze predikované.

V prvé řadě vždy u každého ze zvolených genů uvádím výsledky zarovnávání sekvencí, které slouží jako výchozí stanovisko pro další analýzu genů. Pro lokální zarovnání jsem u všech sekvencí použila Smith-Watermanův algoritmus, nejlepší výsledky jsem dostala při použití skórovací matice NUC44 a penalizace mezery nastavenou na hodnotu 9. Pro výpočet jsem využila funkci swalign z bioinformatického toolboxu Matlabu. Ke globálnímu zarovnání jsem pak použila Needleman-Wunschova algoritmu, nejlepší výsledky jsem dostala při použití skórovací matice NUC44 a penalizace mezery nastavenou na hodnotu 5. Pro výpočet jsem využila funkci nwalign z bioinformatického toolboxu Matlabu. Výstupy zarovnání sekvencí jsou přiloženy jako .xps soubory v elektronické dokumentaci.

Následně jsem pomocí vytvořeného programu v sekvencích vyhledala pozice úseků s vysokou denzitou A a G pro různé velikosti okna W (9; 15; 21; 27; 33) a různé hodnoty prahu (0,7; 0,8; 0,9). Výsledky jsou z důvodu velkého rozsahu uvedeny pouze v elektronické dokumentaci v souboru s názvem analyza.pdf.

Poté jsem vyjádřila procentuální zastoupení úseků bohatých na A a G v celé sekvenci a zvlášť v kódujících a nekódujících oblastech (vycházela jsem z pozic CDS uvedených na stránkách NCBI). Snažila jsem se tak zjistit jejich funkci v genetickém kódu. Dle předpokladu (viz kapitola 1.1.5) by měl být obsah A a G vyšší v kódující části sekvence. Na výsledky má velký vliv podmínka, jak dlouhé úseky s vysokou denzitou A a G již bereme v potaz. Proto u každého genu uvádím tři tabulky, v první započítávám úplně všechny nalezené pozice, v druhé jsem použila podmínku, že nalezený úsek musí být delší než polovina použitého okna a v třetí musí mít nalezená oblast minimálně délku okna použitého pro výpočet denzitních vektorů. Ve všech případech každý nalezený úsek prodloužím na každou stranu o W/2, čímž se snažím eliminovat zkreslení vzniklé při výpočtu denzitních vektorů. Pro každou kombinaci velikosti okna a prahu je v tabulkách uveden nejdříve procentuální obsah A-G bohatých oblastí v celé sekvenci, následně v kódující a pak v nekódující části sekvence. Pro srovnání jsem obdobnou analýzu provedla i u náhodně vygenerované sekvence.

4.1. Analýza vybraných genů

4.1.1. Gen ALB1

Gen kóduje protein albumin, což je jeden z proteinů krevní plazmy. Tvoří až 60 % všech plazmatických bílkovin. Kromě krve se vyskytuje také v dalších tělních tekutinách - jako je

tkáňový a mozkomíšní mok. Je důležitý pro transport různých látek, především steroidů, mastných kyselin a hormonů štítné žlázy. Pomáhá také udržet stálé vnitřní prostředí v organismu. [23]

Zpracovávané sekvence člověka a myši jsou dle údajů z NCBI potvrzené a schválené kurátory, oproti tomu sekvence šimpanze je pouze predikovaná v rámci výpočtu celého genomu, ale není ověřena a potvrzena.

Při lokálním zarovnání byly sekvence člověka a šimpanze shodné z 98 %, více se lišily prakticky jen v intervalu cca od 1500bp do 1650bp. Při lokálním zarovnání sekvence člověka a myši vyšla shoda jen z 56 %, přičemž odlišnosti byly rozptýleny rovnoměrně v celých sekvencích.

Při globálním zarovnání byly sekvence člověka a šimpanze podobné také z 98 %, problematický byl opět především úsek cca od 1500bp do 1650bp, kde se nacházely velké odlišnosti, jinak byly sekvence prakticky shodné. Globální zarovnání pro sekvenci člověka a myši vyšlo obdobně, jako zarovnání lokální, shoda byla 57 % a odlišnosti byly rozptýleny rovnoměrně v průběhu celých sekvencí.

Tab. 4 Procentuální obsah A-G bohatých oblastí v sekvenci genu ALB1 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – bez ohledu na délku nalezeného úseku (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen ALB1 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	33,76/45,25/32,39	32,09/47,81/30,21	28,81/35,79/27,97	28,09/41,26/26,52	11,88/11,42/11,94
0,8	12,39/19,67/11,52	12,93/18,09/12,31	5,95/7,98/5,71	3,63/1,64/3,86	1,68/1,80/1,66
0,9	2,41/2,40/2,41	2,00/1,69/2,04	0,93/0,00/1,04	0,00/0,00/0,00	0,00/0,00/0,00
Gen ALB1 - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	34,39/45,92/33,10	33,59/48,49/31,91	29,75/36,26/29,02	29,64/43,00/28,13	12,46/11,94/12,52
0,8	12,72/20,67/11,83	13,58/20,85/12,77	6,36/8,34/6,14	3,59/1,71/3,80	1,65/1,88/1,63
0,9	2,43/2,51/2,42	1,98/1,77/2,01	0,94/0,00/1,04	0,00/0,00/0,00	0,00/0,00/0,00
Gen ALB1 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	39,46/46,96/38,47	40,27/42,09/40,03	35,24/25,78/36,48	32,82/22,55/34,17	17,58/7,01/18,97
0,8	15,41/14,56/15,52	18,21/17,84/18,25	9,04/6,08/9,43	5,19/1,75/5,65	2,61/1,64/2,74
0,9	3,18/1,86/3,36	1,70/1,09/1,78	1,08/0,00/1,22	0,38/0,00/0,42	0,39/0,00/0,44

Výsledky analýzy procentuálního obsahu A-G bohatý úseků bez ohledu na jejich délku v sekvenci genu ALB1 jsou uvedeny v Tab. 4. Ze získaných dat je patrná podobnost mezi výsledky pro sekvence člověka a šimpanze, u myši se výsledky lehce liší. Pro kombinace nižšího prahu a kratšího okna vychází předpoklad vyššího obsahu A-G bohatých oblastí v kódující části sekvence. Pro vyšší prahy zvláště v kombinaci s dlouhým oknem je obsah A-G bohatých oblastí v kódující části sekvence naopak nižší než mimo ni. U myši výsledky nevychází dle předpokladu. Kromě kombinace krátkého okna (9 a 15) a malé velikosti prahu (0,7) vychází denzita A a G vyšší vždy v nekódující části sekvence.

Tab. 5 Procentuální obsah A-G bohatých oblastí v sekvenci genu ALB1 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W/2 (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen ALB1 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	17,95/25,30/17,07	10,91/13,66/10,58	8,26/7,81/8,32	4,58/2,24/4,85	1,81/0,00/2,02
0,8	5,17/5,46/5,13	3,62/3,88/3,59	1,36/0,00/1,52	0,45/0,00/0,51	0,00/0,00/0,00
0,9	0,86/1,31/0,81	0,25/0,00/0,28	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen ALB1 - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	18,34/26,61/17,40	11,47/16,68/10,88	8,71/8,17/8,77	5,31/2,34/5,65	1,78/0,00/1,98
0,8	5,28/6,45/5,15	3,97/4,05/3,96	1,17/0,00/1,31	0,23/0,00/0,25	0,00/0,00/0,00
0,9	0,84/1,37/0,78	0,37/0,00/0,41	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen ALB1 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	21,79/21,73/21,80	15,03/12,59/15,35	11,15/8,26/11,53	8,56/5,47/8,96	3,53/0,00/3,99
0,8	6,25/4,71/6,45	4,56/3,45/4,70	1,80/1,64/1,82	0,97/0,00/1,10	0,81/0,00/0,92
0,9	0,95/0,71/0,98	0,35/0,00/0,40	0,38/0,00/0,42	0,38/0,00/0,42	0,39/0,00/0,44

Tab. 6 Procentuální obsah A-G bohatých oblastí v sekvenci genu ALB1 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen ALB1 - Homo sapiens					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	6,81/6,01/6,91	4,44/6,67/4,17	2,15/2,13/2,15	0,65/0,00/0,73	0,00/0,00/0,00
0,8	1,39/0,93/1,45	0,55/0,00/0,62	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,10/0,00/0,11	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen ALB1 - Pan troglodytes					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	7,12/6,28/7,21	4,40/6,97/4,12	2,60/2,23/2,64	0,65/0,00/0,72	0,00/0,00/0,00
0,8	1,47/0,97/1,53	0,71/0,00/0,78	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,10/0,00/0,12	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen ALB1 - Mus musculus					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	8,50/4,87/8,97	6,66/5,75/6,78	4,41/1,97/4,73	3,24/0,00/3,67	1,09/0,00/1,24
0,8	1,40/0,99/1,45	1,19/0,00/1,35	1,00/0,00/1,13	0,43/0,00/0,48	0,50/0,00/0,57
0,9	0,29/0,00/0,33	0,35/0,00/0,40	0,38/0,00/0,42	0,38/0,00/0,42	0,00/0,00/0,00

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W/2 v sekvenci genu ALB1 jsou uvedeny v Tab. 5. Opět je patrná podobnost mezi výsledky pro sekvence člověka a šimpanze, u myši se výsledky liší. Pro kombinace nižšího prahu (0,7; 0,8) a kratšího okna (9, 15) vychází předpoklad vyššího obsahu A-G bohatých oblastí v kódující části sekvence, oproti předchozí tabulce je rozdíl v obsahu v a mimo CDS výraznější, téměř deset procent. Pro vyšší prahy zvláště v kombinaci s dlouhým oknem je obsah A-G bohatých oblastí v kódující části

sekvence naopak nižší než mimo ni. U myši výsledky naprosto neodpovídají předpokladu, obsah A a G v CDS je pro všechna okna i prahy nižší než ve zbytku sekvence.

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W v sekvenci genu ALB1 jsou uvedeny v Tab. 6. Opět je viditelná ekvivalence mezilidskou a šimpanzí sekvencí - při použití velikosti okna 15 a prahu 0,7 je obsah A a G v CDS vyšší než mimo ni. Ve všech ostatních případech výsledky neodpovídají předpokladu. Při použití dlouhého okna a vysokého prahu nejsou v sekvenci nalezeny žádné úseky vyhovující podmínce.

4.1.2. Gen HBB

Gen kóduje protein beta podjednotky hemoglobinu. Hemoglobin je součástí červených krvinek a slouží k přenosu kyslíku z plic do tkání a oxidu uhličitého z tkání zpět do plic. Mutace beta podjednotky hemoglobinu způsobuje srpkovitou anémii. [24]

Zpracovávaná sekvence člověka je dle údajů z NCBI potvrzena a schválená kurátory, oproti tomu sekvence šimpanze a myši jsou pouze predikované v rámci výpočtu celých genomů, ale nejsou zcela ověřeny a potvrzeny.

Při lokálním zarovnání byly sekvence člověka a šimpanze shodné z 99 %, tedy prakticky totožné. U lokálního zarovnání sekvence člověka a myši byla shoda jen z 65 %, odlišnosti byly rozptýleny rovnoměrně v celých sekvencích.

Při globálním zarovnání byly sekvence člověka a šimpanze shodné z 91 %, lišily se především na začátku a konci sekvence. Globální zarovnání pro sekvenci člověka a myši vyšlo obdobně, jako zarovnání lokální, shoda byla opět 65 % a odlišnosti byly rozptýleny v průběhu celých sekvencí.

Tab. 7 Procentuální obsah A-G bohatých oblastí v sekvenci genu HBB v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – bez ohledu na délku nalezeného úseku (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen HBB - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	30,14/27,93/30,98	28,39/23,87/30,12	28,14/27,70/28,31	24,60/15,99/27,88	20,30/15,54/22,12
0,8	10,65/15,09/8,95	9,96/6,76/11,19	4,36/0,00/6,02	0,00/0,00/0,00	0,00/0,00/0,00
0,9	2,30/2,03/2,41	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen HBB - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	32,00/35,66/30,33	32,57/34,74/31,58	32,11/43,20/27,08	30,16/30,51/30,00	23,45/25,00/22,75
0,8	11,47/15,26/9,75	11,70/10,85/12,08	5,68/2,57/7,08	3,15/2,57/3,42	0,00/0,00/0,00
0,9	1,61/1,65/1,58	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen HBB - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	24,86/22,30/26,05	24,14/25,00/23,74	20,77/32,88/15,13	16,62/15,77/17,02	10,39/15,54/7,98
0,8	13,11/14,19/12,61	9,96/11,26/9,35	5,09/0,00/7,46	2,08/0,00/3,05	0,00/0,00/0,00
0,9	3,15/2,03/3,68	2,51/0,00/3,68	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

Výsledky analýzy procentuálního obsahu A-G bohatý úseků v sekvenci genu HBB bez ohledu na jejich délku jsou uvedeny v Tab. 7. U tohoto genu výsledky u člověka příliš nevychází dle předpokladu. Procentuální zastoupení A-G bohatých oblastí je v kódujícím úseku sekvence člověka vždy nižší než obsah v nekódující části, s výjimkou hodnoty prahu 0,8 a velikosti okna 9. U šimpanze hodnoty odpovídají předpokladu pro všechny kombinace prahu 0.7 a pro všechny prahy při použití okna velikosti 9. U myši vykazovaly vyšší procentuální zastoupení A-G bohatých úseků v CDS kombinace nízkého prahu 0.7 a hodnoty okna 15 až 33 a také kombinace vyššího prahu 0,8 a nízké hodnoty okna 9 a 15.

Tab. 8 Procentuální obsah A-G bohatých oblastí v sekvenci genu HBB v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W/2 (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen HBB - Homo sapiens					
prah\W	9	15	21	27	33
0,7	13,01/15,77/11,96	8,16/6,31/8,86	8,47/6,98/9,04	7,78/9,23/7,23	0,00/0,00/0,00
0,8	2,80/0,00/3,87	2,99/0,00/4,13	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen HBB - Pan troglodytes					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	13,25/14,34/12,75	9,23/7,72/9,92	9,92/9,38/10,17	9,69/12,50/8,42	2,75/5,88/1,33
0,8	3,73/1,29/4,83	4,36/2,39/5,25	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen HBB - Mus musculus					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	15,90/16,67/15,55	10,32/10,14/10,40	4,94/0,00/7,25	3,01/0,00/4,41	0,00/0,00/0,00
0,8	4,80/3,38/5,46	5,01/0,00/7,35	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	1,15/0,00/1,68	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W/2 v sekvenci genu HBB jsou uvedeny v Tab. 8. Je patrná podobnost mezi výsledky pro sekvence člověka a šimpanze, u myši se výsledky liší. Předpoklad je splněn pouze pro kombinaci velikosti okna 9 a 27 a prahu 0,7 u šimpanze a člověka. U myši pouze pro práh 0,7 a velikost okna 9. Při použití dlouhého okna a vysokého prahu nejsou v sekvenci nalezeny žádné úseky vyhovující podmínce.

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W v sekvenci genu HBB jsou uvedeny v Tab. 9. Výsledky analýzy za použití těchto parametrů neodpovídají předpokladu. Pouze při použití krátkého okna a nízkého prahu jsou v sekvenci vůbec nalezeny dostatečně dlouhé úseky s vysokým obsahem A a G a jejich obsah je vždy vyšší v nekódující oblasti.

Tab. 9 Procentuální obsah A-G bohatých oblastí v sekvenci genu HBB v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen BETA - Homo sapiens					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	5,73/3,83/6,45	1,93/0,00/2,67	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,8	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen BETA - Pan troglodytes					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	6,54/4,60/7,42	3,50/2,57/3,92	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,8	1,15/1,29/1,08	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen BETA - Mus musculus					
prah\W	**9**	**15**	**21**	**27**	**33**
0,7	8,67/8,11/8,93	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,8	2,65/0,00/3,89	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

4.1.3. Gen P53

Tento gen kóduje protein p53, který ovlivňuje expresi mnohých genů, čímž přispívá ke kontrole růstu buněk, oprav DNA, apoptóze (programovaná buněčná smrt) a angiogenezi (tvorbě cévního řečiště). Mutace v tomto genu jsou spojeny s celou řadou lidských nádorů. Alternativním sestřihem tohoto genu vznikají různé izoformy transkripce tohoto genu. [25]

Zpracovávané sekvence člověka a myši jsou dle údajů z NCBI potvrzené a schválené kurátory, oproti tomu sekvence šimpanze je pouze predikovaná v rámci výpočtu celého genomu, ale není ověřena a potvrzena. Pro určení kódujících úseků jsem použila CDS první transkripční varianty genu cellular tumor antigen p53 izoformy a.

Při lokálním zarovnání byly sekvence člověka a šimpanze shodné z 93 %, rozdíly byly největší v úseku cca od 14 700bp do 16 000bp, na jehož konci se u sekvence šimpanze nachází poměrně dlouhý úsek nedefinovaných bází, který rozdílnost sekvencí ještě zvyšuje. Při lokálním zarovnání sekvence člověka a myši byla jejich shoda 63 %, odlišnosti byly rozptýleny rovnoměrně v průběhu celých sekvencí.

Při globálním zarovnání byly sekvence člověka a šimpanze shodné opět z 93 %, přičemž problematický byl znovu výše zmiňovaný úsek. U globálního zarovnání sekvence člověka a myši byla shoda pouhých 47 %, což bylo dáno mimo jiné také poměrně rozdílnou délkou sekvencí.

Výsledky analýzy procentuálního obsahu A-G bohatý úseků v sekvenci genu P53 bez ohledu na jejich délku jsou uvedeny v Tab. 10. U této sekvence se výsledky neodpovídají předpokladu, obsah A-G bohatých úseků je v kódující části u člověka i šimpanze vždy nižší než mimo ni. V sekvenci myši předpoklad splňuje pouze kombinace nízkého prahu 0,7 a krátká velikost okna 9 a 15, překvapivě ale také vysoký práh 0,9 a velikost okna 9.

Tab. 10 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – bez ohledu na délku nalezeného úseku (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen P53 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	41,27/32,49/41,85	43,91/33,84/44,58	48,82/25,21/50,37	46,76/21,32/48,43	27,07/14,13/27,92
0,8	16,47/12,01/16,76	24,57/13,45/25,30	14,83/3,47/15,58	8,44/6,01/8,60	5,76/0,25/6,12
0,9	4,66/5,16/4,63	3,54/0,00/3,77	3,08/0,00/3,28	1,76/0,00/1,88	1,19/0,00/1,27
Gen P53- Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	41,62/27,38/42,45	44,14/29,68/44,98	48,39/22,86/49,87	45,42/17,20/47,05	26,16/11,62/27,00
0,8	16,30/9,03/16,72	24,63/8,26/25,57	14,24/0,96/15,01	8,81/3,75/9,10	5,28/0,29/5,56
0,9	4,67/4,13/4,70	3,59/0,00/3,80	3,32/0,00/3,51	1,98/0,00/2,09	1,71/0,00/1,81
Gen P53 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	36,78/37,51/36,70	35,29/35,98/35,22	34,60/23,44/35,86	31,72/22,25/32,80	19,02/10,66/19,97
0,8	14,84/14,07/14,93	18,43/13,13/19,03	10,26/4,09/10,96	5,35/2,30/5,70	2,26/0,00/2,51
0,9	3,40/4,26/3,31	2,02/0,00/2,25	1,49/0,00/1,66	0,28/0,00/0,31	0,30/0,00/0,33

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W/2 v sekvenci genu P53 jsou uvedeny v Tab. 11. Vybráním pouze delších úseků s vyšším obsahem A a G se procentuální zastoupení v a mino CDS velice změnilo oproti předchozí tabulce. Vyšší procentuální obsah vyšel pouze u sekvence myši při použití délky okna 9 a velikosti prahu 0,7. Ve všech ostatních případech výsledky neodpovídají předpokladu.

Tab. 11 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W/2 (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen P53 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	27,11/17,34/27,76	21,94/13,45/22,50	14,56/9,39/14,90	11,83/6,09/12,21	7,03/1,61/7,39
0,8	9,70/7,02/9,88	7,62/4,15/7,85	4,91/2,62/5,06	3,46/0,00/3,69	1,66/0,00/1,77
0,9	3,17/2,03/3,24	2,31/0,00/2,47	1,62/0,00/1,73	0,21/0,00/0,23	0,00/0,00/0,00
Gen P53 - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	27,12/13,93/27,88	21,96/9,80/22,66	14,58/7,20/15,01	12,49/2,50/13,07	8,01/1,83/8,36
0,8	9,67/5,38/9,92	7,96/2,11/8,30	5,17/0,00/5,47	2,90/0,00/3,07	1,44/0,00/1,53
0,9	3,35/2,31/3,41	2,62/0,00/2,77	2,01/0,00/2,12	0,22/0,00/0,23	0,00/0,00/0,00
Gen P53 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	20,40/22,76/20,13	14,91/12,28/15,21	11,87/7,16/12,41	8,92/1,71/9,74	3,64/1,71/3,86
0,8	7,63/6,65/7,75	5,90/1,53/6,39	2,15/0,00/2,40	1,07/0,00/1,19	0,00/0,00/0,00
0,9	1,05/1,02/1,05	0,44/0,00/0,49	0,30/0,00/0,33	0,00/0,00/0,00	0,00/0,00/0,00

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W v sekvenci genu P53 jsou uvedeny v Tab. 12. Výsledky analýzy při použití těchto parametrů neodpovídají předpokladu. Pouze při použití krátkého okna a nízkého prahu jsou v sekvenci vůbec nalezeny dostatečně dlouhé úseky s vysokým obsahem A a G a jejich obsah je vždy vyšší v nekódující oblasti.

Tab. 12 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen P53 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	11,73/7,70/11,99	8,92/3,72/9,26	6,38/0,85/6,74	5,98/1,61/6,27	1,94/1,61/1,96
0,8	3,46/0,00/3,68	4,10/0,00/4,37	1,46/0,00/1,56	0,34/0,00/0,36	0,36/0,00/0,38
0,9	1,77/0,00/1,88	1,02/0,00/1,09	0,21/0,00/0,23	0,00/0,00/0,00	0,00/0,00/0,00
Gen P53 - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	11,49/5,76/11,82	9,16/0,77/9,65	6,85/0,96/7,19	5,77/1,83/6,00	2,93/1,83/3,00
0,8	3,60/0,00/3,81	4,09/0,00/4,33	1,93/0,00/2,04	0,32/0,00/0,33	0,00/0,00/0,00
0,9	2,14/0,00/2,27	1,36/0,00/1,44	0,22/0,00/0,23	0,00/0,00/0,00	0,00/0,00/0,00
Gen P53 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	10,70/7,59/11,05	7,11/4,60/7,40	4,61/1,79/4,93	1,98/1,71/2,01	0,00/0,00/0,00
0,8	2,20/0,00/2,45	2,23/0,00/2,49	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,36/0,00/0,41	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

4.1.4. Gen CDK2

Protein kódovaný tímto genem patří mezi cyklin-dependentní kinázy (CDK), což jsou enzymy podílející se na regulaci řady významných buněčných procesů např. buněčného cyklu, apoptózy, diferenciace neuronálních buněk atp. Činnost CDK2 je regulována cyklinem A a E a inhibována faktory p21 a p27. Zvýšená exprese tohoto genu byla prokázána u některých rakovinných buněk, například u kolorektálního karcinomu. [26]

Zpracovávané sekvence člověka a myši jsou dle údajů z NCBI potvrzené a schválené kurátory, oproti tomu sekvence šimpanze je pouze predikovaná v rámci výpočtu celého genomu, ale není ověřena a potvrzena. Pro určení kódujících úseků jsem použila CDS první transkripční varianty genu cyklin-dependentní kinázy 2 izoformy 1.

Při lokálním zarovnání byly sekvence člověka a šimpanze shodné z 92 %, největší rozdíl je způsoben úsekem neurčitých bází v sekvenci šimpanze cca od 3436bp do 9791bp. Ve zbylé části jsou sekvence prakticky shodné. U lokálního zarovnání sekvence člověka a myši byla shoda 72 % a odlišnosti byly rozptýleny v celku rovnoměrně v průběhu celých sekvencí.

Při globálním zarovnání byly sekvence člověka a šimpanze shodné pouze z 81 %, přičemž problematický byl znovu výše zmiňovaný úsek a navíc se na konci sekvencí v odlišnostech

projevila rozdílná délka sekvencí. U globálního zarovnání pro sekvenci člověka a myši byla shoda 63 %, v sekvenci se vyskytují i úseky velmi dobré shody, ale převažují úseky dosti rozdílné.

Tab. 13 Procentuální obsah A-G bohatých oblastí v sekvenci genu CDK2 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – bez ohledu na délku nalezeného úseku (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen CDK2 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	37,04/26,98/38,80	38,13/27,54/39,99	37,39/27,09/39,19	38,02/16,95/41,71	24,23/17,95/25,33
0,8	16,35/9,92/17,47	20,11/12,71/21,40	11,67/5,80/12,71	9,10/3,46/10,09	5,55/0,00/6,53
0,9	4,54/0,00/5,34	5,39/1,67/6,04	2,01/0,00/2,37	1,61/0,00/1,90	1,86/0,00/2,19
Gen CDK2 - Pan troglodytes					
práh\W	9	15	21	27	33
0,7	33,51/28,68/34,32	33,26/30,19/33,77	30,21/30,57/30,15	34,97/19,12/37,61	23,46/20,25/23,99
0,8	14,67/11,19/15,25	17,74/13,96/18,37	11,17/6,54/11,94	7,36/3,90/7,94	4,54/0,00/5,30
0,9	4,18/0,00/4,88	5,35/1,89/5,93	1,56/0,00/1,82	1,13/0,00/1,32	0,70/0,00/0,82
Gen CDK2 - Mus musculus					
práh\W	9	15	21	27	33
0,7	31,77/29,78/32,11	31,62/29,68/31,95	30,54/27,86/30,99	28,83/33,62/28,01	14,45/25,84/12,50
0,8	15,01/11,43/15,63	16,65/13,06/17,26	9,52/11,91/9,11	4,50/6,05/4,23	2,60/0,00/3,05
0,9	3,35/1,15/3,72	2,52/4,32/2,21	1,79/0,00/2,09	0,82/0,00/0,96	0,52/0,00/0,61

Výsledky analýzy procentuálního obsahu A-G bohatý úseků v sekvenci genu CDK2 jsou uvedeny v Tab. 13. Pro sekvenci člověka a šimpanze výsledky naprosto neodpovídají předpokladu, v CDS je obsah oblastí A a G nižší pro všechny kombinace velikosti okna a prahu. U sekvence myši je předpoklad splněn pro velikosti okna 27 a 33 a prahu 0,7 a pro velikosti okna 21 při nastavení prahu na hodnotu 0,8.

Tab. 14 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W/2 (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen CDK2 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	22,62/12,60/24,37	18,33/9,48/19,88	14,49/8,92/15,46	12,59/10,37/12,98	6,89/0,00/8,09
0,8	9,08/3,34/10,09	8,96/3,23/9,97	5,84/3,34/6,27	3,08/0,00/3,62	1,71/0,00/2,01
0,9	2,79/0,00/3,28	1,01/0,00/1,19	1,66/0,00/1,95	0,00/0,00/0,00	0,00/0,00/0,00
Gen CDK2 - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	19,67/14,21/20,57	16,40/10,69/17,35	14,69/10,06/15,46	12,52/11,70/12,65	5,33/0,00/6,22
0,8	8,55/3,77/9,34	8,46/3,65/9,26	4,83/3,77/5,01	2,46/0,00/2,87	1,85/0,00/2,16
0,9	2,50/0,00/2,91	0,99/0,00/1,15	1,19/0,00/1,38	0,00/0,00/0,00	0,00/0,00/0,00
Gen CDK2 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	19,85/15,08/20,67	14,59/13,26/14,82	11,16/14,51/10,59	7,87/13,26/6,95	3,35/0,00/3,92
0,8	6,10/5,96/6,13	5,20/7,97/4,73	2,80/2,88/2,78	1,90/0,00/2,22	1,49/0,00/1,75
0,9	1,39/1,15/1,43	0,69/0,00/0,81	0,44/0,00/0,51	0,00/0,00/0,00	0,00/0,00/0,00

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W/2 v sekvenci genu CDK2 jsou uvedeny v Tab. 14. Za použití těchto parametrů výsledky příliš nevychází dle předpokladu, procentuální zastoupení A-G bohatých oblastí je v kódujícím úseku sekvence nižší než mimo něj u všech třech druhů i prakticky u všech kombinací velikosti okna a prahu. Výjimku tvoří pouze sekvence myši a hodnoty prahu 0,8 pro velikost okna 15 a 21.

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W v sekvenci genu CDK2 jsou uvedeny v Tab. 15. Při použití této podmínky se překvapivě vychází obsah A a G v CDS v lidské a šimpanzí sekvenci pro kombinaci velikosti okna 27 a prahu 0,7 vyšší než mimo něj. U sekvence myši je takových výsledků více, především pro nižší hodnoty prahu – 0,7 a 0,8.

Tab. 15 Procentuální obsah A-G bohatých oblastí v sekvenci genu P53 v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen CDK2 - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	11,89/7,58/12,65	8,85/7,02/9,17	8,85/4,68/9,58	5,72/5,80/5,71	2,30/0,00/2,70
0,8	4,14/0,00/4,87	4,71/3,23/4,96	3,03/0,00/3,56	0,88/0,00/1,04	0,00/0,00/0,00
0,9	0,83/0,00/0,98	0,55/0,00/0,65	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen CDK2 - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	11,42/8,55/11,90	8,55/7,92/8,65	8,55/5,28/9,09	4,29/6,54/3,92	2,44/0,00/2,85
0,8	3,74/0,00/4,36	3,83/3,65/3,86	2,46/0,00/2,87	0,95/0,00/1,11	0,00/0,00/0,00
0,9	0,34/0,00/0,40	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen CDK2 - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	10,07/9,70/10,13	5,72/6,24/5,63	4,79/8,17/4,22	4,99/5,19/4,96	0,94/0,00/1,10
0,8	2,21/3,27/2,03	3,28/5,76/2,85	1,27/0,00/1,48	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,49/0,00/0,58	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

4.1.5. Gen SHBG

Tento gen kóduje protein ze skupiny globulinů, proteinů přítomných v krevní plazmě, charakteristických nižší pohyblivostí než albumin. SHB globulin váže steroidy, přispívá tak k přenosu např. pohlavních hormonů, odtud jeho název sex hormone-bindig globulin (globulin vázající pohlavní hormony). Různé transkripční varianty kódují různé izoformy tohoto proteinu. [27]

Zpracovávané sekvence člověka a myši jsou dle údajů z NCBI potvrzené a schválené kurátory, oproti tomu sekvence šimpanze je pouze predikovaná v rámci výpočtu celého genomu, ale není ověřena a potvrzena. Pro určení kódujících úseků jsem použila CDS první transkripční varianty genu kódující sex hormone-binding globulin izoformy 1.

Při lokálním zarovnání byly sekvence člověka a šimpanze shodné z 99 %, zdály se být téměř totožné. U lokálního zarovnání sekvence člověka a myši byla shoda 65 %, odlišnosti byly rozptýleny rovnoměrně v průběhu celých sekvencí.

Při globálním zarovnání sekvence člověka a šimpanze byla shoda pouhých 17 %, což bylo způsobeno především rozdílnou délkou sekvencí. U globálního zarovnání pro sekvenci člověka a myši byla shoda 16 %, také sehrála velkou roli rozdílná délka sekvencí.

Tab. 16 Procentuální obsah A-G bohatých oblastí v sekvenci genu SHBG v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – bez ohledu na délku nalezeného úseku (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen SHBG - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	42,62/35,81/43,08	45,07/31,43/45,98	47,47/29,53/48,66	45,38/27,54/46,57	27,85/4,71/29,40
0,8	17,45/12,16/17,80	23,52/11,41/24,32	16,62/5,21/17,38	10,87/2,56/11,42	8,20/0,00/8,75
0,9	4,53/2,15/4,69	3,84/0,00/4,09	3,53/0,00/3,77	1,35/0,00/1,44	0,79/0,00/0,84
Gen SHBG - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	40,08/35,76/41,62	40,47/28,94/44,59	45,13/33,68/49,22	49,97/35,07/55,28	30,95/5,56/40,01
0,8	18,29/14,47/19,65	21,73/12,50/25,02	16,95/6,02/20,85	10,41/2,66/13,17	8,73/0,00/11,85
0,9	5,51/3,01/6,40	3,77/0,00/5,12	2,19/0,00/2,97	0,00/0,00/0,00	0,00/0,00/0,00
Gen SHBG - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	37,25/35,15/38,60	36,89/26,16/43,82	38,77/25,41/47,39	39,84/20,30/52,45	22,94/4,54/34,82
0,8	19,06/14,03/22,31	22,14/19,14/24,07	10,45/4,79/14,11	7,96/0,00/13,10	3,66/0,00/6,02
0,9	4,92/2,39/6,55	3,20/0,00/5,27	2,07/0,00/3,41	0,00/0,00/0,00	0,00/0,00/0,00

Výsledky analýzy procentuálního obsahu A-G bohatý úseků v sekvenci genu SHBG jsou uvedeny v Tab. 16. U tohoto genu výsledky naprosto nevychází dle předpokladu, procentuální zastoupení A-G bohatých oblastí je v kódujících částech sekvencí vždy nižší než mimo něj u všech třech druhů i u všech kombinací velikosti okna a prahu. Rozdíly jsou navíc mnohdy poměrně velké.

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W/2 v sekvenci genu SHBG jsou uvedeny v Tab. 17. Ani použití podmínky minimální délky nalezených úseků nezměnilo fakt, že obsah A a G u této sekvence vychází vždy vyšší mimo kódující úsek.

Výsledky analýzy procentuálního obsahu A-G bohatý úseků delších než W v sekvenci genu SHBG jsou uvedeny v Tab. 18. Výsledky analýzy při použití těchto parametrů neodpovídají předpokladu. Pouze při použití krátkého okna a nízkého prahu jsou v sekvenci vůbec nalezeny dostatečně dlouhé úseky s vysokým obsahem A a G, avšak jejich obsah je vždy vyšší v nekódující oblasti.

Tab. 17 Procentuální obsah A-G bohatých oblastí v sekvenci genu SHBG v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W/2 (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen SHBG - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	27,23/12,99/28,18	21,51/9,51/22,31	15,04/6,45/15,62	14,44/0,00/15,40	9,95/0,00/10,62
0,8	9,70/4,88/10,02	8,44/4,88/8,68	5,34/0,00/5,69	2,95/0,00/3,14	1,19/0,00/1,26
0,9	2,88/0,00/3,08	1,69/0,00/1,80	0,93/0,00/0,99	0,21/0,00/0,23	0,00/0,00/0,00
Gen SHBG - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	23,01/13,54/26,38	21,36/11,57/24,86	19,78/6,94/24,36	16,10/0,00/21,84	9,16/0,00/12,43
0,8	10,04/6,37/11,35	10,07/5,44/11,73	4,08/0,00/5,53	0,00/0,00/0,00	0,00/0,00/0,00
0,9	2,62/0,00/3,55	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen SHBG - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	21,59/15,18/25,72	18,12/10,97/22,74	15,73/8,58/20,34	9,71/0,00/15,97	5,44/0,00/8,95
0,8	10,84/3,71/15,44	6,73/3,55/8,79	2,20/0,00/3,62	0,00/0,00/0,00	0,00/0,00/0,00
0,9	1,84/0,99/2,40	0,68/0,00/1,12	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

Tab. 18 Procentuální obsah A-G bohatých oblastí v sekvenci genu SHBG v závislosti na velikosti okna W a prahu (v celé sekvenci/v CDS/mimo CDS) – nalezené úseky delší než W (červeně podbarveno pokud je obsah A-G bohatých oblastí větší v CDS než mimo CDS)

Gen SHBG - Homo sapiens					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	12,18/5,38/12,63	10,56/5,38/10,91	9,00/0,00/9,60	7,88/0,00/8,40	3,80/0,00/4,05
0,8	3,72/0,00/3,96	3,82/0,00/4,08	1,68/0,00/1,79	0,92/0,00/0,98	0,36/0,00/0,38
0,9	1,24/0,00/1,32	0,54/0,00/0,57	0,21/0,00/0,23	0,00/0,00/0,00	0,00/0,00/0,00
Gen SHBG - Pan troglodytes					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	14,42/6,83/17,13	13,45/6,48/15,94	9,62/0,00/13,05	9,53/0,00/12,92	4,53/0,00/6,15
0,8	3,38/0,00/4,58	1,89/0,00/2,56	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
Gen SHBG - Mus musculus					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	12,94/6,35/17,20	8,61/2,81/12,35	6,38/0,00/10,49	4,14/0,00/6,82	0,00/0,00/0,00
0,8	2,43/0,00/3,99	2,07/0,00/3,41	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00
0,9	0,55/0,00/0,91	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00	0,00/0,00/0,00

4.2. Testování obsahu A-G bohatých oblastí v náhodné sekvenci

V rámci analýzy jsem měla za úkol zvážit náhodnost úseků bohatých na A a G. K tomuto účelu jsem použila funkce z bioinformatického toolboxu Matlabu s názvem ***randseq.m***. Tato funkce vytvoří náhodnou sekvenci DNA požadované délky. Nechala jsem si vygenerovat sekvenci délky 2500bp a následně provedla analýzu na procentuální obsah A-G bohatých oblastí obdobně jako u zkoumaných genů. Výsledky jsou uvedeny v Tab. 19.

Z vypočtených dat je patrné, že obsah A-G bohatých oblastí je v náhodné sekvenci nižší než u analyzovaných genů. Rozdíl je dobře viditelný především při vyhledávání delších úseků s vysokým obsahem A a G. Můžeme tedy vyvozovat, že námi nalezené úseky nevznikly náhodou, což jen potvrzuje jejich důležitou úlohu v sekvenci.

Tab. 19 Procentuální obsah A-G bohatých oblastí v náhodné sekvenci v závislosti na velikosti okna W a prahu

Náhodná sekvence – bez ohledu na délku A-G bohatých úseků					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	31,52	19,36	19,48	13,32	0,00
0,8	5,60	6,20	1,92	0,00	0,00
0,9	0,40	0,00	0,00	0,00	0,00
Náhodná sekvence – A-G bohaté úseky delší než W/2					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	9,92	5,00	1,28	1,64	0,00
0,8	0,60	0,88	0,00	0,00	0,00
0,9	0,00	0,00	0,00	0,00	0,00
Náhodná sekvence – A-G bohaté úseky delší než W					
práh\W	**9**	**15**	**21**	**27**	**33**
0,7	2,48	1,16	0,00	0,00	0,00
0,8	0,00	0,00	0,00	0,00	0,00
0,9	0,00	0,00	0,00	0,00	0,00

4.3. Vyhodnocení

Úkolem bakalářské práce bylo pomocí vytvořeného programu provést analýzu lidského, šimpanzího a myšího genomu na výskyt adenin-guanin bohatých úseků. Analyzovat celé genomy v rámci bakalářské práce samozřejmě není možné, vybrala jsem si proto pět vzorových genů, vyskytujících se u všech třech vybraných druhů. Snažila jsem se zvolit geny různé délky a z různých chromosomů. Také jsem vybírala především geny průkazně ověřené – v genové databázi NCBI označené jako reviewed (přezkoumána) nebo validated (ověřena). Použité sekvence člověka jsou ověřeny všechny, myší z většiny, avšak použité sekvence šimpanze jsou bohužel pouze predikované.

Přesto je v analýze velmi dobře patrná podobnost mezi geny člověka a šimpanze, ve všech případech se sekvence myši od obou liší výrazně více, než zmíněné dvě sekvence mezi sebou. V prvé řadě jsem pomocí vytvořeného programu vypočítala pozice úseků s vysokým obsahem A a G při použití tří různých prahů – 0,7; 0,8; 0,9 a pěti různých velikostí okna – 9; 15; 21; 27; 33. Chtěla jsem tímto způsobem zjistit, které nastavení je pro vyhledávání nejvýhodnější.

Vzhledem k množství takto získaných dat, bylo nutné je ještě dále nějakým způsobem zpracovat. Proto jsem vypočítala procentuální zastoupení nalezených oblastí v kódující a nekódující části sekvence (vycházela jsem z CDS úseků uvedených na NCBI). Předpokladem bylo, že v exonech se budou úseky bohaté na A a G vyskytovat ve větší míře, nežli v intronech.

Při procentuálním vyjádření má na výsledky velký vliv způsob určování krajních bodů nalezených úseků. Výpočtem denzitních vektorů dochází ke zkreslení, v průběhu výpočtu totiž sekvenci procházíme pomocí okna a průměrujeme hodnoty v něm obsažené. Pokud následně prahováním najdeme pozici s vysokou denzitou, vznikla tato hodnota průměrováním W/2 hodnot indikačního vektoru na každou stranu od počítaného umístění, proto vždy přidávám W/2 pozic na začátek i konec nalezeného úseku s vysokou denzitou. Tímto způsobem se snažím částečně eliminovat vliv zkreslení oknem.

Velikost okna koresponduje se zkreslením výsledků - čím delší okno tím větší zkreslení. Použitím krátkého okna získáme sice přesnější informaci o výskytu kratších úseků s vyšším obsahem A a G, ale opravdu dlouhý úsek obsahující především A a G, ovšem lehce narušený obsahem jiných bází, při použití krátkého okna rozkouskujeme na několik malých úseků. Naopak při použití příliš velkého okna mohou být zase kratší úseky bohaté na A a G přeskočeny, neboť jejich vlivem nedojde k dostatečnému vzrůstu výsledné denzity v okně. Tímto způsobem však můžeme lépe najít dlouhé úseky bohaté na A a G. Proto v práci pracuji s poměrně velkým rozsahem velikostí oken od 9 po 33 a snažím se zjistit, která hodnota je pro analýzu nejvýhodnější.

Výsledky analýzy genů ovšem pro žádné z výše uvedených nastavení neodpovídaly předpokladu. Proto jsem se ještě pokusila vybírat jen delší úseky a zanedbat ty krátké, ve kterých denzita jen místně vzrostla. V prvním případě jsem zvolila jako hranici délku poloviny použitého okna a následně délku celého okna. Ani po tomto opatření se však hledané oblasti ve většině případů nevyskytovaly v kódující části častěji než mimo ni. Občas byl sice pro některé hodnoty velikosti okna a prahu (zpravidla pro kratší okna a nižší prahy) obsah A-G bohatých oblastí v CDS vyšší, avšak rozdíl oproti zbytku sekvence zpravidla nebyl velký. Ze získaných dat se proto nedá dost dobře usuzovat, jaká velikost okna a prahu je pro analýzu nejvhodnější.

Celkově je z výsledků patrná spíše tendence pro vyšší obsah A-G bohatých oblastí v intronech, což popírá předpoklad uvedený také v kapitole 1.1.5. V rámci mé práce jsem však pracovala s pouhými pěti náhodně vybranými geny, pro průkaznější zjištění by bylo nutné genů zpracovat mnohem více.

Na druhou stranu z vypočtených hodnot je patrné, že výskyt A-G bohatých oblastí v sekvencích není pouze náhodný. Úseky bohaté na A a G se vyskytují poměrně často v sekvencích člověka, šimpanze i myši. V průběhu evoluce tady zřejmě byly zachovány. Otázkou však zůstává, zda mají nějakou funkci a jakou.

ZÁVĚR

Cílem bakalářské práce bylo zpracovat podklady a dostupné informace k tématu vyhledávání oblastí bohatých na adenin a guanin, vytvořit program realizující jejich vyhledávání a s jeho pomocí provést analýzu části genomu šimpanze, člověka a myši.

Teoretická část se věnuje složení a struktuře DNA. Zabývá se její replikací, transkripcí a translací. Podrobněji je rozebrán projekt, jehož cílem bylo oskenovat celý genom člověka. Jsou zmíněny i informace o genomu šimpanze a myši. Také stručně popisuji největší internetové genové databáze a nejpoužívanější formát pro ukládání sekvencí - fasta.

Druhá část se již věnuje samotnému problému vyhledávání oblastí bohatých na adenin a guanin. Je nastíněna problematika numerického zpracování sekvenčních dat DNA a následně vybrána konkrétní metoda pro další zpracování. Zvolila jsem metodu vyhledávání v sekvenci na základě výpočtu denzitních vektorů. Sestavila jsem vývojové diagramy programu realizujícího výpočty a napsala pro tuto funkci pseudokód.

V další části popisuji implementaci zvoleného řešení do programovacího prostředí Matlab R2010a. Vysvětluji úlohu jednotlivých funkcí a podrobněji rozebírám i vytvořené uživatelské prostředí aplikace.

Čtvrtou a zároveň poslední částí bakalářské práce je analýza pěti různých genů člověka, šimpanze a myši. Snažila jsem se ověřit předpoklad, že úseky bohaté na adenin a guanin se vyskytují častěji v kódující části genů nežli ve zbytku sekvence. Vyhledala jsem tedy úseky s vysokým obsahem A a G a procentuálně vyjádřila jejich obsah v CDS a mimo CDS. Toto jsem pro každý gen provedla pro různé kombinace velikosti okna a prahu, také jsem vyzkoušela vyhledávat pouze úseky určité délky. Ovšem při žádném z vyzkoušených nastavení parametrů vyhledávání nebyl obsah A-G bohatých oblastí v exonech výrazně vyšší než v intronech. Dokonce se spíše zdálo, že se hledané úseky více vyskytují naopak v intronech, ani to ale nebylo pravidlem. Provedla jsem analýzu pouhých pěti genů, není tedy možné vyvozovat obecné závěry, k tomu by bylo zapotřebí analyzovat mnohem větší objem dat, což by však přesahovalo rozsah této práce. Vytvořené funkce by ovšem zcela jistě byly použitelné i k rozsáhlejší analýze, bylo by pouze nutné více zautomatizovat načítání sekvencí a vypisování výsledků.

LITERATURA

[1] **Snustad, D. Peter a Simmons, Michael J.** *Genetika.* [překl.] Jiřina Relichová. Nakladatelství Masarykova univerzita, Brno: 2009. ISBN 978-80-210-4852-2.

[2] **Šípek jr., Antonín.** *Genetika - Biologie; Váš zdroj informací o genetice a biologii.* [Online] [Citace: 22. říjen 2012.] Dostupné z URL: http://www.genetika-biologie.cz/.

[3] DNA. *Univerzity of Bristol.* [Online] [Citace: 8. říjen 2012.] Dostupné z URL: http://www.chm.bris.ac.uk/webprojects2002/baldelli/www/dna.html.

[4] DNA structure. *Wikimedia commons.* [Online] [Citace: 10. říjen 2012.] Dostupné z URL: http://cs.wikipedia.org/wiki/Soubor:DNA_Structure%2BKey%2BLabelled.pn_NoBB_cs.png.

[5] **Ruiz, Mariana.** DNA replication en. *Wikimedia commons.* [Online] 24. leden 2007. [Citace: 10. Říjen 2012.] Dostupné z URL: http://commons.wikimedia.org/wiki/File:DNA_replication_en.svg.

[6] Codons aminoacids table. *Wikimedia commons.* [Online] 21. únor 2012. [Citace: 13. Říjen 2012.]Dostupné z URL: http://commons.wikimedia.org/wiki/File:Codons_aminoacids_table.png.

[7] **Ruiz, Mariana.** Ribosome mRNA translation en. *Wikimedia commons.* [Online] 29. Září 2009.[Citace: 10. říjen 2012.] Dostupné z URL: http://commons.wikimedia.org/wiki/File:Ribosome_mRNA_translation_en.svg.

[8] **Jaroslav Kypr, Michaela Vorlíčková.** *Konformační vlastnosti řetězců DNA obsahujících guanin a adenin.* 92, Asociace českých chemických společností, Brno: 1998, svazek Chemické Listy. ISSN 1213-7103. [Online] [Citace: 10. Říjen 2012.] Dostupné z URL: http://www.chemicke-listy.cz/docs/full/1998_07_530-537.pdf

[9] DNA sequence. *Wikimedia commons.* [Online] 19. říjen 2007. [Citace: 13. říjen 2012.] Dostupné z URL: http://commons.wikimedia.org/wiki/File:DNA_sequence.svg.

[10] **Svršek, Jiří.** Natura plus. [Online] VIII 2008. [Citace: 24. 10 2012.] Dostupné z URL: http://natura.baf.cz/natura/2000/8/20000803.html. ISSN 1212-6748.

[11] **Vácha, Marek.** Identifikace etických problémů plynoucích z nových poznatků o lidském genomu. *Brno 2006.* [Online] [Citace: 28. 10 2012.] Dostupné z URL: http://is.muni.cz/th/98186/lf_d/PhD29.pdf.

[12] Mus musculus (ID 52) - genome - NCBI. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 25. 10 2012.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/genome/52.

[13] **Petr, Jaroslav.** Nobelova cena za genový knockout. *osel.cz.* [Online] 10. říjen 2007. [Citace: 25. 10 2012.] Dostupné z URL: http://www.osel.cz/index.php?clanek=2994&akce=show2&dev=1.

[14] Pan troglodytes (ID 202) - genome - NCBI. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 5. listopad 2012.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/genome?term=pan%20troglodytes.

[15] **Pokorný, Milan.** Šimpanzí genom přečten. *Český rozhlas.* [Online] 1. září 2005. [Citace: 5. listopad 2012.]Dostupné z URL: http://www.rozhlas.cz/leonardo/veda/_zprava/simpanzi-genom-precten--187926.

[16] **Cvrčková, Fatima.** *Úvod do praktické bioinformatiky.* Nakladatelství Academia, Praha: 2006. ISBN 80-200-1360-1.

[17] **Provazník, Ivo.** *Sekvence a databáze - přednáška předmětu ABIN.* BRNO: VUT, Fakulta elektrotechniky, Ústav biomedicínského inženýrství, 2013.

[18] Accepted Input Formats. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 20. 04 2013.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml.

[19] **Liping, Du, Hongxia, Zhou a Hong, Yan.** OMWSA: detection of DNA repeats using moving window spectral analysis. *Bioinformatics.* [Online] 14. leden 2007. [Citace: 14. 11 2012.] Dostupné z URL: http://bioinformatics.oxfordjournals.org/content/23/5/631.full.

[20] *Bioinformatika – Návod do počítačových cvičení.* BRNO: VUT, Fakulta elektrotechniky, Ústav biomedicínského inženýrství, 2011/2012.

[21] **Maděránková, Denisa a Provazník, Ivo.** *Motive Representation in Nucleotide Densities of Bird's Mitochondrial Gene COX1.* VUT FEKT. Brno: 2012

[22] **Chytil, Jiří.** Vývojové diagramy - 1.díl. *programujte.com.* [Online] 2005. červenc 25. [Citace: 16. listopad 2012.] Dostupné z URL: http://programujte.com/clanek/2005080105-vyvojove-diagramy-1-dil/.

[23] **Podhrázský, Milan.** Kapitola 2 - Pseudokód. *Algoritmy a datové struktury.* [Online] 20. červenec 2002. [Citace: 20. listopad 2012.] Dostupné z URL: http://www.alg.webzdarma.cz/diplomka/kap2/pseudokod.html.

[24] ALB Albumin - gene. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 11. 05 2013.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/gene/213.

[25] HBB hemoglobin - gene. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 11. 05 2013.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/gene/3043.

[26] TP53 tumor protein 53 - gene. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 11. 05 2013.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/gene/7157.

[27] CDK2 cyclin-dependent kinase 2 - gene. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 11. 05 2013.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/gene/1017.

[28] SHBG sex hormone-bindig globulin - gene. *NCBI (National Center for Biotechnology Information).* [Online] [Citace: 11. 05 2013.] Dostupné z URL: http://www.ncbi.nlm.nih.gov/gene/6462.

POUŽITÉ ZKRATKY

DNA	deoxyribonukleová kyselina
RNA	ribonukleová kyselina
mRNA	mediátorová RNA
tRNA	transferová RNA
rRNA	ribozomální RNA
A	adenin (cyklická dusíkatá báze)
C	cytozin (cyklická dusíkatá báze)
G	guanin (cyklická dusíkatá báze)
T	tymin (cyklická dusíkatá báze)
ddATP	2',3'-dideoxyadenozintrifosfát
ddCTP	2',3'-dideoxycytidintrifosfát
ddGTP	2',3'-dideoxyguanozintrifosfát
ddTTP	2',3'-dideoxytymidintrifosfát
NCBI	National Center for Biotechnology Information
EMBL	European Molecular Biology Laboratory Data Library
DDBJ	DNA Data Bank od Japan
CDS	kódující část sekvence (exony genu)

OBSAH ELEKTRONICKÉ DOKUMENTACE

Přiložené soubory:

Vlachynska_Alzbeta_BP.pdf	→ elektronická verze bakalářské práce
Cti_me.txt	→ pokyny pro použití přiložených funkcí

Program:

GUI_denzitni_vektory.m	→ hlavní funkce
denzitni_vektory.m	→ pomocná funkce pro výpočet denzitních vektorů
urceni_pozic.m	→ pomocná funkce pro určení pozic s vysokým obsahem adeninu a guaninu
parseArgs.m	→ pomocná funkce pro vykreslení grafu
samexaxis.m	→ pomocná funkce pro vykreslení grafu
fastaread.m	→ pomocná funkce pro čtení z fasta souboru z bioinformatického toolboxu Matabu
seqtool.m	→ pomocná funkce z bioinformatického toolboxu Matlabu
randseq.m	→ funkce pro generování náhodné sekvence z bioinformatického toolboxu Matlabu

Analýza:

analyza.m	→ skript provádějící analýzu vlivu velikosti okna a prahu na nalezené A-G bohaté úseky pro pět zvolených gen
analyza_CDS.m	→ skript provádějící analýzu na procentuální obsah A-G bohatých oblastí v CDS vybraných genů
analyza.pdf	→ výsledky analýzy vlivu velikosti okna a prahu na nalezené A-G bohaté úseky pro pět zvolených genů
sekvence.rar	→ knihovna analyzovaných sekvencí
zarovnani.rar	→ knihovna výsledků zarovnání analyzovaných sekvencí