

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

Genome structure visualization

Bakalárska práca

2021

Bc. Oleksandr Korotetskyi

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

Genome structure visualization

Bakalárska práca

Študijný program: Informatika
Študijný odbor: 9.2.1. Informatika
Školiace pracovisko: Katedra počítačov a informatiky (KPI)
Školiteľ: doc. Ing. Ján Genči, PhD.
Konzultant:

Košice 2021

Bc. Oleksandr Korotetskyi

Abstrakt v SJ

V tejto bakalárskej práci analyzujem všeobecnú štruktúru genómu rôznych organizmov (eukaryoty, prokaryoty a vírusy) s cieľom porozumieť rozdielom a navrhnúť možné riešenia ich vizualizácie. Aby som pochopil súčasný stav problému, popisujem a porovnávam niektoré populárne existujúce programy, ktoré sú schopné vizualizovať vlastnosti genómu. V ďalšom kroku vykonávam 2D vizualizáciu a analýzu genómu SARS-CoV-2 pomocou niektorých existujúcich a mnou vyvinutých techník. Potom sa výsledky dosiahnuté počas vizualizácie overia a skripty, ktoré som na ich získanie použil, sa skomponujú do samostatnej aplikácie.

Klúčové slová v SJ

Programovanie, bioinformatika, vizualizácia údajov, genóm, covid-19

Abstrakt v AJ

In this bachelor thesis I analyze the general genome structure of different organisms (eukaryotes, prokaryotes and viruses) in order to understand the differences and come up with possible solutions for their visualization. Also, in order to understand the current state of a problematic, I describe and compare some of popular existing programs that are capable of visualizing genome properties. As the next step, I perform 2D visualization and analysis of SARS-CoV-2 genome using some existing techniques and one developed by me. Later, the results achieved during the visualization are verified and the scripts that were used to obtain them being composed into a stand-alone application.

Klúčové slová v AJ

Programming, bioinformatics, data visualization, genome, covid-19

Bibliografická citácia

KOROTETSKYI, Oleksandr. *Genome structure visualization*. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, 2021. ??s. Vedúci práce: doc. Ing. Ján Genči, PhD.

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY
Katedra počítačov a informatiky

**ZADANIE
BAKALÁRSKEJ PRÁCE**

Študijný odbor: **Informatika**

Študijný program: **Informatika**

Názov práce:

Vizualizácia štruktúry genómu
Genome structure visualization

Študent:

Oleksandr Korotetskyi

Školtiteľ:

doc. Ing. Ján Genči, PhD.

Školiace pracovisko:

Katedra počítačov a informatiky

Konzultant práce:

Pracovisko konzultanta:

Pokyny na vypracovanie bakalárskej práce:

1. Oboznámiť sa s aktuálnym stavom v oblasti analýzy všeobecnej štruktúry genómu.
2. Analyzovať existujúce spôsoby prezentácie štruktúry genómu.
3. Navrhnúť nástroj na prezentáciu štruktúry genómu vybraného organizmu.
4. Navrhnutý prototyp implementovať.
5. Prezentovať dosiahnuté výsledky.
6. Spracovať dokumentáciu podľa pokynov vedúceho práce.

Jazyk, v ktorom sa práca vypracuje: slovenský

Termín pre odovzdanie práce: 28.05.2021

Dátum zadania bakalárskej práce: 30.10.2020

.....
prof. Ing. Liberios Vokorokos, PhD.

dekan fakulty

Čestné vyhlásenie

Vyhlasujem, že som záverečnú prácu vypracoval(a) samostatne s použitím uvedenej odbornej literatúry.

Košice, 13.5.2021

.....

Vlastnoručný podpis

Podčakovanie

Na tomto mieste by som rád podčakoval svojmu vedúcemu práce za jeho čas a odborné vedenie počas riešenia mojej záverečnej práce.

Rovnako by som sa rád podčakoval svojim rodičom a priateľom, najmä *Adamovi Galuškovi* a *Sultanu Shaimardanovi* za ich podporu a povzbudzovanie počas celého môjho štúdia.

V neposlednom rade by som sa rád podčakoval spoločnosti *RedBull* a *Ozzy Osbornovi* za energiu pri napísaní tejto práce.

Obsah

Úvod	1
1 Analytická časť	3
1.1 General Genome Structure	3
1.1.1 Nukleotidy: základná podjednotka genómu	3
1.1.2 Priestorová štruktúra nukleovej kyseliny	6
1.1.3 Organizácia eukaryotického genómu	6
1.1.4 Organizácia prokaryotického genómu	9
1.1.5 Organizácia vírusových genómov	10
1.1.6 Gény: umiestnenie a všeobecná štruktúra	12
1.2 Existujúce riešenia pre reprezentáciu údajov o genóme	15
1.2.1 Webové prehliadače genómu	16
1.2.2 Funkcie a vlastnosti	16
2 Syntetická časť	21
2.1 FASTA, GFF a GBK formaty	21
2.2 Analýza a vizualizácia genómu SARS-CoV-2	23
2.2.1 Nucleotides distribution and GC-content	24
2.2.2 Gatesova metóda	25
2.2.3 Metóda 2D Matrix	28
2.2.4 Vylepšenie metódy 2D Matrix	30
2.2.5 Získanie aminokyselín	31
2.2.6 Identifikácia a vizualizácia ORF	33
2.3 Zloženie softvéru	36
3 Vyhodnotenie	39
4 Záver	43
Literatúra	44

Zoznam obrázkov

1.1	Štruktúry pyrimidínov a purínov nachádzajúcich sa v DNA a RNA. Skupiny cukru sú zvýraznené modrou farbou a dusíkaté zásady sú zvýraznené oranžovou farbou. Atómy cukru sú očíslované od 1 do 5. Atómy purínového kruhu sú očíslované od 1 do 9, zatiaľ čo tie pyrimidínového kruhu sú očíslované od 1 do 6.	4
1.2	Párovanie a komplementácia báz DNA. Dva reťazce skrutkovice, ktoré sú šípené v smere 5' až 3', sú antiparalelné. Základne na jednom vlákne špirály sú komplementárne k tým na opačnom vlákne, A vždy párovaný s T a G vždy párovaný s C	5
1.3	Dvojvláknová špirála DNA.	7
1.4	Organizácia eukaryotického genómu.	8
1.5	The nucleosome structure. H2A, H2B, H3 and H4 represent different types of histones.	9
1.6	Porovnanie eukaryotickej a prokaryotickej DNA.	10
1.7	Charakteristika vírusu	11
1.8	Both strands are read in the 5'3'direction. Each strand has three reading frames, depending on which nucleotide is chosen as the starting position.	13
1.9	Organizácia intrónov a exónov počas procesov koherentných s DNA v životnom cykle bunky.	13
1.10	Typická štruktúra párovania intramolekulárnych báz RNA.	14
1.11	Hlavné používateľské rozhranie prehľadávača genómov UCSC, ktoré zobrazuje predvolené stopy v predvolenom poradí pre klaster ľudského génu alfa-globínu.	17
1.12	Užívateľské rozhranie prehliadača genómu Ensembl s predvoleným nastavením stôp anotácií zobrazujúcich klaster génov alfa-globínu. Grafické anotácie sú zobrazené v hlavnej časti tela rozdelené do troch sekcií zhora nadol.	18

1.13 Užívateľské rozhranie prehliadača genómu ryže MSU. Prehľad chro-	
mozómov je zobrazený v hornej časti, regionálne zobrazenie je zo-	
bazené v strede a spodná časť predstavuje podrobné zobrazenie	
pre štyri stopy anotácií.	20
1.14 Prehliadač genómu Rice-Map. Podrobne informácie o jednotlivých	
položkách sa zobrazujú na pravom paneli a interpretujú zdroj úda-	
gov, umiestnenie záznamu, postupnosť a funkciu atď.	20
 2.1 Prvé 45 riadkov anotačného súboru genómu SARS-CoV-2. Pravá	
časť anotácie je na obrázku skrátená.	22
2.2 Prvé 45 riadkov obsahu súboru GBK zodpovedajúcich genómu SARS-	
CoV-2	24
2.3 Diagrama ukazujúca distribúciu SARS-CoV-2 nukleotidov.	26
2.4 Časť reprezentácie genómu SARS-CoV-2 pomocou Gatesovej me-	
tódy. Dlhá sekvencia 33 adenínových nukleotidov je označená čier-	
nym kruhom.	26
2.5 Posledné 143 nukleotidy genómu vírusu SARS-CoV-2 v súbore FASTA	
so sekvenovaným genómom. Je zvýraznená dlhá sekvencia 33 ade-	
nínových nukleotidov.	27
2.6 Časť prvej vizualizácie chromozómu Encephalitozoon Intestinalis.	
Rovnaké sekvencie sú zobrazené v strede.	27
2.7 Porovnanie genómov SARS-CoV-2 (zľava) a SARS-CoV (sprava)	
pomocou Gatesovej metódy.	28
2.8 Zakreslenie DNA do dvojrozmernej matice.	29
2.9 Vizualizácia genómu SARS-CoV-2 pomocou metódy 2D Matrix.	
Najplynulejšia sekvencia genómu je zložená z 29 903 nukleotidov	
a zobrazená matica obsahuje 29 929 pozícií (173 na každej strane),	
čierne štvorčeky v pravom dolnom rohu predstavujú prázdny pries-	
tor, ktorý sa nepoužil na vizualizáciu..	29
2.10 Vizualizácia pôvodného genómu SARS-CoV-2 [A] a rovnakého ge-	
nómu s bodovou mutáciou [B] na druhom nukleotide (T zmenené	
na G) pomocou vylepšenej metódy 2D Matrix.	31
2.11 Kódujúce oblasti genómu SARS-CoV-2 sú medzi ostatnými ORF	
zvýraznené červenou farbou. Zahŕňajú ORF1ab, ORF3a, S proteín,	
M proteín a N proteín. Vizualizácia sa vykonáva pomocou progra-	
mu BioPython	35

Zoznam tabuľiek

1.1	Dvojvláknová špirála DNA	6
1.2	Hlavné funkcie populánych prehľadávačov genómu.	19
2.1	Získané proteínové sekvencie genómu SARS-CoV-2, ktoré sú zlo- žené z viac ako 50 aminokyselín.	33
2.2	Výsledky porovnania medzi získanými proteínovými sekvenciami SARS-CoV-2 pomocou BLAST.	34
2.3	BLAST výsledky vyhľadávania pre ORF SARS-CoV-2.	36

Úvod

Poradie sekvencie DNA a jej variácie sú samotným aspektom, ktorý určuje vývojové procesy organizmu, určuje náchylnosť na rôzne choroby a jedinečne identifikuje každého tvora. Táto oblasť bola vždy na periférii záujmov vedeckej spoločnosti, od objavu v roku 1869 biochemikom Fredricha Mieschera narodeného vo Švajčiarsku. Napríklad Projekt ľudského genómu (HGP), ktorý sa začal 1. októbra 1990 a bol dokončený v apríli 2003, bol jedným z najväčších počinov. bádania v dejinách vedy. Bolo to zamerané na čítanie všetkých sekvencií DNA nášho druhu, *Homo sapiens*. Všetko vo všetkom, HGP nám predstavil schopnosť čítať kompletný genetický plán prírody pre stavbu človeka. Napriek úspešnému dokončeniu projektu však stále existuje množstvo neznámych vlastností DNA, ktoré si vyžadujú dôkladné štúdium.

Väčšina ľudí nevie, čo je bioinformatika, a táto práca je pokusom o jej ponorenie.

Pandémia COVID-19 priniesla nové výzvy pre ľudstvo a zvlášť pre bioinformatiku: po sekvenovaní by sa mal každý genóm správne analyzovať, aby sa lepšie pochopili jeho vlastnosti. Počas analýzy genómu sa často používajú rôzne vizualizačné techniky, aby sa dali dátá ľahko pochopiť. Preto si vizualizácia štruktúry a vlastností genómu SARS-CoV-2 zaslúži osobitnú pozornosť.

Prvá kapitola tejto diplomovej práce je zameraná na analýzu všeobecnej štruktúry genómu rôznych organizmov (eukaryotov, prokaryotov a vírusov) s cieľom pochopiť, ako si ich vizualizovať. Táto kapitola tiež obsahuje prehľad existujúcich riešení predstavujúcich údaje o genóme.

Druhá kapitola tejto práce obsahuje analýzu súborov, ktoré sa používajú na ukladanie údajov SARS-CoV-2, analýzu a implementáciu rôznych 2D vizuálnych techník do genómu koronavírusu. Táto práca navyše navrhuje nový prístup, ako je možné vizualizovať údaje o genóme, aby bolo možné vidieť najmenšie rozdiely v genómoch bez dôkladného skúmania. Druhá kapitola navyše popisuje zloženie softvéru, ktorý je vyrobený zo skriptov, ktoré sa použili na vizualizáciu genómu SARS-CoV-2.

Aj keď existuje niekoľko nástrojov na spracovanie DNA, problém reprezentácie rôznych vlastností genómu, ktoré sa môžu lísiť, pokiaľ ide o počet konkrétnych génov alebo úplné chromozómov (ak sú prítomné), zostáva stále aktuálny. Navyše, spracovanie genómu a jeho vizualizácia si vyžaduje efektívny prístup, pokiaľ ide o veľkosť údajov a výpočtové schopnosti priemerného počítača. Cieľom tejto práce je reprezentovať niektoré kľúčové vlastnosti genómu takým spôsobom.

Formulácia úlohy

V tejto bakalárskej práci by som chcel analyzovať štruktúru genómu organizmov, opísat existujúce riešenia predstavujúcich údaje o genóme. Samotným cieľom je vykonať analýzu, vizualizáciu a komparáciu genómu SARS-CoV-2 pomocou rôznych techník a porovnať získané výsledky s existujúcimi. Ďalším cieľom je zostaviť všetky skripty, ktoré boli použité na analýzu a vizualizáciu, do samostatnej aplikácie.

1 Analytická časť

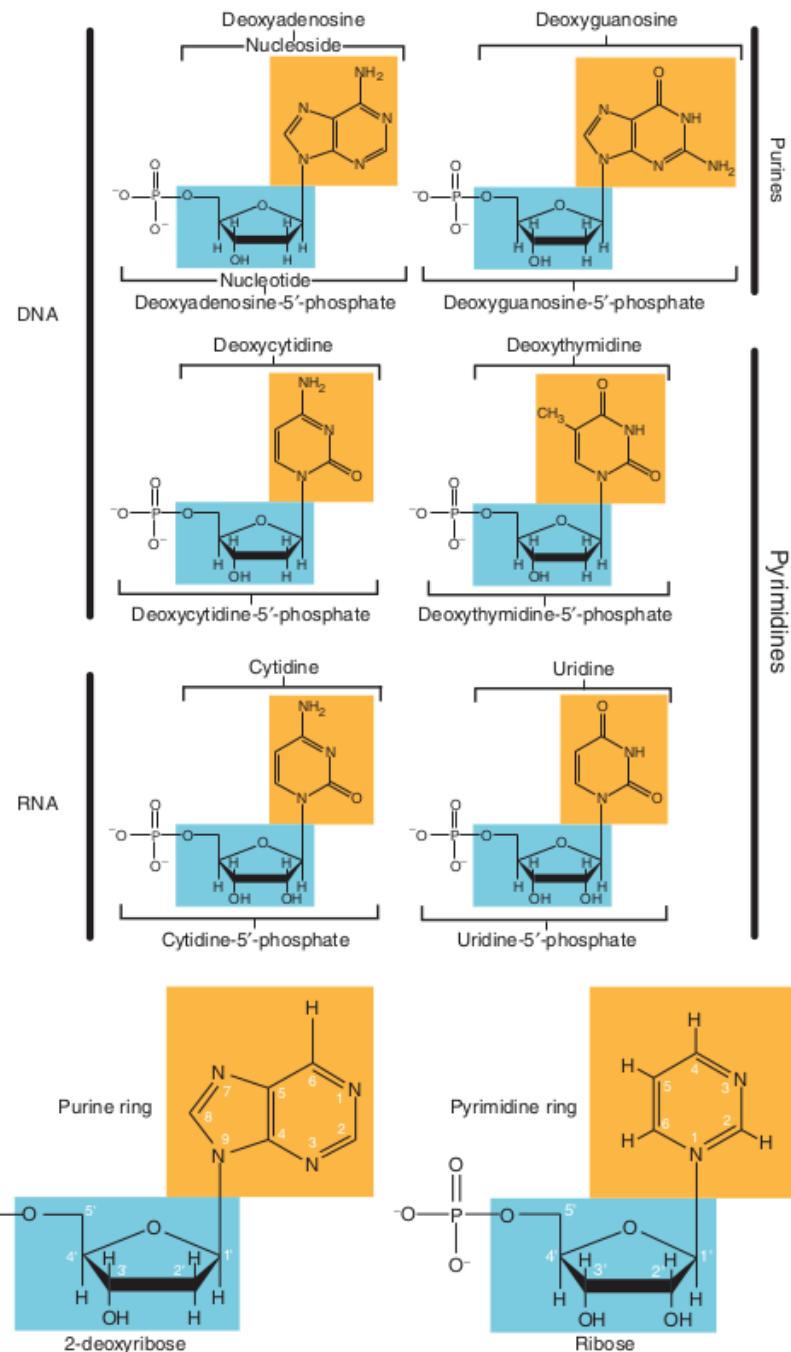
Táto kapitola sa venuje analýze štruktúry genómu rôznych organizmov, alebo inými slovami biologickému pozadiu, ktoré je nevyhnutné pre vytvorenie akéhokoľvek bioinformatického softvéru. Okrem toho, osobitne popisuje existujúce riešenia pre reprezentáciu údajov o genóme s cieľom porozumieť moderným prístupom k vizualizácii štruktúry genómu. Každá zo spomínaných tém je popísaná v jej podsekcii.

1.1 General Genome Structure

U väčšiny eukaryotických a prokaryotických organizmov je dedičným materiálom buď lineárna dvojvláknová molekula DNA (deoxyribonukleová kyselina) alebo kruhová dvojvláknová molekula DNA. Avšak, niektoré extracelulárne formy života môžu používať RNA (ribonukleová kyselina) ako stavebný blok svojho genómu. Napríklad vírusy majú genóm zložený buď z jednovláknovej DNA, dvojvláknovej DNA alebo RNA, v závislosti od typu vírusu. Samotný genóm teda je úplná genetická informácia, alebo inými slovami, všetky jedinečné sekvencie DNA (RNA) organizmu.

1.1.1 Nukleotidy: základná podjednotka genómu

DNA aj RNA sú polymérne molekuly, ktoré sú zložené z lineárnych reťazcov rôznych kombinácií štyroch rôznych podjednotiek, ktoré sa nazývajú nukleotidy. Samotný nukleotid je základnou jednotkou molekúl DNA a RNA, monomérom, ktorý sa však v bunke nachádza nielen ako nosič genetickej informácie, ale tiež ako nosič energie použitej na napájanie enzymatických reakcií [1]. Cukor s piatimi atómami uhlíka, fosfátová skupina a dusíkatá zásada sú tri odlišné zložky, ktoré spolu tvoria celkom zložitú molekulu nukleotidov (obr. 1.1). Kombinácia cukru a bázy sa nazýva nukleozid, zatiaľ čo fosfát-cukor-báza sa nazýva nukleotid.



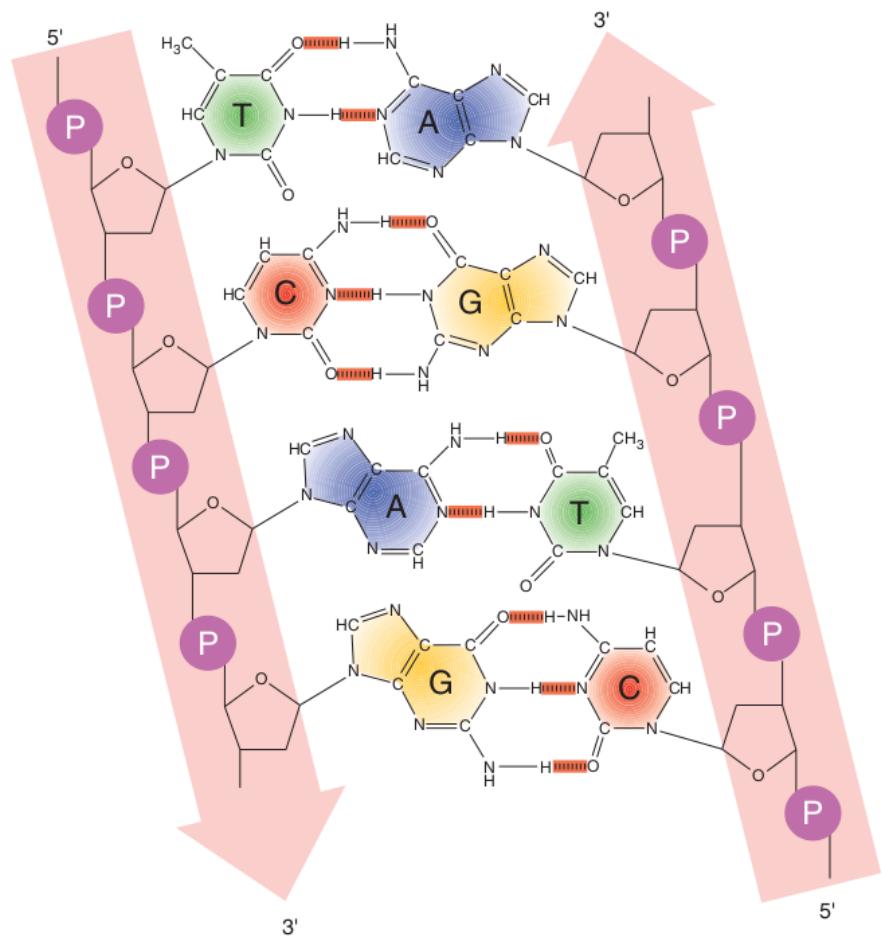
Obr. 1.1: Štruktúry pyrimidínov a purínov nachádzajúcich sa v DNA a RNA. Skupiny cukru sú zvýraznené modrou farbou a dusíkaté zásady sú zvýraznené oranžovou farbou. Atómy cukru sú očíslované od 1 do 5. Atómy purínového kruhu sú očíslované od 1 do 9, zatiaľ čo tie pyrimidínového kruhu sú očíslované od 1 do 6.

Dinukleotid, trinukleotid a polynukleotid sú výrazy zodpovedajúce dvom, trom alebo mnohým nukleotidom navzájom spojených.

Nukleotidom môže byť buď purín s dvojitým kruhom, alebo pyrimidín s jed-

ným kruhom. Guanín (G) a adenín (A) sú bežné puríny pre DNA aj RNA; pyrimidín nazývaný cytozín (C) je tiež prítomný v obidvoch nukleových kyselinách. Pyrimidín uracil (U) je však obmedzený iba na RNA, ktorý je v DNA nahradený tymínom (T).

Sú prípustné iba dve kombinácie párov báz - A párovaný s T (U) a C pár s G (obr. 1.2). Stáva sa to kvôli geometriám nukleotidových báz a relatívnym pozíciám atómov, ktoré sa podieľajú na spojení [2]. Táto vlastnosť robí dve sekvencie polynukleotidov v skrutkovice komplementárnymi.



Obr. 1.2: Párovanie a komplementácia báz DNA. Dva reťazce skrutkovice, ktoré sú šípené v smere 5' až 3', sú antiparalelné. Základne na jednom vlákne špirály sú komplementárne k tým na opačnom vlákne, A vždy párovaný s T a G vždy párovaný s C

Diskrétné nukleotidy sú navzájom spojené cukrom-fosfátovými väzbami, ktoré spájajú fosfátovú skupinu na 5' uhlíku jedného nukleotidu s hydroxylovou skupinou na 3' uhlíku iného nukleotidu. Párovanie báz medzi adenínom a tymínom (uracil) zahŕňa dve vodíkové väzby, ale medzi cytozínom a guanínom tri vodíkové väzby.

1.1.2 Priestorová štruktúra nukleovej kyseliny

Pretože trojrozmerná štruktúra nukleotidu nie je úplne tuhá, je možné, aby DNA mala rôzne priestorové architektúry: A-forma, B-forma, Z-forma a kruhová (tab. 1.1). Poloha bázy vzhľadom na cukor s piatimi atómami uhlíka sa môže meniť rotáciou a týmto spôsobom významne ovplyvňujú trojrozmernú konfiguráciu molekuly a skrutkovice následne, ktorá je viditeľna na obrázku 1.3.

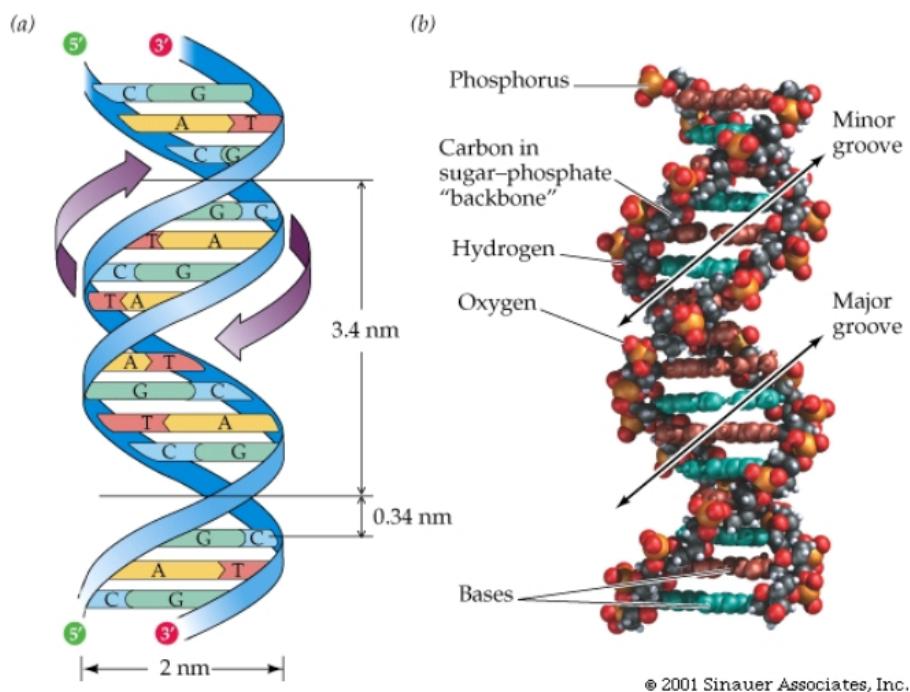
Tabuľka 1.1: Dvojvláknová špirála DNA

Vlastnosti rôznych konformácií dvojitej špirály DNA			
Vlastnosť	B-DNA	A-DNA	Z-DNA
Typ špirály	Pravotočivá	Pravotočivá	Ľavotočivá
Počet párov báz na ťah	10	11	12
Vzdialenosť medzi párami báz (nm)	0.34	0.29	0.37
Vzdialenosť na jedno otočenie (nm)	3.4	3.2	4.5
Priemer (nm)	2.37	2.55	1.84
Hlavná drážka	Široká, hlboká	Úzka, hlboká	Plochá
Vedľajšia drážka	Úzka, plytká	Široká plytká	Úzka, hlboká

Aj keď sú RNA zvyčajne jednoreťazcové, niektoré sekvencií sú schopné tvoriť dvojité špirály. Dvojité špirály RNA sú však zriedkavé a nezdá sa, že sa podieľajú na procesoch súvisiacich s genómom v eukaryotických a prokaryotických organizmoch. Naviac, kruhová DNA môže existovať v niekoľkých formách, vrátane jednostrannej C-DNA, intaktnej dvojvláknovej C-DNA (uzavreté kruhy s oboma prameňmi kovalentne spojené), prezývaná DS-C-DNA (iba jeden prameň kovalentne spojený) a vo forme "zreťazených kruhov", ale ich vlastnosti nie sú opísané v priloženej tabuľke 1.1 .

1.1.3 Organizácia eukaryotického genómu

Eukaryoti sú organizmy, ktorých bunky majú jadro uzavreté v jadrovom obale [3]. V eukaryotických bunkách sa nukleová kyselina nachádza v organele viaza-



Obr. 1.3: Dvojvláknová špirála DNA.

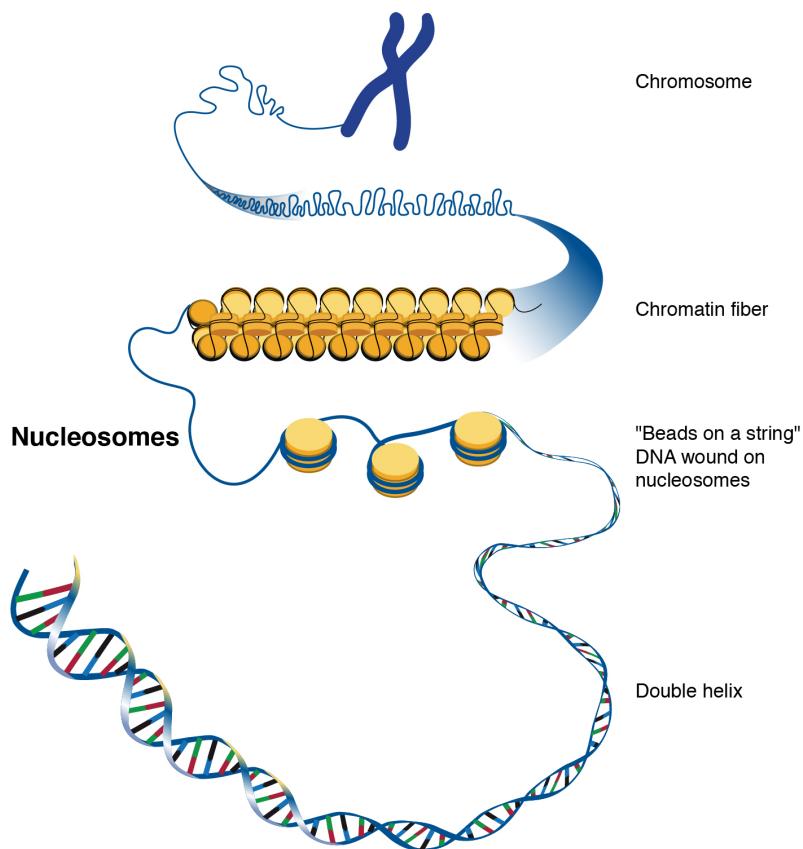
nej na membránu, ktorá sa nazýva jadro. Jadrový genóm je rozdelený na súbor lineárnych molekúl DNA s dvojitou špirálou, z ktorých každá je obsiahnutá v chromozóme. Samotný chromozóm je lineárny reťazec DNA obalený okolo asociovaných proteínov, ktoré dávajú štruktúru spojeným nukleovým kyselinám.

Nie sú známe žiadne výnimky z tohto vzorca: eukaryoty, ktoré boli študované, obsahujú molekuly DNA, ktoré sú vždy lineárne a majú najmenej dva chromozómy. Jediná variabilita na tejto úrovni organizácie eukaryotického genómu je koherentná s počtom chromozómov. Ďalej sa zdá, že biologické vlastnosti organizmu nezávisia od počtu chromozómov. [4]

Konce eukaryotických chromozómov sú tiež koncami lineárnej duplexnej DNA a ako také si vyžadujú špeciálnu štruktúru, ktorá zaistuje ich zachovanie. Dôvod je spojený so spôsobom, akým sa replikuje dvojvláknová DNA [5]. Ak by neexistoval spôsob dokončovania koncov, chromozómy by sa po každom delení buniek skracovali.

Telomery sú špecializované sekvencie nukleových kyselín, ktorých úlohou je chrániť konce chromozómov. Vo väčšine eukaryotov pozostáva telomera z krátkeho opakovania TTAGGG s dĺžkou stoviek jednotiek, ale opakovany segment sa môže medzi jednotlivými druhmi lísiť. Tieto opakovania sa tiež medzi jednotlivými druhmi značne líšia, avšak každý druh si zachováva fixnú priemernú dĺžku telomer vo svojej zárodočnej línii.

Napriek veľkosti jadra (5 - 10 um) je celková dĺžka DNA v ľudskej bunke pri-

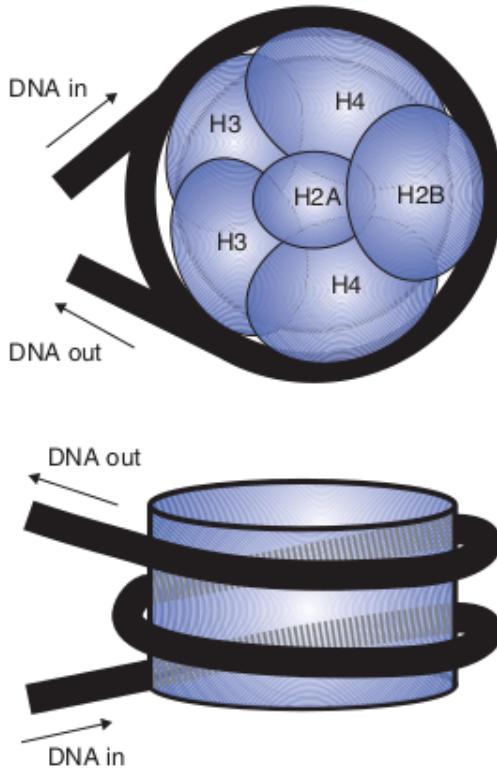


Obr. 1.4: Organizácia eukaryotického genómu.

bližne 2,1 m a môže byť zabalená do bunky vďaka spôsobu, ktorým je nukleová kyselina uložená (obr 1.4).

Genetický materiál vo vírusoch a baktériách pozostáva z reťazcov DNA alebo RNA takmer bez bielkovín. U eukaryotov je však podstatné množstvo bielkovín spojené s DNA za vzniku chromatínu. Na najnižšej úrovni je DNA organizovaná obalením vláken DNA okolo proteínov nazývaných históny, ktoré obsahujú veľké množstvo pozitívne nabitéch aminokyselín arginínu a lyzínu. Tieto aminokyseliny a históny všeobecne hrajú rozhodujúcu štrukturálnu úlohu, čo umožňuje viazať negatívne nabité fosfátové skupiny DNA nukleotidov.

V priemere DNA valcovaná okolo histónov pozostáva z 140 - 150 párov báz, v závislosti od druhu. Takýto komplex DNA a histónov sa nazýva nukleozóm (obr. 1.5). Tieto nukleozómy môžu byť ďalej navíjané do čoraz väčších závitov až do vytvorenia chromozómov. Tesné navinutie DNA však obmedzuje schopnosť buniek dostať sa k DNA a spracovať ju. Namiesto toho, aby bola nukleová kyselina neustále zvinutá, nachádza sa zvyčajne v stave zvanom chromatín, kde sú niektoré segmenty kyseliny pevne navinuté (heterochromatín), zatiaľ čo iné segmenty sú



Obr. 1.5: The nucleosome structure. H2A, H2B, H3 and H4 represent different types of histones.

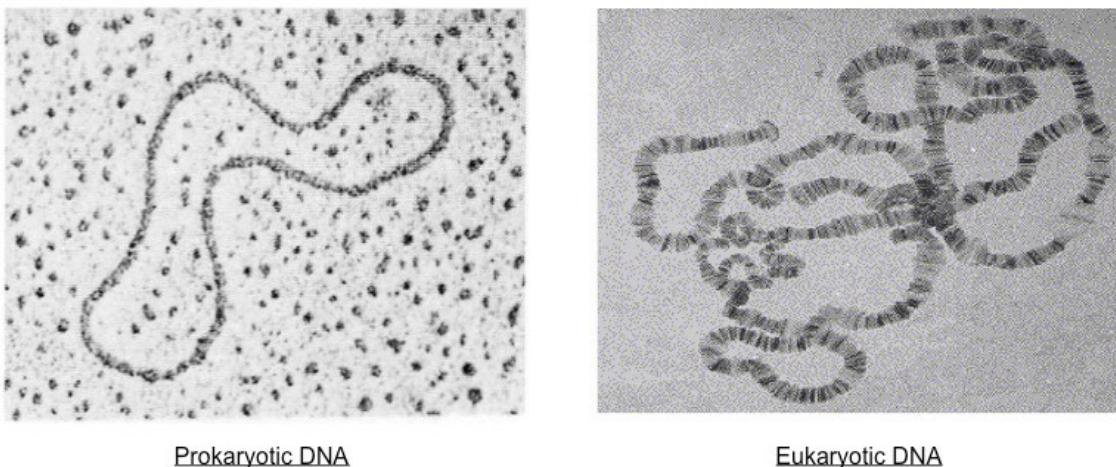
úplne otvorené (euchromatín). Euchromatínová DNA je vysoko dostupná pre molekulárne komplexy používané bunkou, a preto sa s ňou ľahšie manipuluje.

Množstvo a rozsah balenia určuje bunka, ktorá kontroluje, ktoré časti genómu je možné exprimovať a ktoré nie. Ovplyvňuje bunkovú funkciu a javí sa ako hlavná príčina diferenciácie typu buniek pri rovnakej DNA.

1.1.4 Organizácia prokaryotického genómu

Prokaryot je bunkový organizmus, ktorému chýba obalené jadro. Prokaryotické genómy sa veľmi líšia od eukaryotických, najmä čo sa týka fyzickej organizácie genómu v bunke. Aj keď sa slovo „chromozóm“ používa na opis štruktúr DNA a proteínov prítomných v prokaryotických bunkách, jedná sa o nesprávny názov, pretože táto štruktúra nepripomína obyčajný eukaryotický chromozóm. V typickom prokaryote je genóm obsiahnutý v jedinej kruhovej molekule DNA lokalizovanej v nukleoide - ľahko sa sfarbujucej oblasti inak nevýraznej prokaryotickej bunky (obr. 1.6).

Väčšina toho, čo je známe o organizácii DNA v nukleoide, pochádza zo štúdií *E. coli*. Prvou rozoznateľnou vlastnosťou, bolo, že kruhový genóm *E. coli* je zvi-



Prokaryotic DNA

Eukaryotic DNA

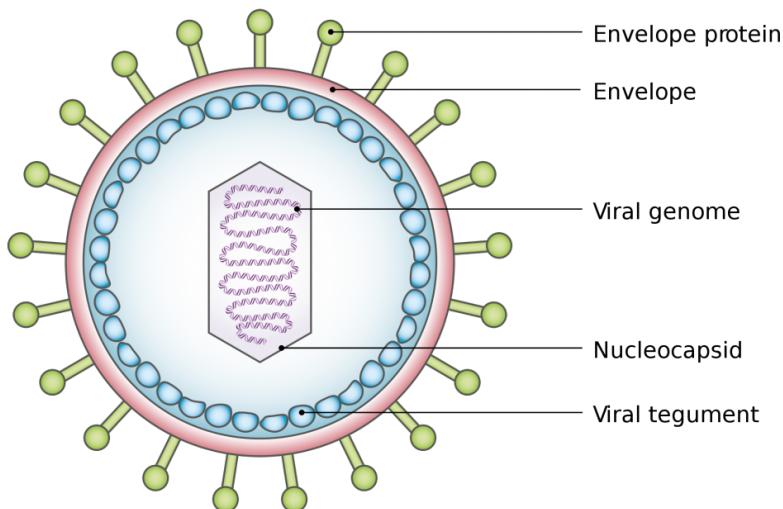
Obr. 1.6: Porovnanie eukaryotickej a prokaryotickej DNA.

nutý. Superzávitnica nastáva, keď sa do dvojitej špirály DNA zavedú ďalšie dvojité špirály (pozitívna nadzávitnica) alebo ak sa odstránia závity (negatívna nadzávitnica). Pri lineárnej molekule sa torzné napätie vyvolané pretočením alebo odtočením okamžite uvoľní rotáciou koncov molekuly DNA, ale kruhová molekula, ktorá nemá konce, nemôže týmto spôsobom uvoľniť napätie. Namiesto toho kruhová molekula reaguje navinutím okolo seba a vytvorí kompaktnejšiu štruktúru. Nadzávitnutie je preto ideálny spôsob, ako zabaliť kruhovú molekulu do malého priestoru.

Napriek konvenčnej prokaryotickej štruktúre genómu sa nachádza čoraz viac lineárnych genómov. Lineárne molekuly majú voľné konce, ktoré musia byť odlíšiteľné od zlomov DNA, takže tieto chromozómy vyžadujú terminálne štruktúry rovnocenné s telomériami eukaryotických chromozómov. U *Borrelia burgdorferi* a *Agrobacterium tumefaciens* sú skutočné konce chromozómov rozlíšiteľné, pretože sa vytvára kovalentná väzba medzi 5' a 3' koncami polynukleotidov v dvojitej špirále DNA a v *Streptomyces coelicolor* sa konce zdajú byť označené špeciálnymi väzbovými proteínmi.

1.1.5 Organizácia vírusových genómov

V prvom rade sú vírusy mimobunkovou formou života. To znamená, že ich životný cyklus a štruktúra sú vo všeobecnosti menej komplikované v porovnaní s ostatnými. Vírusy môžu mať extrémne jednoduchý dizajn, ktorý sa skladá z nukleovej kyseliny obklopanej proteínovým obalom známym ako kapsida (obr 1.7). Kapsida je zložená z menších bielkovinových zložiek, ktoré sa označujú ako kapsoméry. Kombinácia kapsid + genóm sa nazýva nukleokapsid.



Obr. 1.7: Charakteristika vírusu

Vírusy môžu obsahovať aj ďalšie zložky, pričom najbežnejšou je ďalšia membránová vrstva obklopujúca nukleokapsid, ktorá sa nazýva obálka. Obal sa skutočne získava z jadrovej alebo plazmatickej membrány infikovanej hostiteľskej bunky a potom sa modifikuje vírusovými proteínmi nazývanými peploméry. Niektoré vírusy obsahujú vírusové enzýmy, ktoré sú potrebné na infekciu hostiteľskej bunky a sú kódované vo vírusovom genóme. Kompletný vírus so všetkými zložkami potrebnými na infekciu hostiteľských buniek sa označuje ako virión.

Zatiaľ čo bunky obsahujú pre svoj genóm dvojvláknovú DNA, vírusy sa neobmedzujú iba na túto formu. Navyše, ako bolo uvedené na samom začiatku, okrem vírusov dsDNA (dvojvláknová DNA) existujú aj vírusy s jednovláknovou DNA (ssDNA), dvojvláknovou RNA (dsRNA) a jednovláknovou RNA (ssRNA). V tejto poslednej kategórii môže byť ssRNA buď pozitívna v zmysle (+ ssRNA, čo znamená, že môže prepisovať správu, napríklad mRNA), alebo môže byť v negatívnom zmysle (-ssRNA, čo naznačuje, že je komplementárna k mRNA). Niektoré vírusy dokonca začínajú jednou formou nukleovej kyseliny v nukleokapside a potom ju počas replikácie konvertujú do inej formy. Môžu byť navyše viacdielne, čo znamená, že pozostávajú z niekoľkých molekúl RNA.

Všeobecne DNA vírusy majú väčšiu veľkosť ako RNA vírusy a jednovláknové genómy sú menšie ako dvojvláknové. Existuje hypotéza, že jednovláknový vírus je menší, pretože tento typ molekuly je krehkejší ako dvojvláknová molekula. To všeobecne platí pre vírusy ssDNA aj ssRNA. Medzi dvojvláknovými genómami môžu mať buď „malé“, alebo „veľké“ genómy. Jedným z hlavných rozdielov medzi týmito dvoma genómami je mechanizmus replikácie DNA. Malé genómy využívajú aktivity hostiteľskej polymerázy, zatiaľ čo veľké genómy kódujú vlastnú

DNA polymerázu.

1.1.6 Gény: umiestnenie a všeobecná štruktúra

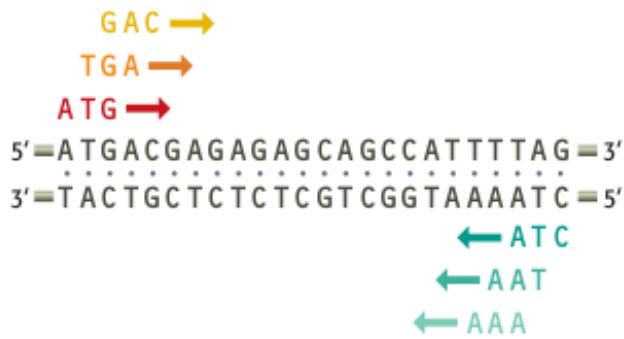
Gén je sekvencia nukleotidov v DNA alebo RNA, ktorá kóduje syntézu génového produktu, buď RNA alebo proteínu, ktoré majú charakteristické vlastnosti. Povaha všetkých týchto špecifických znakov nie je v súčasnosti úplne pochopiteľná, a preto kontrola sekvencie nie je spoľahlivým spôsobom lokalizácie génov [4].

Okrem obvyklých génov sú aj pseudogény prítomné v rôznych genómoch. Pseudogény sú sekvencie genómovej DNA s takou podobnosťou s normálnymi génymi, že sa považujú za nefunkčné kópie alebo za blízkych príbuzných génov [5]. Sú tvorené dvoma spôsobmi:

- Klasické duplikované pseudogény sa tvoria, keď tandemovo duplikované gény akumulujú mutácie tak, že jeden z génov sa stane nefunkčným. Tieto mutácie môžu zabrániť transkripcii a / alebo translácii (procesy potrebné pre syntézu proteínov).
- Spracované pseudogény sa tvoria hromadením mutácií v géne, ktorý bol retrotransponovaný na nové miesto. Vyznačujú sa absenciou intrónov (vid' d'alej), ktoré sú prítomné v rodičovskom géne.

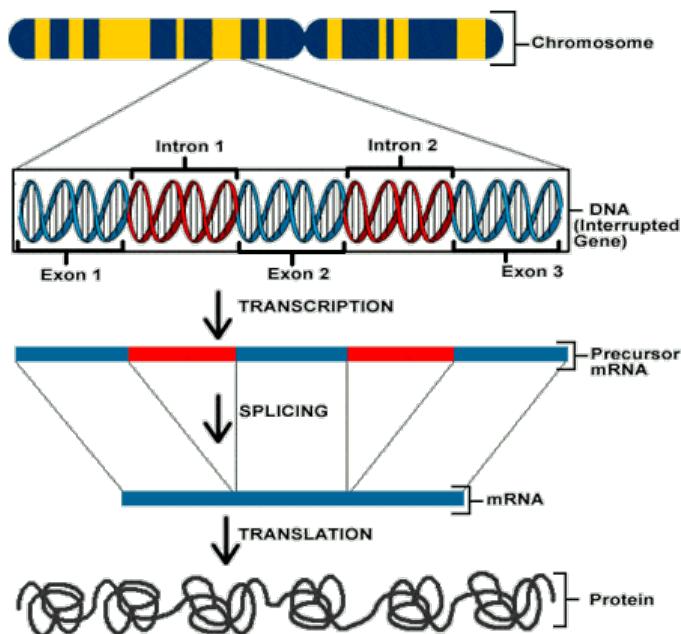
Gény, ktoré kódujú proteíny, zahŕňajú otvorené čítacie rámce (ORF) pozostávajúce zo série kodónov (trinukleotidov), ktoré špecifikujú aminokyselinovú sekvenciu proteínu, ktorý gén kóduje. ORF začína iniciačným kodónom - zvyčajne (ale nie vždy) ATG - a končí terminačným kodónom: TAA, TAG alebo TGA. Hľadanie DNA sekvencie pre ORF, ktoré začínajú ATG a končia terminačným tripletom, je preto jedným zo spôsobov hľadania génov. Analýzu komplikuje skutočnosť, že každá sekvencia DNA má na komplementárnom vlákne šesť čítacích rámcov, tri v jednom smere a tri v opačnom smere (obr. 1.8).

Kľúčom k úspechu skenovania ORF je frekvencia, s akou sa v sekvencii DNA objavujú terminačné kodóny. Ak má DNA náhodnú sekvenciu a obsah GC 50%, potom sa každý z troch terminačných kodónov - TAA, TAG a TGA - objaví v priemere raz za 64 bp. Ak je obsah GC vyšší ako 50%, potom sa terminačné kodóny, ktoré sú bohaté na AT, vyskytujú menej často, ale stále sa bude očakávať jeden za každých 100–200 bp. To znamená, že náhodná DNA by nemala vykazovať veľa ORF dlhších ako 50 kodónov [6]. Väčšina génov je naopak dlhšia ako 50 kodónov: priemerná dĺžka je 317 kodónov pre *Escherichia coli*, 483 kodónov pre *Saccharomyces cerevisiae* a približne 450 kodónov pre človeka [7]. Skenovanie ORF vo



Obr. 1.8: Both strands are read in the 5'3'direction. Each strand has three reading frames, depending on which nucleotide is chosen as the starting position.

svojej najjednoduchšej forme preto berie údaj, napríklad 100 kodónov, ako najkratšiu dĺžku predpokladaného génu a zaznamenáva pozitívne zásahy do všetkých ORF dlhších ako je tátó. With bacterial genomes, simple ORF scanning is an effective way of locating most of the genes in a DNA sequence. The real genes in the sequence cannot be mistaken because they are much longer than 50 codons in length. With bacteria the analysis is further simplified by the fact that the genes are very closely spaced and hence there is relatively little intergenic DNA in the genome (only 11% for *E. coli*). The most of bacterial genes do not overlap.

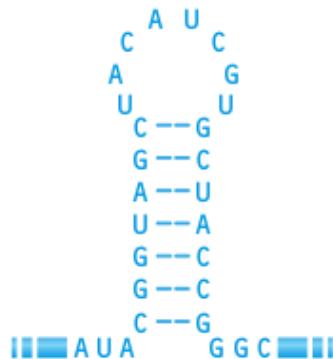


Obr. 1.9: Organizácia intrónov a exónov počas procesov koherentných s DNA v životnom cykle bunky.

Aj keď skenovanie ORF funguje dobre pre bakteriálne genómy, je menej efektívne pri lokalizácii génov v sekvenciach DNA od vyšších eukaryotov. Je to častočne preto, že medzi skutočnými génymi v eukaryotickom genóme je podstatne viac priestoru (napríklad približne 62% ľudského genómu je intergénnych), čo zvyšuje šance na nájdenie falošných ORF. Ale hlavným problémom ľudského genómu a genómov vyšších eukaryotov všeobecne je, že ich gény sú často štiepené intrónmi (nekódujúce oblasti génu), a tak sa v sekvencii DNA nejava ako spojené ORF. Mnoho exónov (kódajúcich oblastí génu) sú kratšie ako 100 kodónov, niektoré pozostávajú z menej ako 50 kodónov a pokračovanie čítacieho rámca do intrónu zvyčajne vedie k terminačnej sekvencii, ktorá sa zdá uzavrieť ORF (obr. 1.9). Inými slovami, gény vyššieho eukaryotu sa neobjavujú v sekvencii genómu, pokiaľ sú dlhé ORF a jednoduché skenovanie ORF ich nedokáže lokalizovať.

Navyše, keďže niektoré vírusy (hlavne eukaryotické) majú vo svojich genómoch štruktúry intrón-exón [8], skenovanie ORF nie je nezvratnou metódou lokalizácie génov medzi nimi.

Skenovanie ORF je vhodné pre gény kódajúce proteín, ale gény pre funkčné RNA, ako sú rRNA a tRNA, neobsahujú otvorené čítacie rámce. Funkčné molekuly RNA však majú svoje vlastné charakteristické znaky, ktoré môžu byť použité na uľahčenie ich objavenia v sekvencii genómu (obr. 1.10).



Obr. 1.10: Typická štruktúra párovania intramolekulárnych báz RNA.

Najdôležitejšou z týchto vlastností je schopnosť zložiť sa do sekundárnej štruktúry, ako je napríklad štvorlístok priatý molekulami tRNA. Tieto sekundárne štruktúry sú držané spolu párovaním báz nie medzi dvoma samostatnými polynukleotidmi, ako napríklad v dvojitej špirále DNA, ale medzi rôznymi časťami toho istého polynukleotidu - párovanie intramolekulárnych báz.

Aby sa mohli vytvoriť intramolekulárne páry báz, musia byť nukleotidové sek-

vencie v dvoch častiach molekuly komplementárne, a na vytvorenie komplexnej štruktúry, ako je štvorlístok, musia byť komponenty týchto párov komplementárnych sekvencií usporiadane v charakteristickom poradí v rámci sekvencie RNA. Tieto vlastnosti poskytujú množstvo informácií, ktoré možno použiť na lokalizáciu génov tRNA v sekvencii genómu.

Rovnako ako tRNA, rRNA a niektoré z malých funkčných RNA tiež prijímajú sekundárne štruktúry, ktoré sú dostatočne zložité na to, aby umožnili identifikáciu ich génov bez veľkých ťažkostí [9]. Menej ľahko sa dajú lokalizovať ďalšie funkčné gény RNA, pretože RNA zaberajú štruktúry, ktoré zahŕňajú relatívne malé párovanie báz alebo párovanie báz nie je v pravidelnom obrazci.

1.2 Existujúce riešenia pre reprezentáciu údajov o genóme

Vďaka rýchlemu vývoju technológií sekvenovania novej generácie sa sekvenovali stájisíce genómov. Všetky údaje o postupnosti, ako aj anotácie sa zhromažďujú v databázach genómu a sú verejne dostupné prostredníctvom webových portálov, ako je portál genómov NCBI a webová stránka databázy genómov EBI.

Systematickou integráciou sekvencií genómu spolu s anotáciami generovanými prostredníctvom mnohých heterogénnych údajov, prehliadač genómu poskytuje jedinečnú platformu pre molekulárnych biológov, aby mohli tieto genomické údaje efektívne a pohodlne prehliadať, vyhľadávať, získavať a analyzovať. Vďaka grafickému rozhraniu pomáha prehliadač genómu používateľom intuitívne extrahovať a sumarizovať informácie z obrovského množstva nespracovaných komplexných údajov.

Prehliadače genómu možno vo všeobecnosti rozdeliť na webové prehľadávače a samostatné aplikácie. Webové prehliadače genómu, ktoré sú zvyčajne vhodnejšie na podporu biologického výskumu vďaka svojej kvalite údajov, flexibilnej dostupnosti a vysokému výkonu.

- Po prvej, špecializované organizácie často zhromažďujú a integrujú vysoko kvalitné anotačné údaje do webových prehľadávačov genómu a poskytujú komunité množstvo aktuálnych informácií.
- Po druhé, používateľia k nim môžu mať prístup kdekolvek pomocou štandardného webového prehľadávača, čím sa vyhnú ďalšiemu úsiliu pri nastavovaní lokálneho prostredia pre inštaláciu aplikácií a prípravu dát.

- Po tretie, webové prehľadávače genómu sa zvyčajne inštalujú na vysoko výkonné servery a môžu podporovať zložitejšie a rozsiahlejšie dátové typy a aplikácie.

1.2.1 Webové prehliadače genómu

V súčasnosti existujú dva typy webových prehľadávačov genómu.

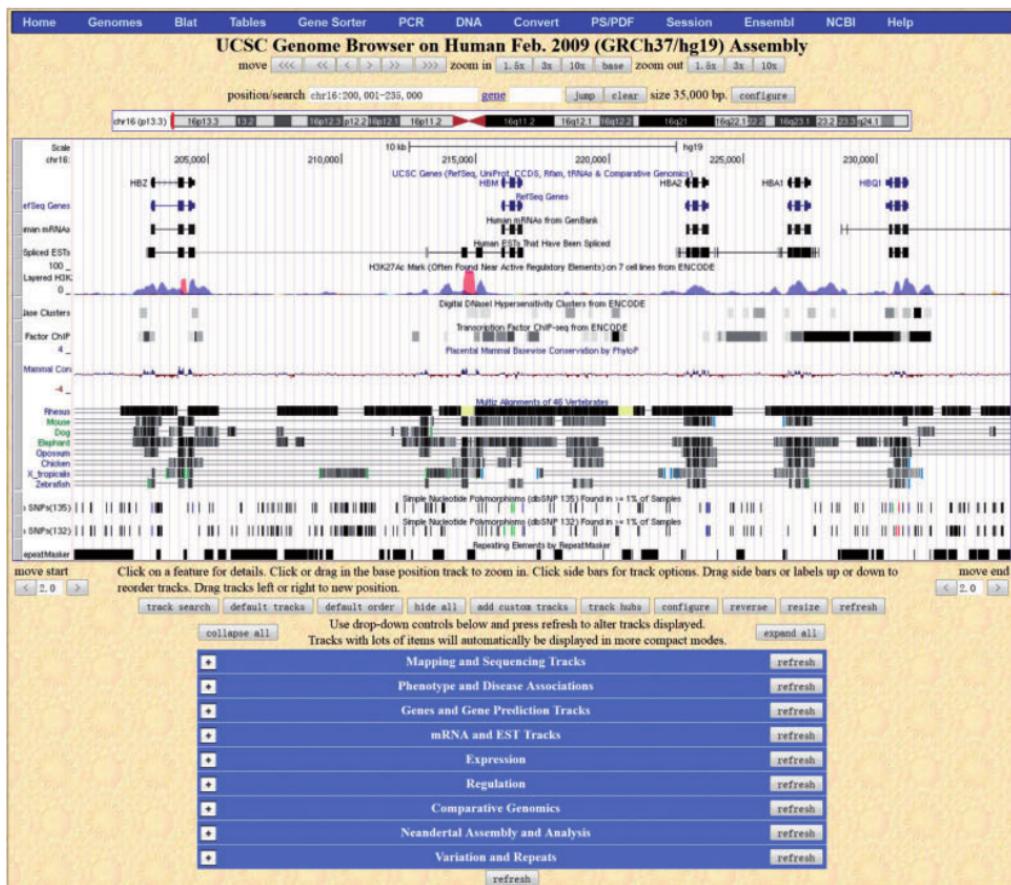
Prvým typom sú prehliadače genómu pre viac druhov, ako napríklad projekt Ensembl [10], databáza genómov UCSC [11] a webová stránka prehliadača NCBI [12]. Tieto prehľadávače genómu podporujú medzidruhovú komparatívnu analýzu. Väčšina z nich obsahuje veľké množstvo anotácií, ktoré zahŕňajú génový model, dôkazy o transkripcii, profily expresie, regulačné údaje, genómovú konverzáciu atď. Každá skupina vopred vypočítaných údajov anotácií sa v prehľadávačoch genómu nazýva stopa. Podstatou prehliadača genómu je zhromaždiť viac stôp pod rovnakou genomickou súradnicou pozdĺž osi Y, aby používatelia mohli ľahko preskúmať konzistenciu alebo rozdielnosť anotačných údajov a urobiť úsudok o vlastnostiach genomickej oblasti.

Druhým typom sú druhovo špecifické prehliadače genómu, ktoré sa zamieriajú hlavne na jeden modelový organizmus a môžu mať pre konkrétny druh viac anotácií. Poháňaný projektom Generic Model Organism Database (GMOD) sa zhromažďujú desiatky softvérových nástrojov otvoreného zdroja na vytváranie a správu biologických databáz genómu. Rámec GBrowse [13] je jedným z najpopulárnejších nástrojov v projekte GMOD. Tabuľka 1.2 uvádzajúca zoznam niekoľkých bežných webových prehľadávačov genómu, medzi ktoré patrí Ensembl, prehľadávač genómov UCSC a GBrowse, ku ktorým má prístup veľké množstvo používateľov na celom svete.

1.2.2 Funkcie a vlastnosti

Webový prehliadač genómu často poskytuje centralizovanú databázu alebo súbor databáz na ukladanie rôznych typov anotačných údajov získaných od niekoľkých organizácií. Výzvou pre všeobecné prehľadávače genómu je, ako správne zobraziť tieto informácie pre rôzne stupne genómu. Keďže je požadovaná veľká genomická oblasť, je potrebné do obrazu zahrnúť obrovské množstvo informácií, ktoré by mohli preťažiť server a sieť. Príliš veľa ťažkých a komplikovaných detailov navyše narúša pozornosť používateľa.

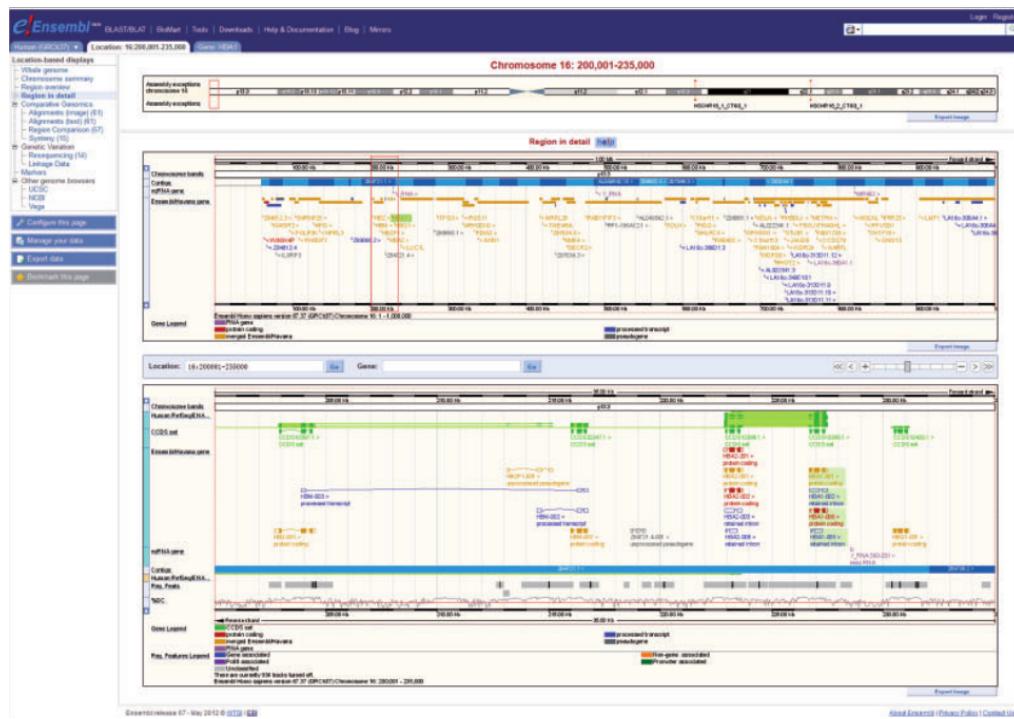
Prehliadač genómu UCSC, ktorý je jedným z veľkých hráčov vo vizualizácii genomických údajov, sa snaží tento problém vyriešiť poskytnutím viacerých zo-



Obr. 1.11: Hlavné používateľské rozhranie prehľadávača genómov UCSC, ktoré zobrazuje predvolené stopy v predvolenom poradí pre klaster ľudského génu alfa-globínu.

brazení tracku (obr. 1.12). Každú stopu je možné zobraziť v rôznych režimoch, napríklad hustú, úplne rozbalenú alebo skrytú. Používateľ môže ísť hlbšie po hustej trati 25 a otvoriť ju v plnom režime. Pre zobrazenie tracku je možné veľa stupníc. Najnižší je jeden chromozóm a najvyššia stupnica je sekvencia párov báz. Môže sa zobraziť hustý pohľad na niektoré tracky, aby sa pri oddialení na veľkú oblasť chromozómu skryli komplikované detaily, aby mal používateľ široký obraz o vybranej oblasti chromozómov.

Prehliadač genómu Ensembl poskytuje rovnaké užívateľské rozhranie pre každý organizmus. Hlavné telo rozhrania obsahuje dva panely (obr. 1.13). Na ľavom paneli je uvedená hlavná ponuka pre zobrazenia založené na polohe na rôznych úrovniach od celého genómu, súhrnu chromozómov až po prehľad oblastí a podrobnejších oblastí. Poskytnuté sú tiež odkazy na komparatívnu genomiku, genetické variácie a sekvenčné markery. Hlavný panel je usporiadaný do troch častí zhora nadol a poskytuje používateľom rôzne stupnice na analýzu genómu. Okrem zobrazenia umiestnenia poskytuje Ensembl samostatné stránky na zobraze-



Obr. 1.12: Užívateľské rozhranie prehliadača genómu Ensembl s predvoleným nastavením stôp anotácií zobrazujúcich klaster génov alfa-globínu. Grafické anotácie sú zobrazené v hlavnej časti tela rozdelené do troch sekcií zhora nadol.

nie rôznych typov informácií usporiadaných do štruktúr so záložkami. Je užitočné mať prehľad o veľkej ploche chromozómu a súčasne hľadať podrobnosti do niekoľkých malých oblastí. Spojenie paralogických génov na jednej stránke by mohlo výrazne podporiť komparatívnu analýzu.

Napríklad prehliadač sekvencií NCBI podporuje používateľov pri prezeraní rôznych oblastí vo vnútri toho istého chromozómu a poskytuje flexibilný navigačný prístup založený na viacerých paneloch s rôznymi farebnými kurzormi označujúcimi príslušné genómové polohy.

Ako už bolo spomenuté vyššie, existujú druhovo špecifické prehliadače genómu, ktoré sú k dispozícii online. Svetlým príkladom takýchto nástrojov je prehliadač genómu ryže MSU a prehliadač genómu Rice-Map [14].

V prehliadači genómu ryže MSU môžu používateelia prehľadávať gén OsSPL14 zadáním výrazu „SPL14“ do vyhľadávacieho poľa. Prehliadač genómu ryže MSU, ktorý je založený na platforme GBrowse, poskytuje anotačné pohľady v rôznych mierkach, vrátane prehľadu chromozómov, regionálneho a podrobného zobrazenia (obr. 1.13). Zväčšené zobrazenie poskytuje používateľom širší obraz, aby mohli pohodlne skontrolovať anotáciu pred a za. Na podrobnom anotačnom plátnе je poskytnutých viac ako 82 anotačných stôp, ktoré zahŕňajú génový model, dôkaz transkriptu, profilovanie expresie, zarovnanie sekvencie, genetický marker,

Tabuľka 1.2: Hlavné funkcie populárnych prehľadávačov genómu.

Vlastnosti	UCSC	Ensemble	GBrowse
Vizualizácia			
Anotácia navigácia	Prehliadanie podľa stránok, umožňujúce pretiahnutie	Prehliadanie podľa stránok	Prehliadanie podobné mape v obmedzenom regióne
Viaceré okná na stránke	-	-	-
Získavanie a analýza údajov			
Systém dopytov	Prehliadač tabuľiek	BioMart	-
Užívateľsky orientovaná analýza	Priame odoslanie údajov	-	Doplňkové nástroje
Strojovo orientované rozhranie	-	Prostredníctvom BioMart	Prostredníctvom BioDAS
Prispôsobenie			
Nahrajte stopy používateľov	+	+	+
Obsah prispievany používateľmi	Obnovenie dát na základe relácie	Mechanizmus osobnej anotácie, záložky a skupiny	-

SNP, pokrytie RNA-Seq a ďalšie genómové znaky. Okrem základných informácií o génovom modeli môžu používateľia skontrolovať tento gén v rôznych vývojových štádiách prostredníctvom rôznych expresných údajov RNA-Seq.



Obr. 1.13: Užívateľské rozhranie prehliadača genómu ryže MSU. Prehľad chro-
mozómov je zobrazený v hornej časti, regionálne zobrazenie je zobrazené v
strede a spodná časť predstavuje podrobne zobrazenie pre štyri stopy anotácií.

V prehľadávači genómu Rice-Map sú rôzne stopy anotácií usporiadané do vi-
zualizačného plátna podobného mape, pričom názov otvorených stôp je uvedený
na pravom paneli (obr. 1.14). Okrem základných anotácií génov existujú aj bohaté
anotácie pre zosúladenie krízových genómov a hodnoty ochrany, ktoré poskytujú
dôležité stopy pre vyšetrovanie tohto génu v iných rastlinách.



Obr. 1.14: Prehliadač genómu Rice-Map. Podrobne informácie o jednotlivých
položkach sa zobrazujú na pravom paneli a interpretujú zdroj údajov, umiest-
nenie záznamu, postupnosť a funkciu atď.

Pretože prehliadače genómu sú schopné poskytnúť používateľovi množstvo rôznych biologicky špecifických informácií, definícia týchto pojmov nie je uvedená v tejto práci.

2 Syntetická časť

Táto kapitola je zameraná na použitie niektorých populárnych postupov vizualizácie 2D genómu na genóm SARS-CoV-2. Pretože téma je mimoriadne zložitá, sú predstavené iba niektoré z existujúcich metód. Komplexný prehľad o týchto metodach, ich analýze a implementácii je obsiahnutý v zodpovedajúcich častiach.

Druhá časť tejto kapitoly je zameraná na zloženie funkčného, ale jednoduchého softvéru, ktorý je schopný vizualizovať genóm SARS-CoV-2 pomocou predtým opísaných techník a knižníc.

2.1 FASTA, GFF a GBK formaty

Najskôr, pochopenie toho, ako sa genomické údaje ukladajú, je klúčom ku správnej vizualizácii. Možné riešenia vizualizácie možno všeobecne rozdeliť do dvoch samostatných kategórií: tie, ktoré používajú samotnú sekvenciu genómu, a tie, ktoré používajú anotácie genómu [15].

Prvá kategória funguje na sekvencií surovej DNA (RNA) a zvyčajne sa používa na hľadanie rôznych vzorov, tandemových opakovaní, bodových mutácií alebo na vizuálne porovnanie genómov príbuzných druhov. Prvotné údaje, buď DNA, alebo aminokyselinové sekvencie, sa zvyčajne ukladajú vo formáte FASTA (prípony súborov .fasta, .fa, .fna)[16].

Zdrojový kód uvedený nižšie demonštruje štruktúru súboru FASTA obsahujúceho sekvenciu DNA SARS-CoV-2.

```
>NC_045512.2 |Coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTCGATCTCTTAGATCT
GTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCCT
CACGCAGTATAATTAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
```

Zdrojový kód 2.1: Prvých 180 nukleotidov z genómovej sekvencie SARS-CoV-2 vo formáte FASTA. Riadok popisu, ktorý sa začína znakom „>“, obsahuje informácie o postupnosti. Za začiatocným riadkom je samotná aktuálna sekvencia v štandardnom jednopísmenovom reťazci znakov.

Druhá kategória medzitým používa predspracované a dobre preštudované údaje [17], ktoré sú získané zo surovej sekvencie. Anotácie genómu obsahujú miesta kódujúcich oblastí genómu, a preto môžu byť užitočné v oblastiach spojených s genetikou, syntézou proteínov, dedičnosťou atď.

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM985889v3
#!genome-build-accession NCBI_Assembly:GCF_009858895.2
##sequence-region NC_045512.2 1 29903
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2697049
NC_045512.2 RefSeq region 1 29903 . + . ID=NC_045512.2:1..29903;Dbxref=taxon:2697049;cc
NC_045512.2 RefSeq five_prime_UTR 1 265 . + . ID=id-NC_045512.2:1..265;gbkey=5'UTR
NC_045512.2 RefSeq gene 266 21555 . + . ID=gene-GU280_gp01;Dbxref=GeneID:43740578;Name=
NC_045512.2 RefSeq CDS 266 13468 . + 0 ID=cds-YP_009724389.1;Parent=gene-GU280_gp01;Dbxref
NC_045512.2 RefSeq CDS 13468 21555 . + 0 ID=cds-YP_009724389.1;Parent=gene-GU280_gp01;Dt
NC_045512.2 RefSeq mature_protein_region_of_CDS 266 805 . + . ID=id-YP_009724389.1:1..180
NC_045512.2 RefSeq mature_protein_region_of_CDS 806 2719 . + . ID=id-YP_009724389.1:180
NC_045512.2 RefSeq mature_protein_region_of_CDS 2720 8554 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 8555 10054 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 10055 10972 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 10973 11842 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 11843 12091 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 12092 12685 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 12686 13024 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 13025 13441 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 13442 13468 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 13468 16236 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 16237 18039 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 18040 19620 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 19621 20658 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq mature_protein_region_of_CDS 20659 21552 . + . ID=id-YP_009724389.
NC_045512.2 RefSeq CDS 266 13483 . + 0 ID=cds-YP_009725295.1;Parent=gene-GU280_gp01;Dbxref
NC_045512.2 RefSeq mature_protein_region_of_CDS 266 805 . + . ID=id-YP_009725295.1:1..180
NC_045512.2 RefSeq mature_protein_region_of_CDS 806 2719 . + . ID=id-YP_009725295.1:180
NC_045512.2 RefSeq mature_protein_region_of_CDS 2720 8554 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 8555 10054 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 10055 10972 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 10973 11842 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 11843 12091 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 12092 12685 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 12686 13024 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 13025 13441 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq mature_protein_region_of_CDS 13442 13480 . + . ID=id-YP_009725295.
NC_045512.2 RefSeq stem_loop 13476 13503 . + . ID=id-GU280_gp01;Dbxref=GeneID:43740578
NC_045512.2 RefSeq stem_loop 13488 13542 . + . ID=id-GU280_gp01-2;Dbxref=GeneID:43740568;i
NC_045512.2 RefSeq gene 21563 25384 . + . ID=gene-GU280_gp02;Dbxref=GeneID:43740568;i
NC_045512.2 RefSeq CDS 21563 25384 . + 0 ID=cds-YP_009724390.1;Parent=gene-GU280_gp02;Dt
```

Obr. 2.1: Prvé 45 riadkov anotačného súboru genómu SARS-CoV-2. Pravá časť anotácie je na obrázku skrátená.

Každá anotácia genómu má 9 povinných polí:

1. ID sekvencie
2. Zdroj
 - Opisuje algoritmus alebo postup, ktorý vygeneroval túto funkciu. Typicky Genescane alebo Genebank
3. Typ vlastnosti
 - Opisuje, o čo ide (mRNA, doména, exón atď.)
4. Začiatok vlastnosti

5. Koniec vlastnosti
6. Skóre
 - Hodnoty podobnosti sekvenčí alebo predpovedí
7. Prameň (+ alebo -)
8. Fáza
 - Označuje, kde vlastnosť začína odkazom na čítací rámec (ORF)
9. Atribúty
 - Zoznam dvojíc značiek a hodnôt oddelených bodkočiarkou, ktorý poskytuje ďalšie informácie o každej funkcií

Tieto údaje sa zvyčajne ukladajú do súborov GFF (General Feature Format). Prípony názovov súborov sú .gff, .gff2. a .gff3 [18].

Okrem anotácií GFF a súborov sekvenčí FASTA sa široko používajú aj súbory GBK (formát Genbank). Formát Genbank umožňuje ukladanie informácií okrem sekvencie DNA / proteínu. Uchopenie obrazovky zobrazuje rôzne podrobnosti, prvá časť obsahuje LOCUS, DEFINÍCIU, PRÍSTUP a VERZIU záznamu a je označená výrazom PÔVOD, konečným detailom je skutočná postupnosť. Týchto päť prvkov je podstatnou súčasťou formátu GenBank.

Nepodstatné časti záznamu obsahujú takzvané metaúdaje a môžu obsahovať podrobnejšie informácie o organizme, krížové odkazy na iné databázy a dokonca aj zoznam publikácií, v ktorých je tento záznam uvedený. Časť záznamu „VLASTNOSTI“ popisuje dôležité charakteristiky sekvencie záznamu, ako je prítomnosť kódujúcich sekvenčí, proteínov atď.

2.2 Analýza a vizualizácia genómu SARS-CoV-2

Na vykonanie samotnej analýzy sa použijú balíky BioPython a DNA Features Viewer.

Na vizualizáciu genómu SARS-CoV-2 sú potrebné predtým opísané súbory s genómovými údajmi. Súbory FASTA aj GBK je možné získať na webových stránkach NCBI (MN908947).

```

LOCUS      MT350282          29903 bp   RNA    linear    VRL 17-APR-2020
DEFINITION Severe acute respiratory syndrome coronavirus 2 isolate
ACCESSION  MT350282
VERSION    MT350282.1
KEYWORDS   .
SOURCE     Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2)
ORGANISM   Severe acute respiratory syndrome coronavirus 2
Viruses; Riboviria; Nidovirales; Cornidovirinae; Coronaviridae;
Orthocoronavirinae; Betacoronavirus; Sarbecovirus.
REFERENCE  1 (bases 1 to 29903)
AUTHORS   Malta,F., Amgarten,D., de Oliveira,D.B.L., Araujo,D.B.,
Machado,R.R.G., Santana,R.A.F., Mangueira,C.L.P., Durigon,E.L. and
Pinho,J.R.R.
TITLE     Direct Submission
JOURNAL   Submitted (16-APR-2020) Clinical Laboratory, Hospital Israelita
Albert Einstein, Av. Albert Einstein, 627/701, São Paulo, SP
05652-000, Brazil
COMMENT   ##Assembly-Data-START##
Assembly Method :: Spades v. 3.11; Geneious v. 2020.1
Coverage   :: 600lx
Sequencing Technology :: Illumina
##Assembly-Data-END##
FEATURES
source      Location/Qualifiers
            1..29903
            /organism="Severe acute respiratory syndrome coronavirus
            2"
            /mol_type="genomic RNA"
            /isolate="SARS-CoV-2/human/BRA/SP02cc/2020"
            /host="Homo sapiens"
            /db_xref="taxon:2697049"
            /lab_host="VERO E6 cells"
            /country="Brazil"
            /collection_date="2020-03-18"
            266..21555
            /gene="ORF1ab"
            join(266..13468,13468..21555)
            /gene="ORF1ab"
            /ribosomal_slippage
            /codon_start=1
            /product="ORF1ab polyprotein"
            /protein_id="QJA41639.1"
            /translation="MESLVPFGNEKTHVQLSLPVLOVRDVLRGFGDSVEEVLSearq
            HKDGTCGLVEVEKGVLpqLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
            TLGVLPVHGIEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDpyEDFQEN
gene        Location/Qualifiers
            266..21555
            /gene="ORF1ab"
            /codon_start=1
            /product="ORF1ab polyprotein"
            /protein_id="QJA41639.1"
            /translation="MESLVPFGNEKTHVQLSLPVLOVRDVLRGFGDSVEEVLSearq
            HKDGTCGLVEVEKGVLpqLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
            TLGVLPVHGIEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDpyEDFQEN
CDS         Location/Qualifiers
            266..21555
            /gene="ORF1ab"
            /codon_start=1
            /product="ORF1ab polyprotein"
            /protein_id="QJA41639.1"
            /translation="MESLVPFGNEKTHVQLSLPVLOVRDVLRGFGDSVEEVLSearq
            HKDGTCGLVEVEKGVLpqLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
            TLGVLPVHGIEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDpyEDFQEN

```

Obr. 2.2: Prvé 45 riadkov obsahu súboru GBK zodpovedajúcich genómu SARS-CoV-2

2.2.1 Nucleotides distribution and GC-content

Analýza sa zvyčajne začína čítaním sekvencie DNA:

Zdrojový kód 2.2: Čítanie sekvencie DNA pomocou BioPython:

```

from Bio.SeqRecord import SeqRecord
from Bio import SeqIO
cov19 = SeqIO.read('MN908947.fna', "fasta")

```

Jednou z najdôležitejších genomických vlastností je GC-content (alebo obsah guanín-cytozínu) [19]. Je to percento dusíkatých báz v molekule DNA alebo RNA, ktoré sú buď guanín (G) alebo cytozín (C). Toto opatrenie udáva podiel báz G a C z implikovaných štyroch celkových báz, tiež zahrňajúcich adenín a tymín v DNA a adenín a uracil v RNA.

Obsah GC sa môže uviesť pre určitý fragment DNA alebo RNA alebo pre celý

genóm. Ak sa jedná o fragment, môže to znamenať obsah GC v individuálnom géne alebo časti génu (doméne), skupine génov alebo génových zhľukov alebo nekódujúcej oblasti [20].

Obsah GC sa zvyčajne vyjadruje ako percentuálna hodnota, niekedy však ako pomer. Percento obsahu GC sa počíta ako:

$$\frac{G + C}{A + T + G + C} * 100\%$$

Distribúciu nukleotidov (A, T, C, G) v DNA Covid19 je možné vypočítať pomocou priloženého kódu.

Zdrojový kód 2.3: Skript na výpočet distribúcie nukleotidov v genóme SARS-CoV-2.

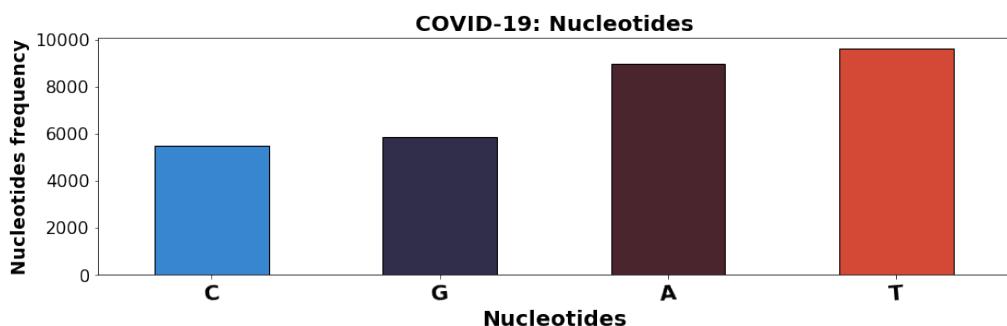
```
#Count the nucleotides frequency in the DNA
DNA = SARS_Cov_2_DNA
nucleotides = {}
for n in DNA:
    if n in nucleotides:
        nucleotides[n] += 1
    else:
        nucleotides[n] = 1

#create a dataframe
nts = pd.DataFrame(data=nucleotides,
                     index=[0]).T.reset_index()
nts = nts.rename(columns={0: 'frequency',
                           'index': 'nucleotides'})
nts = nts.sort_values(by=['frequency'], ascending=True)
```

Prvým pozorovaním je, že frekvencia nukleotidov A (8954) a T (9594) je vyššia ako frekvencia C (5492) a G (5863). Preto je obsah GC 37,97%. Pri komparácii môže byť obsah GC v eukaryotoch, ako sú stavovce, vrátane ľudí, až 60% [21].

2.2.2 Gatesova metóda

2D metódy sú primárne založené na karteziánskom súradnicovom systéme a samotnou reprezentáciou je sada bodiek alebo vektorov zodpovedajúcich rôznym vlastnostiam genómu; Gatesova metóda je typickým príkladom 2D vizualizačných techník, ktoré pracujú so surovými genomickými dátami. Preto sa počas vizualizácie spracováva súbor FASTA.

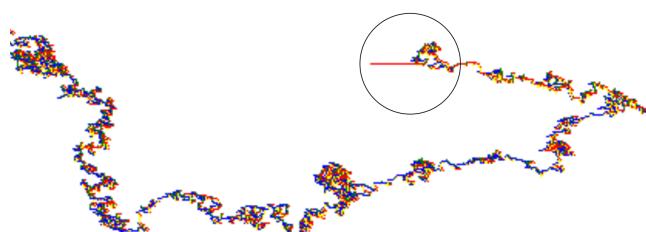


Obr. 2.3: Diagrama ukazujúca distribúciu SARS-CoV-2 nukleotidov.

Je to pochopiteľné a bol som vybraný ako jeden z metód na implementáciu, pretože som ho omylom znova objavil pri tvorbe bakalárskej práce.

Štyri bázy nukleových kyselín sú priradené k štyrom osiam 2D karteziánskeho súradnicového systému. Daná sekvencia sa vyniesie do grafu podľa distribúcie jej báz v príslušnom smere; vo výpočtoch je adenín (A) priradený k negatívnej osi x, cytozín (C) k pozitívnej osi y, guanín (G) k pozitívnej osi x a tymín (T) k negatívnej osi y. Vážený priemer súradníc x a y každého bodu sekvencie dĺžky N predstavuje ľažisko. Euklidovská vzdialenosť medzi začiatkom a centrom hmoty poskytuje kvantitatívny deskriptor grafu, ktorý sa nazýva polomer grafu.

Táto metóda môže byť použitá na hľadanie podobnosti medzi príbuznými genómami a na hľadanie vzorov v konkrétnych. Napríklad po vizualizácii sekvencie genómu SARS-CoV-2 pomocou Gatesovej metódy je sekvencia 33 adenínových (červených) nukleotidov ľahko rozlíšiteľná v celom genóme.



Obr. 2.4: Časť reprezentácie genómu SARS-CoV-2 pomocou Gatesovej metódy.

Dlhá sekvencia 33 adenínových nukleotidov je označená čiernym kruhom.

Po vyšetrení súboru FASTA môžeme vidieť, že na samom konci sekvencie genómu je 33 adenínových nukleotidov, ako je uvedené nižšie (obr. 2.5).

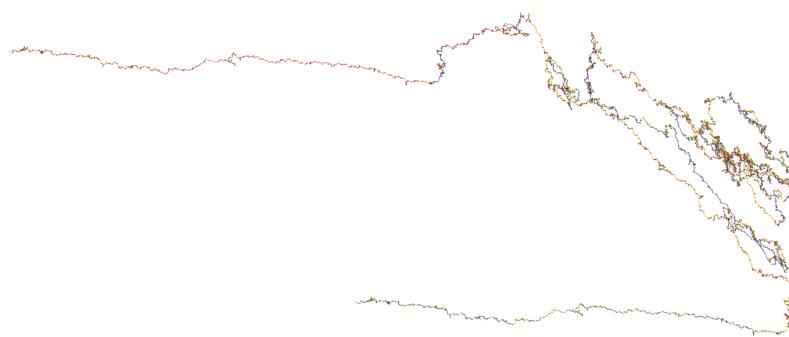
Ďalším príkladom nájdenia vzorov v sekvencii DNA pomocou Gatesovej metódy je pokus o vizualizáciu nukleotidovej sekvencie prvého chromozómu *Encephalitozoon intestinalis*, ktorý sa považuje za najmenší eukaryotický genóm [**smalleu**]. Sekvenciu je možné ľahko získať pomocou zdrojov NCBI (ID CP001942).

```

ACAGTGAAACAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAAT
TTTAGTAGTGCTATCCCCATGTGATTAAATAGCTTCTAGGAGAACAAAAAA
AAAAAAAAAAAAAAAAAAAAAA
  
```

Obr. 2.5: Posledné 143 nukleotidy genómu vírusu SARS-CoV-2 v súbore FASTA so sekvenovaným genómom. Je zvýraznená dlhá sekvencia 33 adenínových nukleotidov.

Po vizualizácii spomínaného chromozómu je prvým pozorovaním to, že obidva konce sekvencie sú takmer identické (obr. 2.6). Jediný rozdiel, s výnimkou bodových mutácií, je ten, že sú zložené z nukleotidov komplementu. Preto to možno považovať za dôkaz použiteľnosti metódy pri vyhľadávaní vzorov.



Obr. 2.6: Časť prvej vizualizácie chromozómu Encephalitozoon Intestinalis. Rovnaké sekvencie sú zobrazené v strede.

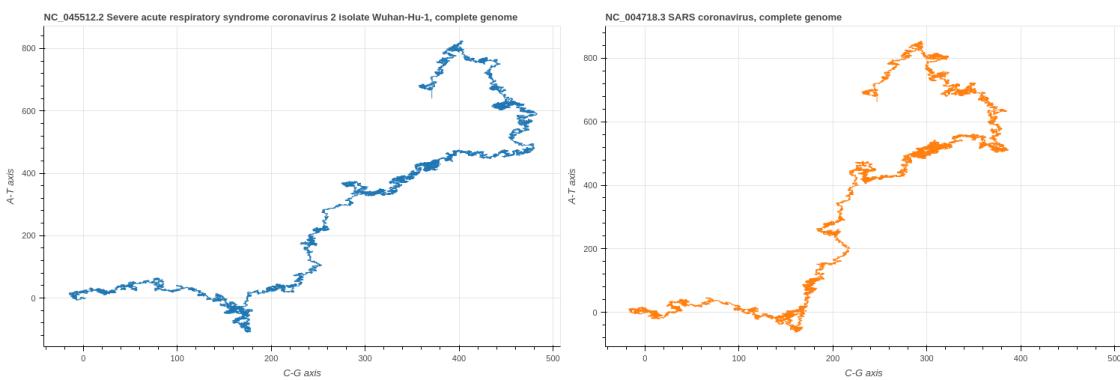
Najbližším príbuzným moderného vírusu + ssRNA SARS-CoV-2 je vírus SARS-CoV, ktorý bol príčinou vypuknutia SARS v roku 2004 [22]. Za jasný príklad použitia tejto metódy možno považovať komprimáciu sekvencií genómu SARS-CoV a SARS-CoV-2.

Samotný rozruch je možné vykonať pomocou Squiggle, dvojrozmernej knižnice vizualizácie sekvencií DNA [23].

Ako je zrejmé z poskytnutého grafu, obidva vírusy sa navzájom podobajú z hľadiska štruktúry nukleotidovej sekvencie [24].

Po spracovaní sekvencií oboch vírusov pomocou algoritmu pairwise2 sa percento podobnosti medzi nimi rovná 83,34%.

Algoritmus PairWise je variantou najlepšieho algoritmu lokálneho zarovnania Smithovho-Watermanovho algoritmu [25]. Všetky tieto algoritmy patria do triedy známej ako algoritmy minimálnej úpravy reťazcov. Bol vybraný na porovnanie sekvencií kvôli spôsobu, akým zarovnáva sekvencie. Hlavné rozdiely medzi PairWise a iným algoritmom zarovnania sú v tom, že okrem bežných trestov, ako sú Gap Opening Penalty (GOP), Gap Extension Penalty (GEP) a Match, Pair-



Obr. 2.7: Porovnanie genómov SARS-CoV-2 (zľava) a SARS-CoV (sprava) pomocou Gatesovej metódy.

Wise predstavil dva nové ďalšie parametre použité pre porovnanie [25].

Samotné porovnanie je možné ľahko vykonať pomocou knižnice BioPython, ako je to demonštrované v priloženom kóde.

Zdrojový kód 2.4: Algoritmus Pairwise2 využívajúci BioPython; COV1.seq a COV2.seq sú sekvencie DNA vírusov SARS-CoV a SARS-CoV-2. Poskytujú sa dva argumenty samotného zarovnania aby sa znížila komplexita a čas zosúlodenia.

```
from Bio import pairwise2

alm = pairwise2.align.globalxx(COV1.seq, COV2.seq,
                               one_alignment_only=True, score_only=True)
print('Similarity (%):', alm / len(COV2.seq) * 100)
```

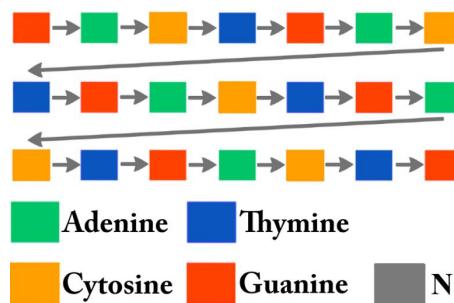
Najdôležitejšou nevýhodou Gatesovej metódy je však degenerácia, čo znamená, že vizualizácia nemusí byť nevyhnutne jedinečná. Napríklad TGAC je štvorec (hore, doprava, dole a vľavo), ale rovnako je tomu aj v GTCA (vpravo, hore, vľavo, dole) a obe sekvencie budú mať vo vizualizačnom grafe rovnakú štruktúru.

2.2.3 Metóda 2D Matrix

Ďalšou zaujímavou metódou na vizualizáciu DNA, ktorá pracuje na sekvenčii DNA, je metóda 2D matice. Cieľom je vykresliť celú sekvenčiu genómu do štvorcového obrázka s preddefinovanou veľkosťou. Plotovanie sa vykonáva z ľavého rohu do pravého rohu až po koniec čiary a potom sa presunie na ľavú stranu nového (obr. 2.8).

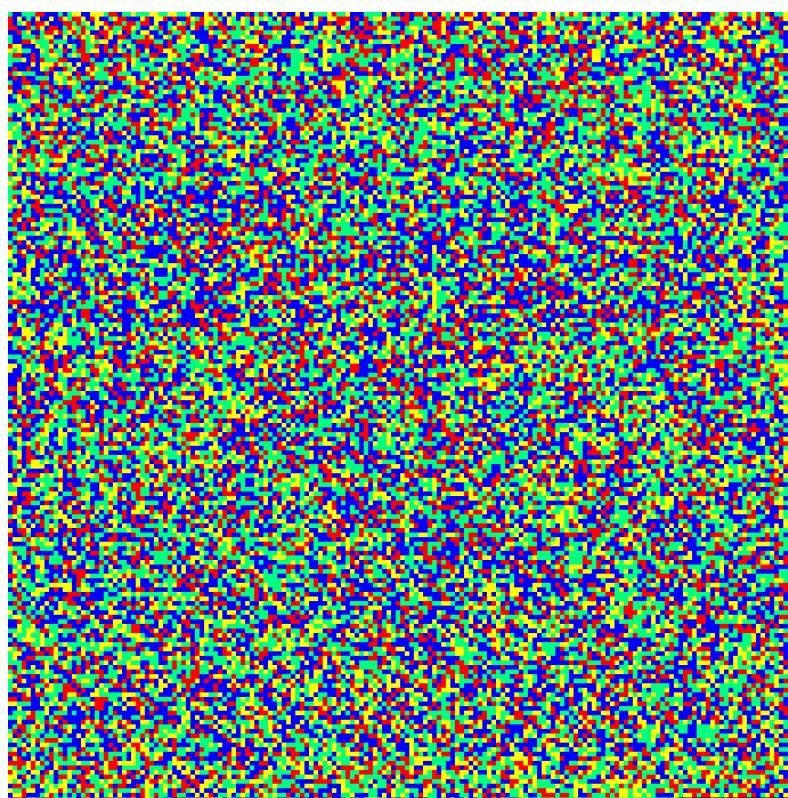
Každý nukleotid je reprezentovaný pixelom (štvorcom) konkrétnej farby.

Táto metóda by mohla byť užitočná pri hľadaní tandemových opakovaní [26] bez podrobného preskúmania strojom, pretože by mohli byť vizuálne zistiteľné.



Obr. 2.8: Zakreslenie DNA do dvojrozmernej matice.

Ako je však viditeľné na obrázku 2.9, pre genóm SARS-CoV-2 nie sú nevyhnutné žiadne významné a zrejmé tandemové opakovania. Dá sa to vysvetliť celkovou malou veľkosťou a zložitosťou genómu.



Obr. 2.9: Vizualizácia genómu SARS-CoV-2 pomocou metódy 2D Matrix. Najplynulejšia sekvencia genómu je zložená z 29 903 nukleotidov a zobrazená matica obsahuje 29 929 pozícií (173 na každej strane), čierne štvorčeky v pravom dolnom rohu predstavujú prázdny priestor, ktorý sa nepoužil na vizualizáciu..

Okrem toho táto metóda vykresluje každý genóm na obrázku pevnej veľkosti, ktorý by sa mohol použiť na jednoznačnú identifikáciu samotnej sekvencie. Avšak v prípade bodových alebo dokonca významných mutácií môže byť rozdiel v vynesenom genóme ľahko odlíšiteľný bez strojového vyšetrenia.

2.2.4 Vylepšenie metódy 2D Matrix

Táto práca navrhuje novú vizualizačnú techniku, ktorá je schopná vizualizovať každý genóm jedinečne pomocou hash-funkcie [27], čo by mohlo byť riešením vyššie uvedenej nevýhody.

Hlavnou myšlienkou vizualizačnej techniky je rekurzívny algoritmus, ktorý rozdeľuje obraz na menšie časti a farbí každú z nich v závislosti na predchádzajúcej. Toto zabráni náhodnému šumu, ktorý by sa mohol objaviť, rozdelením na malé kúsky a samostatným vyfarbením každého z nich. Namiesto toho existujú väčšie regióny, ktoré si zachovávajú určitú kontinuitu, aj keď sa ich základné časti rozchádzajú.

Rekurzívny algoritmus sa skladá z troch funkcií:

- Funkcia, ktorá dosahuje hash nukleotidovej sekvencie pomocou algoritmu sha256 [28].
- Funkcia, ktorá rekurzívne rozdelí počiatočný prázdný obrázok na 1/8 častí. Pre každú z počiatočných oblastí sa to robí rekurzívne 8-krát.
- Funkcia, ktorá zafarbí každú časť podľa hashu a vloží každú časť nad väčšú. Parameter nepriehľadnosti, ktorý je vypočítaný pre každú veľkosť oddielu a slúži na zabránenie prekrývania medzi farbami menších oddielov a väčších (ktoré sú predtým zafarbené).

Zoznam 2.5 ukazuje výstup z konzoly počas vytvárania obrázka s veľkosťou 512x512 pixelov. Hodnoty šírky a výšky predstavujú veľkosť rozdelených obrázkov, ktoré sú zafarbené, a hodnota krytie zobrazuje krytie počas každého prekrytia oddielu. Obrázky generované na ilustráciu tejto metódy používali rovnicu krytie definovanú ako:

$$\text{opacity} = 256 * (\text{level}/2)/2^{(\text{level}-1)}$$

kde úroveň predstavuje súčasnú hĺbku rekurzie. Túto rovnicu je však možné zmeniť tak, aby predstavovala ďalšie schémy vyfarbovania.

Zdrojový kód 2.5: Výstup z konzoly vyrobený počas vizualizácie pomocou vylepsenej metódy 2D Matrix.

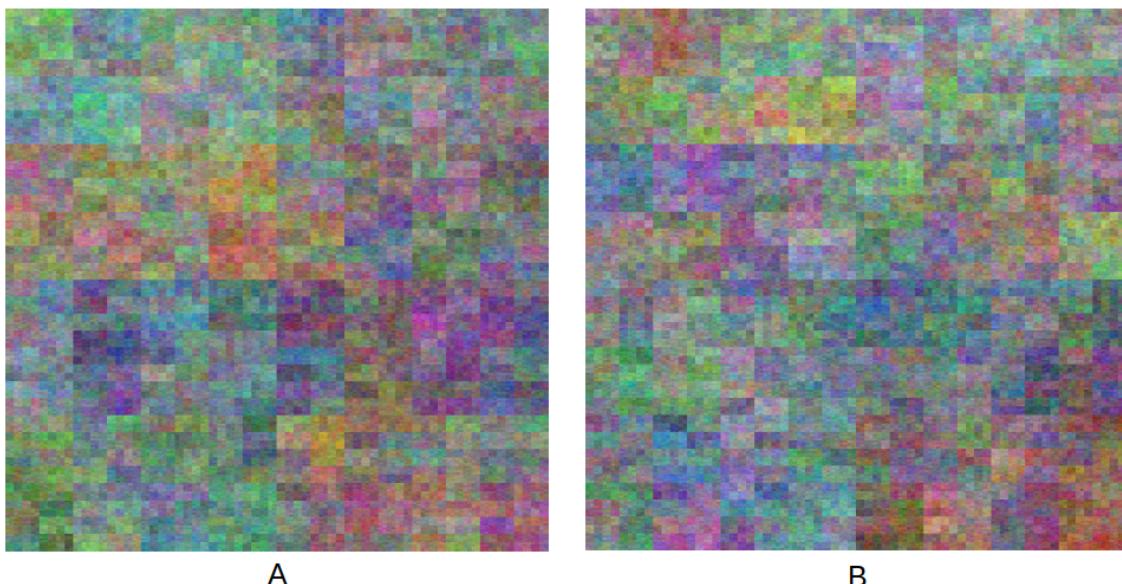
```
seed: 223dd4d5c80b98e69f9f8536afa54b...
width, height: 256.0, 256.0
opacity: 50%
width, height: 128.0, 64.0
```

```

opacity: 37%
width, height: 32.0, 32.0
opacity: 25%
width, height: 16.0, 8.0
opacity: 15%
width, height: 4.0, 4.0
opacity: 9%
width, height: 2.0, 1.0
opacity: 5%

```

Ako je viditeľné na obrázku 2.10, táto metóda umožňuje jednoduché rozlíšenie rôznych genómov voľným okom.



Obr. 2.10: Vizualizácia pôvodného genómu SARS-CoV-2 [A] a rovnakého genómu s bodovou mutáciou [B] na druhom nukleotide (T zmenené na G) pomocou vylepšenej metódy 2D Matrix.

Kvôli rekurzívnej povahy algoritmu môže byť obraz generovaný v menších alebo väčších veľkostiach a pri zachovaní rovnakej úrovne detailov.

Medzi hlavné nevýhody však patrí malé praktické využitie a obmedzenie týkajúce sa veľkosti genómu. V súčasnosti metóda podporuje vizualizáciu genómov, ktoré majú menej ako 262 145 nukleotidov.

2.2.5 Získanie aminokyselín

Pre ďalšiu vizualizáciu musí byť surová sekvencia DNA prevedená na sekvenciu aminokyselín. Existuje 61 kodónov (trinukleotidov) pre 20 aminokyselín a

každý z nich je „načítaný“, aby určil určitú aminokyselinu z 20, ktoré sa bežne nachádzajú v bielkovinách. Každá aminokyselina môže byť napísaná ako písmeno latinskej abecedy. Preto je sekvencia aminokyselín predstavovaná sekvenciou písmen A-V.

Jeden kodón, AUG, špecifikuje aminokyselinu metionín a tiež slúži ako štartovací kodón na signalizáciu začiatku stavby proteínu. Existujú ďalšie tri kodóny, ktoré nešpecifikujú aminokyseliny: UAA, UAG a UGA, ktoré povedia bunke, keď je polypeptid kompletnejší [29]. Celkovo sa táto zbierka vzťahov kodón-aminokyselina nazýva *genetický kód*, pretože umožňuje bunkám „dekódovať“ mRNA do reťazca aminokyselín.

Pred konverziou sekvencie DNA na aminokyselinu je potrebné ju najskôr transkribovať do molekuly mRNA [30] pomocou funkcie transkripcie (). Naštastie s funkciou translate () BioPython prevádzza mRNA na aminokyselinové reťazce (zdrojový kod 2.5). Reťazce sú oddelené znakom *, čo je stop kodón (UAA, UAG a UGA).

Zdrojový kód 2.6: Prepis a preklad pomocou BioPythonu

```
cov_DNA = covid19.seq
cov_mRNA = covid_DNA.transcribe()
cov_aa = covid_mRNA.translate()
```

Som zistil, že genóm SARS-CoV-2 obsahuje 9967 aminokyselín oddelených stop kodónmi * alebo, inými slovami, 775 aminokyselinových reťazcov. Je potrebné spomenúť, že nie všetky aminokyselinové sekvencie sú bielkoviny. Iba sekvencie s viac ako 20 aminokyselinami kódujú funkčné proteíny [31]. Krátke aminokyselinové sekvencie sú oligopeptidy a majú ďalšie funkčné skupiny. Ďalším krokom je filtrovanie získaných sekvencií takým spôsobom, že zostanú len tie dlhé, aby sa sústredili iba na proteíny.

Po odstránení krátkych proteínov iba 5 zvyšných proteínov splňa podmienku dĺžky (zdrojový kód 2.6) a sú uvedené v tabuľke 2.1.

Najjednoduchší spôsob overenia výsledkov je nájsť proteínové sekvencie, ktoré sú už k dispozícii v databázach a ktoré sú najviac podobné získaným proteínovým sekvenciám. Na tieto účely sa použilo vyhľadávanie BLAST.

BLAST (základný vyhľadávací nástroj na lokálne zarovnanie) je algoritmus a program na porovnanie informácií o primárnych biologických sekvenciách, ako sú napríklad aminokyselinové sekvencie proteínov alebo nukleotidy sekvencií DNA a / alebo RNA. Vyhľadávanie BLAST umožňuje výskumníkovi porovnať predmetnú proteínovú alebo nukleotidovú sekvenciu (nazývanú dopyt) s knižnicou alebo databázou sekvencií a identifikovať sekvencie knižnice, ktoré sa po-

Tabuľka 2.1: Získané proteínové sekvencie genómu SARS-CoV-2, ktoré sú zložené z viac ako 50 aminokyselín.

	Aminoacid sequence	Protein length
1	CTIVFKRVCVGSAARLTPCGTGTSTDVVYRAFDIYND...	2701
2	ASAQRSQITLHINELMDLFMRIFTIGTVTLKQGEIKD...	290
3	TNMKIILFLALITLATCELYHYQECVRGTTVLLKEPC...	123
4	AQADEYELMYSFVSEETGTLIVNSVLLFLAFVVFLV...	83
5	QQMFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINL...	63

dobajú dopytovanej sekvencii nad určitou hranicou.

Zdrojový kód 2.7: Filtrovanie aminokyselinových sekvencií a ich ukladanie do dátového rámca

```
Proteins = covid_aa.split('*')

#Remove proteins with less than than 50 amino acids
for i in Proteins[:]:
    if len(i) < 50:
        Proteins.remove(i)

#Store the protein sequences in a pandas dataframe

proteins=pd.DataFrame(Proteins)
proteins['amino acid sequence'] = proteinas[0].apply(str)
proteins['Protein length'] = proteinas[0].apply(len)
proteins.rename(columns={0: "sequence"}, inplace=True)
pro=proteins.drop('sequence', axis=1)
pro=pro.sort_values(by=['Protein length'], ascending=False)
```

Po vyhľadaní reťazca 83 aminokyselín pomocou BLAST výsledky ukázali, že má 100% podobnosť s malým membránovým proteínom Envelope, ktorý patrí do genómu SARS-CoV-2. Výsledky vykonaných BLAST hľadaní ďalších získaných proteínov sú uvedené v tabuľke 2.2. Za zmienku stojí, že najvyššia podobnosť sa zistila medzi ostatnými druhmi koronavírusov.

2.2.6 Identifikácia a vizualizácia ORF

Ďalšia vizualizácia pracuje s predspracovanými údajmi uloženými v súbore anotácií GenBank. Ako bolo spomenuté v prvej kapitole tejto práce, skenovanie ORF nie je úplným dôkazom nájdenia génových polôh, pretože nie každý ORF je gé-

Tabuľka 2.2: Výsledky porovnania medzi získanými proteínovými sekvenciami SARS-CoV-2 pomocou BLAST.

	Dĺžka proteinu	DB:ID	Organism	Protein	Zhoda
1	2701	P0C6X7	Replicase polyprotein 1ab	Replicase polyprotein 1ab	96%
2	290	Q0Q474	Bat coronavirus 279/2005 (BtCoV)	Protein 3	75%
3	123	Q3I5J0	Bat coronavirus Rp3/2004	Protein 7a	89%
4	83	P0DTC4	Human SARS coronavirus (SARS-CoV-2)	Envelope small membrane protein	100%
5	63	Q3I5J1	Bat coronavirus Rp3/2004	Non-structural protein 6	69%

novým začiatkom. Čím je však ORF dlhší, tým je pravdepodobné, že je súčasťou génu. [6]

Identifikácia kódujúcich sekvencií (CDS) je dôležitým krokom vo funkčnej anotácii génov. Typický CDS začína ATG a končí stop kodónom. CDS je sekvencia nukleotidov, ktorá zodpovedá sekvencii aminokyselín v proteíne [32]. Preto bola na identifikáciu kódujúcich oblastí genómu nevyhnutná analýza a získanie aminokyselín (vykonané v časti 2.2.5). Genóm SARS-CoV-2 kóduje až 50 neštrukturálnych, štrukturálnych a doplnkových proteínov [33].

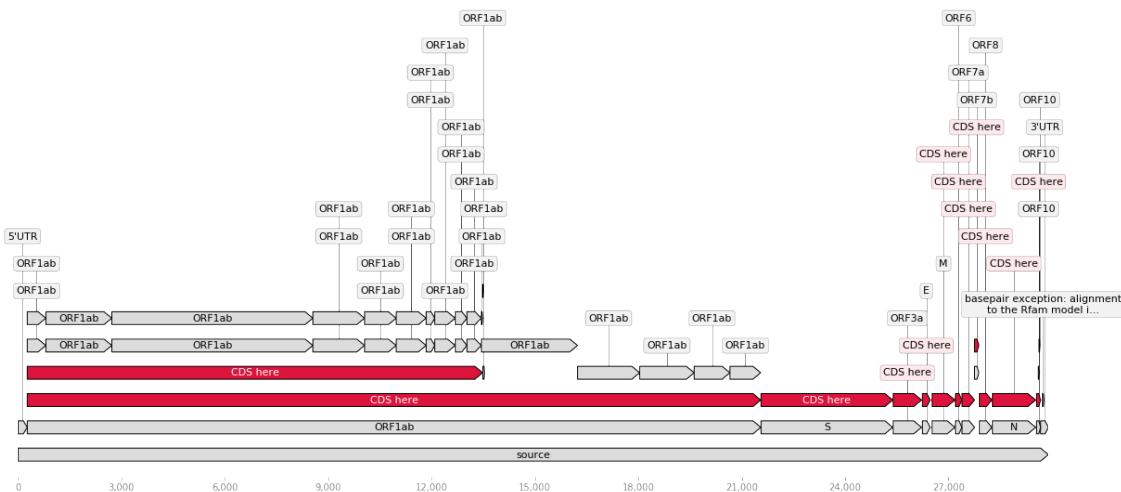
Zdrojový kód 2.7 je výstupný skript pre vizualizáciu, ktorý nájde ORF v genóme SARS-CoV-2 pomocou anotačného súboru genómu. Minimálna dĺžka proteínu je nastavená na 200 aminokyselín, aby sa získali iba dlhé ORF.

Zdrojový kód 2.8: Pozície CDS genómu SARS-CoV-2

```
PKGK MESLVP GFNE KTH...FAV - length 4409, strand 1, 253:13483
CTIV FKR VCGV SAARLT...VNN - length 2701, strand 1, 13449:21555
LEKT TELLFL VMFLLTT...HYT - length 1293, strand 1, 21502:25384
ASAQR SQITLH INELMD...VPL - length 290, strand 1, 25347:26220
SSGLN ELNI ILVFLFGT...LVQ - length 243, strand 1, 26459:27191
RSCCF RFLH NEQT KMSD...TQA - length 433, strand 1, 28231:29533
```

Som zistil, že genóm Covid-19 má 6 ORF s viac ako 200 aminokyselinami. Obrázok 2.11 ukazuje všetky ORF a CDS načítané z genómu SARS-CoV-2 pomocou BioPython.

COVID-19: Coding regions CDS



Obr. 2.11: Kódujúce oblasti genómu SARS-CoV-2 sú medzi ostatnými ORF zvýraznené červenou farbou. Zahŕňajú ORF1ab, ORF3a, S proteín, M proteín a N proteín. Vizualizácia sa vykonáva pomocou programu BioPython

Výsledky porovnania získaných sekvencií s existujúcimi sekvenciami sa usku-

Tabuľka 2.3: BLAST výsledky vyhľadávania pre ORF SARS-CoV-2.

	Dĺžka ORF	DB:ID	Protein	Organism	Zhoda
1	4409	P0C6U8	Replicase polyprotein 1a	Human SARS coronavirus (SARS-CoV)	80%
2	2701	P0C6X7	Replicase polyprotein 1ab	Human SARS coronavirus (SARS-CoV)	96%
3	1293	P59594	Spike glycoprotein	Human SARS coronavirus (SARS-CoV)	76%
4	290	Q0Q474	Protein 3	Bat coronavirus Rp3/2004	95%
5	243	Q0Q472	Membrane protein	Bat coronavirus Rp3/2004	92%
6	433	P59595	Nucleoprotein N	Human SARS coronavirus (SARS-CoV)	91%

točňujú pomocou programu BLAST a sú uvedené v tabuľke 2.3.

2.3 Zloženie softvéru

Softvér vyvinutý v priebehu tejto práce sa javí ako súbor skriptov používaných na vizualizáciu genómu SARS-CoV-2 v skôr opísaných metódach. Na rozdiel od webových prehľadávačov genómu, kde sa výpočty vykonávajú na strane servera, navrhovaný softvér všeobecne predstavuje jednoduchú samostatnú konzolovú aplikáciu. Nevykonávajú sa nijaké významné výpočty, a preto aplikácia nespĺňa náročné systémové požiadavky.

Napriek skôr opísaným moderným riešeniam reprezentujúcim údaje o genóme sa tento softvér javí ako iba jednoduchý vizualizačný nástroj, ktorý nepodporuje rôzne vlastné vizualizačné stopy ani navigáciu. Nedostatok prispôsobenia možno vysvetliť extrémnou zložitosťou vytvárania plne funkčného prehľadávača genómu.

Softvér sa píše hlavne pomocou balíka BioPython v Pythone 3.8. Všetky požadované balíčky sú uvedené v projektovej dokumentácii.

Hlavnou myšlienkou programu je umožniť používateľovi zvoliť si, ktoré informácie týkajúce sa genómu vírusu si chce zobraziť. Skladá sa z 9 modulov, ktoré majú vlastnú úlohu pri samotnom procese vizualizácie:

1. **Main Module** je jadrom programu. Zodpovedá za zabezpečenie navigácie v rámci programu používateľom. Zaoberá sa vstupom a výstupom z konzoly, navrhuje dostupné metódy vizualizácie a získava podrobnosti potrebné na ich výkon.
2. **Sequence Collector** je zodpovedný za stiahnutie všetkých požadovaných sekvencií a súborov anotácií z databázy NCBI.
3. **Statistics Generator** získava štatistické údaje, ako je obsah GC a distribúcia nukleotidov / aminokyselín. Užívateľ si môže zvoliť oblasť genómu, ktorá sa má štatisticky zhromažďovať, a zvoliť si, či chce, aby sa jeho údaje ukladali do konkrétneho súboru.
4. **Gates Visualization** vykoná vizualizáciu pomocou Gatesovej metódy do súboru .png. Užívateľ je schopný upraviť farebné závislosti.
5. **2D Matrix** modul vykreslí vybraný genóm do 2D matice do súboru .png. Veľkosť výstupného obrázka sa počíta automaticky.
6. **2D IMatrix** modul vykreslí genóm do 2D matice vybranej veľkosti pomocou algoritmu hash funkcie do súboru .png. Čím väčšia veľkosť, tým viac pamäte RAM je potrebných.
7. **Protein Plotter** generuje proteínové sekvencie genómu podľa nukleotidovej sekvencie. Užívateľ je schopný odstrániť tie, ktoré nespĺňajú dĺžkové podmienky a je schopný zvoliť si, či výsledky vytlačí alebo ich uloží do súboru .csv.
8. **ORF Plotter** generuje obraz ORF SARS-CoV-2 a pomeru obsahu GC v genóme.

9. **Comparsion** modul vykonáva porovnanie vybraných genómov. Percento podobnosti sa získa pomocou algoritmu pairwise2.

Moduly nie sú schopné vzájomne interagovať, ale každý z nich je možné zavolať z hlavného.

Každý z modulov je popísaný v dokumentácii a je sprevádzaný príkladmi vstupu a výstupu. Každý modul navyše obsahuje rôzne testy za behu, ktoré zabraňujú neočakávanému správaniu programu.

V súčasnosti všetky moduly okrem ORF Plotter podporujú spracovanie rôznych genómov, pretože výkon týchto metód nezávisí od konkrétnych vlastností genómu. Neodporúča sa však používať Proteínový plotter s genómami, ktoré kvôli svojej zložitosti obsahujú zložité intrón-exón štruktúry. V takom prípade sa výsledky môžu výrazne lísiť od skutočných charakteristík genómu.

3 Vyhodnotenie

V prvej časti tejto práce som analyzoval všeobecnú štruktúru genómu rôznych organizmov, DNA a RNA z molekulárneho, biologického a informatického hľadiska.

Priestorová štruktúra DNA a jej rozdiely sú uvedené v tabuľke 1.1.

Genomy eukaryotov, prokaryotov a vírusov boli opísané v zodpovedajúcich častiach. Opis týchto genómov zahrnuje ich jedinečné vlastnosti, ako sú štruktúry intrónu a exónu prítomné hlavne v genómoch eukaryotov a génové zhluky prítomné v genómoch prokaryotov. Gény, ich typy, polohy, funkcie a vzorce, ktoré umožňujú ich nájdenie (ORF), boli analyzované a správne opísané.

Ďalším krokom bola klasifikácia existujúcich riešení pre reprezentáciu údajov o genóme, aby sa dosiahlo pochopenie toho, ako moderný softvér funguje a čo môže navrhnuť používateľom po celom svete. Môžu byť klasifikované ako *webové a samostatné aplikácie*.

Okrem toho boli presne opísané dve kategórie existujúcich prehľadávačov genómu:

- **Species-independent** boli analyzované a porovnané riešenia, ktoré sú schopné vizualizovať akýkoľvek genóm, vrátane prehľadávača genómov Ensemble, prehľadávača UCSC a GBrowse (tabuľka 1.2).
- **Species-specific** riešenia, ktoré sú zamerané na vizualizáciu konkrétnych druhov, vrátane prehliadača genómu MSU a Rice-Map.

Potom sa kvôli vizualizácii vybral genóm vírusu SARS-CoV-2 kvôli extrémnej zložitosti prokaryotických a eukaryotických genómov a kvôli svetovej pandémii.

V druhej časti tejto práce som s cieľom pochopiť, ako a kde sa údaje o genóme ukladajú, som vykonal komplexnú analýzu formátov, ktoré sa zvyčajne používajú na ukladanie informácií súvisiacich s genómom: formát súborov FASTA, GFF a GenBank (GBK) formáty súborov. Ako príklad pre samotnú analýzu boli použité súbory súvisiace so SARS-CoV-2.

Dospel som k záveru, že vizualizáciu genómu je možné vykonať dvoma spôsobmi:

- Prvým spôsobom je vizualizácia nespracovaných údajov, ktoré sú nukleotidovou sekvenciou pre DNA a RNA alebo aminokyselinovou sekvenciou (pomocou informácií v súboroch FASTA).
- Druhým spôsobom je vizualizácia predtým spracovaných a dobre preštu-dovaných údajov, ktoré obsahujú polohy génov, polohy ORF atď. A ktoré sú uložené v súboroch anotácií genómu (súbory GFF a GenBank).

Potom som vykonal analýzu a vizualizáciu genómu SARS-CoV-2 pomocou nasledujúcich metód.

Distribúcia nukleotidov a obsah GC analýza obsahu súboru FASTA ukázala, že genóm SARS-CoV-2 je zložený z 29 903 nukleotidov (základných jednotiek genómu). Distribúcia bola nasledovná: adenín (A) sa objavil v sekvencii genómu 8954-krát, tymín (T) - 9594, cytozín (C) - 5492 a guanín (G) - 5863-krát. Poznaním týchto údajov som vypočítal vlastnosť GC obsahu genómu, ktorá bola nízka a vyzerala ako 37,97 %.

Táto vlastnosť je dôležitá, pretože ukazuje, že genóm SARS-CoV-2 má malý počet ORF (génov), pretože zvyčajne začínajú v genómových oblastiach bohatých na GC.

Gatesova metóda vizualizácie bola vybraná na vykonanie, pretože som ju náhodou znova objavil ked' premýšľal, ako je možné vizualizovať genóm. Spracováva tiež súbor FASTA. Po vykonaní tejto vizualizácie som si všimol, že podivná sekvencia 33 adenínových (A) nukleotidov končí genómom SARS-CoV-2 (obrázok 2.4). Aby som sa ubezpečil, že počas vizualizácie nedošlo k žiadnym chybám, skontroloval som samotnú sekvenciu FASTA a dospel som ku záveru, že vizualizácia bola vykonaná správnym spôsobom.

Napriek degenerácii, ktorá je hlavnou nevýhodou metódy, dokáže vykresliť vzory DNA bez strojového vyšetrenia, čo som preukázal vizualizáciou prvého chromozómu najmenšieho známeho eukaryotického genómu, ktorý patrí do *Encephalitozoon Intestinalis* (obrázok 2.6)..

Aby som pochopil, či je táto metóda vhodná na vizuálne zloženie genómov príbuzných druhov, vizualizoval som vedľa seba genóm SARS-CoV-2 a genóm jeho najbližšieho relatívneho vírusu SARS-CoV-1. Vytvoril som hypotézu, že musia byť veľmi podobné, pretože sa navzájom podobajú z hľadiska dosiahnutej vizualizácie (obrázok 2.7), a aby som to dokázal, spustil som algoritmus pairwise2,

ktorý potvrdil, že ich percento podobnosti sa rovná 83,34%. Vybral som si tento algoritmus, pretože patrí medzi najlepšie algoritmy zarovnávania.

2D Matrix metóda bola vybraná vizualizácia, pretože tandemové opakovania v sekvencii genómu môžu byť vizuálne rozlíšiteľné bez strojového vyšetrenia. Po samotnej vizualizácii sa mi však nepodarilo nájsť ani jedné z nich (obrázok 2.9). Táto metóda tiež vyžaduje súbor FASTA so sekvenciou.

Hlavnou nevýhodou tejto metódy je, že bodové mutácie v genóme, alebo dokonca významné, nie sú takmer postrehnutelné bez komplexnej analýzy získaného obrazu.

Vylepšenie metódy 2D Matrix bolo zavedene mnou, aby sa vyrovnal s nevýhodou predchádzajúcej metódy pomocou hash funkcie.

Dokázal som to porovnaním pôvodného genómu SARS-CoV-2 a rovnakého genómu s bodovou mutáciou na druhom nukleotide (T bol substituovaný G) pomocou mojej metódy (obrázok 2.10).

Získanie aminokyseliny bolo vykonané za účelom získania proteínov SARS-CoV-2 potrebných pre ďalšiu metódu.

Po dosiahnutí proteínov zo sekvencie DNA uloženej v súbore FASTA som pomocou BLAST vyhľadávania porovnal výsledky s existujúcimi proteínmi SARS-CoV-2. Súcit ukázal, že sa neurobili takmer žiadne chyby, pretože pomer podobnosti bol veľmi vysoký. Kvôli môjmu záujmu som ich tiež porovnal s proteínmi relatívnych druhov, ktoré tiež vykazovali vysoký pomer podobnosti (tabuľka 2.2).

Identifikácia a vizualizácia ORF sa uskutočnila pomocou anotačného súboru GenBank, aby sa vizualizovali tie časti genómu SARS-CoV-2, ktoré sú pravdepodobne gény.

Po vyhľadaní umiestnení ORF v súbore GBK a porovnaní ich indexov s predtým získanými proteínmi, po odstránení tých krátkych, som dosiahol 6 kódujúcich sekvencií genómu (génov). Vizualizácia dosiahnutých génov medzi ostatnými ORF je viditeľná na obrázku 2.11.

Na overenie výsledkov bolo vykonané BLAST vyhľadávanie (tabuľka 2.3). Ukázalo sa, že som tieto gény správne zviditeľnil, pretože sú veľmi podobné tým existujúcim.

Ďalším a posledným krokom bolo spojiť všetok použitý kód do jednoduchej samostatnej konzolovej aplikácie, ktorá je schopná vizualizovať genóm SARS-CoV-2 pomocou predtým opísaných postupov. Podrobnejšia architektúra softvéru je popísaná v príslušnej časti.

Z hlavných nevýhod vyvinutého programu môžem pripustiť, že je pomerne jednoduchý, nemá grafické rozhranie a momentálne nie všetky metódy podpo-

rujú vizualizáciu ľubovoľného genómu. Preto je možné vyuvinutý nástroj pre vizualizáciu označiť ako textit druhovo špecifický.

Na záver možno všetky ciele, ktoré boli stanovené v tejto práci, považovať za splnené: analyzuje sa štruktúra genómu spolu s existujúcimi nástrojmi a genóm SARS-CoV-2 sa úspešne vizualizuje. Nie všetky však boli dokončené tak, ako som čakal: tvorba softvéru pre modernú a prispôsobiteľnú vizualizáciu genómu si vyžaduje dlhú pracovnú dobu a nemá zmysel, pretože existujú riešenia, ktoré už majú vynikajúci výkon. Softvér sa preto javil ako triviálna konzolová aplikácia, ktorú však nemožno považovať za skutočný prehliadač genómu, ale ktorá určite splňa stanovený cieľ.

4 Záver

Počas práce na tejto bakalárskej práci som sa ponoril do oblasti bioinformatiky, komplexne som analyzoval štruktúru genómu rôznych organizmov, porovnával existujúce riešenia a vyvíjal nové.

Vyvinutý program pracuje na súboroch FASTA a GenBank a je schopný vizualizovať genóm SARS-CoV-2 a (čiastočne) genómy iných organizmov pomocou rôznych 2D vizualizačných techník. Medzi nimi je aj Vylepšená 2D maticová metóda, ktorú som vyvinul ja.

Budúce práce a vylepšenia sa môžu zamerať hlavne na zavedenie nových vizualizačných techník, vylepšenie existujúcich, pridanie podpory pre ďalšie genómy, aby bol nástroj univerzálny, a na zabezpečenie programu grafickým rozhraním.

Literatúra

1. REECE, Richard J. *Analysis of Genes and Genomes*. New York: Wiley, 2004. ISBN 978-0-470-84380-2.
2. BROWN, T. A. *Introduction to Genetics: A Molecular Approach*. 1. vyd. New York : Garland Science Pub., 2011. ISBN 9780815365099.
3. YOUNGSON, Robert M. *Collins Dictionary of Human Biology*. Glasgow: HarperCollins, 2006. ISBN 978-0-00-722134-9.
4. BROWN, T. A. *Genomes* 3. New York: Garland Science Pub., 2007. ISBN 978-0-470-84380-2.
5. SANDY B. PRIMROSE, Richard Twyman. *Principles of Gene Manipulation and Genomics*. 7. vyd. New York: Wiley, 2006. ISBN 978-1-405-13544-3.
6. SIEBER, Patricia; PLATZER, Matthias; SCHUSTER, Stefan. The Definition of Open Reading Frame Revisited. *Trends in Genetics*. 2018, roč. 34, č. 3, s. 167–170. ISSN 0168-9525.
7. FICKLETT, J. W. Finding genes by computer: the state of the art. 1996, s. 12:316–320.
8. LEWIN, Benjamin. *Genes (3rd ed.)* New York: Wiley, 1987. ISBN 0-471-83278-2.
9. D, Hebenstreit. "Methods, Challenges and Potentials of Single Cell RNA-seq". *Biology*. 2012. Dostupné z DOI: doi:10.3390/biology1030658.
10. HUBBARD T. Barker D., Birney E. The UCSC genome browser database. *Nucleic Acids Research*. 2003, roč. 31, č. 1, s. 51–54.
11. KAROLCHIK D. Baertsch R., Diekhans M. The Ensembl genome database project. *Nucleic Acids Research*. 2002, roč. 30, č. 1, s. 38–41.
12. WOLFSBERG, T. G. Using the NCBI Map Viewer to browse genomic sequence data. *Current Protocols in Human Genetics*. 2011, roč. 18, č. 18, s. 15.

13. DONLIN, Maureen J. Using the Generic Genome Browser (GBrowse). *Current Protocols in Bioinformatics*. 2009, roč. 28, č. 1, s. 9.9.1–9.9.25. Dostupné z DOI: [10.1002/0471250953.bi0909s28](https://doi.org/10.1002/0471250953.bi0909s28).
14. WANG J, Kong L, Zhao S. Rice-Map: a new-generation rice genome browser. *BMC Genomics*. 2011, roč. 12, č. 1, s. 165.
15. MAYYA SEDOVA, Lukasz Jaroszewski. Coronavirus3D: 3D structural visualization of COVID-19 genomic divergence. *Bioinformatics Oxford Academic*. 2020, roč. 36, s. 4360–4362. Dostupné z DOI: [10.1093/bioinformatics/btaa550](https://doi.org/10.1093/bioinformatics/btaa550).
16. PEARSON WR, Lipman DJ. "Improved tools for biological sequence comparison". *Proceedings of the National Academy of Sciences of the United States of America*. 1988, roč. 85, s. 2444–8. Dostupné z DOI: [doi:10.1073/pnas.85.8.2444](https://doi.org/10.1073/pnas.85.8.2444).
17. C. A. STEWARD A. P. Parker, B. A. Minassian. Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med.* 2017, roč. 9, s. 49. Dostupné z DOI: [10.1186/s13073-017-0441-1](https://doi.org/10.1186/s13073-017-0441-1).
18. G. PETREA, M. Pertea. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research*. 2020, roč. 9, č. 304. Dostupné z DOI: [10.12688/f1000research.23297.2](https://doi.org/10.12688/f1000research.23297.2).
19. KAMRAN KARIMI, Daniel M Wuitchik. "Improved tools for biological sequence comparison". *Evol Bioinform Online*. 2018, roč. 14, s. 43. Dostupné z DOI: [10.1177/1176934318788866](https://doi.org/10.1177/1176934318788866).
20. WUITSCHICK, JEFFREY D.; KARRER, KATHLEEN M. Analysis of Genomic G + C Content, Codon Usage, Initiator Codon Context and Translation Termination Sites In Tetrahymena Thermophila. *Journal of Eukaryotic Microbiology*. 1999, roč. 46, č. 3, s. 239–247. Dostupné z DOI: [10.1111/j.1550-7408.1999.tb05120.x](https://doi.org/10.1111/j.1550-7408.1999.tb05120.x).
21. LI Y, Yang X. GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. *Mol Genet Genomics*. 2020, roč. 295, s. 1537–1546. Dostupné z DOI: [10.1007/s00438-020-01719-0](https://doi.org/10.1007/s00438-020-01719-0).
22. DOREMALEN Neeltje, et al. van. Aerosol and Surface Stability of SARS-CoV-2 as Compared With SARS-CoV-1. *The New England Journal of Medicine*. 2020, roč. 382, č. 16, s. 1564–1567. Dostupné z DOI: [10.1056/NEJMc2004973](https://doi.org/10.1056/NEJMc2004973).

23. LEE, Benjamin D. Squiggle: a user-friendly two-dimensional DNA sequence visualization tool. *Bioinformatics*. 2018. Dostupné z DOI: [10.1093/bioinformatics/bty807](https://doi.org/10.1093/bioinformatics/bty807).
24. BO LIAO, Xuyu Xiang; ZHU, Wen. Coronavirus phylogeny based on 2D graphical representation of DNA sequence. *Journal of computational chemistry*. 2006, roč. 27, s. 1196–1202. Dostupné z DOI: [doi:10.1002/jcc.20439](https://doi.org/10.1002/jcc.20439).
25. BIRNEY E. T. Julie D. G., Toby J. PairWise and SearchWise: Finding the Optimal Alignment in a Simultaneous Comparison of a Protein Profile against All DNA Translation Frames. *Nucleic Acids Research*. 1996, roč. 24, č. 14, s. 2730–2739. ISSN 0305-1048. Dostupné z DOI: [10.1093/nar/24.14.2730](https://doi.org/10.1093/nar/24.14.2730).
26. SEAMAN, Josiah; BUGGS, Richard J. A. FluentDNA: Nucleotide Visualization of Whole Genomes, Annotations, and Alignments. *Frontiers in Genetics*. 2020, roč. 11, s. 292. ISSN 1664-8021. Dostupné z DOI: [10.3389/fgene.2020.00292](https://doi.org/10.3389/fgene.2020.00292).
27. WOUTER PENARD, Tim van Werkhoven. On the Secure Hash Algorithm family. 2016, roč. 3, s. 41.
28. CHRISTOPH DOBRAUNIG, Maria Eichlseder; MENDE, Florian. Analysis of SHA-512/224 and SHA-512/256. 2016, roč. 2, s. 37.
29. PÁNEK T. Žihala D., Sokol M. Nuclear genetic codes with a different meaning of the UAG and the UAA codon. *BMC Biology*. 2016, roč. 15, s. 8. Dostupné z DOI: [10.1186/s12915-017-0353-y](https://doi.org/10.1186/s12915-017-0353-y).
30. SHAPIRO E Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews genetics*. 2013, roč. 14, s. 201–213. Dostupné z DOI: [doi:10.1038/nrg3542](https://doi.org/10.1038/nrg3542).
31. SHI, Z.; WANG, L.-F. 26 - Evolution of SARS Coronavirus and the Relevance of Modern Molecular Epidemiology. In: TIBAYRENC, Michel (ed.). *Genetics and Evolution of Infectious Diseases (Second Edition)*. Second Edition. London: Elsevier, 2017, s. 601–619. ISBN 978-0-12-799942-5. Dostupné z DOI: [10.1016/B978-0-12-799942-5.00026-3](https://doi.org/10.1016/B978-0-12-799942-5.00026-3).
32. BAYAT, Ardeshir. Bioinformatics. *BMJ*. 2002, roč. 324, č. 7344, s. 1018–1022. ISSN 0959-8138. Dostupné z DOI: [10.1136/bmj.324.7344.1018](https://doi.org/10.1136/bmj.324.7344.1018).
33. KHAILANY, Rozhgar A.; SAFDAR, Muhamad; OZASLAN, Mehmet. Genomic characterization of a novel SARS-CoV-2. *Gene Reports*. 2020, roč. 19, s. 100682. ISSN 2452-0144. Dostupné z DOI: [10.1016/j.genrep.2020.100682](https://doi.org/10.1016/j.genrep.2020.100682).