

**Technická univerzita v Košiciach**  
**Fakulta elektrotechniky a informatiky**

# **Vizualizácia štruktúry genómumu**

**Bakalárska práca**

**2021**

**Bc. Oleksandr Korotetskyi**

**Technická univerzita v Košiciach  
Fakulta elektrotechniky a informatiky**

## **Vizualizácia štruktúry genómu**

**Bakalárská 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**

Bakalárska práca analyzuje všeobecnú štruktúru genómu rôznych organizmov (eukaryoty, prokaryoty a vírusy) s cieľom porozumieť rozdielom rôznych genómov a navrhnuť možné riešenia ich vizualizácie. Opisuje a porovnáva niektoré populárne existujúce programy, ktoré sú určené na vizualizáciu vlastnosti genómu. Zaobera sa 2D vizualizáciou a analýzou genómu SARS-CoV-2 nielen pomocou niektorých existujúcich, ale aj v práci vyvinutých techník. Výsledky dosiahnuté počas vizualizácie sa overuju a vyvája sa aplikácia na vizualizáciu štruktúry genómu.

## **Klúčové slová v SJ**

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

## **Abstrakt v AJ**

This bachelor thesis analyzes the general genome structure of different organisms (eukaryotes, prokaryotes and viruses) in order to understand the differences of various genomes and to come up with possible solutions for their visualization. Describes and compares some of popular existing programs that are designed for visualization of genome properties. As the next step, the 2D visualization and analysis of SARS-CoV-2 genome are being performed using some existing techniques and one delivered by this thesis. The results obtained during the visualization are being verified and an application to visualize the structure of the genome is being developed.

## **Klúčové slová v AJ**

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

## **Bibliografická citácia**

KOROTETSKYI, Oleksandr. *Vizualizácia štruktúry genómu*. 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. Navrhnuť 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. Libérios Vokorokos, PhD.  
dekan fakulty

## **Čestné vyhlásenie**

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

Košice, 28.5.2021

.....

*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

---

<b>Úvod</b>	<b>1</b>
<b>1 Analýza štruktúry genómu a nástrojov na jej reprezentáciu</b>	<b>3</b>
1.1 Všeobecná štruktúra genómu . . . . .	3
1.1.1 Nukleotidy: základná podjednotka genómu . . . . .	3
1.1.2 Priestorová štruktúra nukleovej kyseliny . . . . .	6
1.1.3 Organizácia eukaryotického genómu . . . . .	7
1.1.4 Organizácia prokaryotického genómu . . . . .	10
1.1.5 Organizácia vírusových genómov . . . . .	11
1.1.6 Gény: umiestnenie a všeobecná štruktúra . . . . .	12
1.2 Existujúce riešenia pre reprezentáciu údajov o genóme . . . . .	16
1.2.1 Webové prehliadače genómu . . . . .	17
1.2.2 Funkcie a vlastnosti . . . . .	17
<b>2 Vizualizácia a analýza genómu SARS-CoV-2</b>	<b>23</b>
2.1 FASTA, GFF a GBK formaty . . . . .	23
2.2 Analýza a vizualizácia genómu SARS-CoV-2 . . . . .	25
2.2.1 Distribúcia nukleotidov a GC-content . . . . .	26
2.2.2 Gatesova metóda . . . . .	27
2.2.3 Metóda 2D Matrix . . . . .	30
2.2.4 Vylepšenie metódy 2D Matrix . . . . .	32
2.2.5 Získanie aminokyselín . . . . .	34
2.2.6 Identifikácia a vizualizácia ORF . . . . .	36
2.3 Navrh a implementácia softvéru . . . . .	41
<b>3 Vyhodnotenie</b>	<b>44</b>
<b>4 Záver</b>	<b>48</b>
<b>Literatúra</b>	<b>49</b>

<b>Zoznam príloh</b>	<b>52</b>
<b>Documentácia</b>	<b>54</b>
Úvod . . . . .	54
Inštalácia . . . . .	54
Popis spustenia a činnosti aplikácie . . . . .	55
Architektúra aplikácie . . . . .	61
Záver . . . . .	65

# Zoznam obrázkov

---

1.1	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 . . . . .	4
1.2	Štruktúry nukleotidov 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. . . . .	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	Obidva reťazce sa čítajú v smere 5'3'. Každý reťazec má tri čítacie rámce, v závislosti od toho, ktorý nukleotid je vybraný ako štartová pozícia. . . . .	13
1.9	Organizácia intrónov a exónov v DNA počas transkripcíí a translácií. . . . .	14
1.10	Typická štruktúra párovania intramolekulárnych báz RNA. . . . .	15
1.11	Hlavné používateľské rozhranie prehľadávača genómov UCSC, ktoré zobrazuje predvolené stopy v predvolenom poradí pre klaster ľudského génu alfa-globínu. . . . .	18
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. . . . .	19

---

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. . . . .	21
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. . . . .	21
2.1 Prvé 45 riadkov anotačného súboru genómu SARS-CoV-2. Pravá	
časť anotácie je na obrázku skrátená. . . . .	24
2.2 Prvé 45 riadkov obsahu súboru GBK zodpovedajúcich genómu SARS-	
CoV-2 . . . . .	26
2.3 Diagrama ukazujúca distribúciu SARS-CoV-2 nukleotidov. . . . .	28
2.4 Časť reprezentácie genómu SARS-CoV-2 pomocou Gatesovej me-	
tódy. Dlhá sekvencia 33 adenínových nukleotidov je označená čier-	
nym kruhom. . . . .	28
2.5 Posledné 143 nukleotidy genómu vírusu SARS-CoV-2 v súbore FASTA	
so sekvenovaným genómom. Je zvýraznená dlhá sekvencia 33 ade-	
nínových nukleotidov. . . . .	29
2.6 Časť prvej vizualizácie chromozómu Encephalitozoon Intestinalis.	
Rovnaké sekvencie sú zobrazené v strede. . . . .	29
2.7 Porovnanie genómov SARS-CoV-2 (zľava) a SARS-CoV (sprava)	
pomocou Gatesovej metódy. . . . .	30
2.8 Zakreslenie DNA do dvojrozmernej matice. . . . .	31
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.. . . . .	31
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. . . . .	34
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 . . . . .	40

# Zoznam tabuľiek

---

1.1	Dvojvláknová špirála DNA . . . . .	6
1.2	Hlavné funkcie populánych prehľadávačov genómu. BioMart a Bio-Das sú komunitné projekty, ktoré poskytujú jediný prístupový bod k distribuovaným výskumným údajom . . . . .	20
2.1	Získané proteínové sekvencie genómu SARS-CoV-2, ktoré sú zložené z viac ako 50 aminokyselín. . . . .	36
2.2	Nejaké výsledky porovnania medzi získanými proteínovými sekvenciami SARS-CoV-2 pomocou BLAST. . . . .	37
2.3	Nejaké výsledky BLAST vyhľadávania pre ORF SARS-CoV-2. . . . .	39

# Úvod

---

Sekvencia DNA a jej variácie sú samotným aspektom, ktorý určuje vývojové procesy organizmu, určuje náchylnosť na rôzne choroby a identifikuje každého tvora. Táto oblasť bola vždy v centre záujmu vedeckej komunity, od objavenia v roku 1869 biochemikom Fredrichom Mieschera. Napríklad Projekt Ľudského Genómu, ktorý sa začal 1 októbra 1990 a bol dokončený v apríli 2003, bol jedným z najväčších počinov bádania v dejinách vedy. Projekt bol zameraný na určenie všetkých sekvenčí DNA nášho druhu *Homo sapiens*.

V konečnom dôsledku, Projekt Ľudského Genómu umôžnil nám čítanie kompletného genetického plánu prírody pre stavbu človeka. Avšak čítanie genetického kódu neznamená pochopenie jeho štruktúry, a preto, napriek úspešnému dokončeniu projektu stále existuje množstvo neznámych vlastností DNA, ktoré si vyžadujú dôkladné štúdium.

Pandémia COVID-19 priniesla nové výzvy pre ľudstvo a upriamila pozornosť vedeckej komunity na bioinformatiku: po sekvenovaní by sa mal každý genóm byť správne analyzovaný, aby jeho vlastnosti sa boli pochopiteľné. Počas analýzy genómu sa často používajú rôzne vizualizačné techniky, aby sa dalo rozlísiť rôzne vzory, vykonať vizuálne porovnanie medzi rôznymi údajmi a poskytnuť informácie zrozumiteľným spôsobom. Preto si vizualizácia štruktúry a vlastností genómu SARS-CoV-2 zaslúži osobitnú pozornosť.

Prvá kapitola tejto bakalárskej práce je zameraná na analýzu všeobecnej štruktúry genómu rôznych organizmov (eukaryoty, prokaryoty a vírusy) s cieľom najst vhodne univerzálné spôsoby ich vizualizácie. Táto kapitola tiež obsahuje prehľad, porovnanie a analýzu niektorých existujúcich programov určených na reprezentáciu údajov o genóme s cieľom pochopiť moderné prístupy ku riešeniu danej problematiky.

Druhá kapitola tejto práce obsahuje analýzu formátov, ktoré sa používajú na ukladanie údajov o genomoch (na príklade SARS-CoV-2), analýzu a implementáciu rôznych techník pre 2D vizualizáciu genómu koronavírusu. Táto kapitola navyše navrhuje nový spôsob vizualizácie sekvenčí genómu, ktorý umôžní iden-

tifikovať najmenšie rozdiely v genómoch bez dôkladného skúmania. Druhá kapitola navyše opisuje štruktúru a zostavenie softvéru, určeného na vizualizáciu genómu SARS-CoV-2 pomocou použitých metód.

Tretia kapitola tejto práce obsahuje komplexný prehľad vykonanej práce vrátane vyhodnotenia výsledkov a dosiahnutia stanovených cieľov, analyzy výhod a nevýhod využívania riešenia na vizualizáciu genómu.

Táto práca je pokusom o využitie aplikovanej bioinformatiky na vizualizáciu štruktúry genómu moderného koronavírusu.

## **Formulácia úlohy**

Úlohou tejto bakalárskej práce je:

Vykonať všeobecnú analýzu štruktúry genómov rôznych druhov organizmov. To znamená, nájsť a popísť všetky vlastnosti štruktúry a vzory v genómoch rôznych organizmov, ktoré by mohli byť užitočné v kontexte vizualizácie štruktúry genómu.

Analyzovať a porovnať existujúce moderne nástroje na vizualizáciu štruktúry a reprezentáciu vlastností genomu.

Vizualizovať štruktúru genómu SARS-CoV-2 pomocou rôznych techník a porovnať získané výsledky s existujúcimi údajmi.

Navrhnuť a vytvoriť softvér na vizualizáciu štruktúry genómu pomocou použitých metód, spracovať dokumentáciu ku nemu.

# 1 Analýza štruktúry genómu a nástrojov na jej reprezentáciu

---

Táto kapitola sa venuje analýze štruktúry genómu rôznych organizmov, alebo inými slovami biologickému pozadiu, ktoré umôžní vytvorenie bioinformatického softvéru. Lebo bez pochopenia toho, ako je organizovaný genóm, nedá sa aj navrhnuť nástroj na jeho vizualizáciu.

Okrem toho, táto kapitola osobitne popisuje a porovnáva existujúce riešenia pre reprezentáciu údajov o genóme s cieľom porozumieť moderným prístupom k vizualizácii štruktúry genómu.

Každá zo spomínaných tém je popísaná v zodpovedajúcej podsekcií.

## 1.1 Všeobecná štruktúra genómu

U väčšiny 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

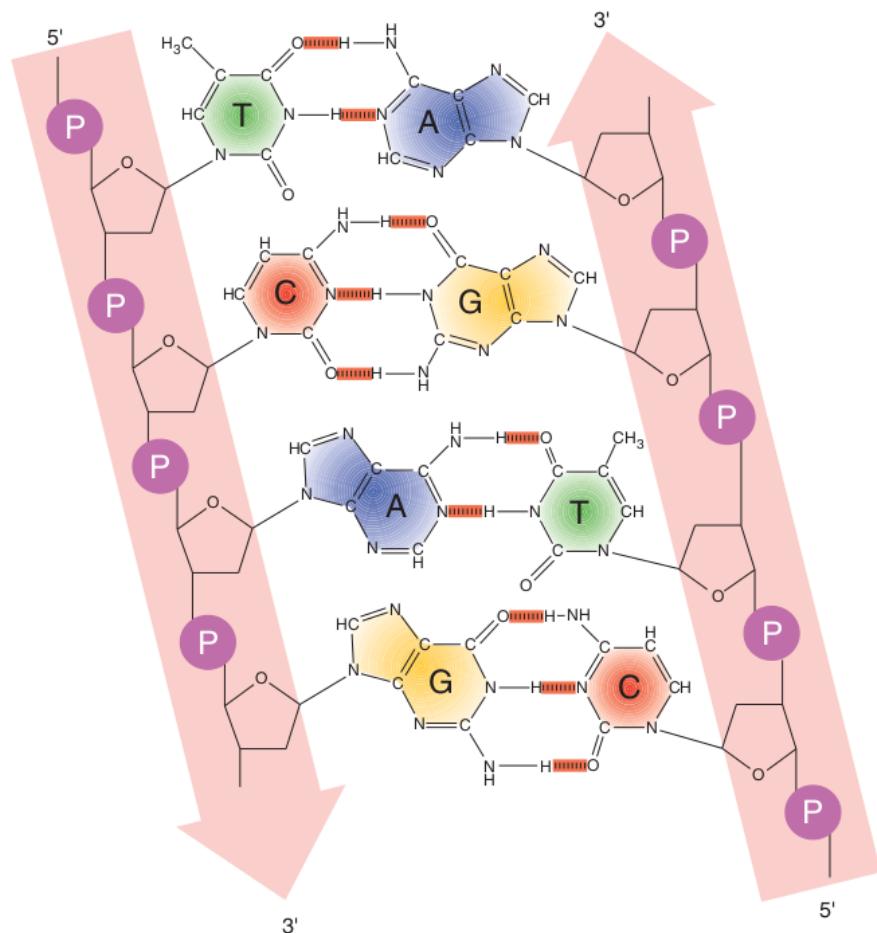
DNA aj RNA sú polymérové molekuly, ktoré sú zložené z lineárnych reťazcov rôznych kombinácií štyroch typov nukleotidov.

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].

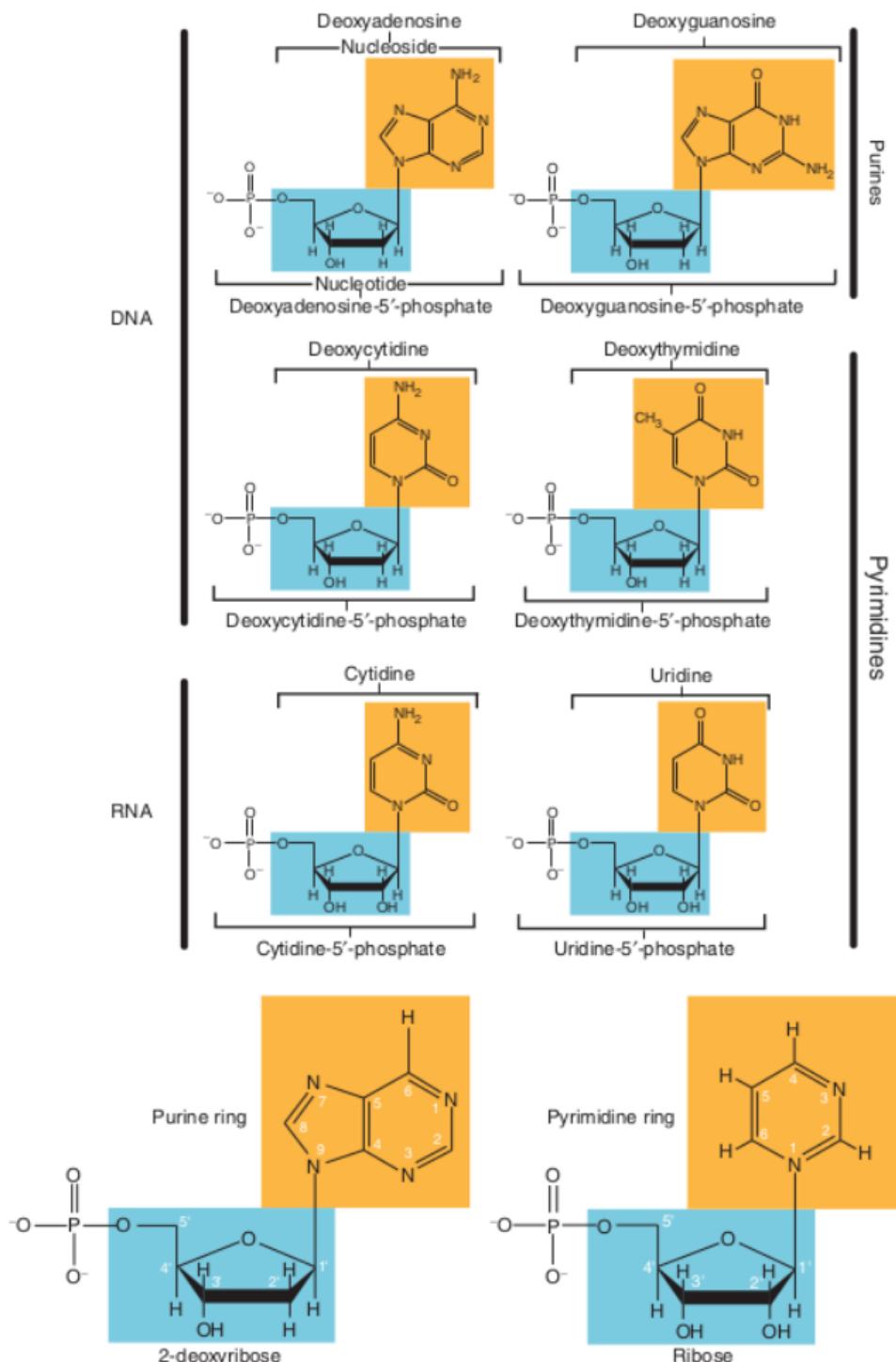
Cukor s piatimi atómami uhlíka, fosfátová skupina a dusíkatá zásada sú tri odlišné zložky, ktoré spolu tvoria celkom zložitú molekulu nukleotidov. Kombinácia cukru a dusíkatéj zásady sa nazýva nukleozid, zatiaľ čo fosfát-cukor-zásada sa nazýva nukleotid.

Nukleotidom môže byť buď purín s dvojitým kruhom, alebo pyrimidín s jedným kruhom (obr. 1.2). 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 párovaný s G (obr. 1.1). Stáva sa to kvôli geometriám nukleotidových báz a relatívnym pozíciam atómov, ktoré sa podielajú na spojení [2]. Táto vlastnosť robí dve sekvencie polynukleotidov v skrutkovnicie DNA (RNA) komplementárnymi.



Obr. 1.1: 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



Obr. 1.2: Štruktúry nukleotídov 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, troma alebo mnohým nukleotidom spojených navzájom.

Každý koniec molekuly DNA je označený číslom: jeden koniec sa označuje ako 5') a druhý koniec sa označuje ako 3'. Čísla 5' a 3' označujú počet atómov uhlíka v molekule cukru deoxyribózy, ku ktorým sa viaže fosfátová skupina.

Diskrétne 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.

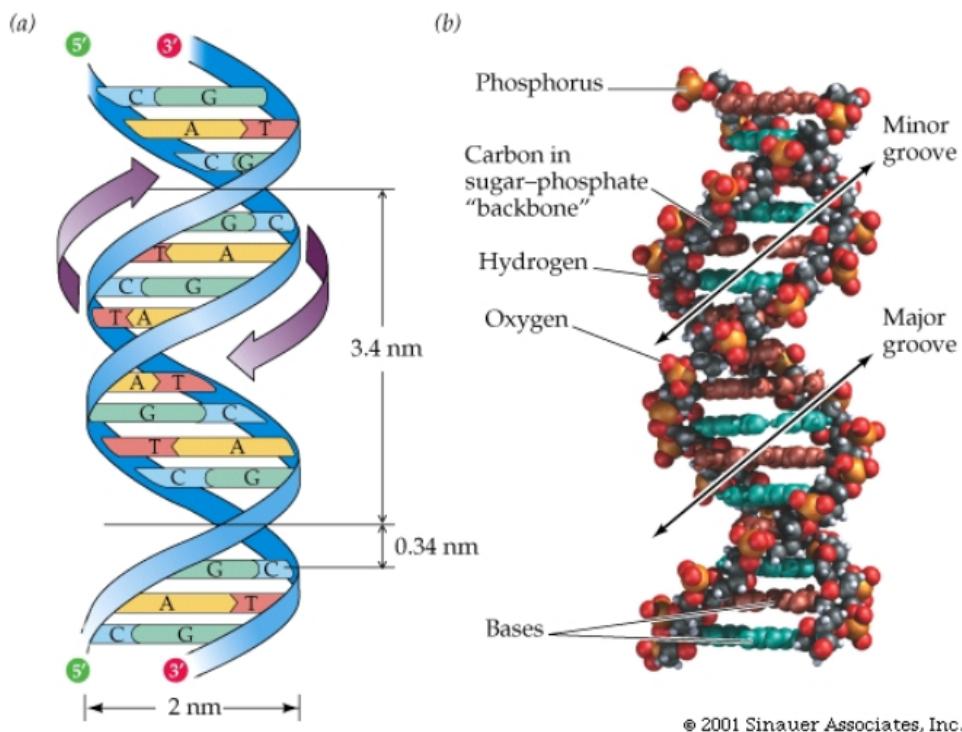
### 1.1.2 Priestorová štruktúra nukleovej kyseliny

Pretože trojrozmerná štruktúra nukleotidu nie je úplne tuhá, je možné, aby DNA mala rôzne priestorové architektúry: B-forma, A-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-forma	A-forma	Z-forma
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, plynútka	Široká plynútka	Úzka, hlboká

Aj keď sú RNA zvyčajne jednorefazcové, niektoré sekvencií sú schopne 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



Obr. 1.3: Dvojvláknová špirála DNA.

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" [2], ale ich vlastnosti nie sú opísané v priloženej tabuľke 1.1 .

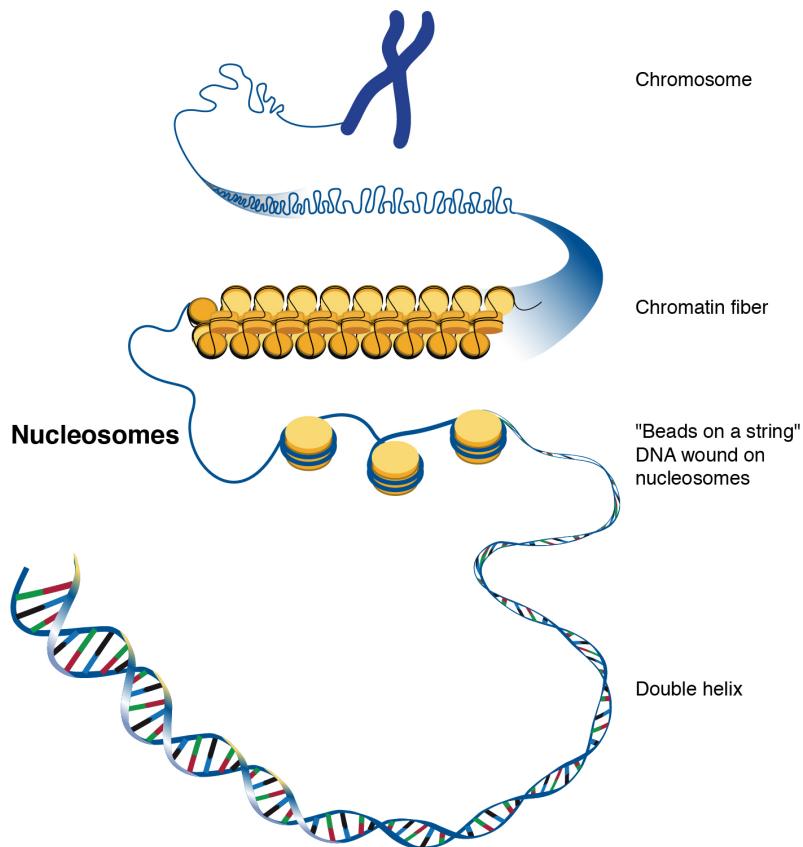
### 1.1.3 Organizácia eukaryotického genómu

Eukaryoti sú organizmy, bunky ktorých majú jadro uzavreté v jadrovom obale [3]. V eukaryotických bunkách sa nukleová kyselina nachádza v organele viazané 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 spojenej s nimi nukleovéj kyseline.

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

Konce eukaryotických chromozómov sú tiež koncami lineárnej duplexnej DNA



Obr. 1.4: Organizácia eukaryotického genómu.

a ako také si vyžadujú špeciálnu štruktúru, ktorá by zaisťovala 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.

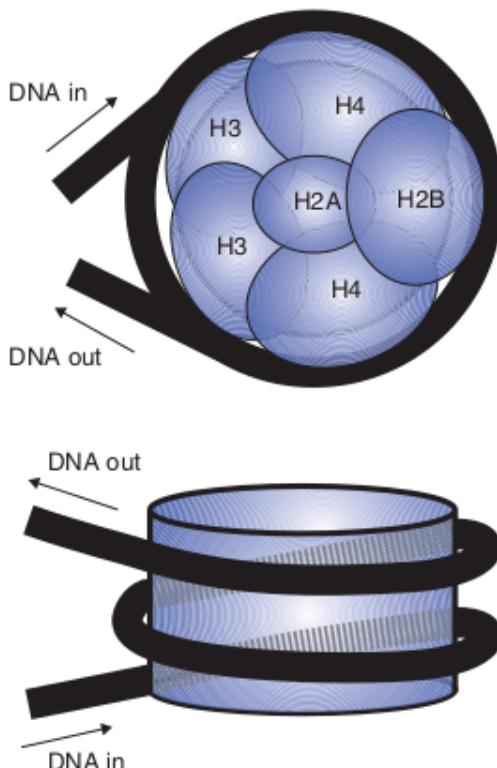
Teloméry sú špecializované sekvencie nukleových kyselín, ktorých úlohou je chrániť konce chromozómov. Vo väčšine eukaryotických buniek pozostáva telomer z krátkeho opakovania nukleotidov TTAGGG s dĺžkou stoviek jednotiek, ale opakovaný segment sa môže medzi jednotlivými druhmi lísiť. Tieto opakovania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

Napriek veľkosti jadra (5 - 10 um) je celková dĺžka DNA v ľudskej bunke približne 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 ktoré sú spojené s DNA na udržanie štruktúry. Takýto komplex bielkovín a DNA

sa nazýva chromatín.

Na najnižšej úrovni je DNA organizovaná obalením vlákien DNA okolo proteínov nazývaných históny, ktoré obsahujú veľké množstvo pozitívne nabitéch aminokyselín arginínu a lyzínu. Tieto aminokyseliny ako súčasť histónov všeobecne hrajú rozhodujúcu štrukturálnu úlohu, čo umožňuje viazať negatívne nabité fosfátové skupiny nukleotidov.



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

V priemere DNA namotaná 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, ktoré sú súčasťou chromatínu, 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 úplne otvorené (euchromatín). Euchromatínová DNA je vysoko dostupná pre molekulárne komplexy používané bunkou, a preto sa s ňou ľahšie manipuluje.

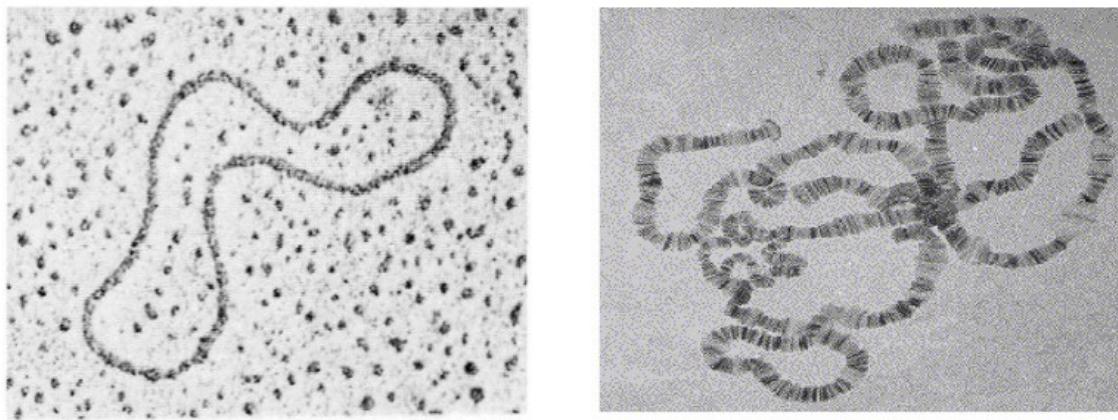
Množstvo a rozsah navinutie DNA určuje bunka, ktorá kontroluje, ktoré časti genómu je možné využívať na syntézu proteinov a ktoré nie. Toto navinutie DNA

vplyvňuje bunkovú funkciu a javí sa ako hlavná príčina diferenciácie typu buniek pri rovnakej DNA.

Okrem jadrového genómu eukaryotické bunky zvyčajne obsahujú mitochondriálny genóm, ktorý je uložený v mitochondriách a je podstatne kratší ako jadrový a nezávislý od neho. Tento genóm obsahuje informácie iba o samotných mitochondriách. Vo väčšine študovaných organizmov obsahujú mitochondrie iba kruhové molekuly DNA; v niektorých rastlinách sú súčasne prítomné kruhové aj lineárne molekuly [4].

#### 1.1.4 Organizácia prokaryotického genómu

Prokaryoty su bunkové organizmy, ktorým chýba obalené jadro. Prokaryotické genómy sa veľmi líšia od eukaryotických, najmä čo sa týka fyzickej organizácie genómu v bunke. 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).



Prokaryotic DNA

Eukaryotic DNA

Obr. 1.6: Porovnanie eukaryotickej a prokaryotickej DNA.

Väčšina toho, čo je známe o organizácii DNA v nukleoide, pochádza zo štúdií *E. coli* [4]. Prvou rozoznateľnou vlastnosťou, bolo, že kruhový genóm *E. coli* je zvinutý.

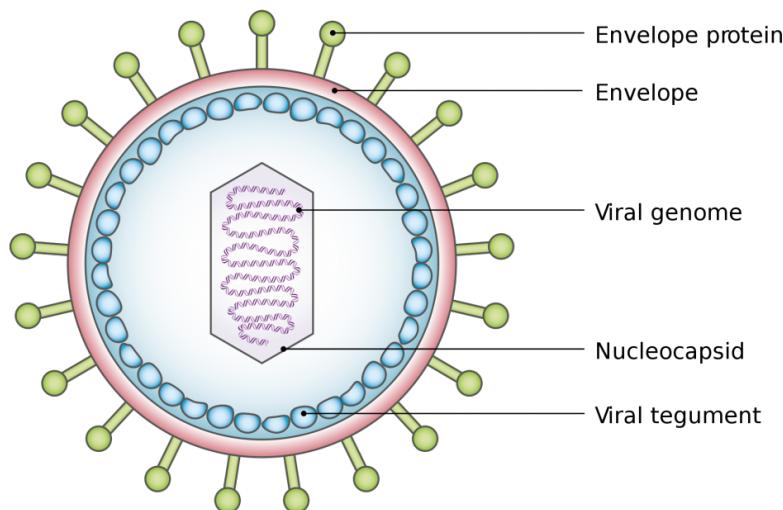
Superzávitnica nastáva, keď sa do dvojitej špirály DNA zavedú ďalšie dvojité špirály (pozitívna nadzávitnica) alebo ak sa odstránia závity (negatívna nadzávitnica). Pri lineárnej molekule sa torzné napätie vyvolané pretočením alebo odtočením okamžite uvoľní rotáciou koncov molekuly DNA, ale kruhová molekula, ktorá nemá konce, nemôže týmto spôsobom uvoľniť napätie. Namiesto toho kruhová molekula reaguje navinutím okolo seba a vytvorí kompaktejšiu štruktúru.

Nadzávitnutie je preto ideálny spôsob, ako zabaliť kruhovú molekulu do malého priestoru.

Napriek konvenčnej prokaryotickej štruktúre genómu sa nachádza čoraz viac lineárnych genómov [5]. Lineárne molekuly majú voľné konce, ktoré musia byť odlišiteľné 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 proteínmami.

### 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 inými organizmami. Vírusy môžu mať extrémne jednoduchú štruktúru, ktorá 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).

SARS-CoV-2 služí ako príklad typického +ssRNA virusu.

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 [5]. To všeobecne platí pre vírusy ssDNA aj ssRNA.

Medzi dvojvláknovými genómami môžu mať bud' „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 génnimi, 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 translácií (procesy potrebné pre syntézu proteínov).
- Spracované pseudogény sa tvoria hromadením mutácií v géne, ktorý bol

retrotransponovaný na nové miesto. Vyznačujú sa absenciou intrónov, ktoré sú prítomné v rodičovskom géne.

Gény, ktoré kódajú proteíny, zahŕňajú otvorené čítacie rámce (ORF) pozostávajúce zo série kodónov (trinukleotidov), ktoré špecifikujú aminokyselinovú sekvenčiu 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 triplétom, 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).



Obr. 1.8: Obidva reťazce sa čítajú v smere 5'3'. Každý reťazec má tri čítacie rámce, v závislosti od toho, ktorý nukleotid je vybraný ako štartová pozícia.

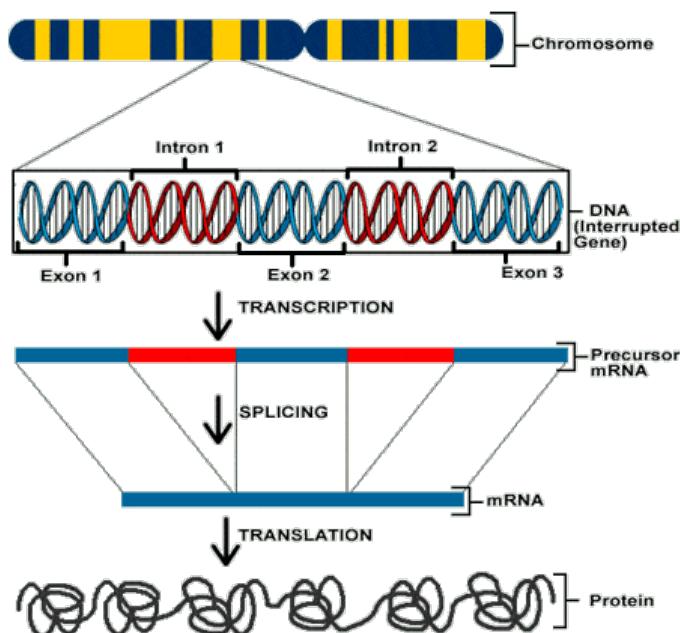
Kľúčom k úspechu skenovania ORF je frekvencia, s akou sa v sekvenции DNA objavujú terminačné kodóny. Ak má DNA náhodnú sekvenčiu a obsah GC 50%, potom sa každý z troch terminačných kodónov - TAA, TAG a TGA - objaví v priemere raz za 64 bp. Ak je obsah GC vyšší ako 50%, potom sa terminačné kodóny, ktoré 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 svojej najjednoduchšej forme preto berie údaj, napríklad 100 kodónov, ako najkratšiu dĺžku predpokladaného génu a zaznamenáva pozitívne zásahy do všetkých ORF dlhších ako je táto.

V prípade bakteriálnych genómov je jednoduché skenovanie ORF efektívny spôsobom lokalizácie väčšiny génov v sekvenции DNA. Skutočné gény v sekvenции sa nemôžu myliť, pretože sú oveľa dlhšie ako 50 kodónov. U baktérií je analýza

ďalej zjednodušená skutočnosťou, že gény sú veľmi blízko seba vzdialené, a preto je v genóme relatívne málo intergénnej DNA (iba 11 % pre *E. coli*). Väčšina bakteriálnych génov sa neprekryva.

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 génmi v eukaryotickom genóme je podstatne viac priestoru (napríklad približne 62% ľudského genómu su intergénne), 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, obr. 1.9), a tak sa v sekvencii DNA nejava ako spojité ORF. Mnoho exónov (kódujúcich oblastí génu, obr. 1.9) sú kratšie ako 100 kodónov, niektoré pozostávajú z menej ako 50 kodónov a pokračovanie čítacieho rámca do intrónu zvyčajne viedie k terminačnej sekvencii, ktorá sa zdá uzavrieť.



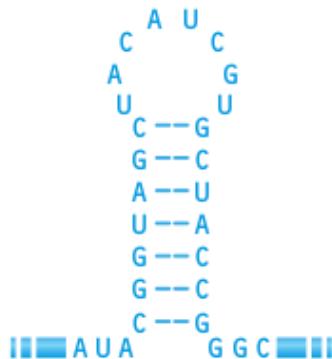
Obr. 1.9: Organizácia intrónov a exónov v DNA počas transkripcíí a translácií.

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 kódujúce proteín, ale gény pre funkčné RNA, ako sú rRNA a tRNA, neobsahujú otvorené čítacie rámce. Funkčné mole-

kuly 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 priyatý 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 sekvencie 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 ľažkostí [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.

Na zhrnutie, je môžne vyvodiť závery o tom, ako presne je možné vizualizovať štruktúru genómu, a to, že najjednoduchšie vizualizácie sa môžu týkať nasledujúcich oblastí:

- Vizualizácia nukleotidovej sekvencie za účelom vyhľadávania vzorov a vyhodnotenia ich štatistickej distribúcie v retazci DNA.

- Vizualizácia pozícií génov s cieľom analyzovať oblasti genómu, ktoré kodujú určité proteíny.
- Vizualizácia štruktúry chromozómov, pozícií intrónov a exónov ak sa ide o genómy eukaryotov.

Pretože genóm SARS-CoV-2 je vírusový, ďalšia vizualizácia by mohla pokrývať iba prvé dve možnosti.

## **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 prvé, špecializované organizácie často zhromažďujú a integrujú vysoko kvalitné anotačné údaje do webových prehľadávačov genómu a poskytujú komunité množstvo aktuálnych informácií.
- Po druhé, používateelia k nim môžu mať prístup kdekoľvek 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**

Webový softvér je väčšinou používaný na vizualizáciu rôznych vlastností genómu. Desktopové aplikácie určene na visualizáciu genomu sú menej používané, pretože táto úloha obyčajne vyžaduje veľké výpočtové schopnosti, ktoré nedokáže poskytnúť každý počítač, a preto nebudu popísané.

V súčasnosti existujú dva typy webových prehľadávačov genómu.

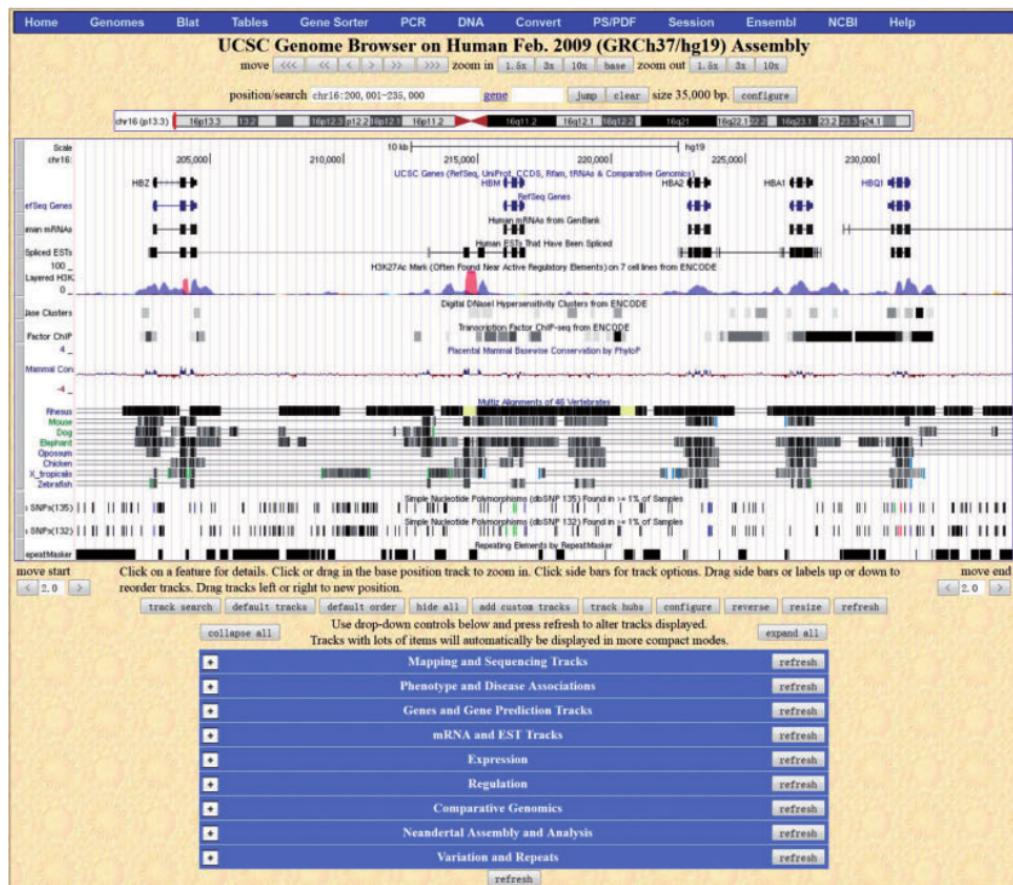
Prvým typom sú prehliadače genómu pre viac druhov, ako napríklad projekt Ensembl [10], databáza genómov UCSC [11] a webová stránka prehliadača NCBI [12]. Tieto prehľadávače genómu podporujú medzidruhovú komparatívnu analýzu. Väčšina z nich obsahuje veľké množstvo anotácií, ktoré zahŕňajú génový model, dôkazy o transkripcíí, 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. Keď 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-

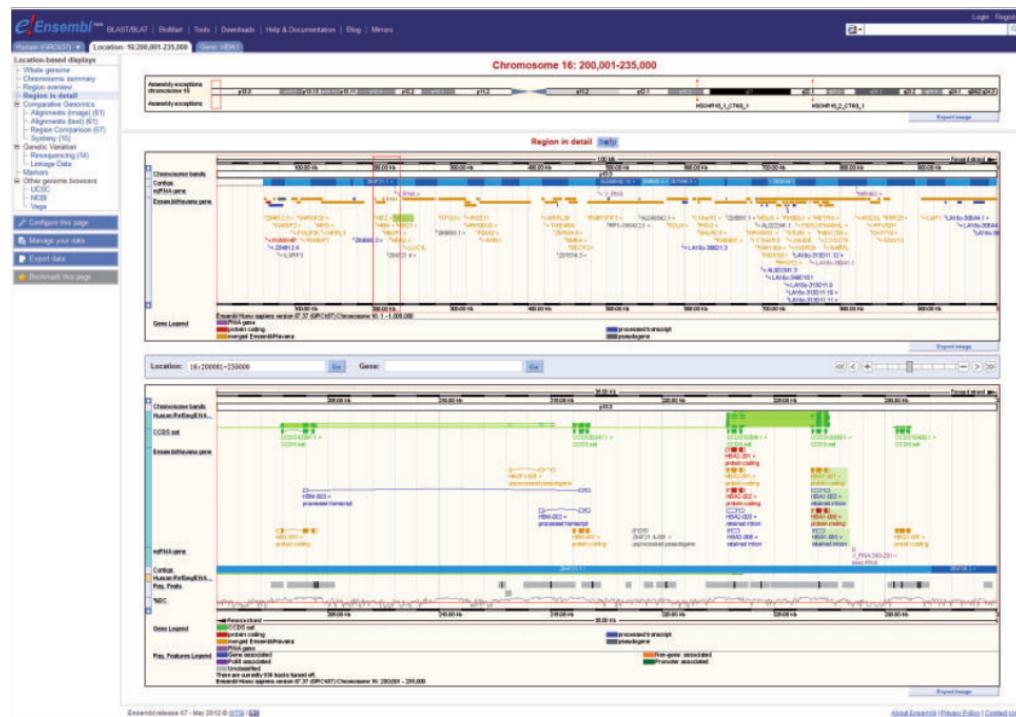


Obr. 1.11: Hlavné používateľské rozhranie prehľadávača genómov UCSC, ktoré zobrazuje predvolené stopy v predvolenom poradí pre klaster ľudského génu alfa-globínu.

brazení tracku (obr. 1.11). 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. 1.12).

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.



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.

Okrem zobrazenia umiestnenia poskytuje Ensembl samostatné stránky na zobrazenie rôznych typov informácií usporiadaných do štruktúr so záložkami. Je užitočné mať prehľad o veľkej ploche chromozómu a súčasne hľadať podrobnosť 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. Všetky populárne prehliadače genómu sú porovnané v tabuľke 1.2.

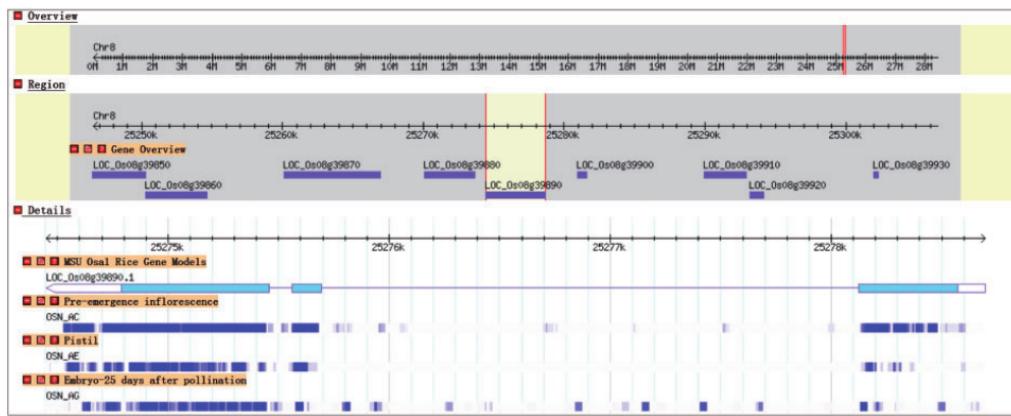
Ako už bolo spomenuté vyššie, existujú druhovo špecifické prehliadače genómu, ktoré sú k dispozícii online. Svetlým príkladom takýchto nástrojov je prehliadač genómu ryže MSU a prehliadač genómu Rice-Map [14].

V prehliadači genómu ryže MSU môžu používatelia 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

Tabuľka 1.2: Hlavné funkcie populárnych prehľadávačov genómu. BioMart a BioDas sú komunitné projekty, ktoré poskytujú jediný prístupový bod k distribuovaným výskumným údajom

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	-	Doplnkové nástroje
Strojovo orientované rozhranie	-	Prostredníctvom BioMart	Prostredníctvom BioDAS
Prispôsobenie			
Nahrajte stopy používateľov	+	+	+
Obsah prispievaný používateľmi	Obnovenie dát na základe relácie	Mechanizmus osobnej anotácie, záložky a skupiny	-



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

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

V prehľadávači genómu Rice-Map sú rôzne stopy anotácií usporiadane do vizualizačného 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 krížových 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, umiestnenie 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.

Na zhrnutie vykonanéj analýzy, je možné dospieť k záveru, že moderné a populárne nástroje na vizualizáciu genómu majú tendenciu byť zložité, založené na webe a už poskytujú používateľovi všetky potrebné informácie.

Ich zložitosť je pre človeka, ktorý nie je zvyknutý pracovať v oblasti bioinformatiky, veľkou nevýhodou. Preto, vývoj softvéru na vizualizáciu štruktúry genómu bude zameraný na poskytovanie jednoduchej a ľahko použiteľnej aplikácií, ktorú sa dá ľahko nainštalovať a spustiť bežnému používateľovi.

## 2 Vizualizácia a analýza genómu SARS-CoV-2

---

Táto kapitola je zameraná na použitie niektorých postupov vizualizácie 2D genómu na genóm SARS-CoV-2. Pretože téma je mimoriadne zložitá, sú predstavené iba niektoré z existujúcich metód. Komplexný prehľad týchto metódach, ich analýza a implementácia je obsiahnutá v zodpovedajúcich sekciách.

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

### 2.1 FASTA, GFF a GBK formaty

Najskôr, pochopenie toho, ako sa genomické údaje ukladajú, je kľúčom 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
ATTAAAGGTTATACCTTCCCAGGTAACAAACCAACCAACTTCGATCTCTTAGATCT
GTTCTCTAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCCT
```

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 začiatočným 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 (obr. 2.1). 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..186
NC_045512.2 RefSeq mature_protein_region_of_CDS 806 2719 . + . ID=id-YP_009724389.1:186
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..186
NC_045512.2 RefSeq mature_protein_region_of_CDS 806 2719 . + . ID=id-YP_009725295.1:186
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:43740578
NC_045512.2 RefSeq gene 21563 25384 . + . ID=gene-GU280_gp02;Dbxref=GeneID:43740568;t
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 vlastností
  - Opisuje, o čo ide (mRNA, doména, exón atď.)
4. Začiatok vlastnosti
5. Koniec vlastnosti
6. Skóre
  - Hodnoty podobnosti sekvencií 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 názvov súborov sú .gff, .gff2. a .gff3 [18].

Okrem anotácií GFF a súborov sekvencií FASTA sa široko používajú aj súbory GBK (formát Genbank, obr. 2.2). 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 sekvencií, proteínov atď.

## 2.2 Analýza a vizualizácia genómu SARS-CoV-2

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

```

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="MESLVPFGNEKTHVQLSLPVLOQRDVLVRGFGDSVEEVLSearq
HLKD6TCGLVEVEKGVLpqLEQPYVFIKRSDARTAPHGHVMVELVAELEG1QYGRSGE
TLGVLVPHVGIEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDpyEDFQEN
gene
CDS

```

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

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).

### 2.2.1 Distribúcia nukleotidov a 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 zahŕňajúcich 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.

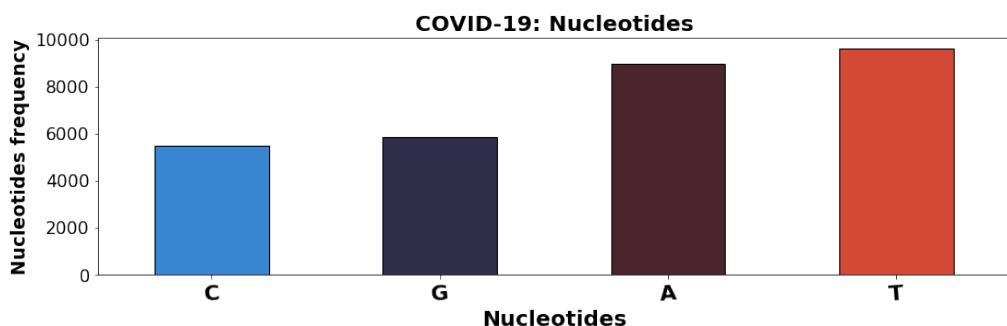
```
#Count the nucleotides frequency in the DNA
DNA = SARS_Cov_2_DNA
nucleotides = {}
for n in DNA:
    if n in nucleotides:
        nucleotides[n] += 1
    else:
        nucleotides[n] = 1

#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), čo je viditeľné na obr. 2.3 . 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



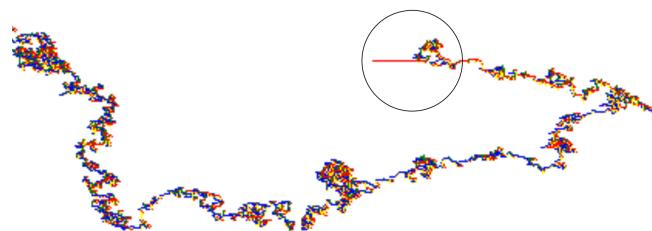
Obr. 2.3: Diagrama ukazujúca distribúciu SARS-CoV-2 nukleotidov.

vlastnostiam genómu; Gatesova metóda je typickým príkladom 2D vizualizačných techník, ktoré pracujú so surovými genomickými dátami. Preto sa počas vizualizácie spracováva súbor FASTA.

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 vynesie 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 ťažisko. Euklidovská vzdialenosť medzi začiatkom a centrom hmoty poskytuje kvantitatívny deskriptor grafu, ktorý sa nazýva polomer grafu.

Táto metóda môže byť použitá na hľadanie podobností 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 me-

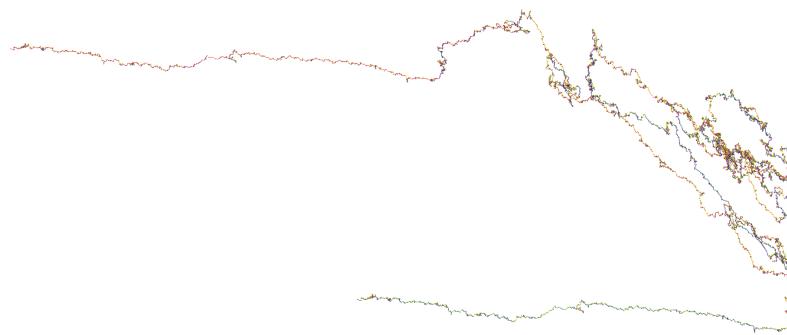
```

ACAGTGAACAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAAT
TTTAGTAGTGCTATCCCCATGTGATTAAAGCTCTTAGGAGAATGACAAAAAAAAAAAAA
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.

tôdy 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).

Po vizualizácii spomínaného chromozómu je prvým pozorovaním to, že obidva konca 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.



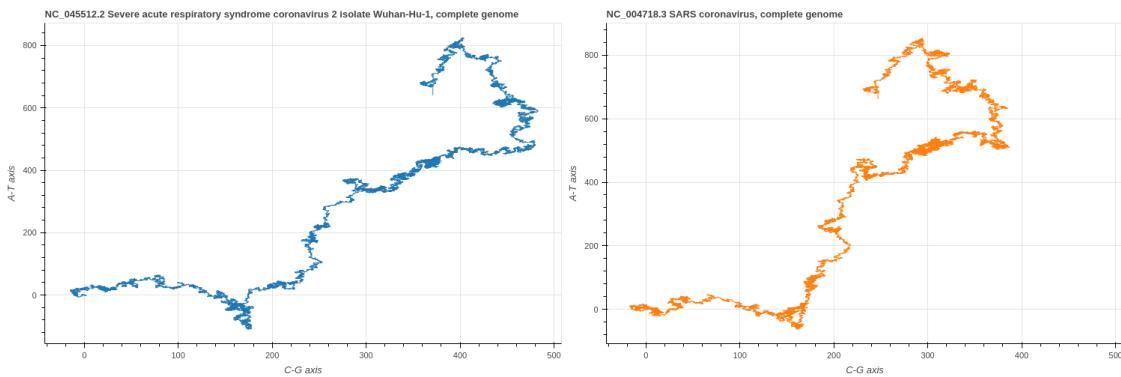
Obr. 2.6: Časť prvej vizualizácie chromozómu *Encephalitozoon Intestinalis*. Rovnaké sekvencie sú zobrazené v strede.

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.

Ako je zrejmé z poskytnutého grafu, obidva vírusy sa navzájom podobajú z hľadiska štruktúry nukleotidovej sekvencie [23].

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 [24]. 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,



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

ako sú Gap Opening Penalty (GOP), Gap Extension Penalty (GEP) a Match, Pair-Wise predstavil dva nové ďalšie parametre použité pre porovnanie [24].

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 zosúladenia.

```
from Bio import pairwise2

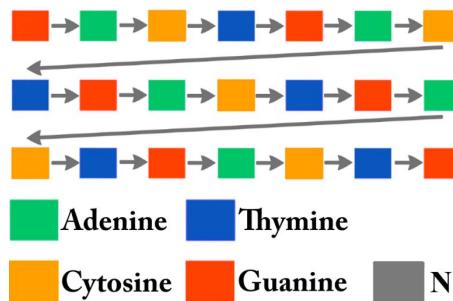
alm = pairwise2.align.globalxx(COV1.seq, COV2.seq,
                               one_alignment_only=True, score_only=True)
print('Similarity (%): ', alm / len(COV2.seq) * 100)
```

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

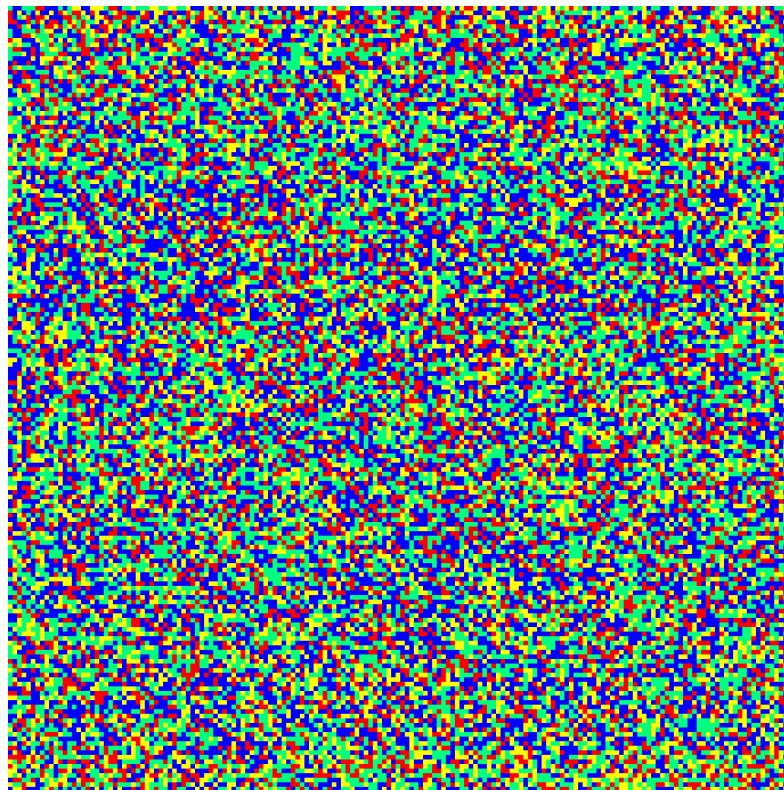
Ďalšou zaujímavou metódou na vizualizáciu DNA, ktorá pracuje na sekvenčii DNA, je metóda 2D matice. Cieľom je vykresliť celú sekvenciu genómu do štvorcového obrázka s preddefinovanou veľkosťou. Plotovanie sa vykonáva z ľavého rohu do pravého rohu až po koniec čiary a potom sa presunie na ľavú stranu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

Každý nukleotid je reprezentovaný pixelom (štvorcom) konkrétnej farby.



Obr. 2.8: Zakreslenie DNA do dvojrozmernej matice.

Táto metóda by mohla byť užitočná pri hľadaní tandemových opakovania [25] bez podrobnejšieho preskúmania strojom, pretože by mohli byť vizuálne zistiteľné. 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 pozícii (173 na každej strane), čierne štvorčeky v pravom dolnom rohu predstavujú prázdny priestor, ktorý sa nepoužil na vizualizáciu..

Okrem toho táto metóda vykresľuje každý genóm na obrázku pevnej veľkosti, ktorý by sa mohol použiť na jednoznačnú identifikáciu samotnej sekvencie. Av-

šak 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 [26], 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 štyroch hlavných funkcií, ktoré sú popísane a uvedene v pseudokode:

- Funkcia, ktorá dosahuje hash nukleotidovej sekvencie pomocou algoritmu sha256 [27].

```
DEFINE FUNCTION geHash(filename):  
    SET hash TO hashlib.sha256()  
    OPEN filename as file:  
        FOR lump IN iter(lambda: file.read(READ_MAX), b''):  
            hash.update(lump)  
    RETURN hash.hexdigest()
```

- Funkcia, ktorá rekurzívne rozdelí počiatočný prázdný obrázok na 1/8 častí. Pre každú z počiatočných oblastí sa to robí rekurzívne 8-krát.

```
DEFINE FUNCTION getBlockSize(imgSize, depth):  
    SET (w, h) TO imgSize  
    WHILE depth > 0:  
        SET h TO h / 4  
        SET w TO w / 2  
        depth -= 1  
        IF depth < 1:  
            BREAK  
        SET h TO h / 2  
        SET w TO w / 4  
        SET depth TO depth - 1
```

```
RETURN (w, h)
```

- Funkcia, ktorá vytvoí a zafarbí každú časť podľa hashu.

```
DEFINE FUNCTION drawLayer(imgSize, depth, mode):
    SET (w, h) TO getBlockSize(imgSize, depth)
    IF w < 1 OR h < 1:
        RETURN False
    SET layer TO EMPTY_LAYER
    DRAW EMPTY_LAYER
    SET x, y TO 0
    WHILE y < imgSize:
        WHILE x < imgSize:
            DRAW EMPTY_LAYER (x, y, imgH, imgW)
            COLOR LAYER
            SET x TO w + x
            SET y TO h + x
    RETURN layer
```

- Funkcia, ktorá vloží každú časť nad väčšú do jedného obrázku. Parameter nepriehľadnosti je vypočítaný pre každú veľkosť oddielu a slúži na zabránenie prekrývania medzi farbami menších oddielov a väčších (ktoré sú predtým zafarbené).

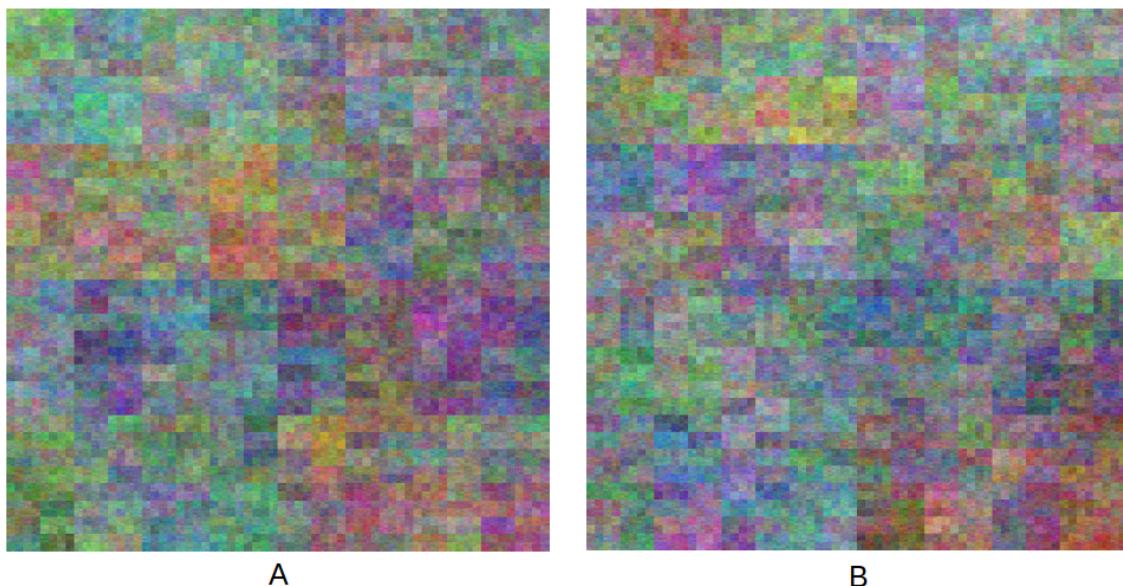
```
DEFINE FUNCTION visualize(filename, mode, ssize):
    SET depth TO 1
    SET layers TO []
    SET size TO (ssize, ssize)
    SET image TO drawLayer(size, depth, mode)
    WHILE 1:
        SET depth TO depth + 1
        APPEND NEW LAYER (depth, layers)
        IF layers[-1] EQUALS False:
            BREAK
        SET opacity TO 256 * (depth / 2) / 2 ** (depth - 1)
        MERGE LAYERS (opacity, layers)
```

Obrázky generované na ilustráciu tejto metódy používali rovnicu krytie definovanú ako:

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

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

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 kvôli začiatocným rôzmeram obrazu (512x512 px).

## 2.2.5 Získanie aminokyselín

Pre ďalšiu vizualizáciu musí byť surová sekvenčia DNA prevedená na sekvenčiu 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 sekvenčia aminokyselín predstavovaná sekvenčiou 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 [28]. 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 [29] pomocou funkcie transkripcie (). Naštastie s funkciou translate () BioPython prevádzza mRNA na aminokyselinové reťazce (zdrojový kód 2.5). Reťazce sú oddelené znakom \*, čo je stop kodón (UAA, UAG a UGA).

Zdrojový kód 2.5: 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 [30]. 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 spĺňa podmienku dĺžky (zdrojový kód 2.5) a sú uvedené v tabuľke 2.1.

Zdrojový kód 2.6: Filtrovanie aminokyselinových sekvencií a ich ukladanie do dátového rámca

```
Proteins = covid_aa.split('*')  
  
for i in Proteins[:]:  
    if len(i) < 50:  
        Proteins.remove(i)  
  
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)
```

Najjednoduchší spôsob overenia výsledkov je nájsť proteínové sekvencie, ktoré sú už k dispozícii v databázach a ktoré sú najviac podobné získaným proteíno-

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	ASAQRSQLHINELMDLFMRIFTIGTVTLKQGEIKD...	290
3	TNMKILFLALITLATCELYHYQECVRGTTVLLKEPC...	123
4	AQADEYELMYSFVSEETGTLIVNSVLLFLAFVVFLV...	83
5	QQMFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINL...	63

vým 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 sekvencí DNA a / alebo RNA. Vyhľadávanie BLAST umožňuje porovnať predmetnú proteínovú alebo nukleotidovú sekvenciu (nazývanú dopyt) s knižnicou alebo databázou sekvencí a identifikovať sekvencie knižnice, ktoré sa podobajú dopytovanej sekvencii nad určitou hranicou.

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 druhami 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 génovým začiatkom. Čím je však ORF dlhší, tým je pravdepodobnejšie, 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 [31].

Na vyhľadanie umiestnenia génov je teda potrebné porovnať údaje o ORF poskytnuté v súbore genbank s aminokyselinami získanými skôr. Preto bola na identifikáciu kódujúcich oblastí genómu nevyhnutná analýza a získanie aminokyselín (vykonané v časti 2.2.5).

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

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

Ak sa počiatočný bod ORF v nukleotidovej sekvencii zhoduje so začiatkom aminokyseliny, potom tento ORF je CDS (gén). Toto je môžne urobiť pomocou nasledujúceho kódzu:

```

for strand, nuc in [(+1, seq), (-1, seq.reverse_complement())]:
    for frame in range(3):
        trans = str(nuc[frame:]).translate(trans_table))
        trans_len = len(trans)
        aa_start = 0
        aa_end = 0
        while aa_start < trans_len:
            aa_end = trans.find("*", aa_start)
            if aa_end == -1:
                aa_end = trans_len
            if aa_end - aa_start >= min_protein_length:
                if strand == 1:
                    start = frame + aa_start * 3
                    end = min(seq_len, frame + aa_end * 3 + 3)
                else:
                    start = seq_len - frame - aa_end * 3 - 3
                    end = seq_len - frame - aa_start * 3
                answer.append((start, end, strand, trans[
                    aa_start:aa_end]))
            aa_start = aa_end + 1

```

Genóm SARS-CoV-2 kóduje až 50 neštrukturálnych, štrukturálnych a doplnkových proteínov [32].

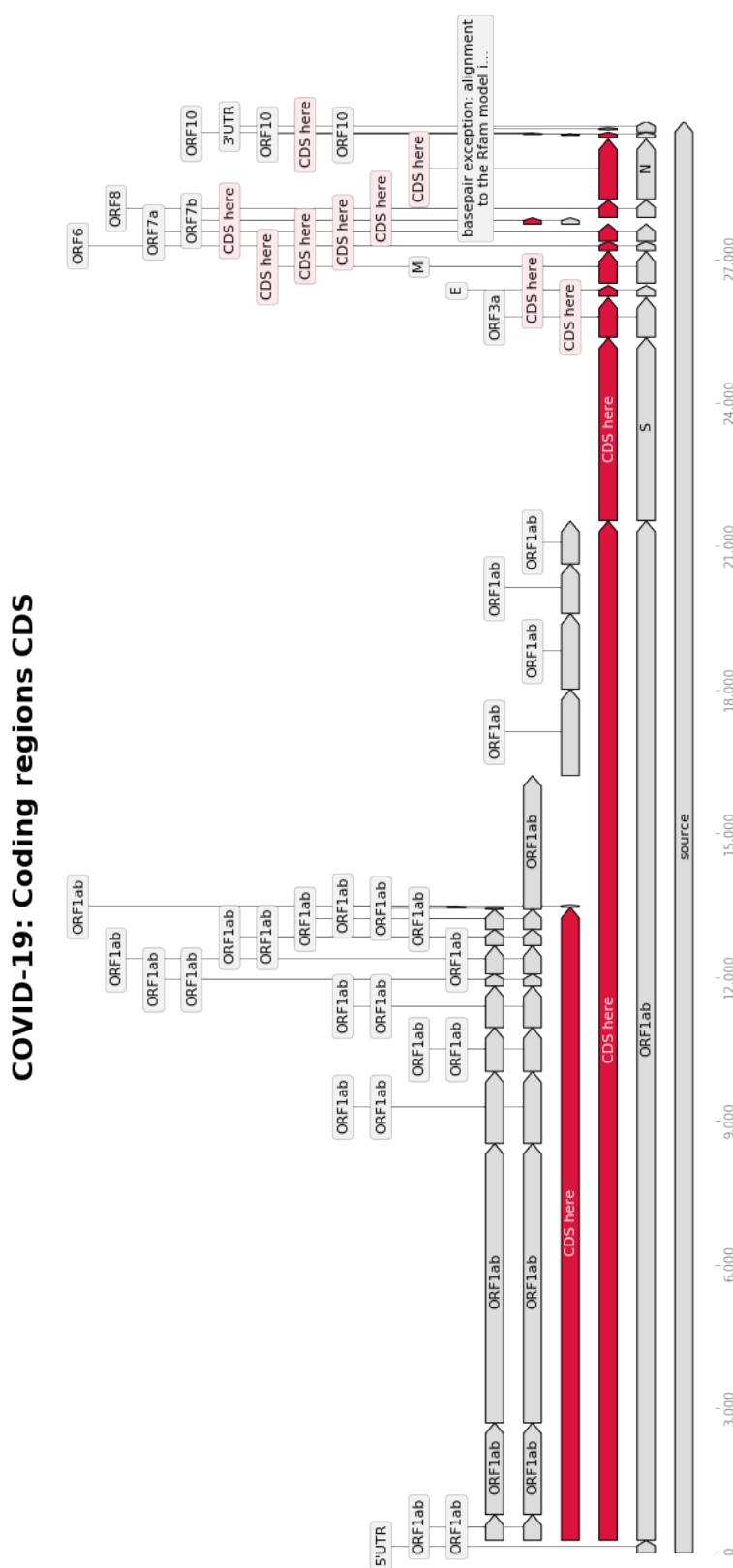
Bolo zistene, že genóm Covid-19 má 6 ORF s viac ako 200 aminokyselinami. Obrázok 2.11 ukazuje všetky ORF a CDS načítané z genómu SARS-CoV-2 pomocou BioPython. Ako bolo ukázane skôr, každa DNA ma 6 môžnych pozície pre začiatok čítania gena, čo je viditeľné na obrázku. Každá stopa reprezentuje jednu môžnu variantu načítania ORF.

Výsledky porovnania získaných sekvencií s existujúcimi sekvenciami sa uskutočňujú pomocou programu BLAST a sú uvedené v tabuľke 2.3.

Je zrejmé, že získané CDS majú vysoké percento podobnosti s druhmi príbuznými k SARS-CoV-2. Každý zo získaných CDS mal 100% zhodu s tými, ktoré existujú v genóme SARS-CoV-2 a preto, visualizáciu ORF je možné považovať za úspešnu.

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

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



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

## 2.3 Navrh a implementácia softvéru

Softvér vyvinutý v priebehu tejto práce sa javí ako zhrnutie kodu použitého na vizualizáciu genómu SARS-CoV-2 v skôr opísaných metódach do samostatnej aplikácií.

Na rozdiel od webových prehľadávačov genómu, kde sa výpočty vykonávajú na strane servera, navrhovaný softvér všeobecne predstavuje samostatnú konzolovú aplikáciu. Nevykonávajú sa nijaké významné výpočty, a preto aplikácia ne splňa náročné systémové požiadavky.

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

Vyvinutý softvér je napísaný podľa paradigmy funkčného programovania. Tento prístup k písaniu softvéru bol zabezpečený predovšetkým nízkou zložitosťou vyvinutého riešenia.

Softvér je napísaný hlavne pomocou balíka BioPython v Pythone 3.8. Všetky požadované balíčky pre správnu činnosť softvéru sú uvedené v projektovej dokumentácii.

Hlavnou myšlienkovou programu je umožniť používateľovi zvoliť si, ktoré informácie týkajúce sa genómu vírusu si chce zobraziť. Vyvinutý softvér podporuje dva režímy expluatácie: *verbose* a *quiet*.

Počas prvého režimu používateľ visualizuje genomy pomocou konzolovej navigácie programu. Tento režim poskytuje popis vykonávanych akcií a umožňuje zadávať požadované data počas behu programu.

Počas druhého režimu sa ide o využitie programu pomocou argumentov príkazového riadku, čo umožnuje ďalšiu automatizáciu a presmerovanie výstupu, čo nie je možné počas použitia webových aplikácií.

Program skladá sa z 8 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

sekvenčí a súborov anotácií z databázy NCBI pre vizualizáciu genómu SARS-CoV-2.

3. **Statistics Generator** získava štatistické údaje, ako je obsah GC a distribúcia nukleotidov / aminokyselín. Užívateľ si môže zvolať oblasť genómu, ktorá sa má štatisticky analyzovať.
4. **Gates Visualization** vykonáva vizualizáciu pomocou Gatesovej metódy do súboru -Gates.png. Užívateľ je schopný zvolať oblasť genómu ktorú chce vizualizovať.
5. **2D Matrix** modul vykreslí vybraný genóm do 2D matice do súboru .png. Veľkosť výstupného obrázka sa počíta automaticky.
6. **2D HMatrix** modul vykreslí genóm do 2D matice vybranej veľkosti pomocou algoritmu hash funkcie do súboru .png.
7. **ORF Plotter** generuje obraz distribúcie ORF a pomeru obsahu GC v genóme.
8. **Comparison** modul vykonáva porovnanie vybraných genómov. Percento podobnosti sa získa na základe algoritmu pairwise2.

Moduly nie sú schopné vzájomne interagovať, ale každý z nich je využívany s hlavného počas visualizácie.

Každý z modulov je sprevádzaný príkladmi vstupu a výstupu popísanymi v dokumentácii. Navyše, každý modul obsahuje rôzne run-time testy, ktoré zabranujú neočakávanému správaniu programu.

V súčasnosti všetky moduly podporujú spracovanie rôznych genómov, pretože výkon týchto metód nezávisí od konkrétnych vlastností genómu. Neodporúča sa však používať ORF plotter s genómami, ktoré kvôli svojej zložitosti obsahujú zložité intrón-exón štruktúry a sú väčšie ako 50000 nukleotidov kvôli extrémnej dĺžke a môžnym chýbam počas vyhľadávaní ORF. V takom prípade sa výsledky môžu výrazne lísiť od skutočných charakteristík genómu a program bude potrebovať náročné systémové požiadavky.

Softvér je určený pre platformu Linux. Program bol vyvinutý a otestovaný na notebooke ASUS X556UQ, ktorý ma nasledujúce charakteristiky:

- RAM: 8 GB
- CPU: Intel® Core™ i5-6198DU CPU @ 2.30GHz × 2

- *Grafická karta:* Nvidia Geforce 940MX
- *HDD:* 1 TB
- *OS:* Linux Mint 20 Cinnamon (kernel v. 5.4.0-66-generic)

### 3 Vyhodnotenie

---

V prvej časti tejto práce bola analyzovaná všeobecná štruktúra 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 intrónov a exónov prítomné hlavne v genómoch eukaryotov a génové clustery prítomné v genómoch prokaryotov. Gény, ich typy, pozície, funkcie a vzory, ktoré umožňujú ich nájdenie (ORF), boli analyzované a detalne 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 pre vizualizáciu dát genómu. Môžu byť klasifikované ako *webové a samostatné aplikácie*.

Okrem toho boli presne opísané dve kategórie existujúcich prehľadávačov genómu:

- **Nezávislé od druhu** riešenia, ktoré sú schopné vizualizovať akýkoľvek genóm, vrátane prehľadávača genómov Ensemble, prehľadávača UCSC a GBrowse (tabuľka 1.2).
- **Druhovo špecifické** riešenia, ktoré sú zamerané na vizualizáciu konkrétnych druhov, vrátane prehliadača genómu MSU a Rice-Map.

Potom na vizualizáciu sa 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, formáty súborov GFF a GenBank (GBK). 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 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 pozície génov, pozície 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 bola náhodou znova objavena počas premýšľania, 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). Pre ubezpečenie, že počas vizualizácie nedošlo k žiadnym chybám, samotná sekvencia FASTA bola skontrolovaná 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)..

Pre pochopenie, či je táto metóda vhodná na vizuálne zostavéniu genómov príbuzných druhov, vizualizoval som 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,

protože patrí medzi najlepšie algoritmy porovnávania.

**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.

**Metóda 2D Hashed Matrix** bola vyvinutá počas písania tejto práce, pre riešenie nevýhody predchádzajúcej metódy pomocou hash funkcie.

Jej funkčnosť bola dokázana porovnaním pôvodného genómu SARS-CoV-2 a rovnakého genómu s bodovou mutáciou na druhom nukleotide (T bol substituovaný G) pomocou tejto metódy (obrázok 2.10).

**Získanie aminokyselin** 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 umiestnených ORF v súbore GBK a porovnaní ich indexov s predtým získanými proteínmi a po odstránení 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 všetký použité kódy do 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 neexistuje možnosť automaticky stahovať genomy iných organizmov okrem SARS-CoV-2, ale všetky metódy

podporujú vizualizáciu ľubovoľného genómu. Preto je možné vyvinutý nástroj pre vizualizáciu označiť ako *druhovo nezávislý*.

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 bolo očakávané na začiatku: tvorba uplné nového softvéru pre modernú a prispôsobiteľnú vizualizáciu genómu si vyžaduje dlhú pracovnú dobu a nemá zmysel, pretože už existujú riešenia, ktoré majú vynikajúci výkon a splnia súčasné potreby. 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, analyzoval a porovnával existujúce riešenia na vizualizáciu údajov a štruktúry genomu a vyvíjal nové.

Vyvinutý program pracuje so súbormi fórmatu FASTA a GenBank a je schopný vizualizovať genóm SARS-CoV-2 a genomy iných organizmov pomocou rôznych 2D vizualizačných techník a poskytovať základe štatistické údaje o genóme. Taktiež program využivá na vizualizáciu vylepšenú 2D maticovú metódu, ktorá umožňuje identifikovať najmenšie rozdiely v genómoch a bola vyvinutá v rámci predloženej práce.

Budúce vylepšenia vyvinutéj aplikácie sa môžu zamerať hlavne na vylepšenie existujúcich vizualizačných techník, využitie ďalších, vizualizáciu ďalších vlastností genomov, prípadne pridanie podpory pre väčšie genómy, a na pridanie grafického rozhrania ku programu.

# 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.
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.

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.
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.
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.
18. G. PETREA, M. Pertea. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research*. 2020, roč. 9, č. 304.
19. KAMRAN KARIMI, Daniel M Wuitchik. "Improved tools for biological sequence comparison". *Evol Bioinform Online*. 2018, roč. 14, s. 43.
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.
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.
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.
23. 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.
24. 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.
25. 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.

26. WOUTER PENARD, Tim van Werkhoven. On the Secure Hash Algorithm family. 2016, roč. 3, s. 41.
27. CHRISTOPH DOBRAUNIG, Maria Eichlseder; MENDE, Florian. Analysis of SHA-512/224 and SHA-512/256. 2016, roč. 2, s. 37.
28. 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.
29. 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.
30. 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.
31. BAYAT, Ardeshir. Bioinformatics. *BMJ*. 2002, roč. 324, č. 7344, s. 1018–1022. ISSN 0959-8138.
32. 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.

# Zoznam príloh

---

**Príloha A** Dokumentácia ku programu na vizualizáciu štruktúry genómu

**Príloha B** Program na vizualizáciu štruktúry genómu

**Príloha C** CD médium – záverečná práca v elektronickéj podobe

**Technicka Univerzita v Košiciach**

Fakulta elektrotechniky a informatiky

**Dokumentácia k programu na vizualizáciu štruktúry  
genómu**

Príloha A k bakalárskej práci

## Úvod

Túto dokumentáciu možno považovať za používateľskú príručku k použitiu programu a ako krátku systémovú príručku, ktorá popisuje funkčnosť a architektúru programu.

Dokumentácia je rozdelená na niekoľko častí, ktoré sa venujú zodpovedajúcim tématam: inštalácia, scenáre spustenia a použitia, architektúra aplikácie.

## Inštalácia

Aplikácia bola vyvinutá pre použitie hlavne na platformách Unix / Linux, a preto môže pokus o jej inštaláciu na platformu Windows viesť k neočakávanému správaniu programu.

### Python 3.8

Aplikácia je napísaná v Pythone 3.8, a preto on je nevyhnutný pre použitie programu. V systémoch založených na Debiane je možné ho nainštalovať pomocou nasledujúcich príkazov v termináli (štandardnom príkazovom riadku):

```
$ sudo apt-get update  
$ sudo apt-get install python3.8
```

Pre systémy založené na Fedora by sa mal použiť nasledujúci príkaz:

```
$ sudo dnf install python3
```

### Pip

Ďalším krokom je inštalácia pip (správca balíkov python), ktorý sa použije na inštaláciu ďalších závislostí. Je to môžne urobiť pomocou nasledujúceho príkazu pre distribúcie založené na Debiane:

```
$ sudo apt-get install python3-pip
```

A pre systémy založené na Fedora musia byť použité:

```
$ curl "https://bootstrap.pypa.io/get-pip.py" -o "get-pip.py"  
$ python get-pip.py
```

## Knižnice

Softvér pracuje na populárnych knižniciach, ktoré poskytujú používateľovi rozšírené funkcie pre sprácovanie genomov a ine účely. Aby bolo možné aplikáciu používať, je potrebné nainštalovať nasledujúcich 12 balíkov:

- `bio==0.4.1`
- `biopython==1.78`
- `matplotlib==3.4.2`
- `numpy==1.19.5`
- `pandas==1.2.4`
- `pillow==8.2.0`
- `pyparsing==2.4.7`
- `requests==2.25.1`
- `seaborn==0.11.1`
- `urllib3==1.26.4`
- `bcbio-gff`
- `dna_features_viewer`

Úplný zoznam požadovaných balíkov sa nachádza v súbore `requirements.txt`, ktorý je umiestnený v koreňovom adresári programu.

Samotnú inštaláciu knižníc je možné vykonať v termináli v koreňovom adresári programu pomocou jedného z nasledujúcich príkazov, ktoré sa môžu lísiť v závislosti od systému:

```
$ pip install -r requirements.txt  
$ pip3 install -r requirements.txt
```

## Popis spustenia a činnosti aplikácie

Tento vizualizačný nástroj podporuje dva režimy vykonávania: *verbose* a *quiet*. Aplikácia v režime *verbose* poskytuje používateľovi komentáre a prostredie na triviálnu interakciu, zatiaľ čo režim *quit* je vhodnejší na účely rýchlejšej vizualizácie a automatizácie.

Oba režimy majú rovnaké funkcie, a preto sa líšia iba v tom, ako používateľ povie programovi, čo má robiť. V režime *verbose* používateľ komunikuje s programom prostredníctvom vstupu a výstupu konzoly, zatiaľ čo v režime *quit* používa iba argumenty príkazového riadku.

Táto časť popisuje rôzne scenáre vykonávania programu v režime *verbose* a sprevádzané príkazom na vykonanie rovnakej akcie iba pomocou argumentov príkazového riadku v režime *quiet*.

Na spustenie aplikácie v režime *verbose* by sa mal použiť jeden z nasledujúcich príkazov:

```
$ python3 Main.py  
$ python3 Main.py -m v
```

Po spustení aplikácie sa objavi hlavné menu nástroja:

```
+-----+  
| --- Welcome to the Visualizer! --- |  
+-----+  
Choose the option:  
1. Download SARS-CoV-2 genome sequence & associated files  
2. Plot sequence statistics  
3. Gates' visualization  
4. 2D Matrix visualization  
5. Improved 2D Matrix visualization  
6. Plot ORFs  
7. Compare genomes  
8. Exit  
Choice:
```

Užívateľ je schopný voliť rôzne možnosti zadáním zodpovedajúceho im čísla. Na ukončenie práci s nástrojom je potrebné stlačiť zadať „q“ pri hoci akom vstupe, na ktorý program čaká.

## Scenár 1

Po výbere prvej možnosti sa v prípade úspechu zobrazia nasledujúce správy.

```
Necessary files are being downloaded...  
Done!
```

Všetky potrebné súbory (SARS-CoV-2.fasta a SARS-CoV-2.gb) sa úspešne stiahli. Všetky ostatné súbory na vizualizáciu genómom musí používateľ pridať ručne do adresára data.

Rovnaký scenár je možné vykonať v režime *quiet* bez akýchkoľvek programovoých správ pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -d
```

## Scenár 2

Po výbere druhej možnosti program požiada používateľa, aby si vybral postupnosť z tých, ktoré sa nachádzajú v adresári data.

```
Choose the sequence to plot the statistics of:  
1. alteromonas.fasta  
2. SARS-CoV-2.fasta  
3. ebola.fasta  
Choice: 2
```

Ďalším krokom je určenie intervalu, o ktorom sa musia štatistické údaje zobrazovať:

```
Specify the interval (0 for the entire genome)  
Start: 1223  
End: 0
```

Nuly predstavujú defaultné hodnoty, zatiaľ čo program vytvára tento výstup:

```
Frequencies of nucleotides on the interval [1223;29903]:  
A: 8626  
T: 9251  
G: 5577  
C: 5226  
Total: 28680  
GC-content on interval [1223;29903]: 0.3767%
```

Zobrazuju sa základné štatistické údaje analyzovanej sekvencie, ktoré by mohli byť užitočné.

Rovnaký scenár je možné vykonať v režime *quiet* pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -s --pos 1223 0
```

## Scenár 3

Po výbere tretej možnosti program požiada používateľa, aby si vybral postupnosť z tých, ktoré sa nachádzajú v adresári data.

```
Choose the sequence to visualize using Gates' method:  
1. alteromonas.fasta  
2. SARS-CoV-2.fasta  
3. ebola.fasta  
Choice: 3
```

Ďalším krokom je určenie intervalu sekvencie genómu, ktorý sa bude vizualizovať:

```
Specify the interval (0 for the entire genome)  
Start:      0  
End:   8000  
Done!
```

Generovaný obrázok je uložený v adresári out.

Rovnaký scenár je možné vykonať v režime *quiet* pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -g ebola.fasta --pos 0 8000
```

#### Scenár 4

Po výbere štvrtej možnosti program požiada používateľa, aby si vybral postupnosť z tých, ktoré sa nachádzajú v adresári data.

```
Choose the sequence to visualize using 2D Matrix method:  
1. alteromonas.fasta  
2. SARS-CoV-2.fasta  
3. ebola.fasta  
Choice: 3  
Done
```

Táto vizualizácia nepodporuje volbu intervalu a vygenerovaný obrázok sa ukláda do adresára out.

Rovnaký scenár je možné vykonať v režime *quiet* pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -x ebola.fasta
```

#### Scenár 5

Po výbere piatej možnosti program požiada používateľa, aby si vybral postupnosť z tých, ktoré sa nachádzajú v adresári data.

```
Choose the sequence to visualize using 2D HMatrix method:
```

1. alteromonas.fasta
2. SARS-CoV-2.fasta
3. ebola.fasta

```
Choice: 2
```

Ďalším krokom je zadavanie veľkosti vytvoreného obrázka a po ňom program vygeneruje nasledujúci výstup:

```
Specify the size of image (preferably a power of 2, >= 512):
```

```
Size (px): 2048
```

```
seed: f9fa11164acc370f5c187a286c25dcffe0b93363c68ce5d658d83e
```

```
w, h: 1024.0, 512.0
```

```
w, h: 256.0, 256.0
```

```
opacity: 50%
```

```
w, h: 128.0, 64.0
```

```
opacity: 37%
```

```
w, h: 32.0, 32.0
```

```
opacity: 25%
```

```
w, h: 16.0, 8.0
```

```
opacity: 15%
```

```
w, h: 4.0, 4.0
```

```
opacity: 9%
```

```
w, h: 2.0, 1.0
```

```
opacity: 5%
```

```
Done
```

Vygenerovaný obrázok sa ukláda do adresára out.

Rovnaký scenár je možné vykonať v režime *quiet* bez výstupu konzoly pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -i ebola.fasta -S 2048
```

## Scenár 6

Po výbere šiestej možnosti program požiada používateľa, aby si vybral anotačný súbor genómu, ktorý sa má použiť, z tých, ktoré sa nachádzajú v adresári data.

```
Choose the annotation file to visualize ORFs:
```

1. ebola.gb
2. SARS-CoV-2.gbk

```
Choice: 2
```

```
Done
```

Vygenerovaný obrázok sa úspešne ulkláda do adresára out.

Rovnaký scenár je možné vykonať v režime *quiet* bez výstupu konzoly pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -o SARS-CoV-2.gbk
```

### Scenár 7

Po výbere siedmej možnosti program požiada používateľa, aby vybral sekvencie genómu na porovnanie so sekvenciami, ktoré sa nachádzajú v adresári data.

```
Choose the first sequence to compare:  
1. alteromonas.fasta  
2. SARS-CoV-2.fasta  
3. ebola.fasta  
Choice: 2  
Choose the second sequence to compare:  
1. alteromonas.fasta  
2. SARS-CoV-2.fasta  
3. ebola.fasta  
Choice: 3  
Similarity (%): 72
```

Po samotnom porovnaní program zadá percento podobnosti.

Rovnaký scenár je možné vykonať v režime *quiet* bez výstupu konzoly pomocou nasledujúceho príkazu:

```
$ python3 Main.py -m q -c SARS-CoV-2.fasta ebola.fasta
```

### Použitie režimu Quiet

Pre ziskanie dokladnej informácií o režime *quiet*, používateľ môže zadať nasledujúci príkaz:

```
$ python3 Main.py -h
```

Tento príkaz zobrazuje všetky podporované argumenty príkazového riadku:

```
usage: Main.py [-h] [-m {q,v}] [-d] [-g GATES] [-o ORF] [-s]
                [-x MATRIX] [-i HASH] [-c COMP COMP] [-S SIZE]
                [-p POS POS] [-a] [-n NAME]
```

**optional arguments:**

---

```

-h, --help                  show this help message and exit
-m {q,v}, --mode {q,v}
                           Execution mode: quiet / verbose
-d, --download              Download SARS-CoV-2 genome associated
                            files
-g GATES, --gates GATES
                           Perform Gates' visualization. Parameter
                           is an input sequence filename
-o ORF, --orf ORF          Plot ORFs of the genome. Parameter is
                           an input sequence filename
-s, --stat                  Obtain genome statistical data including
                           the distribution of nucleotides
                           and a GC-content
-x MATRIX, --matrix MATRIX
                           Plot the nucleotide sequence into a 2D
                           matrix. Parameter is the input sequence
                           filename
-i HASH, --hmatrix HASH
                           Plot the nucleotide sequense into Hashed
                           2D matrix. Parameter is input sequence
                           filename
-c COMP COMP, --compare COMP COMP
                           Compare specified genome sequences using
                           the pairwise2 algorithm
-S SIZE, --size SIZE        Size of the picture side in pixels
-p POS POS, --pos POS POS
                           Start and end positions of the nucleotide
                           sequence to perform an action (0 for default)
-a, --all                   Perform all possible actions but comparison
                           in the default mode
-n NAME, --name NAME        Input sequnce filename

```

## Architektúra aplikácie

Vyvinutý program predstavuje samostatnú konzolovú aplikáciu, ktorá sa skladá z 8 modulov umiestnených v koreňovom adresári programu. Obsah adresára je uvedený nižšie:

- **Main.py**

- `Comparison.py`
- `GatesVisualization.py`
- `MatrixVisualization.py`
- `HMatrixVisualization.py`
- `ORFPlotter.py`
- `SeqCollector.py`
- `StatGenerator.py`
- `requirements.txt`

Počas chodu programu sa vytvárajú dva ďalšie adresáre so súbormi, ak neexistujú: `data` a `out`.

Prvý adresár obsahuje stiahnuté sekvencie a program ho považuje za zdrojový adresár všetkých sekvencií genómu a súborov anotácií genómu, s ktorými program pracuje. Preto, aby bolo možné vizualizovať a pracovať s vlastnými genómami, ich súbory musia byť vložené do adresára `data`.

Druhý slúži na uloženie všetkých výstupných obrázkov formátu `.png`, ktoré program vytvorí. Preto, aby si užívateľ mohol pozrieť vykonané vizualizácie, musí ich vyhľadať v adresári `out`.

### Modules description

`Main.py` je hlavný modul programu, ktorý je zodpovedný za použitie zvyšných modulov na vykonanie zadanej úlohy. Zaoberá sa vstupom a výstupom z konzoly, navrhuje dostupné metódy vizualizácie a získava podrobnosti potrebné na ich výkon. Obsahuje nasledujúce funkcie:

- `verifyArgs()` – overuje a kontroluje argumenty príkazového riadku, ak je program spustený v režime *quiet*. Ak sa vyskytne chyba, program sa zastaví.
- `welcomeBanner()` – ak je zapnutý režim *verbose*, zobrazí üvítací banner".
- `mainMenu()` – ak je režim *verbose* zapnutý, zobrazí hlavné menu aplikácie a požiada používateľa, aby vybral možnosť pokračovania; skontroluje vstup používateľa. Vráti číslo vybratého scenára.

- **v0btainFiles(msg)** – ak je zapnutý režim *verbose*, vypíše všetky súbory vo formáte FASTA a požiada používateľa, aby si jeden vybral. Parameter `msg` predstavuje správu, ktorá sa má zobraziť. Vráti názov vybraného súboru, ak je prítomný, v opačnom prípade sa zobrazí chybové hlásenie a funkcia vráti hodnotu `None`.
- **v0btainFiles2(msg)** – ak je zapnutý režim *verbose*, vypíše všetky súbory vo formáte GenBank a požiada používateľa, aby si jeden vybral. Parameter `msg` predstavuje správu, ktorá sa má zobraziť. Vráti názov vybraného súboru, ak je prítomný, v opačnom prípade sa zobrazí chybové hlásenie a funkcia vráti hodnotu `None`.
- **v0btainInterval()** – ak je režim *verbose* zapnutý, požiada používateľa, aby určil interval sekvencie genómu, na ktorom má vykonať akciu. Vráti pozície `start` a `end` po ich overení.
- **v0btainSize()** – ak je zapnutý režim *verbose*, požiada používateľa, aby určil veľkosť pre vygenerovanie štvorcového obrázku. Vráti veľkosť strany obrázka v pixeloch.
- **main()** – vykoná hlavný cyclus programu a určuje, ktorú akciu má vykonať podľa argumentov príkazového riadku a vstupu používateľa. V režime *quiet* končí program po vykonaní akcie, zatiaľ čo v režime *verbose* znova zobrazí hlavné menu aplikácie.

`SeqCollector.py` je zodpovedný za stiahnutie všetkých požadovaných sekvencií a súborov anotácií z databázy NCBI pre vizualizáciu genómu SARS-CoV-2. V tejto chvíli nepodporuje stiahovanie súborov spojených s inými genómami. Obsahuje nasledujúce funkcie:

- **downloadFiles()** – vytvorí adresár `data`, ak neexistuje, a stiahne (vyžaduje sa internetové pripojenie) sekvenciu genómu a anotačné súbory SARS-CoV-2.

`StatGenerator.py` získava štatistické údaje, ako je obsah GC a distribúcia nukleotidov / aminokyselín. Užívateľ si môže zvoliť oblasť genómu, ktorá sa má štatisticky analyzovať. Obsahuje nasledujúce funkcie:

- **getStats(filename, mode, start, end)** – overuje typ `filename`. Overuje `start` a `end` pozície. Funkcia zastaví vykonávanie programu v chybových prípadoch a zobrazí príslušné chybové hlásenie. Vypisuje štatistické údaje

o zadanom intervale sekvencie genómu a poskytne používateľovi ďalšie komentáre v režime *verbose*.

`GatesVisualization.py` vykonáva vizualizáciu pomocou Gatesovej metódy do súboru `-Gates.png`. Užívateľ je schopný zvoliť oblasť genómu ktorú chce vizualizovať. Obsahuje nasledujúce funkcie:

- `visualize(filename, mode, start, end)` – overuje typ `filename`. Overuje `start` a `end` pozície. Funkcia zastaví vykonávanie programu v chybových prípadoch a zobrazí príslušné chybové hlásenie. Vykonáva Gatesovu vizualizáciu určeného intervalu sekvencie genómu.
- `save(outFileName, image)` – vytvorí adresár `out` ak neexistuje, a uloží vygenerovaný obrázok `outFileName` do adresára `out`.

`MatrixVisualization.py` kreslí zvolený genóm pomocou generácie 2D matice do súboru `-Matrix.png`. Veľkosť výstupného obrázka sa počíta automaticky. Obsahuje nasledujúce funkcie:

- `visualize(filename)` – overuje typ `filename`. Funkcia zastaví vykonávanie programu v chybových prípadoch a zobrazí príslušné chybové hlásenie. Vykoná vizualizáciu sekvencie špecifikovaného genómu jeho vykreslením do 2D matice.

`HMatrixVisualization.py` kreslí genóm do 2D matice vybranej veľkosti pomocou algoritmu hash funkcie do súboru `-Hmatrix.png`. Veľkosť výstupného obrázka môže byť zádana používateľom. Obsahuje nasledujúce funkcie:

- `save(outFileName, image)` – vytvorí adresár `out`, ak neexistuje, a uloží vygenerovaný obrázok `outFileName` do adresára `out`.
- `drawLayer(imgSize, depth, mode)` – kreslí farébne bloky na základe `getRandomColor()` vo veľkosti `getBlockSize(imgSize, depth)` pre jednotlivé vrstvy. Vráti obrázok aktuálnej vrstvy. Poskytuje používateľovi ďalšie komentáre v režime *verbose*.
- `getHash(filename)` – počíta hash sekvencie genómu `filename`.
- `getRandomColor()` – vráti n-ticu náhodných hodnôt farieb vo formáte RGB.
- `getBlockSize(imgSize, depth)` – počíta a vracia `width` a `height` podľa `imgSize` a `depth`. S každou iteráciou cyklu sa každá strana bloku delí na polovicu alebo na štvrtiny, v závislosti od `depth`.

- **visualize(filename, mode, ssize)** – skontroluje typ filename a zastaví vykonávanie programu v prípade chýb. Vykonáva 2D Hashed Matrix vizualizáciu určenej sekvencie genómu rekurzívnym spôsobom. Upraví seed na generovanie náhodných čísel na základe funkcie getHash(filename). Zlúči vrstvy veľkosti size vytvorené funkciou drawLayer(size, depth, mode) v závislosti od definovej opacity. Poskytuje používateľovi ďalšie komentáre v režime *verbose*.

**ORFPlotter.py** generuje obraz distribúcie ORF a pomeru obsahu GC v genóme do súboru -ORFs.png . Obsahuje nasledujúce funkcie:

- **visualize(filename)** – overuje príponu filename. Funkcia zastaví vykonávanie programu v chybových prípadoch a zobrazí príslušné chybové hlásenie. Vykoná vizualizáciu súboru s anotáciami určeného genómu zobrazením ORF.
- **save(outFileName, plt)** – vytvorí adresár out ak neexistuje, a uloží vygenerovaný plot plt do adresára out.

**Comparison.py** vykonáva porovnanie zvolených genómov. Percento podobnosti sa získa na základe algoritmu pairwise2. Obsahuje nasledujúce funkcie:

- **compare(filename1, filename2, mode)** – overuje typy súborov filename1 a filename2 a zastaví vykonávanie programu v prípade chyby a poskytne používateľovi príslušnú správu. Vykonáva porovnanie vybraných sekvencií genómu pomocou algoritmu pairwise2. Poskytuje používateľovi ďalšie komentáre v režime *verbose*.

## Záver

Táto dokumentácia predstavuje komplexný prehľad softvéru, ktorý bol vyvinutý počas bakalárskej práce „Vizualizácia štruktúry genómu“.

Aplikácia pracuje v dvoch možných režimoch a umožňuje používateľovi vizualizovať a analyzovať genómy rôznych organizmov pomocou sady vopred určených techník.

Inštalácia, vykonanie, technické aspekty a scenáre použitia boli podrobne popísané v príslušných častiach.

Na záver by sa ďalšie vylepšenia mohli zamerať na aplikovanie objektovo-orientovanej paradigmy programovania na architektúru programu a na pridanie nových funkcionálít.