Mendelova univerzita v Brně
Provozně ekonomická fakulta

Analýza a indentifikace lokálních struktur DNA

Diplomová práca

Vedúci práce:
prof. RNDr. Ing. Jiří Šťastný, CSc.

Bc. Michal Petrovič

12. mája 2024

Čestné prehlásenie

Prehlasujem, že som prácu **Analýza a indentifikace lokálních struktur DNA** vypracoval samostatne a všetky použité zdroje a informácie uvádzam v zozname použitej literatúry. Súhlasím, aby moja práca bola zverejnená v súlade s § 47b zákona č. 111/1998 Zb., o vysokých školách v znení neskorších predpisov a v súlade s platnou *Směrnicí o zveřejňování závěrečných prací*.

Som si vedomý, že sa na moju prácu vzťahuje zákon č. 121/2000 Zb., autorský zákon, a že Mendelova univerzita v Brne má právo na uzatvorenie licenčnej zmluvy a použitie tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

Ďalej sa zaväzujem, že pred spísaním licenčnej zmluvy o použití diela inou osobou (subjektom) si vyžiadam písomné stanovisko univerzity, že predmetná licenčná zmluva nie je v rozpore s oprávnenými záujmami univerzity a zaväzujem sa uhradiť prípadný príspevok na úhradu nákladov spojených so vznikom diela, a to až do ich skutočnej výšky.

Miesto a dátum .....

podpis

Poďakovanie

Chcel by som poďakovať vedúcemu mojej práce prof. RNDr. Ing. Jiřímu Šťastnému, CSc. za cenné rady a vedenie tejto práce. Ďalej by som chcel poďakovať prof. Mgr. Václavovi Brázdovi, Ph.D. a Mgr. et Mgr. Martinovi Bartasovi, Ph.D. za cenné rady, informácie a odborné testovanie aplikácie. Taktiež by som chcel poďakovať Mgr. Simone Vasekovej za rady a pomoc pri korektúre textu. Zvláštne poďakovanie patrí mojej rodine, priateľom a všetkým, ktorí ma podporovali pri písaní tejto práce.

Abstract

Analysis and Identification of Local DNA Structures

This thesis focuses on designing and implementing new analysis features into the web application DNA Analyser. The application is designed for the identification and analysis of DNA structures. The goal is to expand the platform with tools specialized for identifying Z-DNA conformations and CpX islands. The thesis also addresses restructuring, modernizing, and deploying the application on a production server using container technology.

Keywords: DNA analysis, Z-DNA, CpG islands, DNA structures, bioinformatics, dockerization, modular architecture.

Abstrakt

Analýza a identifikácia lokálnych štruktúr DNA

Diplomová práca sa zameriava na návrh a implementáciu nových analýz do webovej aplikácie DNA Analyser. Aplikácia slúži na identifikáciu a analýzu DNA štruktúr. Cieľom práce je rozšíriť platformu o nástroje špecializované na identifikáciu Z-DNA konformácií a CpX ostrovčekov. Práca sa zaoberá aj reštrukturalizáciou, modernizáciou a nasadeniu aplikácie na produkčný server pomocou technológie kontajnerov.

Kľúčové slová: DNA analýza, Z-DNA, CpG ostrovčeky, DNA štruktúry, bioinformatika, dockerizácia, modulárna architektúra.

Obsah

1 Úvod

Vývoj nástrojov zameraných na bioinformatiku, ktoré slúžia na analýzu sekvencií DNA sa stal kľúčovým pre biomedicínsky výskum. Takéto nástroje poskytujú jedinečné možnosti pre lepšie pochopenie mechanizmov medzi génmi a chorobami. Medzi dôležité aspekty patria lokálne štruktúry DNA, ako sú Z-DNA a CpG ostrovčeky. Špecifické štruktúry zohrávajú významnú úlohu v regulácii génovej expresie a súvisiacich biologických procesoch.

Cieľom diplomovej práce je rozšíriť funkcionalitu aplikácie DNA Analyser o nové analýzy na identifikáciu Z-DNA a CpG ostrovčekov. Aplikácia bola vytvorená na prácu s DNA a RNA sekvenciami, no nezahŕňa pokročilé metódy na identifikáciu spomenutých lokálnych štruktúr. Práca sa zameriava na implementáciu nových nástrojov do aplikácie, ktoré umožnia vedeckej komunite efektívne skúmať špecifické štruktúry DNA.

Práca a jej kapitoly postupne predstavia teoretický rámec DNA a jej analýzy, súčasné prístupy a nástroje, ktoré sa používajú a podrobne opíše návrh a implementáciu nových funkcií. Zároveň budú prezentované porovnania na reálnych dátach, spolu s diskusiou o význame a možnostiach ďalšieho vývoja.

2 Cieľ práce

Cieľom tejto práce je navrhnúť a implementovať nové analýzy do aplikácie DNA Analyser. Aplikácia slúži pre identifikáciu a analýzu DNA štruktúr.

Zámerom je rozšíriť existujúcu platformu o nástroje špecializované na hľadanie Z-DNA konformácií a CpX ostrovčekov. Práca sa zameriava na tieto špecifické ciele:

- **Implementácia Z-DNA Hunter:** Vyvinúť a integrovať Z-DNA Hunter pre identifikáciu Z-DNA konformácií. Tento nástroj umožní efektívne detegovať štruktúry Z-DNA, ktoré sú kľúčové pre pokročilý genetický výskum.
- **Implementácia CpX Hunter:** Implementovať CpX Hunter, ktorý poskytuje možnosť identifikácie CpX ostrovčekov. Tento modul prispieva k dôležitej analýze miest metylácie v DNA.
- **Modernizácia architektúry:** Práca zahŕňa rozdelenie monolitického repozitára do modulárnejšej architektúry a dockerizáciu služieb pre zabezpečenie lepšej škálovateľnosti.
- **Vylepšenie CI/CD:** Opraviť a optimalizovať CI/CD pipeline pre zjednodušenie vývoja a nasadzovania nových verzií. Tento proces zahŕňa implementáciu automatizovaných testov, ktoré zabezpečia kvalitu kódu a integrácie pri každej zmene.
- **Nasadenie na produkčný server:** Aplikácia sa nasadí na produkčný server, aby bola prístupná výskumníkom DNA, ktorí ju využijú pri svojom výskume.

Celkovo tak práca prináša modernizáciu a rozšírenie DNA Analyser aplikácie, čím sa zlepší jej použiteľnosť a prispôsobí sa aktuálnym potrebám genetického výskumu.

3 Nukleové kyseliny

Nukleové kyseliny sú biomakromolekulové látky zodpovedné za organizáciu a reprodukciu živej hmoty. Vo svojich makromolekulách nukleové kyseliny uchovávajú a prenášajú genetickú informáciu bunky a zabezpečujú prepis danej informácie do špecifickej štruktúry bielkovín [1].

Nukleové kyseliny sa delia na dva typy: ribonukleové kyseliny (RNA) a deoxyribonukleové kyseliny (DNA). Oba typy kyselín majú rozdielnu štruktúru a funkciu [2].

3.1 Ribonukleová kyselina (RNA)

Ribonukleová kyselina, známa skratkou RNA, je makromolekulárna zlúčenina zložená z ribonukleotidov. Každý ribonukleotid sa skladá z nukleových báz (adenín, cytozín, guanín alebo uracil), monosacharidu ribózy a jednej alebo viacerých fosfátových skupín [3]. Hrá kľúčovú úlohu v genetickom kódovaní, dekódovaní, regulácii a expresii génov [4].

Je známe množstvo variácií RNA štruktúr. Existujú tri hlavné typy, ktoré sa vyskytujú v bunkách všetkých organizmov:

1. mRNA (messenger RNA): Hrá kľúčovú úlohu v proteínovej syntéze. Slúži ako prenosový molekulárny vzor pre sekvenciu aminokyselín, ktoré sa majú zostaviť do proteínu.
2. rRNA (ribosomal RNA): Je základným stavebným prvkom ribozómov, buniek zodpovedných za syntézu proteínov.
3. tRNA (transfer RNA): Zabezpečuje prenos konkrétnych aminokyselín na ribozóm, kde sa následne pripoja k rastúcemu polypeptidovému reťazcu v procese nazývaný translácia [3, 5].

RNA vírusy sú skupinou vírusov, ktoré obsahujú RNA ako genetický materiál. Niektoré RNA vírusy, ako napríklad HIV a SARS-CoV-2, používajú takzvanú reverznú transkriptázu na replikáciu svojich RNA génov. Reverzná transkriptáza je enzým, ktorý dokáže prepísať RNA vírusu do DNA, ktorá sa začlení do genómu bunky [6, 7].

3.2 Deoxyribonukleová kyselina (DNA)

Deoxyribonukleová kyselina (DNA) je molekulárny plán života, ktorý obsahuje genetické inštrukcie potrebné pri vývoji a fungovaní všetkých známych živých organizmov a mnohých vírusov. DNA sa skladá z dvoch dlhých stočených vlákien, ktoré tvoria dvojzávitnicovú špirálu. Táto dvojzávitnica je pospájaná bázovými pármi pozostávajúcimi zo štyroch typov molekúl, známych ako nukleotidy: adenín (A), tymín (T), guanín (G) a cytozín (C). Špecifické sekvencie týchto bázových párov tvoria gény. Tie určujú štruktúru proteínov a reguláciu ich tvorby, čo vedie k rôznym funkciám a vlastnostiam v organizmoch a medzi organizmami [8].

Objav štruktúry DNA je pripisovaný Jamesovi Watsonovi a Francisovi Crickovi v roku 1953 na základe röntgenových difrakčných snímok, ktoré vyhotovila Rosalind Franklinová. Model, ktorý navrhli Watson a Crick, často označovaný ako dvojitá špirála, poskytol základ pre pochopenie spôsobu uchovávania a kopírovania genetickej informácie. Za svoju prácu získali v roku 1962 Nobelovu cenu a dodnes patrí medzi najvýznamnejšie objavy v oblasti biológie [9].

Replikácia DNA je nevyhnutný proces, ktorý prebieha vo všetkých živých organizmoch a slúži na prenos genetickej informácie z jednej generácie na druhú. Jedná sa o semikonzervatívny proces, čo znamená, že každé vlákno pôvodnej molekuly DNA slúži ako šablóna na vytvorenie komplementárneho vlákna. V tomto procese zohráva kľúčovú úlohu enzým DNA polymeráza. Jeho úloha spočíva v adícií nových nukleotidov do rastúceho vlákna DNA, čím je zabezpečená presná replikácia genetickej informácie [10].

Informácie v DNA je využívaná pri tvorbe proteínov prostredníctvom dvojstupňového procesu, pozostávajúceho z transkripcie a translácie. Pri transkripcii enzým nazývaný RNA-polymeráza prečíta sekvenciu DNA génu a syntetizuje komplementárne vlákno messenger RNA (mRNA). Táto mRNA potom slúži ako predloha pre transláciu. Kedy sa sekvencia nukleotidov premení na sekvenciu aminokyselín a vytvorí proteín. Táto ústredná dogma molekulárnej biológie, prvýkrát navrhnutá práve Francisom Crickom, znázorňuje, akým spôsobom je využívaná genetická informácia v DNA na tvorbu proteínov, ktoré regulujú biologické funkcie [11].

Sekvenovanie DNA je proces určovania presného poradia nukleotidov v molekule DNA. Zahŕňa akúkoľvek metódu alebo technológiu, ktorá sa používa na určenie poradia štyroch báz. Schopnosť sekvenovať DNA viedla k mnohým prelomovým objavom v biológii vrátane pochopenia genetických chorôb, vývoja liečby a forenzných aplikácií. V rámci projektu ľudského genómu, ktorý vznikol na základe medzinárodnej spolupráce, sa v roku 2003 podarilo úspešne sekvenovať celý ľudský genóm, čím bol získaný neoceniteľný zdroj informácií pre genetický výskum [12, 13].

Naše poznatky o DNA sa neustále rozširujú, a tým aj jej využitie. Techniky ako napríklad CRISPR-Cas9 umožňujú presnú úpravu génov, čím sa otvárajú možnosti liečby genetických porúch a mnohých ďalších aplikácií. Tieto výkonné nástroje však vyvolávajú aj etické a spoločenské otázky, ktoré treba starostlivo zvážiť, pri pokračovaní v skúmaní a manipulácii s plánom života [14].

3.3 Porovnanie RNA a DNA

RNA (ribonukleová kyselina) aj DNA (deoxyribonukleová kyselina) sú kritickými zložkami bunkových procesov a zohrávajú dôležitú úlohu pri prenose genetickej informácie. Napriek podobným úlohám vykazujú kľúčové rozdiely v štruktúre, funkcii a stabilite:

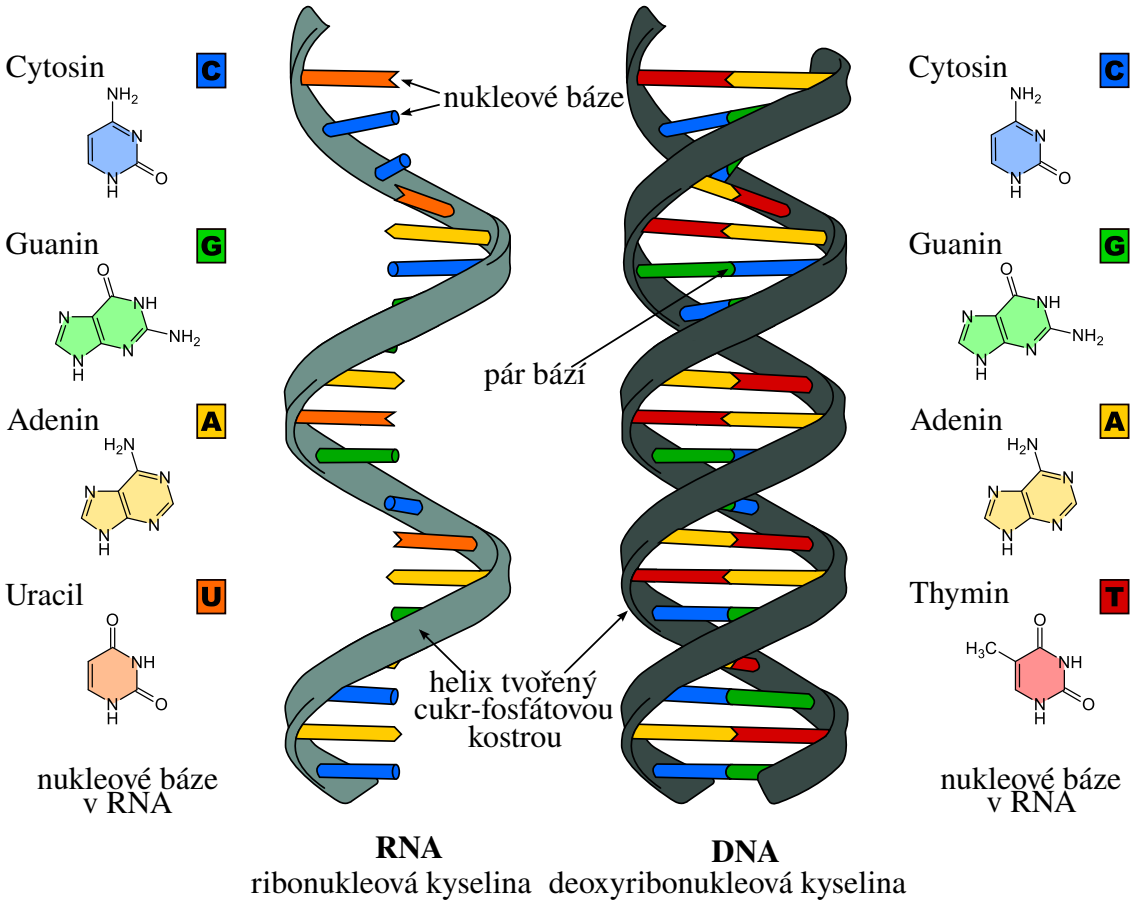
- **Chemická štruktúra:** DNA je dvojvláknová molekula, tvoriaca dvojitú špirálu, zatiaľ čo RNA je za bežných okolností jednovláknová. Daný štrukturálny rozdiel

ovplyvňuje stabilitu a funkčnosť týchto molekúl. Dvojvláknová štruktúra DNA zaisťuje stabilitu a redundanciu, čím je zabezpečený presný prenos genetickej informácie. Na druhej strane jednoreťazcová povaha RNA umožňuje jej skladanie do zložitých štruktúr, čo poskytuje možnosť vykonávať širšiu škálu funkcií [8].

- **Zložka cukru:** DNA obsahuje deoxyribózový cukor, pričom RNA obsahuje ribózový cukor. "Deoxy" v slove deoxyribóza znamená, že molekula DNA neobsahuje atóm kyslíka, ktorý je naopak prítomný v ribózovom cukre RNA molekuly. Tento rozdiel má za následok vyššiu stabilitu molekuly DNA oproti molekule RNA, a to najmä v alkalických podmienkach.
- **Dusíkaté bázy:** V DNA sú prítomné bázy adenín (A), guanín (G), cytozín (C) a tymín (T), zatiaľ čo RNA využíva A, G, C a uracil (U) namiesto tymínu. Táto variabilita v zložení báz ovplyvňuje spôsob, akým dané molekuly interagujú s ďalšími bunkovými zložkami.
- **Funkcia:** Hlavnou funkciou DNA je dlhodobé uchovávanie genetickej informácie. Funguje ako predloha pre všetky proteíny a iné molekuly, ktoré bunka produkuje. Na druhej strane funkcie RNA sú rozmanitejšie a zahŕňajú syntézu bielkovín (mRNA), katalyzáciu biochemických reakcií (ribozómy), štrukturálne úlohy v ribozóme (rRNA) a ďalšie.
- **Umiestnenie:** DNA sa nachádza predovšetkým v jadre bunky, pričom RNA je prítomné v jadre aj v cytoplazme bunky. Dané rozmiestnenie odráža ich rozdielne úlohy v bunke. DNA zostáva v chránenom prostredí, zatiaľ čo RNA sa pohybuje, aby plnila svoje funkcie [15].

Je dôležité poznamenať, že z týchto všeobecných pravidiel existujú výnimky. Napríklad niektoré vírusy využívajú ako genetický materiál RNA alebo niektoré molekuly RNA môžu mať dvojvláknové štruktúry [16].

Na obrázku číslo 1, môžeme vidieť grafické porovnanie štruktúry DNA a RNA a ich nukleových bází.



Obrázok 1: Porovnanie štruktúry RNA a DNA a ich nukleových bází [17].

3.4 Štruktúra dvojzávitnice

Dvojitá špirála nukleových kyselín je základná štruktúra, ktorú môže tvoriť deoxyribonukleová kyselina (DNA) aj ribonukleová kyselina (RNA). Jedná sa o ústredný pojem molekulárnej biológie. Ako sa už spomínalo v sekcii číslo 3.2 bola objavená Jamesom Watsonom a Francisom Crickom v roku 1953 [9].

Dvojitá špirála pozostáva z dvoch vlákien nukleových kyselín, ktoré sa vinú v opačných smeroch a vytvárajú špirálovú štruktúru. Vlákna sa skladajú z nukleotidov, ktoré sa skladajú zo sacharidu (deoxyribóza v DNA a ribóza v RNA), fosfátovej skupiny a jednej z dusíkatých báz (adenín, tymín, cytozín a guanín v DNA a adenín, uracil, cytozín a guanín v RNA) [15].

Nukleotidové bázy z oboch vlákien interagujú v strede špirály a vytvárajú vodíkové väzby, ktoré zabezpečujú stabilitu špirálovej štruktúry. Párovanie báz je špecifické: adenín (A) sa v molekule DNA páruje s tymínom (T) a v molekule RNA s uracilom (U), cytozín (C) sa páruje s guanínom (G) v molekule DNA aj molekule RNA. Dané komplementárne párovanie báz je podstatou replikácie a transkripcie, ktoré sú základom života [8].

V štruktúre dvojitej špirály je tiež prítomná hlavná a vedľajšia drážka, ktoré tvoria priestory medzi dvoma vláknami DNA. Tieto drážky zohrávajú dôležitú úlohu pri interakcii DNA s proteínmi počas procesov transkripcie a replikácie. Hlavná drážka je širšia ako vedľajšia drážka a obe predstavujú miesto pre naviazanie proteínov [18].

Každé vlákno DNA má dva konce označované ako 3' (tri-prim) a 5' (päť-prim). Orientáciu vlákna definujú práve dané konce. Sú určené tým, ako je pripojený fosfát a cukor. Na 5' konci je fosfátová skupina viazaná na piaty uhlík sacharidu. Koniec 3' viaže voľnú hydroxilovú skupinu k tretiemu uhlíku. Konce sú dôležité pri replikácii DNA, pretože DNA polymeráza dokáže syntetizovať nové vlákno len od 5' ku 3' koncu. Vedie to k tvorbe kontinuálneho predného vlákna a diskontinuálneho oneskoreného vlákna počas replikácie [18].

Objav dvojšpirálovej štruktúry DNA bol kľúčovým momentom vo vede. Otvoril cestu mnohým pokrokom v oblasti genetiky a pochopenie jej štruktúry a funkcie je nevyhnutné pre mnohé oblasti biologického výskumu. Je to dôkaz zložitosti a elegancie života na molekulárnej úrovni.

3.4.1 B-DNA

B-DNA je najbežnejšia forma DNA v živých organizmoch a väčšina ľudí sa odkazuje práve na ňu, ak hovorí o štruktúre DNA. Prvýkrát ju predstavili James Watson a Francis Crick v roku 1953, ktorí deklarovali, že DNA existuje ako pravotočivá dvojitá špirála so špecifickou schémou párovania báz: adenín (A) s tymínom (T) a guanín (G) s cytozínom (C) [9].

Štruktúra B-DNA sa vyznačuje pravotočivou dvojitou špirálou s priemerom približne 2 nanometre. Tvorí ju okolo 10,5 párov báz na jeden závit, pričom každý pár báz je od seba vzdialený zhruba 0,34 nanometra. Sacharidovo-fosfátové kostry oboch vlákien tvoria vonkajšiu stranu špirály, pričom bázy sú orientované dovnútra špirály a tvoria špecifické vodíkové väzby, ktoré zabezpečujú stabilitu spojenia vláken. [19].

Dvojzávitnica B-DNA má hlavnú a vedľajšiu drážku rôznej šírky. Tieto drážky slúžia ako väzobné miesta pre proteíny viažuce DNA, ktoré zohrávajú kľúčovú úlohu v biologických procesoch, ako je replikácia, transkripcia a oprava DNA. Širšia hlavná drážka umožňuje proteínom interpretovať sekvenciu DNA bez toho, aby došlo k oddeleniu dvoch vláken, čo poskytuje mechanizmus regulácie génov [8].

Objasnenie štruktúry B-DNA bolo zásadným momentom v dejinách vedy, ktorý otvoril oblasť molekulárnej biológie. Umožnilo nám to pochopiť, ako sa uchováva a prenáša genetická informácia, a tvorí základ techník, ktoré sú dnes zásadné pre biotechnológiu a medicínu [15].

V ukážke DNA kódu číslo 1 je zobrazený úsek ľudskej B-DNA sekvencie v textovej reprezentácii nukleotidov.

```
...CCTTAGGGCAGCCCTAGGCGCAGCGGTGCAAGGAGAGCCACATTTACCCCTGGCGCTGCACGGCCC
TGAGGCTGGGCAAGGCTGTCCACCCCGCTGTCAGAACCCCAGCAGGGAAGGTGTCCAGAAGGCAGTCCT
GGAACCCTGCACAGAGGCCCAGCGGGCACAAGGTTGGGGCAGCTCTGTTCCCAGCAGGCCGAGCCCGGG
TGGCTGGAGAGGGAGCTCTGGAAGGTCAGCCTAGGGGCCGTCGGCCCCTGCAGACCCTGTGCCCAGCCC
AGCATCCCGGGGAGCTCCCTCCACATGCTCATCTCACGAGGTTCCTGCTGCACTCAGAGTGGAGGA...
```

Kód 1: Ľudský B-DNA kód vzorku GRCh38.p14 Chr8 [20]

3.4.2 A-DNA

A-DNA je jednou z mnohých ďalších foriem, ktoré môže molekula DNA nadobudnúť. Túto konkrétnu formu DNA prvýkrát navrhla Rosalind Franklinová v 50. rokoch 20. storočia a štrukturálne sa líši od známejšej B-DNA, čo je forma, ktorú DNA zvyčajne nadobúda za fyziologických podmienok [21]. Táto forma je opísaná v sekcii číslo 3.4.1.

A-DNA je pravotočivá dvojitá špirála podobná B-DNA, ale má kratšiu a širšiu štruktúru. Špirála má 11 párov báz na otočku v porovnaní s 10 pármi báz v B-DNA a dané bázy sa nakláňaj vzhľadom na os špirály. Výsledkom týchto štrukturálnych rozdielov je hlboká a úzka hlavná drážka a plytká a široká vedľajšia drážka, na rozdiel od užšej vedľajšej drážky a širšej hlavnej drážky v štruktúre molekuly B-DNA [22].

Za normálnych fyziologických podmienok sa A-DNA sa zvyčajne nevyskytuje, ale môže sa tvoriť v dehydrovaných vzorkách DNA a RNA alebo v hybridných duplexoch DNA--RNA. A-DNA bola pozorovaná v niektorých komplexoch proteín-DNA a môže zohrávať úlohu pri rozpoznávaní DNA proteínmi. Predpokladá sa tiež, že ovplyvňuje formovanie Z-DNA, ďalšej alternatívnej štruktúry DNA, počas procesu transkripcie [23], opísanej v sekcii číslo 3.4.3.

Objav A-DNA pomohol objasniť plastickosť molekuly DNA a jej potenciálnu úlohu v biologických procesoch, ktorá je naďalej predmetom výskumu. Pochopenie rôznych foriem, ktoré môže DNA nadobúdať, je kľúčové v oblasti štrukturálnej biológie. Môže ovplyvniť spôsob interakcie DNA s proteínmi a inými molekulami, a tým zohrávať úlohu pri regulácii genetických funkcií.

3.4.3 Z-DNA

Z-DNA, ľavotočivá dvojšpirálová štruktúra DNA, je jedinečná forma, ktorá sa odchyľuje od pravotočivej štruktúry B-DNA navrhnutej Jamesom Watsonom a Francisom Crickom [24]. Jedná sa o vysoko polymorfnú biomolekulu s množstvom identifikovaných alternatívnych štruktúr, z ktorých mnohé sú spojené so špecifickými bunkovými funkciami [25]. Z-forma DNA, ktorá je pre túto prácu kľúčová, sa líši od predchádzajúcich dvoch foriem tým, že je ľavotočivá a obsahuje lokálne zákruty reťazca DNA. Najmä sa však objavuje pri výskyte špecifických sekvencií, kedy sa striedajú purínové a pyrimidínové bázy (pu-py) [26]. Zvyčajne ide o sekvencie:

$$(GC)^n \; alebo \; (GT)^n$$

Štruktúra molekuly Z-DNA je diametrálne odlišná od bežnejšej B-DNA. Z-DNA vzniká za určitých podmienok, kedy sú DNA sekvencie bohaté na cytozín a guanín záporne stočené. Táto štruktúra je stabilizovanou syn konformáciou guanínových báz. Špirála sa vinie doľava zig-zag, odtiaľ pochádza jej názov Z-DNA. Hlavná drážka je úzka a hlboká, zatiaľ čo vedľajšia drážka je široká a plytká [26].

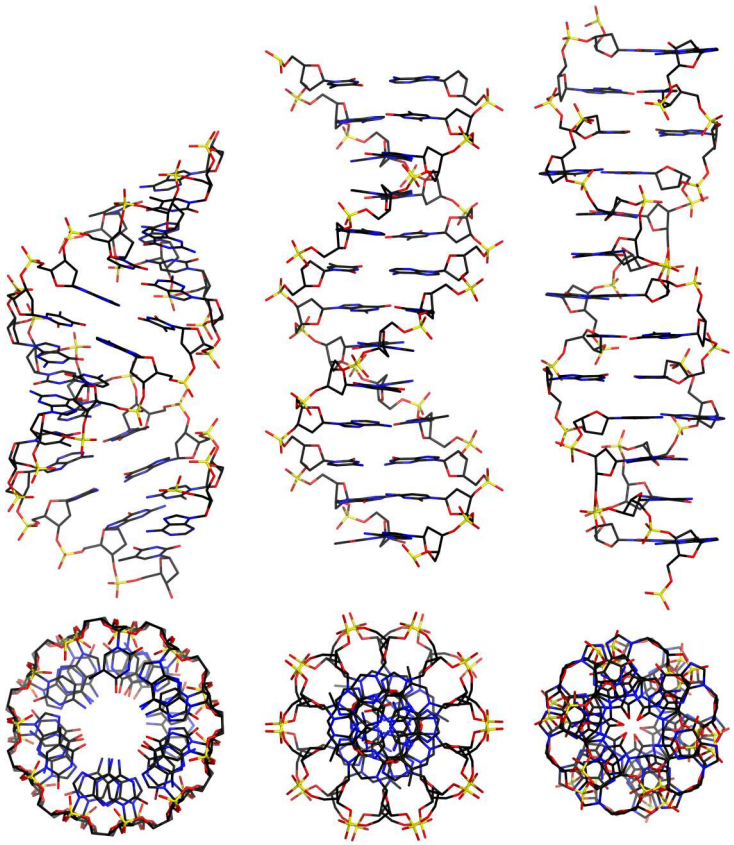
Tvorbu Z-DNA ovplyvňujú rôzne faktory vrátane sekvencie, superhelixovej hustoty a prítomnosti určitých proteínov. Na predpovedanie vzniku Z-DNA v prirodzene sa vyskytujúcich sekvenciách bol vyvinutý počítačom podporovaný termodynamický prístup. Tento prístup využíva termodynamické parametre na predpovedanie potenciálu tvorby Z-DNA, čím poskytuje cenný nástroj na pochopenie biologických úloh Z-DNA [24].

Z-DNA zohráva významnú úlohu v bunkových procesoch organizmov. Predpokladá sa, že Z-DNA sa môže podieľať na regulácii transkripcie, keďže sa často nachádza v oblastiach aktívnych génov [25]. Prechod z B-DNA na Z-DNA a späť je dynamický proces, ktorý môže byť ovplyvnený podmienkami prostredia a prítomnosťou určitých proteínov. Napríklad komplex podobný kubánu, kyselina europium-L-asparágová, môže pri fyziologickom pH rozlišovať medzi štruktúrami DNA a selektívne stabilizovať DNA, ktorá nie je B-formy, ale destabilizovať ostatné. Tento komplex dokáže za podmienok s nízkym obsahom solí pri fyziologickej teplote konvertovať B-formu DNA na Z-formu a tento prechod je reverzibilný, podobne ako pri RNA polymeráze, ktorá mení rozvitú DNA na Z-DNA a po transkripcii ju konvertuje späť na B-DNA [27, 28, 29].

Analýza štruktúr Z-DNA v genomických sekvenciách je dôležitou súčasťou pochopenia ich biologických funkcií. Termogenomické analýzy boli použité na identifikáciu sekvencií s vysokou náchylnosťou na prijatie ľavotočivej Z-DNA. Je dôležité zdôrazniť, že tieto analýzy identifikujú sekvencie s vysokým potenciálom výskytu Z-DNA, nie so stopercentnou určitosťou výskytu danej konformácie v skúmanom géne [25].

Záverom možno povedať, že Z-DNA je fascinujúca a jedinečná forma DNA, ktorá zohráva významnú úlohu pri regulácii génovej expresie. Štúdium Z-DNA naďalej poskytuje cenné poznatky o komplexnom svete štruktúry a funkcie DNA [26].

Obrázok číslo 2 zobrazuje porovnanie štruktúr A-formy, B-formy a Z-formy DNA.

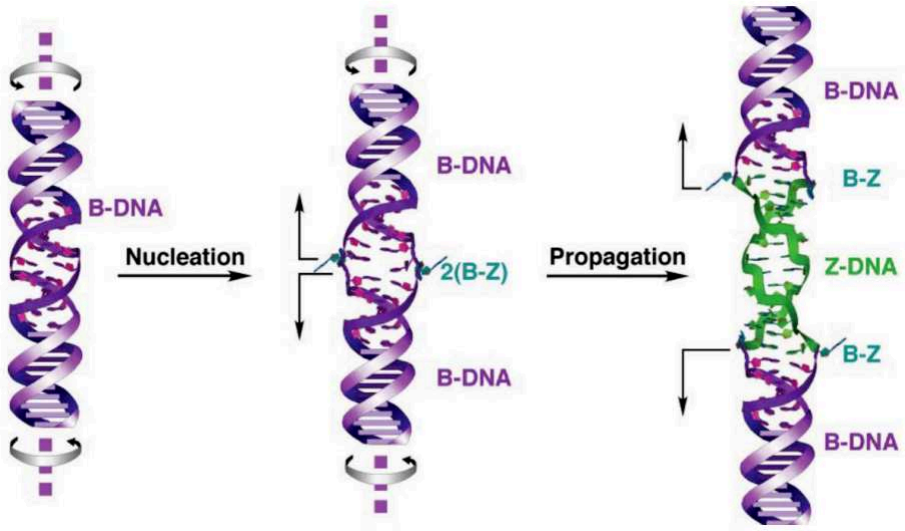


Obrázok 2: A-DNA (vľavo), B-DNA (v strede), a Z-DNA (vpravo) konformácie DNA [30].

V roku 2010 bol prostredníctvom počítačovej simulácie objasnený presný atomárny mechanizmus vzniku Z-DNA [31]. Tieto zistenia boli neskôr experimentálne potvrdené pomocou testov smFRET. Jedná sa o techniku využívajúcu fluorescenciu na meranie vzdialeností na molekulárnej úrovni [32].

Transformácia DNA z pôvodnej B-formy na Z-DNA je spôsobená vznikom dvoch B-Z spojov, medzi ktorými sa vytvára úsek Z-DNA. Tento transformovaný úsek sa často označuje ako zip [31]. Celý proces sa delí do dvoch fáz, ako je zobrazené na obrázku číslo 3:

1. Nukleácia - uvoľnenie vodíkových väzieb
2. Propagácia - posun B-Z spojov smerom k opačným koncom



Obrázok 3: Vznik Z-DNA z pôvodnej B-DNA pomocou B-Z spojov [10].

Predpokladá sa, že prítomnosť Z-DNA v ľudskom genóme je spojená s výskytom mnohých ochorení. Medzi tieto ochorenia sa radí napríklad Alzheimerova choroba, lupus, Mendelské choroby a pravdepodobne aj niektoré formy onkologických ochorení [33, 34].

V ukáže DNA kódu číslo 2 je zobrazený ľudský genóm s viacerými štruktúrami Z-DNA konformácií.

```
CCTTAGGGACACACACACACACACGCGCGCGCGCGCGCGCGCACACACATACACCACACACACAGCCCTTAG
GGCAGCCCTGGGTGTGCACGCGCGCGCGCGCGCGTGTGTGTGTGCGTGTTACAAACACCTTAGGGCAGCCCTCC
TTAGGGCAGCCCTGCGCGCACGCGCGCCCACGCACGCACGCGCACGCCGCACGGGCCTGCCTTAGGGCAGCCCT
CCGCGCGCGCGCGCGCCATTGTGTGGCTGGACTCGGCCGCCCCTGTGGCCTTAGGGCAGCCCTCCCCTTAGGGC
AGCCCTCCGTGGGTGCGTGCGTGTTCCTCCTTAGGGCAGCCCTCCTCAGAACCCCAGCAGGGAAGGTGTCCAGC
```

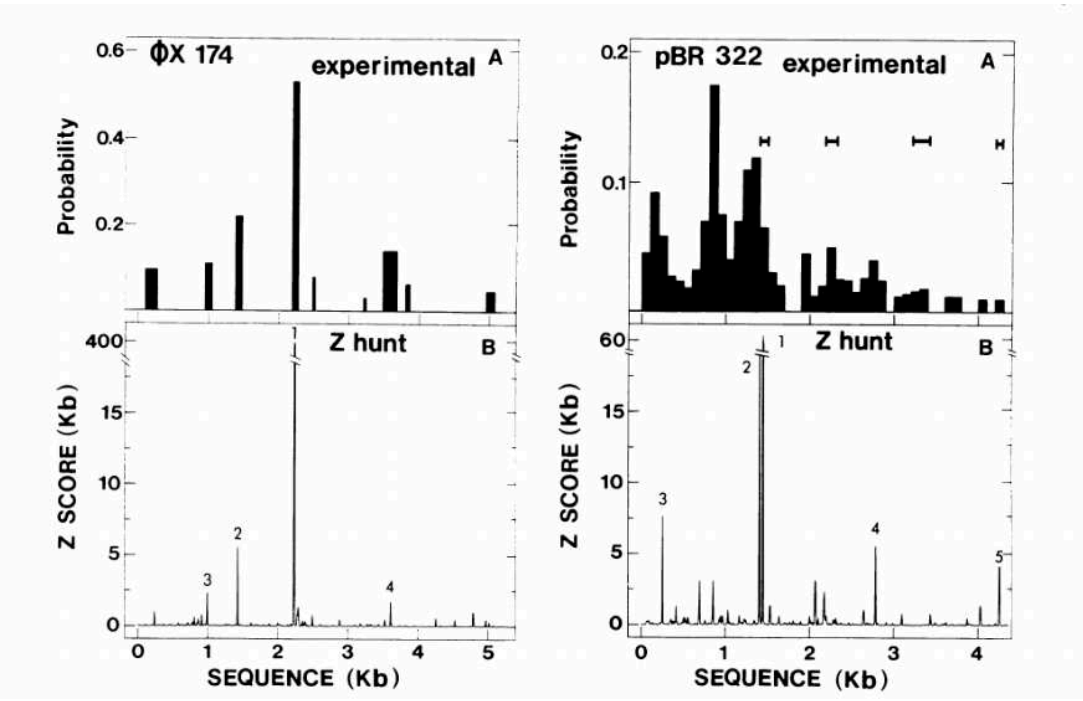
Kód 2: Ukážka Z-DNA štruktúr v ľudskom genóme, ktoré boli predikované a experimentálne overené [35]

3.5 Analýza DNA

Pokrok v analýze DNA štruktúry, konkrétne v štúdii Z-DNA, je výrazne podporovaný využitím počítačov. Z-DNA, ktorá je ľavotočivou formou DNA, je ovplyvnená sekvenciou a mierou negatívneho supercoilingu, teda rozťahovania molekuly DNA, ktorým prechádza. Na skúmanie dlhých sekvencií prirodzene sa vyskytujúcej DNA a identifikáciu kombinácií nukleotidov, ktoré majú vysokú predispozíciu k tvorbe Z-DNA, bol vytvorený počítačový program nazvaný Z-hunt. Tento program využíva štatisticko-mechanický model, ktorý je založený na empiricky stanovených energetických parametroch prechodu z B-DNA do Z-DNA [24]. Ukážku výsledkov programu Z-hunt zobrazuje obrázok číslo 4

V rámci rozsiahlej štúdie boli identifikované a zmapované možné sekvencie Z-DNA vo viac ako milión bázových párov ľudskej DNA, zahrňujúc 137 kompletných génov. K tomu bol využitý počítačový program Z-Hunt-II, ktorý aplikuje striktnú termodynamickú vyhľadávaciu stratégiu na mapovanie výskytu ľavotočivej Z-DNA v genomických sekvenciách. Analýza odhalila, že potenciálne sekvencie tvoriace Z-DNA sa v ľudských génoch vyskytujú pomerne náhodne, pričom sú väčšinou lokalizované na 5' koncoch génov [28].

Pokročilé biomedicínske výpočtové centrum (ABCC) systematicky identifikovalo genómové oblasti, o ktorých sa predpokladá, že tvoria každú kandidátsku štruktúru non-B DNA, vrátane štruktúr Z-DNA, a umiestnilo ich do non-B databázy. Táto databáza poskytuje komplexný zoznam predpovedí alternatívnych štruktúr DNA a umožňuje výskumníkom prelínať informácie o non-B DNA s pozíciami známych zmien v genóme [36].



Obrázok 4: Graf znázorňujúci potenciálne miesta Z-formy DNA oproti experimentálnemu určeniu. Graf A znázorňuje experimentálny odhad Z-DNA a graf B zobrazuje výsledok pravdepodobnosti za použitia programu Z-hunt [24].

Využívaný nástroj n*on-B DNA Motif Search Tool* zahŕňa širokú škálu funkcií na predpovedanie štruktúr, ktoré nie sú súčasťou bežnej štruktúry B-DNA. Medzi tieto štruktúry patrí napríklad ľavotočivá Z-DNA. Algoritmus je navrhnutý tak, aby identifikoval striedavé purín-pyrimidínové sekvencie, ktoré pravdepodobne prechádzajú z pravotočivej B-DNA do ľavotočivej Z-DNA konformácie [36].

Aktualizovaný algoritmus Z-DNA v *non-B DB v2.0* vyhľadáva sekvencie s dĺžkou najmenej 10 nukleotidov, pričom osobitne vylučuje AT/TA dinukleotidov, pre ich nestabilitu

pri tvorbe Z-DNA. Implementácia algoritmu prešla z jazyka Perl do jazyka C aby dosiahla vyššiu rýchlosť [36].

Tento nástroj, hoci je jeden z najpoužívanejších, má veľkú nevýhodu v tom, že nie je prispôsobiteľný, a ako už bolo spomenuté vylučuje kombinácie AT/TA bázových párov. Daná skutočnosť predstavuje problém, vzhľadom na to, že Z-DNA sa môže formovať aj s bázovými pármi AT/TA, a teda existuje možnosť, že dôjde k vynechaniu významných výsledkov [37]. Rovnako sa v algoritme vyskytuje chyba, ktorá spôsobí, že nenastane detekcia Z-DNA konformácie na začiatku prípadne na konci sekvencie [38].

3.6 CpG ostrovčeky a ich epigenetika

CpG ostrovčeky (CpG Islands) sú krátke úseky DNA sekvencie, ktoré sú charakterizované vysokou početnosťou CG dinukleotidov s ohľadom na iné regióny. Typicky sa nachádzajú v promotéroch génov a sú významné pre reguláciu génovej expresie. Približne 70% génových promotérov je v ľudskom genóme spájaných s CpG ostrovčekmi, čo poukazuje na ich dôležitosť pri riadení iniciácie transkripcie [39, 40].

3.6.1 Charakteristika CpG ostrovčekov

CpG ostrovčeky sú zvyčajne dlhé 200 až 2000 bázových párov s vysokou koncentráciou CG dinukleotidov. V porovnaní s inými oblasťami v génoch sú menej metylované, čo prispieva k väčšej otvorenosti a tým sa uľahčí aktivácia transkripcie danej časti DNA [39].

3.6.2 Význam metylácie

Metylácia je biochemická modifikácia DNA. Proces zahŕňa pridanie metylovej skupiny ($-CH_3$) k cytozínovej báze v sekvencii DNA, konkrétne na CpG dinukleotidov. Redkcia je katalyzovaná enzýmami nazývanými DNA metyltransferáty [41].

Metylácia CG v miestach CpG ostrovčekov môže viesť k utlmeniu expresie génov. Toto je kľúčový mechanizmus pre deaktiváciu chromozónu X, imprintovanie a utíšenie génov pre zárodočnú líniu [41]. Nepravidelné vzorce metylácie, ako metylované CpG ostrovčeky, sú považované za abnormality spájané s rôznymi ochoreniami, medzi ktoré sa radia aj nádorové ochorenia. Napríklad hypermetylácia, teda zvýšenie miery metylácie, môže viesť k zníženiu aktivity génov, ktoré za normálnych okolností napomáhajú potláčať karcinogenézu. Zneaktívnenie takéhoto génu vedie k nekontrolovateľnej proliferácií nádorových buniek [42, 43]. Naopak hypometylácia, teda zníženie miery metylácie, vedie k aktivácií onkogénov, čo môže mať za následok vznik určitých druhov zhubných ochorení [44].

3.6.3 Výskum a klinické dôsledky

Pochopenie mechanizmu, ktorý reguluje metyláciu CpG ostrovčekov, má signifikantný význam pre základnú vedu aj klinickú prax. Otvára nové cesty pre vývoj cielených epi-

genetických terapií, ktoré môžu upravovať a regulovať aberantné metylačné vzorce, ktoré sa spájajú s chorobnými stavmi [45, 42].

3.6.4 Ukážka CpG ostrovčku v sekvencii

V nasledujúcej vzorovej DNA sekvencii číslo3 o veľkosti 700 nukleotidov sa nachádza jeden ostrovček (označený červenou farbou). Aby bola určitá sekvencia DNA označená za CpG ostrovček musí spĺňať parametre ako sú napríklad minimálny podiel G a C 50%, dĺžka minimálne 200 nukleotidov a iné. Dĺžka vzorového CpG ostrovčeka predstavuje 398 nukleotidov.

3.7 Analýza a vyhľadávanie

Detekcia a analýza CpG ostrovčekov je momentálne dostupná niekoľkými výpočtovými modelmi. Každý z nich je dizajnovaný na identifikáciu regiónov bohatých na CG páry. Aj keď sú tieto nástroje považované za užitočné, každý z nich obsahuje niekoľko limitácií, ktoré obmedzujú ich využitie.

3.7.1 CpG Island Searcher

Nástroj *CpG Island Searcher* identifikuje ostrovčeky na základe pozorovaných C a G nukleotidov. Daná implementácia sa riadi prísnymi kritériami ako pomer pozorovaných a očakávaných CpG dinukleotidov musí byť väčší ako 0.6 a obsah G a C nukleotidov musí byť viac ako 50% v skúmanej sekvencii. Aj keď sa tento nástroj so svojimi striktnými požiadavkami javí ako jeden z najlepších, v súčasnej dobe už nie je k dispozícií. Momentálne je dostupných niekoľko nových implementácií rovnakého algoritmu, ale každý z nich vyžaduje znalosť programovania [40].

Nástroje na vyhľadávanie CpG ostrovčekov, ktoré sa odvíjajú od algoritmu navrhnutého v priloženej štúdii [40] je napríklad Python skript od Michala Urbana [46] alebo implementácia v Cython-e od Lucasa Nella [47].

3.7.2 MethPrimer

MethPrimer je nástroj určený na navrhovanie primerov (krátkych úsekov DNA) pre bisulfitové PCR a metylačné analýzy. Funguje tak, že po prečítaní sekvencie DNA identifikuje CpG ostrovčeky, teda oblasti s vysokou koncentráciou dinukleotidov CG. Tieto ostrovčeky majú často regulatívny význam, keďže metylácia v ich oblasti môže významne ovplyvniť expresiu génov.

V kontexte bisulfitovej PCR, ktorá sa používa na detekciu metylácie DNA, MethPrimer navrhuje primery na amplifikáciu týchto CpG ostrovčekov po konverzii bisulfitom. Táto chemická úprava mení nemetylované cytosíny na uracil, zatiaľ čo metylované cytosíny ostávajú nezmenené, čo umožňuje identifikovať miesta metylácie.

Aj keď je MethPrimer užitočný nástroj, niekedy generuje nesprávne výsledky alebo nesprávne identifikuje CpG ostrovčeky. Preto sa odporúča dôkladne skontrolovať a ručne vybrať primery, aby sa zabezpečila presnosť v identifikácii CpG ostrovčekov a následnej metylačnej analýze [48].

3.7.3 EMBOSS CpGplot

Aplikácia *EMBOSS CpGplot* patrí medzi balík nástrojov *EMBOSS*. Balík nástrojov EMBOSS je vysoko konfigurovateľný na analýzu DNA a na vyhľadávanie regiónov bohatých na CG dinukleotidy. Napriek svojej flexibilite, môže byť výkonnosť nástroja pre jednotlivé genómy nekonzistentná. Jedná sa hlavne o atypické alebo menej charakterizované sekvencie. Táto variabilita obmedzuje jeho účinnosť v porovnávacích genomických výskumoch, kde sa vyžaduje rovnaká citlivosť u rôznych druhoch genómu [49].

```
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGCGCGCG
CGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG
CGCGCGCGCCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGC
GCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAA
```

Kód 3: Ukážka CpG ostrovčeku v DNA sekvencii. CpG ostrovček je označený červenou farbou [50].

3.8 DNA Analyser

DNA analyser je webová aplikácia, špecifickejšie rich internet application (RIA), ktorá je hostovaná na serveroch Biofyzikálneho ústavu Akadémie vied Českej republiky. Aplikácia zabezpečuje širokú škálu analýz DNA sekvencií. Integruje viaceré nástroje, ktoré slúžia na rôzne aspekty výskumu genómov, od identifikácie špecifických štruktúr až po prognózu funkčných a významných miest v sekvencii. Táto platforma slúži ako základný zdroj pre výskumníkov, ktorí pracujú na pochopení komplexnej genetickej informácie a jej dôsledkoch v biologických systémoch. Táto aplikácia bola vyvinutá vďaka spolupráci Ústavu informatiky na Mendelovej univerzite v Brne a Biofyzikálneho ústavu Akadémie vied Českej republiky [51].

Aplikácia a jej nástroje sú implementované ako Singe-Page Application (SPA) s použitím REST API. Používa technológie Vue.js pre frontend a Java s použitím Spring frameworku pre backend. Táto kombinácia umožňuje nezávislé fungovanie na hardware klienta, pretože všetky výpočty sú vykonávané na strane serveru. Aplikácia podporuje rôzne druhy

importov sekvencií, ako je import priamo z databázy NCBI, upload súboru alebo manuálne zadanie DNA/RNA sekvencie. Vďaka jej architektúre umožňuje analýzy od krátkych úsekov DNA až po celé chromozómy [52, 53].

3.8.1 G4Hunter Web Application

Webová aplikácia G4Hunter je nástroj vyvinutý na uľahčenie predikcie G-quadruplexov (G4) v sekvenciách DNA a RNA. G4 je špeciálna štruktúra, ktorá vzniká keď sa guanínové bázy spoja a vytvoria štvorvláknovú štruktúru. Ide o webovú verziu pôvodného algoritmu G4Hunter, ktorá vylepšuje používateľskú skúsenosť a jednoduchosť použitia. Dopĺňa grafické rozhranie a odstraňuje potrebu lokálnej inštalácie software-u [54, 52].

Aplikácia poskytuje používateľsky prívetivé prostredie na rozbor sekvencií a integruje sériu nástrojov, ktoré pomáhajú pri identifikácii a analýze potencionálnych sekvencií, kde môže dôjsť k tvorbe quadruplexov. Je navrhnutá tak, aby vyhovovala začínajúcim aj skúsenejším výskumníkom a to tým, že poskytuje jednoduchý analytický postup, ako je napríklad nastavovanie parametrov algoritmu. Analýza beží na pozadí a po dokončení ponúkne výsledok a grafickú vizualizáciu [52].

Umožňuje zobrazenie výsledkov v niekoľkých formátoch. Medzi ne patrí interaktívna heatmap a detailné popisy sekvencie. Tieto vizualizácie napomáhajú používateľovi jednoducho interpretovať dáta a identifikovať významné regióny v sekvencii. Aplikácia ponúka export výsledkov v CSV formáte pre prípadné ďalšie spracovanie. Taktiež existuje možnosť exportovania do bedGraph [52].

Validačné testy potvrdzujú, že webová aplikácia poskytuje rovnaké výsledky ako originál G4Hunter napísaný v jazyku Python2. Prínos webovej verzie je okrem spomenutých výhod aj vo vylepšenej rýchlosti a použiteľnosti. Taktiež umožňuje spracovávať väčšie dátové sety a komplexnejšie dotazy, čím uľahčuje širší výskum v oblasti genetiky [52]

3.8.2 G4Killer Web Application

Jedná sa o nástroj, ktorý priamo súvisí s G4Hunter spomenutým v sekcii 3.8.1. Slúži na zjednodušenie výskumu quadruplexov. Umožňuje racionálne dizajnovanie mutácií, ktoré znižujú sklon k tvorbe G4. Aplikácia G4Killer využíva zavedený algoritmus G4Hunter na analýzu a návrh minimálnych mutačných zmien, ktoré sú potrebné na zníženie skóre G4Hunter v sekvencii [55].

Používateľovi je umožnené zadať do nástroju jednu alebo viac sekvencií DNA a určiť cieľové maximálne skóre G4Hunter, ktoré by mutácie nemali prekročiť. Aplikácia na základe analýzy navrhne mutácie, ktoré dosiahnu predom určené skóre s minimálnym počtom zmien, čím sa zachová integrita pôvodnej sekvencie v čo najvyššej možnej miere. Výsledky môžu obsahovať viacej možností mutácie danej sekvencie, ktorá obsahuje opakujúce sa alebo podobné stopy bohaté na G. Výsledky sú vizualizované v intuitívnom rozhraní, ktoré umožňuje porovnanie a výber optimálnej zmutovanej sekvencie [55].

G4Killer funguje na základe breadth-first search (BFS) algoritmu, ktorý má za cieľ efektívne preskúmať všetky možné mutácie a zabezpečiť minimálnu mutačnú cestu. Okrem osobitnej implementácie, sa taktiež implementoval priamo do nástroju G4Hunter, čo umožňuje plynulý prechod od predikcie k návrhu mutácie. Nástroj výrazne uľahčuje a urýchľuje prácu hlavne výskumníkov zaoberajúcich sa genomickou stabilitou, transkripčnou reguláciou a terapeutickým potenciálom quadruplexov [55].

3.8.3 R-Loop Tracker Web Application

Nástroj pre detekciu a analýzu R-loop štruktúr v genomických DNA sekvenciách sa nazýva R-Loop Tracker. R-loop je trojvláknová štruktúra nuklovej kyseliny, ktorá sa skladá z hybridu RNA-DNA a posunutej jednovláknovej DNA. Táto štruktúra je významná vzhľadom na jej úlohu pri zvýšenej mutagenéze a iných bunkových dysfunkciách súvisiacich s ochoreniami ľudstva [53].

Nástroj podporuje dva modely pre detekciu. Zameriava sa na identifikáciu guanínových zoskupení idúcich za sebou. Dané zoskupenia sú kritické pre tvorbu R-loop formácie. Výsledky sa prezentujú pomocou interface-u, ktorý zahŕňa heatmap-u lokácií R-loop štruktúr a detailnú anotáciu každého detekovaného výsledky. Rovnako ako G4Hunter, aj tento nástroj umožňuje export výsledkov do CSV a bedGraph formátov na prípadné ďalšie spracovanie [53].

Tento komplexný nástroj sa nezameriava iba na detekciu, ale pomáha aj pri analýze R-loop, pričom dáva do súvisu genomické údaje s potenciálnymi biologickými dôsledkami [53].

3.8.4 Palindrome Analyser

Palindrome analyser je komplexný webový nástroj, vytvorený medzi prvými v aplikácii DNA Analyser. Je určený k identifikácií a rozboru invertovaných opakovaní (palindrómov) v nukleotidových sekvenciách. Palindrómy sú schopné vytvárať kľúčové štruktúry DNA, napríklad krížové formy (cruciform). Takéto štruktúry majú významný vplyv na replikáciu génov, expresiu a stabilitu nukleozómov. Okrem iného sa podieľajú na vzniku onkologických ochorení, neurodegeneratívnych poruchách a iných chorôb. [56].

Analyzátor poskytuje detailné informácie o vlastnostiach invertovaných opakovaní vrátane ich dĺžky, veľkosti slučiek, nepresností alebo energie vyžadovanej na vytvorenie cruciform-u. User-friendly prostredie ponúka interaktívne grafické znázornenie distribúcie a podrobných charakteristík nájdených palindromov. Podporuje pokročilé možnosti vyhľadávanie a triedenia na základe dĺžky repetície, dĺžky slučky a počtu nezhodných nukleotidov. Tento nástroj má schopnosť paralelizovať svoje výpočty, čo výrazne skracuje čas na spracovanie veľkých genomických sekvencií [56].

3.8.5 p53 Predictor

Prediktor proteinu p53 je špecializovaný komponent nástroja Palindrome Analyser, predstaveného v sekcii 3.8.4, ktorý je určený na zhodnotenie podobnosti sekvencií DNA k proteinu p53. Tento protein je kľúčový nádorový supresor, ktorý sa podieľa na regulácii bunkového cyklu a stabilite genómu. Nástroj využíva referenčnú sekvenciu DNA, ktorá je zobrazená v priloženom DNA kóde 4. Táto sekvencia je známa svojimi optimálnymi väzbovými vlastnosťami s proteinom p53. Je teda klúčová pre jeho funkciu pri inhibícií progresie bunkového cyklu v reakcii na poškodenie DNA [57].

```
GG(A/G)CATGCCCGGGCATG(T/C)CC zjednodušené ako GGACATGCCCGGGCATGTCC
```

Kód 4: Referenčná sekvencia DNA pre väzbu s proteínom p53 [57].

Modul umožňuje používateľom zadávať DNA sekvencie s dĺžkou maximálne 200 nukleotidov s následným výpočtom väzbovej afinity k p53. Taktiež dochádza k identifikácií čiastkovej sekvencie v rámci väčších vstupov vykazujúcich najvyššiu afinitu. Táto funkcia je obzvlášť užitočná pre výskumníkov, ktorí študujú mutácie v rámci väzobných miest p53 a ich vplyv na reguláciu génov [57].

Efektívny a používateľsky prívetivý nástroj poskytuje dôležité informácie o tom, ako môžu genetické odchýlky vo väzobných miestach ovplyvniť schopnosť proteínu regulovať cieľové gény, a tým modulovať bunkovú funkciu a progresiu ochorení [57].

4 Architektúra aplikácie DNA analyser

V tejto kapitole sa nachádza detailný popis a analýza architektúry existujúcej webovej aplikácie. Slúži na hlboké pochopenie pôvodnej infraštruktúry, dizajnu a funkcionality. Dôležitou súčasťou uvedeného textu je aj technologické stack (zoznam technológií) aplikácie. Významná časť je venovaná deskripcii pôvodných technológií a spôsobu komunikácie medzi nástrojmi a súčasťami, ktoré sú popísané v časti 3.8.

Pochopenie originálneho stavu aplikácie je dôležité pre rozširovanie funkcionality a na identifikáciu stávajúcich problémov, ktoré negatívne ovplyvňujú bezproblémový vývoj a pridávanie nových potrebných nástrojov pre výskumníkov.

4.1 Frontend

Ako už bolo spomenuté v sekcii 3.8, aplikácia sa skladá z dvoch hlavných častí. Z pohľadu používateľa je najdôležitejšia práve frontend, teda všetko to, čo užívateľ vidí. Predkladaná kapitola bližšie približuje frontend danej aplikácie.

Frontend (FE) DNA analyzéru bol navrhnutý predovšetkým tak, aby poskytoval dynamické prostredie, ktoré je súčasťou interakcie s používateľmi. Táto časť aplikácie je vytvorená pomocou JavaScript-ového frameworku Vue.js vo verzii 2 [56, 52]. Tento framework je populárnou frontendovou knižnicou, ktorá umožňuje vytvárať interaktívne používateľské rozhranie s dôrazom na deklaratívne renderovanie a zloženie komponentov [58]. Tejto verzii bola ukončená podpora 31.12.2023 a odporúča sa prechod na novšiu verziu [59]. Verzia 3 prináša vylepšenia ako podporu pre TypeScript, kompozičné API, optimalizovaný rendering a mnoho ďalších. Taktiež je označovaný za rýchlejší, menší a viac udržovaný framework [60].

4.1.1 Dynamická úprava menu

Úspešnosť spustenia webovej aplikácie spočíva v korektnom kontaktovaní serveru a následne v správnej odpovedi s možnosťou analýz. Pri prvom načítaní FE odosiela požiadavok na server na zistenie aktuálne dostupných analýz. Navigačné menu sa dynamicky upraví na základe odpovede, aby zobrazovalo len tie možnosti, ktoré sú pre užívateľa dostupné. Daný mechanizmus garantuje, že používateľské rozhranie vždy poskytne len tie alternatívy, ktoré sú pre používateľa spustiteľné.

4.1.2 Monitoring dokončenia analýz

Aplikácia podporuje simultánne spustenie viacerých analýz, pričom beh analýzy môže trvať od niekoľkých sekúnd až po hodiny, preto vyžaduje opakované dotazovanie na server. Pre minimalizovanie zbytočného sieťového toku a vyššiu šancu na úspešné spracovanie požiadavky bez preťaženia serveru je využitá metóda Exponential Backoff. Daná metóda sa často využíva v distribuovaných systémoch a sieťových aplikáciách na efektívne spracovanie asynchrónnych volaní, kedy sa očakáva vysoká variabilita v trvaniach odpovede

alebo dostupnosti služieb [61]. Daný prístup je obzvlášť užitočný v prípadoch, kedy server spracováva dlhotrvajúce úlohy, ako napríklad analýzu DNA. Použitie Exponential Backoff zaisťuje, že systém zostáva reaktívny a efektívny aj pri vysokom zaťažení alebo pri potenciálnych výkyvoch vo výkone serveru.

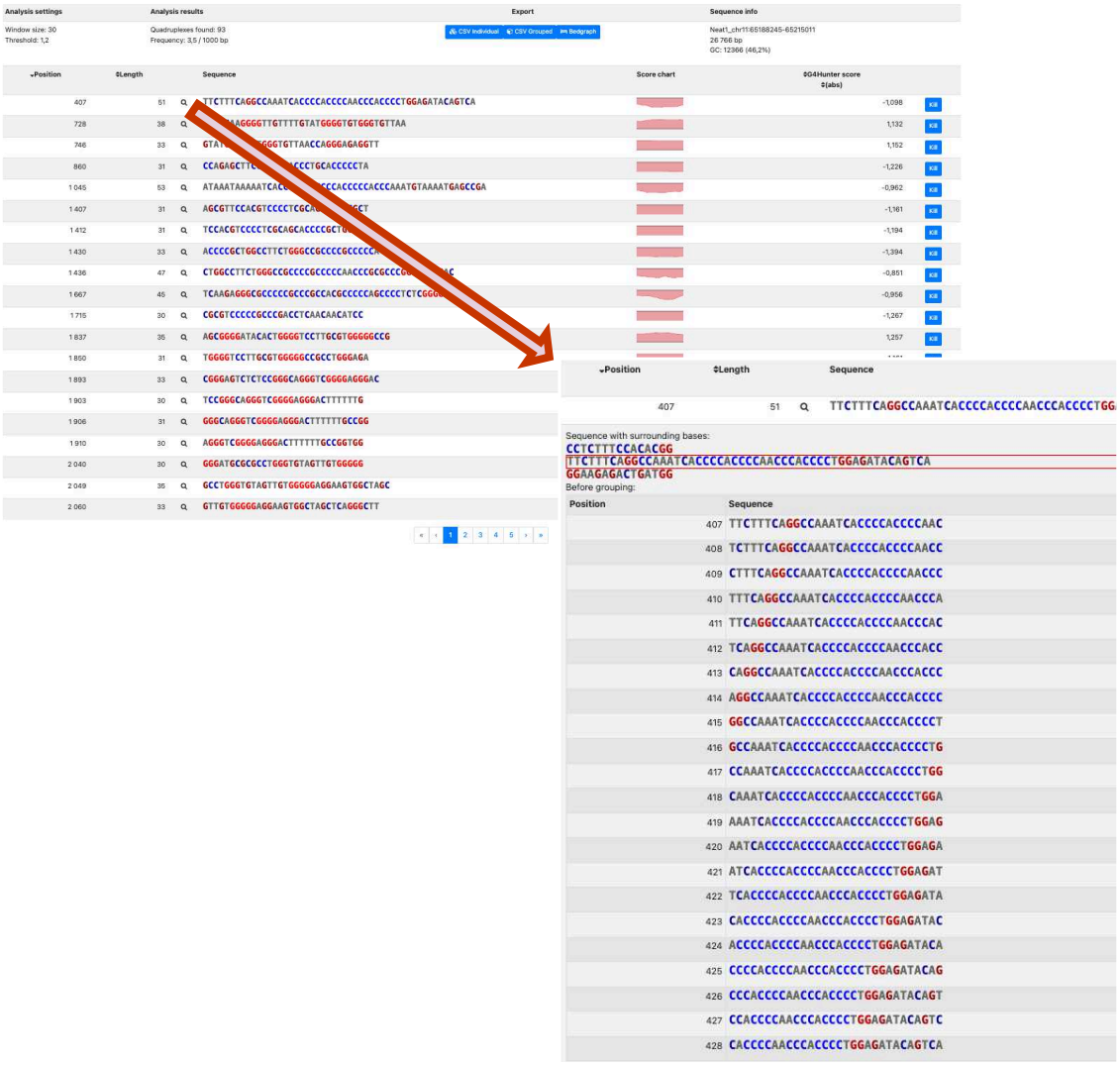
4.1.3 Asynchronita frontendu

Asynchrónnosť je základným prvkom pre bezproblémové používanie aplikácie, hlavne pri práci s údajmi, ktorých spracovanie si vyžaduje určitý čas. V rámci aplikácie DNA Analyser je asynchrónnosť implementovaná pomocou framework-u Vue.js, ktorý mimo iného uľahčuje načítavanie komponentov a údajov. Daná technika umožňuje používateľom pokračovať v interakcii s aplikáciou, zatiaľ čo sa údaje postupne načítavajú alebo aktualizujú na pozadí, bez potreby obnovenia celej stránky či zamrznutia pri renderovaní výsledkov.

Asynchrónne načítanie údajov je dôležité najmä pre zobrazenie veľkého množstva výsledkov analýzy, ktorých spracovanie môže trvať dlhší čas. Vue.js dokáže tento proces efektívne riadiť prostredníctvom rôznych techník a funkcií, ako je napríklad lazy-loading. Táto metóda načíta moduly a komponenty iba vtedy, kedy je to absolútne nevyhnutné. Vďaka čomu sa čas načítania a záťaž na strane klienta znižujú [58].

Ďalšou dôležitou technikou, ktorá sa využívaná v aplikácii na správu veľkých objemov údajov, je stránkovanie. Týmto sa nielen zlepšuje práca s údajmi, ale aj minimalizuje čas potrebný na načítanie alebo zobrazenie výsledkov analýzy. V kombinácii s asynchrónnym načítaním, stránkovanie umožňuje pohodlne a bezproblémovo prechádzať medzi výsledkami bez dlhého čakania [58, 52, 55].

Využitie daných techník zobrazuje obrázok číslo 5. Pri analýze obecne malej sekvencii DNA s 26 766 bázovými pármi vzniká niekoľko strán výsledkov, kedy sa každý z výsledkov delí na ďalšie väčšie celky. Početnosť dát si vyžaduje aplikovanie techník asynchrónneho načítavania a stránkovania.



Obrázok 5: Ukážka využitia techník asynchrónneho načítavania a stránkovania.

4.1.4 Reaktívnosť aplikácie

Reaktivita je kľúčovou funkciou vo frameworku Vue.js. Výrazne vylepšuje interaktivitu a odozvu webových aplikácií. Reaktivita, v kontexte danej aplikácie, je využívaná na zlepšenie používateľského zážitku (user experience) a zefektívnenie spracovania dynamických údajov.

Reaktívne funkcie Vue.js zabezpečujú, že akékoľvek zmeny v údajoch vykonané na pozadí, sa okamžite premietnu do používateľského rozhrania. Zmeny teda nevyžadujú manuálne obnovenie stránky. Daný prístup je prínosný pre:

- Výsledky analýzy: Nové výsledky analýzy sa automaticky zobrazia v používateľskom rozhraní, dynamicky sa aktualizujú grafy, tabuľky a textové súhrny. Podrobnejšie

informácie o grafoch popisuje sekcia 4.1.5.

- Aktualizácia stavu: Priebeh analýzy sa zobrazuje v reálnom čase, vďaka čomu sú používatelia informovaní bez akejkoľvek manuálnej kontroly.

Aplikácia reaktívne spracováva rôzne prvky používateľského rozhrania na základe interakcie alebo aktualizácie údajov:

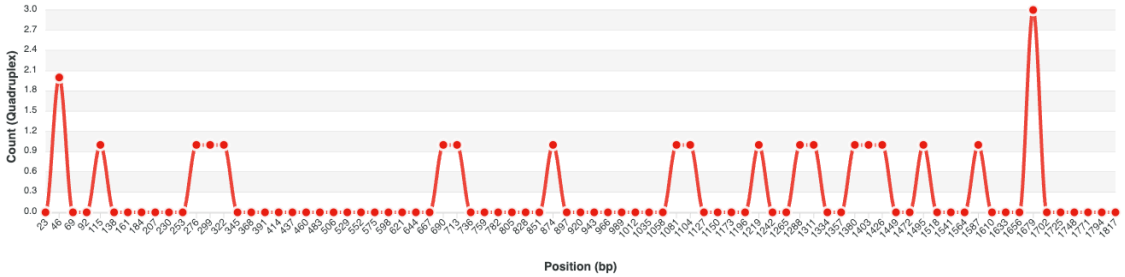
- Podmienený rendering: Pomocou Vue.js sa zapína a vypína zobrazovanie prvkov v závislosti od stavu údajov. Takéto podmienené formátovanie sa používa na zobrazovanie ďalších parametrov analýzy na základe predchádzajúcich atribútov.
- Interaktívna vizualizácia: Zmeny parametrov vizualizácie výsledkov sa prejavia okamžite.

4.1.5 Grafy

Grafické reprezentácie dát sú v bioinformatických aplikáciách nevyhnutnou súčasťou efektívneho sprostredkovania komplexných dátových modelov a výsledkov analýz. V uvedenej aplikácii sa pre každý rozbor využíva niekoľko typov grafov. Každá analýza je vytvorená tak, aby došlo k zjednoteniu formátovania grafov. Grafické zobrazenia dát zlepšujú interpretovateľnosť a dostupnosť výsledkov sekvencie DNA. Medzi používané grafy v aplikácii patria základný čiarový graf, tepelné mapy (heatmap) a minigrafy. Každý z nich využíva ovládacie prvky reaktívnosti a slúži na jedinečný účel v stratégii prezentácie výsledkov.

Čiarové grafy

Čiarové grafy sú využívané pri každej analýze na zobrazenie počtu výsledkov na rôznych pozíciách pozdĺž celej analyzovanej sekvencie DNA. Daný typ grafu je obzvlášť užitočný pri identifikácií oblastí s vysokým počtom výsledkov. Môže dôjsť k identifikácií významných oblastí pre určitú analýzu, čo poskytuje jednoduchú vizuálnu metódu analýzy údajov. Osa X obsahuje rovnomerne rozmiestnené pozície v sekvencii na základe ich dĺžky a osy Y. Osa Y prezentuje počet výsledkov, ktoré pretínajú, teda obsahujú danú pozíciu. Na obrázku 6 je zobrazený graf pre testovaciu sekvenciu a analýzu G4Hunter. Daný graf umožňuje filtráciu výsledkov pomocou kliknutia na vybranú pozíciu. Prípadne existuje možnosť výberu skupiny výsledkov danej sekvencie.



Obrázok 6: Ukážka počtu výsledkov naprieč pozíciami DNA sekvencie.

Heatmap

Heatmapy sú používané pri reprezentácii veľkých súborov údajov v kompaktnej vizuálnej forme. Farby predstavujú hodnoty. Sú obzvlášť účinné pre zobrazenie vzorov alebo gradientov v sekvencii DNA. Dané zobrazenia poskytujú možnosť zistiť rozdiely v úrovniach expresii génov alebo mutačné ohniská. Vizualizácia tak isto pomáha rýchlo určiť a posúdiť celkové rozloženie a oblasti záujmu v rámci komplexných dát. Rovnako ako predchádzajúci graf, ponúka možnosť filtrovania výsledkov na základe výberu ohniska [62]. Na obrázku 7 je ukážka heatmapy.

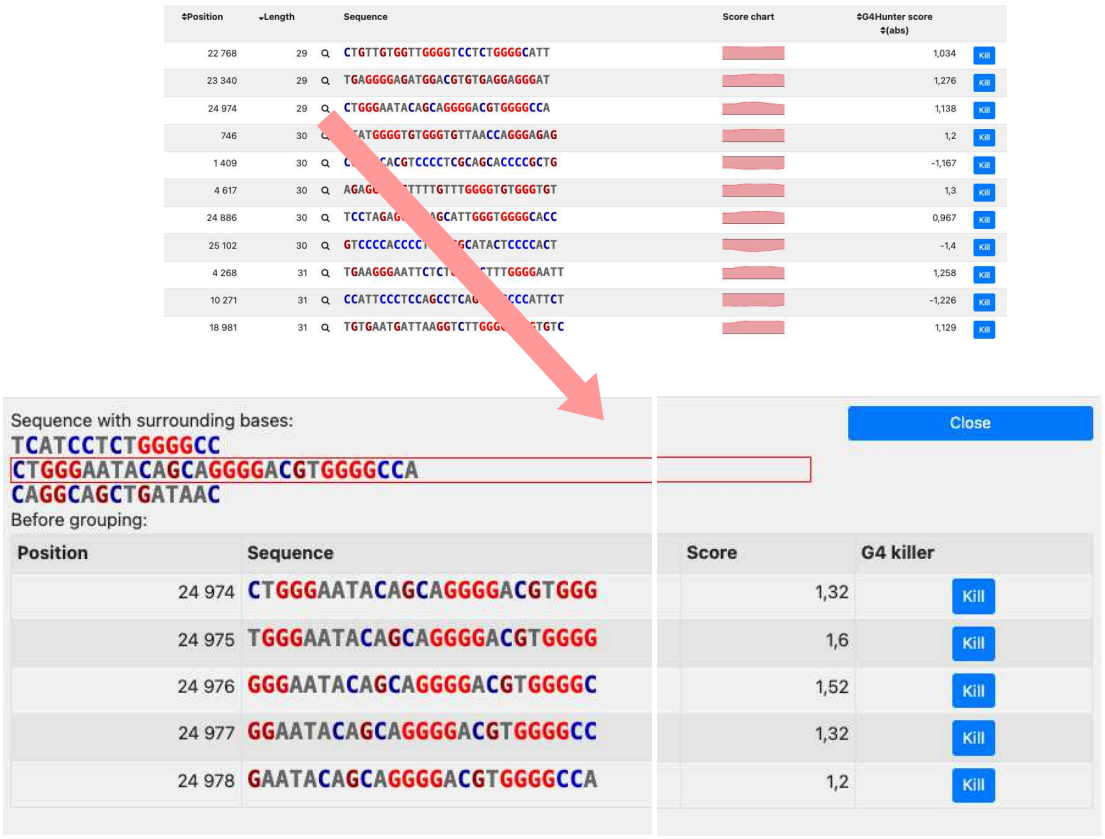
Obrázok 7: Ukážka tepelnej mapy výsledku analýzy naprieč pozíciami DNA sekvencie.

Minichart

Komponenta Minichart je jednoduchá reprezentácia výsledkov špeciálne navrhnutá pre G4Hunter analýzu. Táto vizualizácia poskytuje zhodnotenie každého výsledku bez potreby detailného skúmania každého jedného hitu v analýze. Minichart je určený na zobrazovanie údajov z jednorozmerných polí. Je vhodný na rýchle zobrazenie prehľadov rozloženia údajov vo vybranom výsledku, čo je v tomto prípade každý výstup danej analýzy.

Graf daného typu je vyobrazený na obrázku 8. Rozpísanie minichartu v stĺpci *Score chart* nastane po kliknutí na tlačidlo lupy. Graf zobrazuje dátové body pozdĺž vodorovnej osi, pričom každý bod predstavuje hodnotu zo súboru údajov. Tieto údaje sú spojené čiarami a tvoria súvislý graf. Obsahuje referenčnú nulovú čiaru, ktorá umožňuje rýchlu identifikáciu kladných a záporných hodnôt.

Vďaka implementácii daného komponentu do aplikácie, je umožnené používateľom skúmať menšie segmenty bez nutnosti prechádzať z aktuálneho zobrazenia. Daný prístup výrazne uľahčuje pracovný postup analýzy.



Obrázok 8: Ukážka Minichartu a dát, z ktorých sa daný Minichart vytvoril.

4.1.6 Grafické prvky rozhrania

Medzi ďalšie prvky v DNA Analyser patria strategicky použité grafické komponenty, ktoré poskytujú intuitívne a efektívne UI. Dané prvky napomáhajú pri navigácii, sprostredkovávajú dôležité správy a vizuálne organizujú komplexné genetické dáta. Nasledujúca časť podrobne opisuje úlohu tlačidiel, tooltipov, varovných a chybových správ a farebného kódovania genetickej sekvencie.

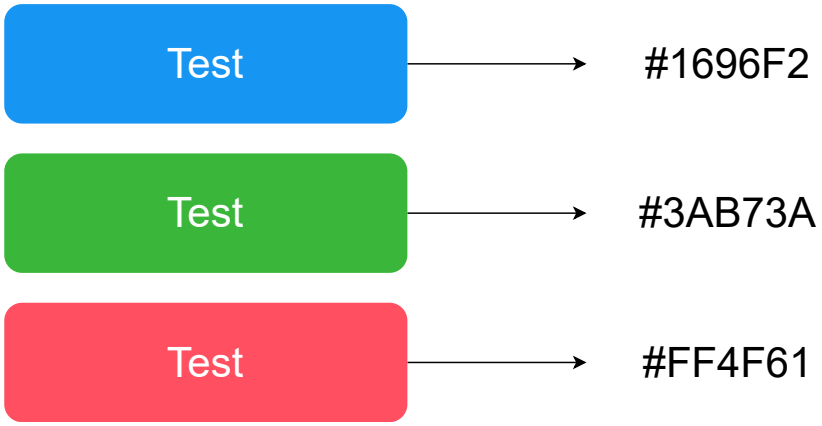
Tlačidlá

V celom systéme sú umiestnené tlačidlá na viditeľnej pozícií a pri každej analýze plnia rovnakú úlohu. Slúžia na spustenie analýzy, export údajov a na ďalšie funkcie v rámci analýz, výsledkov a grafov. V aplikácii existujú viaceré druhý tlačidiel, farebne rozlíšené podľa funkcionality:

- Zelené: slúžia na spustenie analýzy a import údajov.
- Modré: využívajú sa na dynamické zmeny v UI, ako napríklad otvorenie informačnej tabuľky, prepínanie strán, prepínanie grafov a iné.

- Červené: využíva sa pri funkciách, na ktoré by mal užívateľ klásť väčšiu obozretnosť, napríklad vymazanie alebo uzatvorenie analýzy.

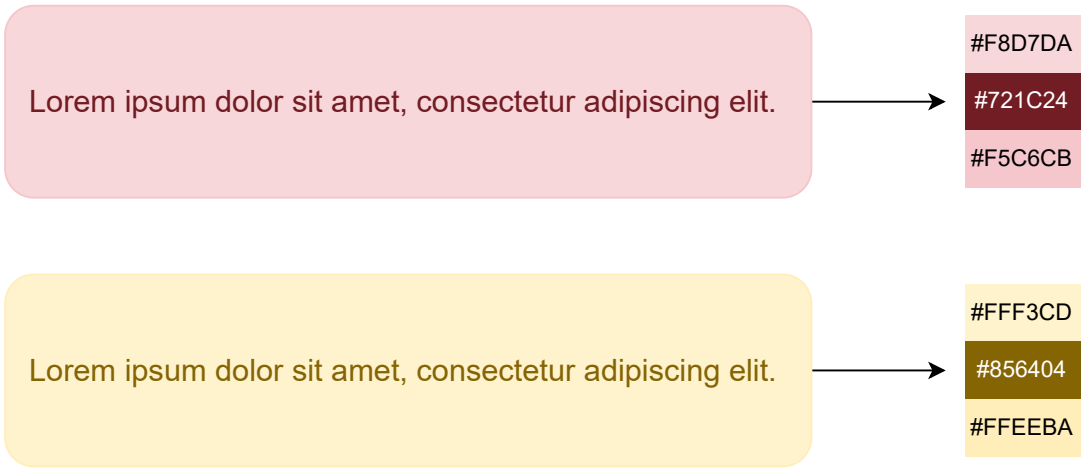
Obrázok 9 ukazuje typy tlačidiel využívaných v aplikácií. Dizajn týchto tlačidiel zostáva rovnaký ale môže sa meniť ich funkcia v závislosti od skupiny použitia.



Obrázok 9: Ukážka typu tlačidiel v aplikácii DNA Analyser.

Upozornenia a chyby

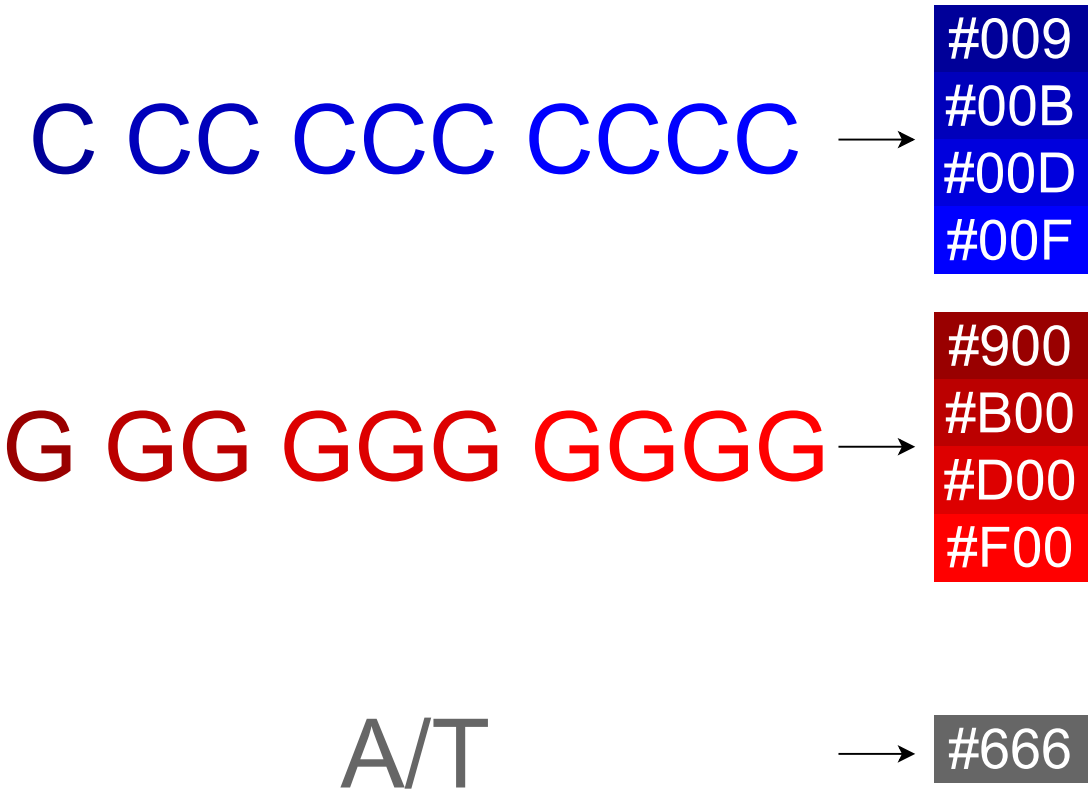
Varovné správy upozorňujú používateľov na možné problémy ako napríklad chýbajúce parametre alebo nulové výsledky z analýzy. Taktiež poskytujú odporúčania pri zadávaní parametrov mimo štandardizované rozsahy. Nabádajú výskumníkov k náprave parametrov pred začatím analýzy. Chyby sa zobrazujú na viditeľnom mieste s červeným podfarbením. Označujú zlyhanie v procese alebo nekonzistentnosť údajov. Poskytujú podrobný opis pre nápravu alebo pomoc pri riešení problému. Obrázok 10 poskytuje ukážku varovných správ aj s ich farebnou schémou.



Obrázok 10: Farebná schéma chybových a varovných správ.

Farebné kódovanie DNA sekvencie

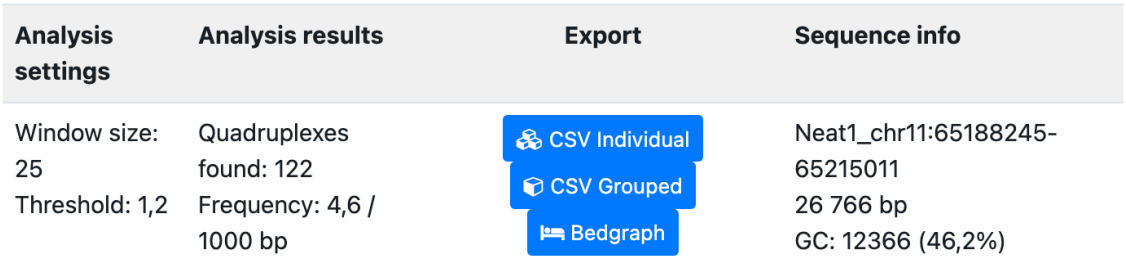
Aplikácia využíva vlastnú farebnú schému sekvencie DNA. Každý nukleotid, prípadne zoskupenie nukleotidov, je reprezentovaná jedinečnou farbou. Schéma umožňuje rýchle vizuálne rozlíšenie potrebných skupín, čím zvyšuje čitateľnosť dlhých sekvencií vo výsledkoch. Výskumníci dokážu na základe toho ľahko identifikovať vzory v rôznych segmentoch DNA. Daná farebná schéma je použitá v celom rozhraní aplikácie. Príklad farebnej schémy je zobrazený na obrázku číslo 11.



Obrázok 11: Farebná schéma rozlíšenia skupín z DNA sekvencie.

Analýzy a výsledky

Pre každú dokončenú analýzu aplikácia ponúka jednoduché zhrnutie, ktoré obsahuje parametre výpočtového modelu, zhrnutie výsledkov, možnosti pre export a informácie o sekvencii. Ide o dôležitú vlastnosť hlavne kvôli rýchlemu rozlíšeniu analýz, keďže aplikácia dovoľuje spustenie viacerých s rôzne nastavenými parametrami pre jednu sekvenciu. Príklad zhrnutia je zobrazený na obrázku12.



Obrázok 12: Zhrnutie analýzy s jej parametrami, výsledkami, možnosťami a informáciami o sekvencii.

Na záver možno konštatovať, že rozhranie aplikácie DNA Analyser využíva množstvo grafických prvkov, ktoré sú zamerané na zefektívnenie práce s výsledkami. Implementované prvky umožňujú bezproblémovo interpretovať genetické údaje.

4.2 Backend

Backend aplikácia využíva rôzne a moderné technológie, ktoré zabezpečujú dostatočný výkon, bezpečnosť a škálovateľnosť. V predloženom texte sú popísané dané technológie a rovnako ich súčasti.

4.2.1 Popis technológií

Daná podkapitola podrobne rozoberá jednotlivé využité technológie v aplikácii DNA Analyser.

Java 11

Základná logika backendu je napísaná v jazyku Java 11. Jedná sa o robustný a široko používaný programovací jazyk. Je známy svojou kompatibilitou s mnohými platformami a rozsiahlym ekosystémom [63]. Na zefektívnenie vývoja aplikácií, napísaných v jazyku Java, sa používa framework Spring, ktorý poskytuje podporu pre *dependecy injection*, správu transakcií a modularizáciu. Kľúčové funkcie, ako napríklad Spring Boot, zjednodušujú nastavenie aplikácie a umožňujú developerom sústrediť sa na bussines logic [64, 65].

PostgreSQL 12

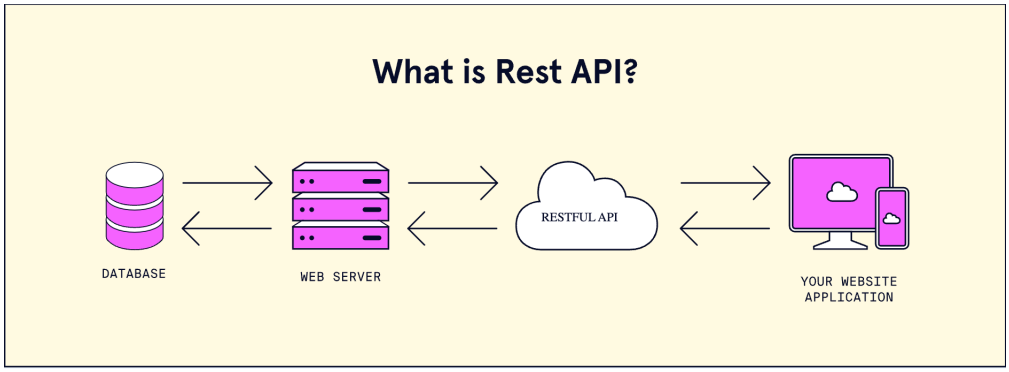
Na trvalé ukladanie údajov sa v aplikácii využíva technológia PostgreSQL vo verzii 12. Jedná sa o výkonný systém pre relačné databázy. Vďaka silnému súladu so štandarmi SQL je vhodný na komplexné dotazy a analýzy. V DNA Analyser sa PostgreSQL databáza využíva hlavne na ukladanie informácii ohľadom sekvencií a výsledkov analýz. Navyše plní funkciu uchovávania informácií o registrovaných a prihlásených používateľoch [66].

H2

Okrem PostgreSQL je v aplikácii využitá aj relačná databáza H2. Jedná sa o pamäťovo nenáročnú relačnú databázu, ktorá je optimalizovaná na výkon a spracovanie veľkých objemov údajov. Databáza slúži na uchovávanie sekvencií DNA. Dokáže efektívne ukladať veľké súbory údajov in-memory aj v perzistentnom režime. Flexibilita, ktorú poskytuje, umožňuje backendu zvládnuť vysokú potrebu zdrojov pri analýze sekvencií DNA. Jednoduchá konfigurácia databázy a jej integrácia s aplikáciami Java zabezpečujú bezproblémovú prevádzku v rámci existujúcej architektúry aplikácie [67]. Tým, že backend využíva výhody oboch typov databáz, dosahuje spoľahlivú rovnováhu medzi všestrannosťou a vysokým výkonom.

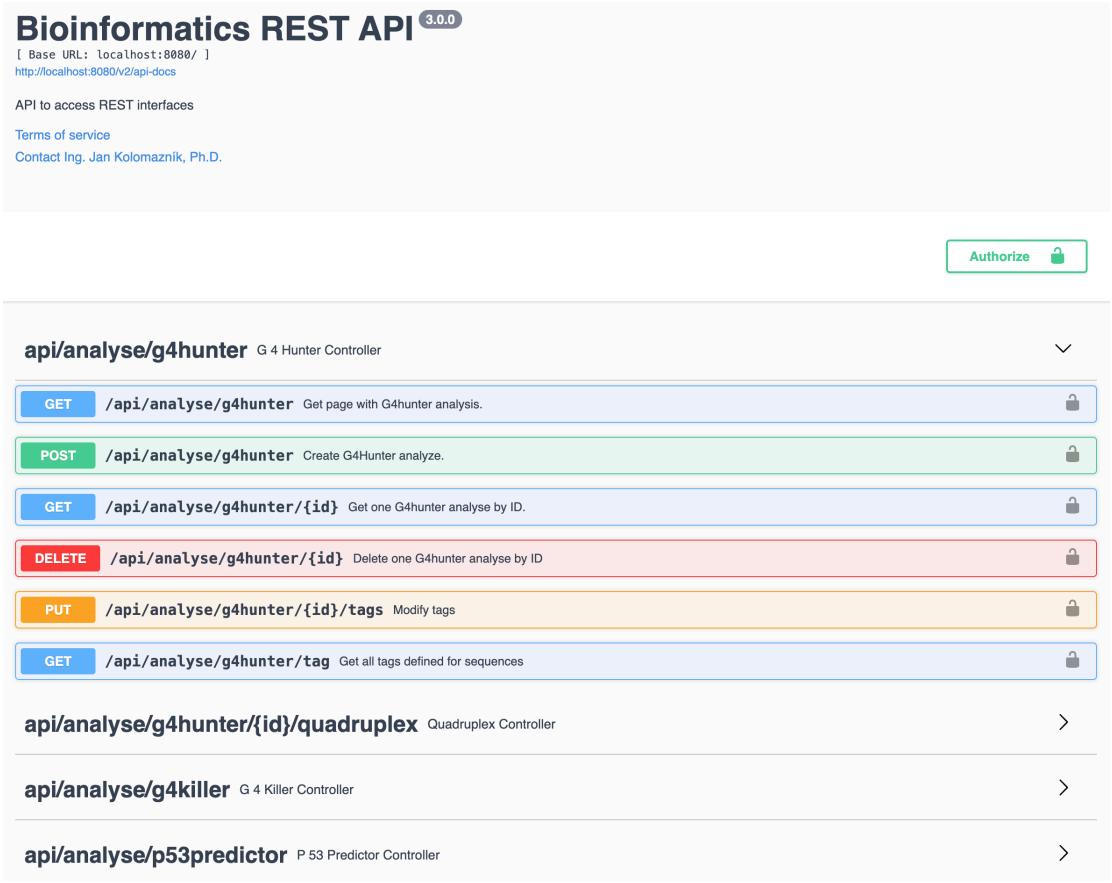
REST API

Rozhranie API RESTful (Representational State Transfer Application Programming Interface) slúži ako komunikačná vrstva medzi backendom a frontendom. Poskytuje koncové body, takzvané endpointy, na ktoré frontend (klient) posiela požiadavky (request) a backend (server) ich následne spracováva a odosiela odpoveď (response). Princípy REST zabezpečujú bezstavovú komunikáciu, štandardizované formáty údajov (JSON - JavaScript Object Notation) a jasné oddelenie záujmov medzi klientom a serverom. Daná štruktúra API zjednodušuje integráciu s rôznymi frontendmi a externými službami [68]. Obrázok 13 znázorňuje komunikácia pomocou REST API.



Obrázok 13: Schéma komunikácie klient-server pomocou REST API [69].

Aby bolo API prístupnejšie a ľahšie udržiavateľné, backend integruje Swagger pre dokumentáciu API. Swagger automaticky generuje interaktívnu dokumentáciu API na základe anotácií v Java kóde, čím umožňuje vývojárom skúmať endpointy, kontrolovať vstupné parametre a testovať odpovede priamo z webového prehliadača. Dokumentácia výrazne zjednodušuje zapájanie a integráciu s klientami [70]. Obrázok 14 približuje daný systém aplikácie DNA Analyser.



Obrázok 14: Prostredie dokumentácie Swagger aplikácie DNA Analyser.

Flyway

Flyway je systém na riadenie verzií databázy, správu zmien schémy a migrácie. Nástroj zabezpečuje konzistentnosť štruktúry databázy v rôznych vývojových prostrediach a fázach nasadenia. Udržiava históriu zmien prostredníctvom prírastkových migračných skriptov SQL [71].

Groovy

Groovy je agilný a dynamický programovací jazyk platformy Java. V aplikácii sa jazyk Groovy používa na spracovanie špecifických skriptovacích úloh [72]. Vylepšuje automatizované testovanie a zvyšuje produktivitu vývojárov. V rámci predkladanej aplikácie sa využíva pre CI/CD, testovanie a vytvorenie Docker kontajnerov.

4.3 Repozitár

Aplikácia bola vyvíjaná niekoľkými developermi od roku 2016 [56]. Nebolo by to uskutočniteľné bez využitia systému na verziovanie. Jedná sa o systém *GitLab*. Pre zdrojový

kód je vytvorený repozitár, ktorý uchováva záznamy o zmenách v kóde, nahlásené chyby a plánované vylepšenia. Na vývoj sa využíva aj systém vetvenia, kedy sa dá ľahko implementovať alebo opraviť časť kódu. Po dokončení vetvy môže následne dôjsť k spojeniu s hlavnou vetvou, kde sa nachádza aktuálna verzia pripravená pre produkciu. V rámci jedného repozitáru sú obsiahnuté všetky časti aplikácie. Zdrojový kód sa nachádza na voľne prístupnej adrese https://git.pef.mendelu.cz/bioinformatics/dna-analyser.

4.4 Docker

Docker je platforma, ktorá umožňuje vývojárom automatizovať balenie, nasadzovanie a spúšťanie aplikácií v kontajneroch. Kontajner obsahuje všetky závislosti, ktoré aplikácie potrebuje na spustenie. Zabezpečuje taktiež konzistentné správanie v rôznych prostrediach [73].

Kľúčovými vlastnosťami Docker-u sú:

- **Izolácia:** Aplikácia a jej závislosti sú izolované v kontajneri, čo zabraňuje vzájomnému ovplyvňovaniu služieb a zároveň umožňuje, aby na tom istom hostiteľskom stroji bežalo viacero aplikácií bez konfliktov.
- **Prenositeľnosť:** Kontajnery sú navrhnuté tak, aby sa dali spustiť v akomkoľvek prostredí, ktoré podporuje Docker. Môže sa jednať o pracovný počítač vývojára, lokálny server alebo cloudovú infraštruktúru. Daná skutočnosť poskytuje flexibilitu a konzistenciu.
- **Škálovateľnosť:** Docker umožňuje horizontálne škálovanie aplikácií s minimálnym úsilím. Je možné využiť orchestračné nástroje, ako sú Kubernetes alebo Docker Swarm.
- **Efektívne využitie zdrojov:** Kontajnery nie sú náročné, zdieľajú jadro hostiteľského systému a rýchlo sa spúšťajú, čím sú efektívnejšie ako virtuálne stroje.
- **Podpora mikroslužieb:** Docker je ideálny systém pre architektúru mikroslužieb. Podporuje nasadenie a správu nezávislých služieb pri zachovaní bezproblémovej komunikácie medzi nimi.

Docker Compose je nástroj, ktorý umožňuje vývojárom definovať a spravovať multi-kontajnerové aplikácie. Na definovanie a spravovanie sa používa YAML súbor [73]. V kóde 1 je možné vidieť naviazanie kontajnerov na jednotlivé moduly. Docker Compose umožňuje:

- Definovať služby, siete, zväzky v konfiguračnom súbore, čím sa zjednodušuje nastavenie aplikácie.
- Spustiť celý balík aplikácií jedným príkazom, čo zvyšuje efektivitu vývoja.
- Spravovať rôzne prostredia (vývojové, testovacie, produkčné) prepísaním nastavení v konfiguračnom súbore.

```yaml
version: '3.0'
services:
  gateway:
    image: dnaanalyser/gateway:${TAG:-latest}
    depends_on:
      - backend
    ...
  backend:
    image: dnaanalyser/dna-analyser:${TAG:-latest}
    depends_on:
      - postgres
    ...
  postgres:
    image: postgres:13

  pgadmin:
    image: dpage/pgadmin4:latest
    depends_on:
      - postgres
volumes:
    ...
```

Kód 1: YAML konfigurácia DNA Analyser aplikácie pre Docker Compose.

Docker Swarm je natívny orchestračný nástroj Docker-u. Umožňuje správu distribuovaných kontajnerov v rámci skupiny (cluster) počítačov. Medzi jeho funkcie patria:

- **Správa skupiny počítačov:** Jednoduché nastavenie a údržba uzlov v skupine Docker pre vysokú dostupnosť.
- **Škálovanie služieb:** Automatické zvyšovanie alebo znižovanie na základe dopytu, udržiavanie vyváženého pracovného zaťaženia.
- **Load Balancing:** Zabudovaný load balancer (vyvažovanie záťaže) zabezpečuje, že prichádzajúce požiadavky sú efektívne smerované na príslušné kontajnerované inštancie.
- **Deklaratívny model:** Služby a požadované stavy sú definované deklaratívnym spôsobom, čím sa zabezpečuje konzistentné nasadenie.

V aplikácii DNA Analyser zabezpečujú nástroje ako Docker Compose a Docker Swarm efektívne nasadenie a orchestráciu modulárnych komponentov. Odstraňujú taktiež potrebu inštalovania všetkých závislostí pre vývojárov iného komponentu. Tým pádom nie je nutné aby vývojári frontendu zdĺhavo inštalovali a nastavovali závislosti backendu a databázy, a naopak.

4.5 Modulárny systém

V softvérovom inžinierstve existuje prístup, nazývaný modulárny systém, ktorý rozdeľuje aplikáciu na menšie, nezávislé jednotky. Dané jednotky sa nazývajú moduly. Moduly sú navrhnuté tak, aby vyájomne spolupracovali, ale môžu fungovať aj nezávisle v rámci väčšieho systému [74]. V prípade aplikácie DNA Analyser modulárna architektúra zahŕňa hlavné komponenty:

- **Gateway:** Funguje ako vstupný bod pre všetky požiadavky klientov a distribuuje ich príslušným modulom.
- **Backend:** Obsahuje bussiness logic aplikácie a spracováva údaje. Funguje ako jadro na vykonávanie operácií požadovaných aplikáciou.
- **Frontend:** Používateľské rozhranie, pomocou ktorého používatelia komunikujú s aplikáciou.
- **PostgreSQL:** Databázový modul, ktorý poskytuje trvalé ukladanie údajov a zabezpečuje konzistenciu a spoľahlivosť dát.

Všetky uvedené moduly sú zabalené pomocou Docker Compose aplikácie, ktorá umožňuje efektívne nasadenie a správu. Daný modulárny systém zabezpečuje, že komponenty možno vyvíjať, udržiavať a škálovať. Moduly pritom zostávajú navzájom úzko integrované, čo poskytuje robustnú a flexibilnú architektúru.

4.5.1 Automatické nasadenie - CI/CD

Automatická integrácia (Continuous Itegration - CI) a automatické nasadenie (Continuous Delivery - CD) sú základnými postupmi pri vývoji moderného softvéru. Sú obzvlášť cenné pri správe modulárnych architektúr. CI/CD automatizuje proces nasadzovania softvéru a zabezpečuje, aby sa zmeny kódu integrovali, testovali a nasadzovali rýchlo a spoľahlivo. V rámci CI/CD je možné nastaviť vlastné pravidlá, ktoré spúšťajú jednotlivé prvky CI/CD.

- **CI:** Zahŕňa automatické testovanie všetkých zmien kódu od viacerých vývojárov v spoločnom repozitári. Pomáha včas identifikovať problémy s integráciou a zabezpečí, aby nové funkcionality alebo opravy nenarušili existujúcu funkčnosť. Je dostupných viacero systémov, ktoré organizujú tieto testy. Pri DNA Analyser sa používa GitLab CI/CD. Testy sa spúšťajú vždy, keď sa zaznamená nová zmena v kóde.
- **CD:** Zabezpečuje, že každá zmena kódu, ktorá prejde fázou CI, sa automaticky nasadí do produkčného prostredia. Daný postup minimalizuje oneskorenie medzi vývojom a nasadením, čím urýchľuje tempo inovácií a opráv.

V kontexte Docker Compose a Docker Swarm je možné nakonfigurovať CI/CD úlohy na vytváranie Docker images a ich nasadzovanie vo viacerých prostrediach. Táto integrácia výrazne znižuje zložitosť a potenciálne chyby spojené s manuálnym nasadzovaním.

Aplikácia DNA Analyser využíva CI/CD na zefektívnenie aktualizácií svojich modulárnych komponentov. Dané nastavenie uľahčuje nezávislý vývoj a údržbu a tiež zabezpečuje úzku integráciu modulov.

5 Metodika práce

Kapitola obsahuje detailný popis metodiky použitej pri refaktoringu kódu aplikácie, vylepšenie a pridanie nových potrebných analýz do DNA Analyser aplikácie. Podrobne vysvetľuje jednotlivé rozhodnutia autora práce, čím sa zabezpečí logickosť, transparentnosť a reprodukovateľnosť postupu.

Aplikácia DNA Analyser si vyžaduje refaktorovanie na plnú modulárnu architektúru. Spravovanie, vývoj a rozširovanie v jednom repozitári je zložité kvôli neprehľadnosti modulov. Ďalším z riešených problémov je nefunkčnosť CI/CD v pôvodnom repozitári a neprehľadnosť vykonávaných úloh. Zároveň sa aplikácia rozširuje o dve nové analýzy podľa požiadaviek výskumných pracovníkov.

5.1 Štruktúra repozitárov

Z pôvodného repozitára, ktorý obsahuje všetky moduly dokopy, je nutné vyextrahovať moduly do osobitných, nových repozitárov. Každý modul bude existovať separátne vo svojom vlastnom repozitári, čím sa docieli väčšia prehľadnosť v kóde, identifikovaných chybách a zabezpečí sa plná modularita systému. Rovnako sa vylepší aj testovanie každého samostatného modulu. Vďaka tomuto prístupu bude môcť byť modul vyvíjaný bez závislosti na ostatných moduloch. Zahŕňa to vyčlenenie nižšie uvedených modulov:

- gateway
- backend
- frontend
- redirect

5.2 Implementácia nových nástrojov

Na základe požiadavok výskumnej skupiny z Biofyzikálneho ústavu Akademie Věd ČR v Brně (IBP) budú implementované dva nové nástroje. Systém sa tak rozšíri o analýzy pre vyhľadávanie Z-DNA štruktúr a CpG ostrovčekov. Pri danom množstve analýz bude konkurecie schopný s ostatnými analyzátormi DNA.

Nový nástroj *CpX Hunter* bude určený na vyhľadávanie CpG ostrovčekov v sekvencii DNA. Bude využívať algoritmus navrhnutý v štúdii *Comprehensive analysis of CpG islands in human chromosomes 21 and 22* [40] s využitím metódy sliding window, pomocou ktorej sa vyhľadávajú maximálne možné veľkosti CpG ostrovčekov. Aplikovaný algoritmu zahŕňa aj krátke sekvencie, ktoré nesplňujú dané kritériá, ale nachádzajú sa medzi dvomi CpG ostrovčekmi. Pre rozšírenie daného algoritmu sa počíta aj s vyhľadávaním ostrovčekov mimo CpG. Jedná sa o CpT, CpA, CpC. Zabezpečí sa tým analyzovanie a hľadanie súvislostí v sekvencii DNA aj mimo striktne definovaných CpG ostrovčekov.

Ďalším pridaným nástrojom bude *Z-DNA Hunter*. Daná analýza bude slúžiť na vyhľadávanie štruktúr Z-DNA a môže pomôcť výskumným pracovníkom pochopiť, kde a s akou pravdepodobnosťou sa bude formovať. Analýza bude využívať pozmenený algoritmus navrhnutý v článku *Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools* [36]. Zmeny budú spočívať v oprave chýb spomenutých v sekcii 3.5. Taktiež bude algoritmus rozšírený o variabilný bodový systém daných nukleotidov a výpočtu percentuálneho skóre, ktoré bude predstavovať pravdepodobnosť formovania Z-DNA konformácie.

5.3 Oprava CI/CD

Existujúce CI/CD v pôvodnom repozitári sú nefunkčné minimálne od septembra 2023. Príčinou je zle nastavený proces a zneaktívnenie potrebných funkcionalít pre vykonávanie všetkých úloh. Extrakciou modulov a prepísaním skriptu pre CI/CD sa znovu aktivuje proces automatického zabalenia a nasadenia aplikácie na testovací a produkčný server. Nefunkčnosť CI/CD je spôsobená aj tým, že virtuálny server GitLab Runner je v súčastnosti neaktívny. Na opravu je potrebné vytvoriť a nakonfigurovať nový virtuálny stroj, na ktorom budú prebiehať automatické pipeline-ny.

CI/CD bude obsahovať úlohy pre automatizované testovanie modulu, vytvorenie Docker obrazov, zverejnenie obrazov a nasadenie, resp. stiahnutie obrazov na testovacom a produkčnom serveri.

5.4 Dockerizácia FE

V súlade so zmenou na plne modulárnu architektúru bude vytvorený Docker obraz aj pre frontend. Daný krok zabezpečí, že frontend bude môcť byť nasadený nezávisle od ostatných častí systému, čo prispeje k vyššej flexibilite a efektívnosti správy aplikácie.

Docker obraz frontendu bude vytvorený pomocou kofigurácie *Dockerfile* súboru, ktorý bude definovať všetky potrebné závislosti, nastavenia a konfiguračné súbory potrebné pre jeho beh. Dockerfile poskytne všetky inštrukcie nutné pre vytvorenie štandardizovaného prostredia, čo zahŕňa aj webový server Nginx, ktorý servuje statické súbory a spracováva klientské požiadavky. Následne bude Docker obraz zverejnený vo verejnom registre https://hub.docker.com/ podobne ako všetky ostatné moduly.

5.5 Odstránenie nepotrebných modulov

Na zvýšenie efektívnosti vývoja a nasadenia budú odstránené moduly gateway a redirect. Keďže prácu týchto modulov dokáže nahradiť služba traefik v Docker Compose, nutnosť modulov pre smerovanie požiadaviek na správny modul odpadá. Traefik[1] je služba, ktorá je vytvorená za účelom smerovanie požiadaviek v prostredí Docker. Dokáže presmerovávať nezabezpečené spojenie na šifrované a uľahčuje prácu s SSL certifikátom Let's

[1]https://traefik.io/traefik/

Encrypt[2]. Odstránením daných modulov odpadá nutnosť udržiavania kódu pre smerovanie a tým sa uľahčí vývoj aplikácie DNA Analyser. Traefik taktiež pridá schopnosť load balancing pre moduly.

5.6 Aktualizácia Docker Compose

Zmenou štruktúry aplikácie a pridania nových služieb vzniká požiadavka na zmenu zavádzacieho postupu v Docker Compose súbore. Do súboru bude nevyhnutné definovať novú službu pre FE aplikácie a nastaviť proxy smerovanie aby mohli FE, BE a databáza medzi sebou komunikovať.

5.7 Výber technológií

Ako bolo uvedené v sekcii číslo 4, aplikácia využíva technológie ako Vue.Js pre FE a Java s frameworkom Spring pre BE. Pre uchovanie konzistencie budú nové analýzy implementované v rovnakých programovacích jazykoch a frameworkoch.

5.8 Testovanie

Pri implementácii nových analýz je nutné vytvoriť testy na overenie funkčnosti algoritmov. Dané testy budú implementované po vzore existujúceho systému pre Z-DNA a CpG analýzy.

Okrem automatizovaných testov bude vyžadované aj akceptačné testovanie na testovacom serveri. Testovací server je spustený na url adrese mendelu.cz[3]. Na danom serveri sa budú testovať nové analýzy, aktualizovaný Docker Compose ale aj správne smerovanie novej služby Traefik.

5.9 Nasadenie

Nasadenie novej verzie aplikácie na produkčný server bude vyžadovať zásah do aktuálne bežiacej verzie. Vďaka architektúre aplikácie budú zachované dáta sekvencií a výsledky analýz pre stávajúcich používateľov. Nasadenie bude vyžadovať aktualizovaný Docker Compose a správne nastavenie ciest ku zväzkom s aktuálnymi dátami.

[2]https://letsencrypt.org/
[3]https://bioinformatika.pef.mendelu.cz/

6 Implementácia

Táto kapitola sa zaoberá technickými detailmi implementácie projektu, kde podrobne rozoberáme kroky, ktoré boli vykonané na rozdelenie pôvodného monolitického repozitára a optimalizáciu jednotlivých modulov pre lepšiu modularitu a škálovateľnosť. Diskutuje sa tu o odstránení a osamostatnení modulov, ako aj o zavedení robustných CI/CD procesov pre backend a frontend, ktoré sú nevyhnutné pre efektívny vývoj a udržateľnosť aplikácie. Okrem toho sa kapitola venuje aj implementácii špecifických funkčných modulov ako Z-DNA Hunter a CpX Hunter, ktoré zlepšujú analytické schopnosti systému. Vysvetľuje sa tiež proces dockerizácie a nasadenia týchto modulov, čo umožňuje ich jednoduchšie a rýchlejšie nasadzovanie v produkčných prostrediach.

6.1 Odstránenie a osamostatnenie modulov

V rámci reštrukturalizácie aplikácie sa rozdelil pôvodný monolitický repozitár, ktorý zahŕňal backend, frontend a gateway. Toto rozdelenie bolo realizované s cieľom zlepšiť modularitu, udržateľnosť a škálovateľnosť systému.

BE a FE aplikácie boli vyňaté a oddelené do individuálnych repozitárov. Daný krok umožnil nezávislé spravovanie zdrojových kódov, čo vedie k efektívnejšiemu vývojovému cyklu a k lepšej integrácii cielených nástrojov pre každú časť aplikácie. Osamostatnením jednotlivých modulov sa taktiež zjednodušil proces CI/CD, čím sa znižuje riziko konfliktov v kóde a urýchľuje vývoj.

V súvislosti so zjednodušením procesu vývoja sa odstránil modul gateway. Pôvodný modul sa nahradil modernejšou technológiou Traefik. Odstránením modulu sa znížila potreba udržiavať a opravovať prípadné chyby.

Repozitáre sa stále nachádzajú v rámci pracovnej skupiny dna-analyser na adresách:

- Backend: https://git.pef.mendelu.cz/bioinformatics/backend
- Frontend: https://git.pef.mendelu.cz/bioinformatics/frontend-vue-2

6.2 CI/CD

V pôvodnom repozitári nebol implementovaný funkčný CI/CD, hlavne z dôvodu nefunkčnosti GitLab Runner a nesprávnosti postupu zostavenia a zverejnenia modulov. V predkladanej práci došlo k implementácií nových nastavení CI/CD pre BE aj FE a k vytvoreniu nových GitLab Runners. Virtuálne stroje sú rozdelené podľa funkcionality a špecializácie:

- **BioRunner (rancher01)** - master
- **BioRunner (rancher02)** - worker
- **BioRunner (rancher03)** - worker
- **BioRunner (rancher04)** - worker, cache

Runner s tagom *cache* je určený na úlohy, ktoré vyžadujú zdieľanie dát. Medzi tieto úlohy patrí zostavenie aplikácie a zverejnenie do docker registry. Runner-y s označením *worker* sa používajú pre testovanie a analýzu kódu.

Rovnako boli vytvorené nové úložiská pre BE aj FE aplikácie v Docker Hub:

- Backend: https://hub.docker.com/r/dnapef/backend
- Frontend: https://hub.docker.com/r/dnapef/frontend

Backend CI

Pre BE sa zaviedli nové CI/CD kroky, ktoré zahŕňajú kompiláciu java kódu, spustenie testov, vytvorenie docker obrazu a následné publikovanie do nového úložiska Docker registry. Proces je rozdelený do viacerých fáz: build, verification, publish a cleanup. Takýmto spôsobom sa zabezpečilo, že každý nový kód prechádza automatickým testovaním a je zverejnený iba po úspešnom overení funkčnosti. Na konci každej pipeline sa vykoná fáza cleanup, ktorá vyčistí runner od dočasných súborov a nepotrebných docker obrazov, čím predchádza zaplneniu disku vo virtuálnom stroji.

Frontend CI

Samostatné kroky CI/CD pre FE sú rozdelené do fáz build, test a publish. Pri každom commite do hlavnej vetvy je obraz vytvorený a zavedený do Docker hubu, čím je vždy pripravený najnovší obraz na nasadenie.

6.3 Z-DNA Hunter

Implementácia modulu Z-DNA Hunter spočívala v navrhnutí pozmeneného algoritmu. Originál algoritmu využíval bodový systém, ktorý bol nemenný pre každý dinukleotid. Tieto hodnotenia sú zobrazené v tabuľke číslo 1.

Dinukleotidy	Skóre
GC, CG	25
GT, TG, CA, AC	3
Ostatné	-

Tabuľka 1: Bodové hodnotenie dinukleotidov podľa algoritmu non-B_gfa [38].

Systém počítania skóre spočíva v princípe, že sa pre každý nukleotid vypočíta skóre na základe toho, aký dinukleotid tvorí v rámci sekvencie DNA. Príklad výpočtu skóre pre alternujúce GC:

```
 G  C  G  C  G  C  G  C  G  C
 |  |  |  |  |  |  |  |  |
+25 |  |  |  |  |  |  |  |
   +25 |  |  |  |  |  |  |
      +25 |  |  |  |  |  |
         +25 |  |  |  |  |
            +25 |  |  |  |
               +25 |  |  |
                  +25 |  |
                     +25 |
                        +25 = 225 / 2 = 112.5
```

Podľa daného algoritmu sa počíta skóre iba pre sekvencie, ktoré spĺňajú minimálnu dĺžku a neobsahujú diklueotidy iné ako uvedené v tabuľke číslo 1. Nevýhodou tejto analýzy je nemožnosť porozumenia výsledkom, keďže sa skóre neudáva vzhľadom na dĺžku analyzovanej sekvencie, resp. nájdenej pravdepodobnej Z-DNA štruktúry. Ako možno vidieť v tabuľke číslo 2, bodové hodnotenie dvoch sekvencií pozostávajúcich z rovnakých dinukleotidov, ale s rozdielnou dĺžkou, má rozdielne skóre.

Sekvencia	Dĺžka	Skóre
$(GC)^5$	10	112.5
$(GC)^6$	12	137.5
$(GT)^{50}$	100	148.5
$(GC)^5 + (GT)^5$	20	138.5

Tabuľka 2: Porovnanie bodového skóre štyroch sekvencií podľa algoritmu non-B_gfa [38].

Daná skutočnosť môže vo výskumníkovi evokovať, že dlhšia sekvencia má väčšiu pravdepodobnosť tvorby Z-DNA konformácie. Táto úvaha nie je správna, keďže alternujúca CG sekvencia má vysokú pravdepodobnosť formovania Z-DNA bez ohľadu na dĺžku daného výseku sekvencie [75]. Druhým prípadom je stav, kedy sa objaví alternujúci GT dinukleotid s dĺžkou 100. Pri počítaní dostávame najvyššie skóre, aj keď daná sekvencia má skoro 10 krát menšiu pravdepodobnosť sformovania Z-DNA konformácie. To vytvára potrebu normalizovania skóre na základe dĺžky skúmaného okna, kedy budú referenčnými hodnotami práve najvyššie zadané hodnoty pre dinukleotid.

Na základe pripomienok z IBP bolo nutné zmeniť výpočtový model a to z dôvodu, že v súčasnej dobe prebiehajú výskumy zameriavajúce sa na oblasti v genóme formujúce Z-DNA štruktúru, ktoré nie sú z veľkej časti tvorené GC opakovaniami. Z daného dôvodu bol pozmenený bodovací systém tak, aby umožnil výskumníkom zvoliť si pravdepodobnosti pre dinukleotidy:

- GC, CG => $score_{GC}$
- GT, TG, AC, CA => $score_{GT\_AC}$

- AT, TA => $score_{AT}$
- AA, AG, TT, TC, GA, GG, CT, CC => nemožné formovať Z-DNA

Pridaním univerzálnych premenných pre dinukleotidy do algoritmu dostaneme výpočtový model bodového skóre, ktorý môžu výskumníci využívať pri experimentovaní s rôznymi parametrami. Navrhnutý algoritmus na detekciu miest v sekvencii DNA s pravdepodobnosťou formovanie Z-DNA štruktúry je možné vidieť v algoritme číslo 1. Popísaný postup zaisťuje využívanie premenných na hodnotenie dinukleotidov a zároveň počíta nie len so skóre ako v pôvodnom algoritme, ale tiež s percentuálnou hodnotou, ktorá odráža skutočný stav nájdeného okna sekvencie DNA. Vytvorený algoritmus dopĺňa aj atribút *threshold*, vďaka ktorému sa vyfiltrujú výsledky pod touto hranicou.

Algoritmus 1: *Procedúra pre nájdenie Z-DNA miesta s percentuálnym skóre.*

```
1:  procedure GetResults(sequence, minSize, threshold, gc, gtac, at)
2:      Initialize empty list 'allResults'
3:      i ← 0
4:      score ← 0
5:      len ← 1
6:      while i < sequence.size do
7:          scorePair ← getSubScore(sequence[i], sequence[i+1])
8:          if scorePair > 0 then
9:              len ← len + 1
10:             score ← score + scorePair
11:         else
12:             if len ≥ minSize then
13:                 kvScore ← score / 2
14:                 maxScore ← (len-1) * max(gc, gtac, at) / 2
15:                 scorePerc ← (kvScore / maxScore) * 100
16:                 if scorePerc ≥ threshold then
17:                     start ← i - len + 2
18:                     seq ← subsequence(sequence, start - 1, len)
19:                     Add result to 'allResults'
20:                 end if
21:             end if
22:             Reset score and len to 0 and 1, respectively
23:         end if
24:         i ← i + 1
25:     end while
        return allResults
26: end procedure
```

Vďaka navrhnutému algoritmu sa dostávame do optimálnejších číselných hodnôt, čo poskytne výskumníkovi objektívnejšie informácie ohľadom pravdepodobnosti, že nájdená

subsekvencia môže formovať Z-DNA. V tabuľke číslo 3 sú uvedené výsledky podľa upraveného algoritmu. Na výpočet boli použité rovnaké parametre ako v originálnom algoritme uvedené v tabuľke číslo 1. Pomocou percentuálneho skóre môže výskumník jednoduchšie určiť, či má pre neho subsekvenčné okno význam na ďalšie skúmanie.

Sekvencia	Dĺžka	Skóre	Maximálne skóre	Skóre [%]
$(GC)^5$	10	112.5	112.5	100 %
$(GC)^6$	12	137.5	137.5	100 %
$(GT)^{50}$	100	148.5	1237.5	12 %
$(GC)^5+(GT)^5$	20	138.5	237.5	58.32 %

Tabuľka 3: Porovnanie bodového skóre štyroch sekvencií podľa navrhnutého algoritmu.

Pre uchovanie informácií ohľadom nájdeného okna sekvencie boli pridané parametre, ktoré zobrazujú percentuálny podiel záujmových dinukleotidov G/C a G/T. Výpočet sa vykonáva po nájdení výsledku a prebieha podľa algoritmu číslo 2.

Algoritmus 2: *Procedúra na počítanie obsahu dinukleotidu v sekvencii.*

```
1: procedure CalculateRichness(seq, len, dinucleotide)
2:     pattern ← dinucleotide or reverse(dinucleotide)
3:     count ← count matches of pattern in seq
4:     richness ← (count / len) * 100
5:     return richness
6: end procedure
```

6.3.1 Implementácia BE

BE analýzy bol implementovaný do existujúceho kódu DNA Analyser aplikácie. Ako uvádza sekcia číslo 4, BE systému využíva jazyk Java a framework Spring.

Ako prvé bolo nutné vytvoriť triedu, ktorá bude reprezentovať analýzu. Táto trieda rozširuje existujúcu triedu Analyse a definuje parametre analýzy s prednastavenými hodnotami. S využitím anotácií sa automaticky budú dáta pridávať do databázy. Vytvorenie tabuľky je zabezpečené pomocou migrácie Flyway.

Implementácia si vyžadovala vytvoriť endpointy, na ktorých aplikácia prijíma požiadavky. Endpointy kopírujú štruktúru ostatných analýz a pridávajú podporu pre základnú komunikáciu pomocou REST API. Pre Z-DNA Hunter boli vytvorené endpointy:

- $GET/api/analyse/zdna$: Získajú sa spustené a dokončené analýzy.
- $POST/api/analyse/zdna$: Spustenie analýzy. Vyžadujú sa parametre v tele požiadavky vo formáte ako je zobrazené v kóde číslo 2. Daný request skontroluje parametre a spustí analýzu.

```
{
  "minSequenceSize": 10,
  "score_at": 0.5,
  "score_gc": 2,
  "score_gtac": 1,
  "sequence": "3fa85f64-5717-4562-b3fc-2c963f66afa6",
  "tags": [
    "string"
  ],
  "threshold": 50
}
```

Kód 2: Telo požiadavky pre spustenie analýzy Z-DNA Hunter.

- $GET/api/analyse/zdna/\{id\}/analysis$: Vráti detaily analýzy podľa zadaného ID.
- $GET/api/analyse/zdna/\{id\}/zdnas$: Vráti nájdené možné Z-DNA konformácie v sekvencii.
- Ostatné: Taktiež boli vytvorené endpointy pre ostatné funkcie na zjednotenie novej analýzy v aplikácií. Medzi tieto funkcie patrí získanie heatmapy, tagov, exportovanie výsledkov do csv a bedGraph a ostatné.

Celkovo bolo implementovaných 12 koncových bodov potrebných na spustenie, vymazanie a získanie údajov analýzy a výsledkov. Obrázok 15 obsahuje výpis všetkých vytvorených endpointov pre Z-DNA Hunter.

Výsledky analýzy sú ukladané do PostgreSQL databázy. Obsahujú ID, pozíciu začiatku, dĺžku okna sekvencie, obsah GT a GC, bodové skóre a percentuálne skóre. Automaticky sa ukladá stredná pozícia výsledku.

V rámci vypracovania analýzy boli vytvorené testy v jazyku Groovy s niekoľkými testovacími sekvenciami. Spúšťajú sa pri každom zostavení aplikácie v rámci CI/CD.

Obrázok 15: Endpointy analýzy Z-DNA Hunter s popisom v aplikácii DNA Analyser.

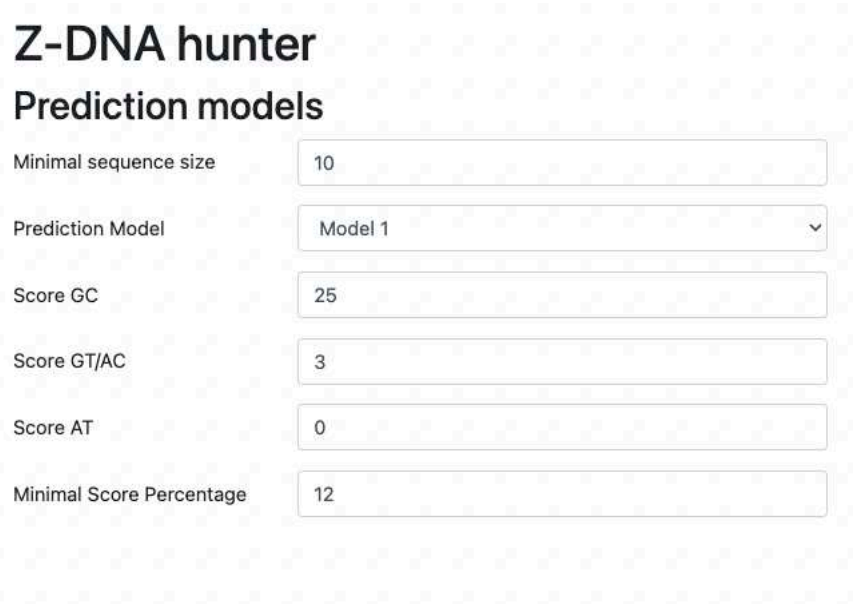
6.3.2 Implementácie FE

Implementácia FE pre Z-DNA Hunter vyžadovala vytvorenie niekoľkých stránok a komponentov. Ako hlavná šablóna bola využitá rovnaká schéma, ako má každá analýza. To znamená, že na pravej strane sa nachádza zoznam dostupných sekvencií v aplikácii, na ľavej strane je umiestnený formulár pre danú analýzu. V spodnej časti sa otvárajú dokončené analýzy a ich výsledky.

Formulár pre analýzu sa skladá z parametrov, ktoré vyžaduje BE. Používateľ má dostupné bodové hodnotenia nukleotidov, minimálnu dĺžku okna a minimálne percentuálne skóre, podľa ktorého sa budú filtrovať výsledky. Pre zjednodušenie práce používateľa boli implementované dva predvolené modely. Nastavenia predikčných modelov sú dostupné v tabuľke číslo 4. V rámci formuláru boli implementované aj varovania a chybové hlášky, ktoré používateľa upozorňujú na nepovolené hodnoty alebo na hodnoty mimo doporučený rozsah. Poradie a dizajn parametrov formuláru analýzy je možné vidieť na obrázku 16.

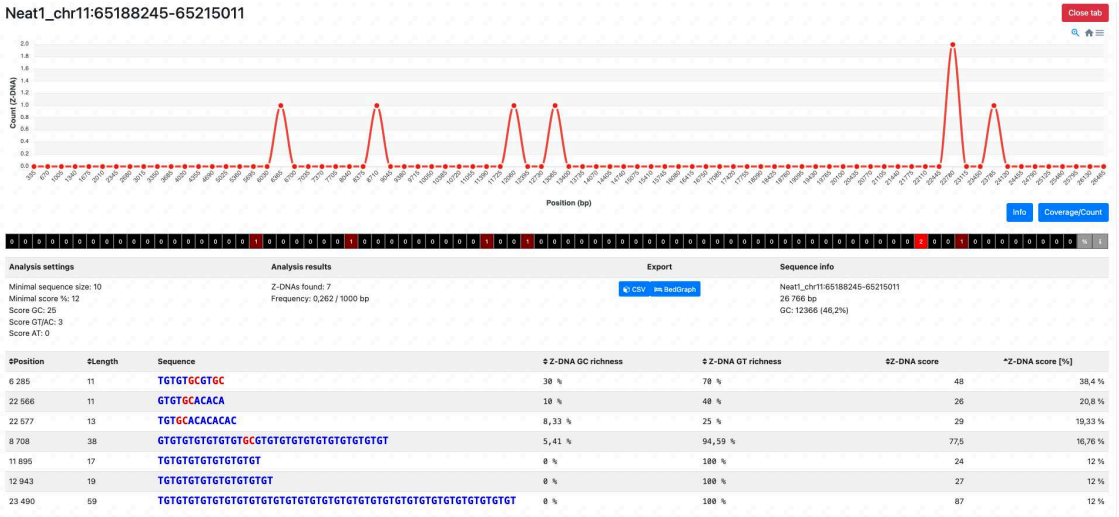
Parameter	Model 1	Model 2
Score GC	25	2
Score GT/AC	3	1
Score AT	0	0.5
Minimal Score Percentage	12	50

Tabuľka 4: Predvolené hodnoty parametrov pre predikčný model 1 a 2.



Obrázok 16: Formulár analýzy Z-DNA Hunter v aplikácii DNA Analyser.

Pri implementácii komponentu, ktorý sa využíva na zobrazenie výsledkov bola použitá stávajúca šablóna z ostatných analýz. Používa sa komponent na zobrazenie čiarového grafu a heatmapy pre grafické znázornenie výsledkov. Pod grafickým znázornením sa nachádzajú informácie ohľadom analýzy, výsledkoch a sekvencie. Jednotlivé výsledky sú vizualizované v tabuľke, kde sú zobrazené dáta nájdených sekvenčných okien. Bola upravená farebná schéma sekvencie vo výsledkovej tabuľke aby odpovedala najzaujímavejším častiam v Z-DNA. Dinukleotidy sa farebne odlišujú. Červenou farbou je vyobrazený dinukleotid GC/CG a všetky ostatné dinuleotidy sú modrej farby. Zabezpečí sa tým rýchle rozpoznanie významných častí. Na obrázku 17 je možné vidieť výsledky z analýzy s predvolenými hodnotami Modelu 1.



Obrázok 17: Výsledky analýzy Z-DNA Hunter pre gén Neat1 na chromozóme 11.

6.4 CpX Hunter

CpG ostrovčeky obsahujú nadmerné zastúpenie dinukleotidov CG, ktoré sú považované za funkčne dôležité časti genómu. Definujú a regulujú promotérové oblasti rôznych genómov. Regulačná funkcia bola naznačená aj u iných dinukleotidov, medzi ktoré patria CT, CA a CC. Kvôli uvedeným skutočnostiam sa v rámci predkladanej práce rozšíril pôvodný algoritmus navrhnutý v článku *Comprehensive analysis of CpG islands in human chromosomes 21 and 22* [40]. Upravený algoritmus berie do úvahy všetky možné kombinácie a preto sa nástroj nazýva CpX Hunter. CpX ostrovčeky sú definované ako oblasti s vysokou frekvenciou dinukleotidov CX v genómovej DNA, kde ”X” predstavuje akýkoľvek nukleotid.

Algoritmus pracuje s využitím posunu okna a spájaním menších ostrovčekov vo viacerých krokoch. Začína sa nastavením vyhľadávacieho okna o veľkosti 200 báz na začiatku sekvencie. V okne sa následne vypočíta percentuálny podiel G a C nukleotidov a pomer pozorovaných a očakávaných CG dinukleotidov pre danú sekvenciu. Okno sa posúva po 1 bázovom páre vpred, až kým sekvencia nespĺňa kritériá pre CpG ostrovček. V momente, kedy sa nájde dané okno, posúva sa o ďalších 200 báz a proces sa opakuje. Tento proces posunu o 200 báz sa iteruje, pokiaľ okno nenapĺňa definované kritériá. Ak posledné okno neodpovedá kritériám, posúva sa o 1 bázový pár späť, dokým nie sú kritéria splnené. Okrem vyhľadávania CpG ostrovčekov algoritmus taktiež identifikuje a spája ostrovčeky, ktoré sú od seba vzdialené o menej ako 100 nukleotidov, čím vytvára rozsiahlejšie regulačné oblasti. Grafické znázornenie algoritmu je zobrazené na obrázku 18. Parametre, ktoré musí sekvenčné okno dodržať aby sa jednalo o CpG ostrovček sú:

- Veľkosť okna: 200 nukleotidov
- Obsah GC (GC): 50 %

- Pomer pozorovaných CpG a očakávaných CpG(Obs_{CpG}/Exp_{CpG}): 60 %

Vzorce pre výpočet obsahu GC:

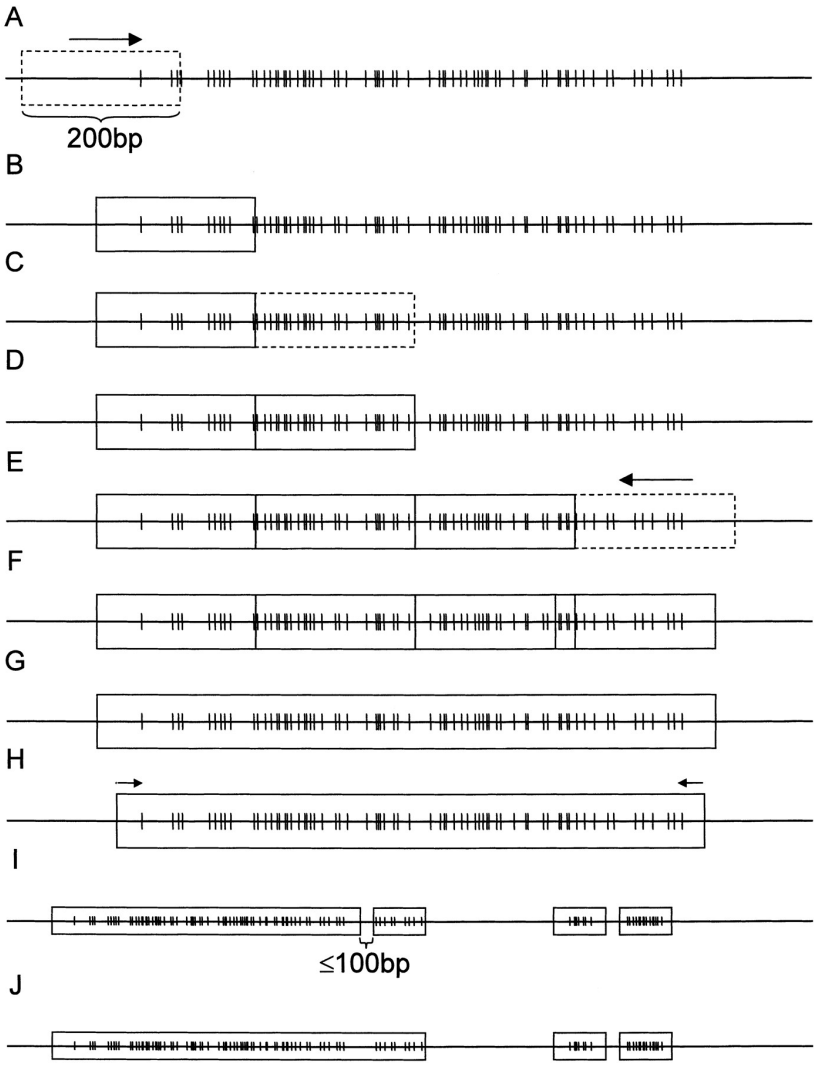
$$GC = (Count_G + Count_C)/Length_{seq} \tag{1}$$

Vzorec pre výpočet pozorovaných CpG:

$$Obs_{CpG} = Count_{CpG}/Length_{seq} \tag{2}$$

Vzorec pre výpočet očakávaných CpG [76]:

$$Exp_{CpG} = (GC/2)^2 \tag{3}$$



Obrázok 18: Postup vyhľadávania CpG ostrovčekov v sekvencii [40].

Pri implementácií bol originálny algoritmus upravený tak, aby bolo možné variabilné zadávanie parametrov, čím sa rozšírili možnosti experimentovania. Navrhnutý postup je zobrazený v algoritme číslo 3. Algoritmus vyhľadá CpX ostrovčeky v sekvencii. Podobne ako originálny algoritmus začína na prvej pozícii sekvencie a vytvára sa okno s dĺžkou potrebnou pre minimálnu veľkosť ostrovčeka. Následným posúvaním, skracovaním a rozširovaním sa hľadá čo najväčšie možné okno, ktoré odpovedá zadefinovaným parametrom. Na záver algoritmu sa využíva procedúra na spájanie ostrovčekov. Každé vytvorené okno pri každej zmene prepočítava informácie ohľadom pozorovaných CpX a očakávaných CpX.

Algoritmus 3: *Procedúra na nájdenie CpX ostrovčekov*

```
1:  procedure FindIslands(record)
2:      found_islands ← empty list
3:      record_position ← 0
4:      while True do
5:          window ← CpXWindow (record, record_position, dinuckleotid)
6:          island_continues ← find_island_shifts(window)
7:          if not island_continues then
8:              break
9:          end if
10:         extend_island_window_shifts(window)
11:         rollback_until_island(window)
12:         if not shrink(window) then
13:             record_position ← window.window_begin + 1
14:             continue
15:         end if
16:         found_islands.append(window)
17:         record_position ← window.window_end
18:     end while
19:     if found_islands is empty then
20:         return empty list
21:     else
22:         return merge_islands(found_islands)
23:     end if
24: end procedure
```

6.4.1 Implementácia BE

Rovnako ako predchádzajúci nástroj, CpX Hunter bol implementovaný v jazyku Java 11 a kopíruje štruktúru nástroja Z-DNA Hunter.

Boli vytvorené dve triedy, ktoré reprezentujú analýzu. Prvá hlavná trieda *CpgAnalyzer* obsahuje štruktúru a logiku pre obsluhu sekvencie, hľadania a manažovanie sekvenčných

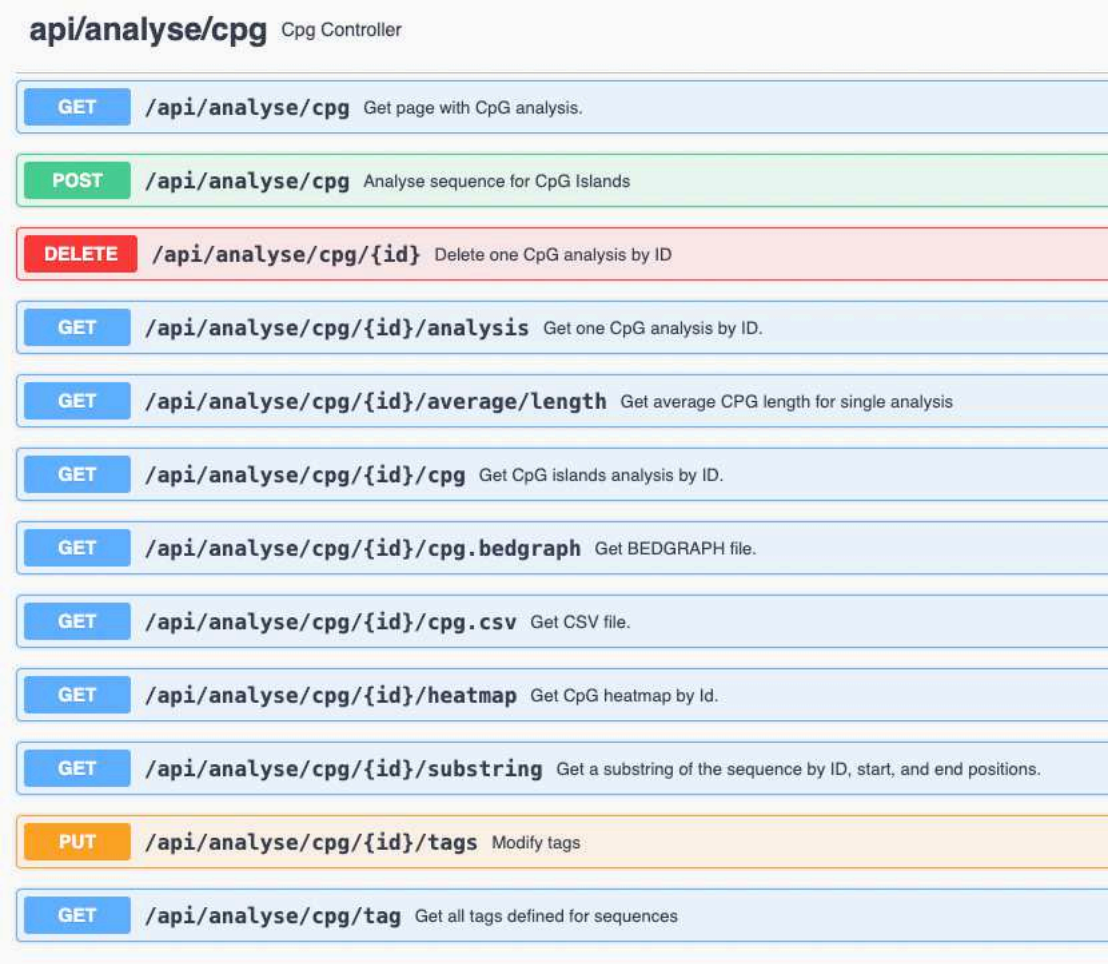
okien. Nachádza sa tu implementovaný algoritmus podľa algoritmu číslo 3 aj s jeho obslužnými funkciami. Druhou vytvorenou triedou analýzy bolo *CpgWindow*. Obsahuje metódy na výpočet a získanie informácií ohľadom sekvenčného okna. Rovnako obsahuje aj metódy zmenšovania a zväčšovania okna z pravej aj z ľavej strany.

Výsledky sa ukladajú odlišne od nástroja Z-DNA Hunter. Výsledkové okná môžu dosahovať veľkých dĺžok, preto bolo potrebné efektívne zabezpečiť, aby sa výsledky neukladali do databázy s celým sekvenčným oknom. Do výsledkovej tabuľky v databáze sa ukladá iba snippet sekvenčného okna o veľkosti 20 nukleotidov. Ak používateľ potrebuje získať celé sekvenčné okno, môže použiť implementovanú funkciu na získanie subsekvencie na základe pozície v danom sekvenčnom súbore.

Celkovo bolo implementovaných 12 endpointov pre obsluhu cez REST API. Endpointy sú rovnaké ako pri nástroji Z-DNA Hunter až na malé odlišnosti. Telo požiadavky pre spustenie sekvencie obsahuje obsahuje parametre potrebné pre analýzu. Parametre sú uvedené v kóde číslo 3. Zoznam všetkých implementovaných endpointov s ich popisom je zobrazený na obrázku 19.

```
{
  "firstNucleotide": "C",
  "minGcPercentage": 0.5,
  "minIslandMergeGap": 100,
  "minObservedToExpectedCpG": 0.6,
  "minWindowSize": 200,
  "secondNucleotide": "G",
  "sequence": "3fa85f64-5717-4562-b3fc-2c963f66afa6",
  "tags": [
    "string"
  ]
}
```

Kód 3: Telo požiadavky pre spustenie analýzy CpX Hunter.



Obrázok 19: Endpointy analýzy CpX Hunter s popisom v aplikácii DNA Analyser.

6.4.2 Implementácia FE

Implementácia prebiehala podobne ako v novej analýze Z-DNA Hunter. Formulár pre analýzu pozostáva z parametrov, ktoré sú vyžadované na komunikáciu cez REST API. Pre lepšiu používateľskú skúsenosť boli nastavené predvolené hodnoty, ktoré odpovedajú odporúčaniam z originálneho algoritmu. Keďže analýza vyhľadáva iba ostrovčeky pozostávajúce z dinukleotidov CX, parameter s názvom *First Nucleotide* je zablokovaný na hodnote ”C”. Parametre aj s ich predvolenými hodnotami sú dostupné v tabuľke číslo 5.

Parameter	Predvolená hodnota
Minimal window size	200
Minimal CX percentage	0.5
Minimal observed/expected CpX	0.6
Maximal island merge gap	100
First Nucleotide	C - disabled
Second Nucleotide	G

Tabuľka 5: Predvolené hodnoty parametrov na analýzu CpX Hunter.

Pri vizualizácií výsledkov sa používajú podobné časti ako v Z-DNA Hunter. Keďže sa do výsledkov ukladá iba krátka časť sekvenčného okna, bolo pridané funkčné tlačidlo lupy, pomocou ktorého si dokáže používateľ prezrieť celé odhalené sekvenčné okno. Na získanie celého výsledku sa používa implementovaný endpoint */cpg/id/substring* s parametrami *start* a *end*. Na obrázku číslo 20 je možné vidieť výsledky analýzy s predvolenými hodnotami, kde červená šípka značí odkrytie celého sekvenčného okna.

Obrázok 20: Využitie funkčného tlačidla lupy za účelom odkrytia celého sekvenčného okna analýzy CpX Hunter pre gén Neat1 na chromozóme 11.

6.5 Dockerizácia modulov

BE

Keďže sa aplikácia rozdelila z jedného repozitára do samostatných, bolo nutné zmeniť a vytvoriť Dockerfile. V pôvodnom repozitári BE aplikácie obsahoval nefunkčný Dockerfile, ktorý pracoval s nesprávnou verziou Java JRE 8 a nepoužíval výhody build systému založeného na Gradle.

V rámci novej implementácie sa vytvoril aktualizovaný konfiguračný súbor kontajnera. Použila sa správna verzia Java, JDK 11. Rovnako sa zaviedla dvojstupňová výstavba (multi-stage build), čo výrazne redukuje veľkosť konečného obrazu tým, že oddeľuje prostredie potrebné pre kompilovanie kódu a prostredia potrebného na spustenie modulu. V prvom kroku sa aplikácia zostaví vrátane všetkých jej závislostí, zatiaľ čo v druhom kroku sa z výstupu nakopírujú iba nevyhnutné súbory do finálneho obrazu.

Ďalej sa využili príkazy COPY pre efektívnejšie spracovanie závislostí a zdrojových kódov. Rozdelením modulu na jej logické časti (použité knižnice, metadáta a triedy) sa zjednodušuje proces nasadzovania a zároveň sa znižuje riziko chýb pri konfigurácii prostredia. Pridaním príkazov VOLUME a EXPOSE sa poskytujú potrebné prostriedky a informácie, ako Docker kontajner komunikuje s vonkajším svetom.

FE

Pre FE aplikácie sa implementoval Dockerfile, ktorý zabezpečuje efektívne a bezpečné nasadenie v produkčnom prostredí. V pôvodnom repozitári nebol Docker pre FE využívaný. Zavedením dockerizácie odpadá nutnosť manuálne zostavovať statické súbory a vkladať ich ručne na produkčný server. Nový implementovaný prístup umožňuje automatizované a konzistentné vytváranie obrazov aplikácie.

V prvej etape vytvárania obrazu sa využíva oficiálny obraz Node.js vo verzii 14. Tento obraz je využitý ako základ, na ktorom sa zostavia statické súbory. Inštalovanie potrebných závislostí prebieha nakopírovaním súborov `package.json` a `package-lock.json` a spustením príkazu `npm ci`. Daný prístup zabezpečí konzistentnosť medzi vývojovým a produkčným prostredím a tiež odpadá nutnosť inštalovania na lokálnom stroji. Posledný príkaz v danej fáze zostaví statické súbory aplikácie.

Druhá fáza (production stage) využíva obraz Nginx, ktorý je optimalizovaný pre nasadenie webových aplikácií. Statické súbory vytvorené v predchádzajúcej etape sa skopírujú na Nginx serveri. Daná konfigurácia zabezpečí, že po spustení kontajnera bude FE aplikácie dostupný na porte 80.

Implementovaná štruktúra Dockerfile poskytuje robustný, bezpečný a ľahko replikovateľný spôsob, ako distribuovať a spravovať FE aplikácie.

6.6 Nasadenie

Aby bolo nasadenie konzistentné medzi všetkými prostrediami (testovací a produkčný server) bol vytvorený zavádzací Docker Compose skript pre Docker Swarm. Súbor compose.yml spúšťa služby získané z registru Docker Hub. Medzi tieto služby patrí:

- **treafik:** traefik:v2.4
- **backend:** dnapef/backend:$TAG:-latest
- **frontend:** dnapef/frontend:$TAG:-latest
- **postgres:** postgres:13

Každá zo služieb má nastavené vlastné environmentálne premenné, ktoré sú potrebné na správne zavedenie služieb. Každá službu si vyžadovala nastaviť obslužné pravidlá reverzného proxy Traefik, aby dokázali BE a FE medzi sebou komunikovať. To bolo dosiahnuté nastavením url adries a pravidiel pre smerovanie požiadaviek.

- **backend:** Host(‘domain‘) && PathPrefix(‘/api/‘)
- **frontend:** Host(‘domain‘) && PathPrefix(‘/‘)

Rozdiel medzi zavádzacím skriptom pre testovací a produkčný server spočíva iba v rozdielnom zadefinovaní Host domény. Služby boli vytvorené a nasadené na daných umiestneniach:

- **Testovací server:** https://bioinformatika.pef.mendelu.cz
- **Produkčný server:** https://bioinformatics.ibp.cz

7 Diskusia

Pri prehliadaní dostupnej literatúry a publikovaných článkov bolo identifikovaných niekoľko prístupov k vyhľadávaniu Z-DNA a CpG štruktúr. V predkladanej práci sa implementovali vyhľadávacie moduly pre webovú aplikáciu, ktoré integrujú najpoužívanejšie algoritmy s viacerými úpravami a vylepšeniami.

Prístup používaný v nástroji Z-DNA Motifs v rámci platformy non-B DNA Motif Search Tool, popísaný v sekcii číslo 3.5, obsahuje nedostatky, ktoré táto práca odstránila. Medzi spomínané nedostatky patrí nemožnosť upraviť parametre analýzy a nefunkčnosť vyhľadania Z-DNA konformácii na začiatku a na konci sekvencie. Aplikácia non-B DNA síce ponúka možnosť skenovania väčšej sekvencii, no aktualizovaná aplikácia DNA Analyser vyniká v použiteľnosti, rýchlosti, správnosti a v lepšej interpretovateľnosti výsledkov. V tabuľke číslo 6 je uvedené porovnanie aplikácií. Ako si je možné všimnúť, oba nástroja ponúkajú mnoho analýz a výhod. Ak sa zameriame na Z-DNA, tak konkurenčný nástroj nedokáže nájsť všetky oblasti, ktoré môžu konformovať Z-DNA. Na porovnanie a overenie nájdených výsledkov, tak odpovedajú všetky okrem posledného okna v sekvencii.

	DNA Analyser	non-B DNA Motif
Dostupné analýzy	7	7
Vstup	Plain Fasta NCBI Batch NCBI	Plain Fasta
Veľkosť vstupu	2048 MiB	20 MiB
Výstup	Interaktívny graf Interaktívna tabuľka CSV bedGraph	GFF GPH PNG TSV
Z-DNA analýza na kóde 5	7	6
Napojiteľné na API	Áno	Nie
Spustenie viacerých analýz naraz	Áno	Nie

Tabuľka 6: Porovnanie funkčnosti nástrojov pre Z-DNA Analýzu.

```
CGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGC
GCGCGCGCGCGCGCGCGCGCG GCGCGCGCGCGCG GCGCGCGCGCGCGCGCGCGCG GCGCGCGCG
CGCGCG GCGCGCGCGCGCGCG GCGCGCGCGCG GCGCGCGCGCG GCGCGC
```

Kód 5: Kód na testovanie Z-DNA analýzy s vyznačenými správnymi výsledkami podľa modelu 1.

Keďže Z-DNA štrutkúra a CpG ostrovček spolu súvisia, tak ďalšiu výhodu predstavuje implementovaná analýza CpX Hunter. Ako už bolo vyššie spomenuté, analýza využíva

upravený algoritmus od Takai a Jones [40]. Keďže pôvodný algoritmus už nie je dostupný, prebehlo testovanie oproti nástroju implementovaného v jazyku Python [46]. Podľa tabuľky číslo 7 konštatujeme, že nový implementovaný nástroj CpX Hunter je niekoľkonásobne rýchlejší[4] ako doteraz používaný skript so zachovaním správnych výsledkov. Ďalšou výhodou je výpis interaktívnych výsledkov, zobrazenie grafov a export z aplikácie. Taktiež dopĺňa jednotnú aplikáciu s množstvom analýz, takže výskumníkovi je uľahčená práca v rámci jedného ekosystému.

Sequence	**NC_000913.3**	**NC_060946.1**	**NC_060925.1**
Time [s]	4.655 (Python) 1.494 (CpX Hunter)	176.745 (Python) 2.063 (CpX Hunter)	889.505 (Python) 8.333 (CpX Hunter)
Results	1650 (Python) 1650 (CpX Hunter)	9478 (Python) 9478 (CpX Hunter)	26377 (Python) 26377 (CpX Hunter)

Tabuľka 7: Porovnanie rýchlosti a správnosti výsledkov medzi CpX Hunter z DNA Analysor a nástroju v pythone.

7.1 Ďalší vývoj

Na ďalší vývoj aplikácie by bolo vhodné zvážiť implementáciu ďalších nástrojov pre analýzu DNA, akými sú SNP (single nucleotide polymorphisms) a STR (short tandem repeats). Ďalej by bolo vhodné pridanie modulov pre pokročilé epigenetické analýzy, ktoré by zahŕňali štúdium metylácie DNA a histónových modifikácií.

Z pohľadu analýzy Z-DNA Hunter by bolo vhodné implementovať ďalší výpočtový model. Súčasné hodnotenie na základe percentuálneho skóre má vysoké množstvo false positives. Pridanie analýzy, ktorá využíva hlbokú neurálnu sieť (DNN) pre detekciu Z-DNA, by mohlo viesť k objasneniu správnosti výsledkov. Obsahovala by veľké množstvo false negatives, keďže sa dá predpokladať, že by sa učila iba na určitých typoch buniek [77]. Implementácia takéhoto typu analýzy by pre používateľa predstavovala výhodu v rýchlom porovnaní výsledkov.

[4]Testované na serveri: Intel Xeon Gold 6230, 80 cores, RAM: 92 GB

8 Záver

V predkladanej diplomovej práci došlo k rozšíreniu funkcionality webovej aplikácie DNA Analyser o nové analýzy zamerané na Z-DNA konformácie a CpX ostrovčeky. Vďaka tomu bolo umožnené efektívnejšie skúmanie lokálnych štruktúr DNA. Hlavným cieľom bolo vyvinúť a implementovať nástroje na identifikáciu špecifických sekvencií v rámci už existujúcej aplikácie a zjednodušiť tak použitie pre vedeckú komunitu.

V teoretickej časti práce sa predostrel prehľad dôležitosti Z-DNA a CpX ostrovčekov v biomedicínskom výskume a analyzovali sa dostupné prístupy na ich identifikáciu. Na základe toho sa navrhli a implementovali nové analytické nástroje, ktoré sa úspešne integrovali do aplikácie DNA Analyser. Tieto nástroje umožňujú presné vyhľadávanie a vizualizáciu týchto sekvencií, čím prinášajú používateľom dostupný nástroj na lepšie pochopenie ich biologických funkcií. Taktiež sa upravil spôsob fungovania z pohľadu nasadenia a zjednodušil sa budúci vývoj aplikácie použitím moderných prístupov.

Počas testovania sa overila správnosť a účinnosť nových metód na viacerých dostupných dátach, čo potvrdilo ich schopnosť identifikovať Z-DNA a CpX ostrovčeky v DNA sekvenciách.

V aktualizovanej aplikácii stále ostáva priestor na ďalšie vylepšenia. Medzi možnosti rozšírenia patrí integrácia nových nástrojov, ktoré identifikujú iné typy štruktúr, alebo prispôsobenie algoritmov na využívanie iných modelov výpočtu a hodnotenia. Taktiež by bolo vhodné zamerať sa na refaktoring kódu. Používateľské rozhranie by sa mohlo vylepšiť pre ešte intuitívnejšie a efektívnejšie prostredie pre analýzu.

V konečnom dôsledku táto práca prispieva k rozvoju bioinformatických nástrojov tým, že poskytuje užitočný zdroj pre výskumníkov, ktorí skúmajú a analyzujú lokálne štruktúry DNA. Rozšírená aplikácia DNA Analyser ponúka komplexnú sadu nástrojov.

Na záver možno konštatovať, že DNA je molekulárnym základom života a slúži ako nositeľ genetickej informácie, ktorá určuje vývoj, fungovanie a rozmanitosť života. Naše chápanie a manipulácia s DNA naďalej prináša revolúciu v biológii, medicíne a mnohých ďalších oblastiach a sľubuje vzrušujúce možnosti do budúcnosti.

9 Literatúra

[1] Richard J Roberts. *Nucleic acid | chemical compound*. Encyclopædia Britannica, feb. 2020. URL: https://www.britannica.com/science/nucleic-acid (cit. 13. 06. 2023).

[2] Steve Minchin a Julia Lodge. "Understanding biochemistry: structure and function of nucleic acids". In: *Essays in Biochemistry* 63 (okt. 2019), s. 433–456. DOI: 10.1042/EBC20180038. URL: https://pubmed.ncbi.nlm.nih.gov/31652314/ (cit. 13. 06. 2023).

[3] Yao Wan a Kunal Chatterjee. *RNA | definition, structure, types, & functions*. Encyclopædia Britannica, júl 2018. URL: https://www.britannica.com/science/RNA (cit. 13. 06. 2023).

[4] Daniel Holoch a Danesh Moazed. "RNA-mediated epigenetic regulation of gene expression". In: *Nature Reviews Genetics* 16.2 (jan. 2015), s. 71–84. DOI: 10.1038/nrg3863. URL: https://www.nature.com/articles/nrg3863 (cit. 13. 06. 2023).

[5] David Wang a Aisha Farhana. *Biochemistry, RNA structure*. PubMed, 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK558999/ (cit. 13. 06. 2023).

[6] Liguo Zhang et al. "Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues". In: *Proceedings of the National Academy of Sciences of the United States of America* 118 (máj 2021). DOI: 10.1073/pnas.2105968118. URL: https://pubmed.ncbi.nlm.nih.gov/33958444/ (cit. 13. 06. 2023).

[7] Susan Payne. "Introduction to RNA Viruses". In: *Viruses* (2017), s. 97–105. DOI: 10.1016/B978-0-12-803109-4.00010-6. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7173417/ (cit. 13. 06. 2023).

[8] Bruce Alberts, A Johnson a J Lewis. *Molecular Biology of the Cell*. 4. vyd. Garland, 2002. ISBN: 9780815332183.

[9] J. D. WATSON a F. H. C. CRICK. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171 (apr. 1953), s. 737–738. DOI: 10.1038/171737a0. URL: https://www.nature.com/articles/171737a0 (cit. 13. 06. 2023).

[10] Herve Seligmann. *DNA Replication*. BoD – Books on Demand, aug. 2011. ISBN: 9789533075938.

[11] Francis Crick. "Central Dogma of Molecular Biology". In: *Nature* 227 (aug. 1970), s. 561–563. DOI: 10.1038/227561a0. URL: https://www.nature.com/articles/227561a0 (cit. 14. 06. 2023).

[12] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431 (okt. 2004), s. 931–945. DOI: 10.1038/nature03001. URL: https://www.nature.com/articles/nature03001 (cit. 14. 06. 2023).

[13] Geoff Spencer a International Human Genome Sequencing Consortium. *International Human Genome Sequencing Consortium-Describes Finished Human Genome Sequence*. Genome.gov, 2013. URL: https://www.genome.gov/12513430/2004-release-ihgsc-describes-finished-human-sequence (cit. 14. 06. 2023).

[14] Jennifer A. Doudna a Emmanuelle Charpentier. "The new frontier of genome engineering with CRISPR-Cas9". In: *Science* 346 (nov. 2014), s. 1258096–1258096. DOI: 10.1126/science.1258096. URL: https://www.science.org/doi/10.1126/science.1258096 (cit. 14. 06. 2023).

[15] Jeremy M Berg et al. *Biochemistry*. 8. vyd. New York Macmillan Education, 2015. ISBN: 9781319153939.

[16] Y. Grace Chen a Sun Hur. "Cellular origins of dsRNA, their recognition and consequences". In: *Nature Reviews Molecular Cell Biology* 23 (nov. 2021), s. 286–301. DOI: 10.1038/s41580-021-00430-1. URL: https://www.nature.com/articles/s41580-021-00430-1 (cit. 14. 06. 2023).

[17] Sponk. *Čeština: Rozdíly mezi DNA a RNA*. Wikimedia Commons, apr. 2014. URL: https://commons.wikimedia.org/w/index.php?title=File:Difference_DNA_RNA-CS.svg (cit. 14. 06. 2023).

[18] Chris R Calladine et al. *Understanding DNA*. Elsevier, mar. 2004. ISBN: 9780080474663.

[19] H R Drew et al. "Structure of a B-DNA dodecamer: conformation and dynamics." In: *Proceedings of the National Academy of Sciences* 78 (apr. 1981), s. 2179–2183. DOI: 10.1073/pnas.78.4.2179. (Cit. 14. 06. 2023).

[20] *Genome Data Viewer - NCBI*. www.ncbi.nlm.nih.gov, máj 2023. URL: https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.40 (cit. 18. 06. 2023).

[21] ROSALIND E. FRANKLIN a R. G. GOSLING. "Molecular Configuration in Sodium Thymonucleate". In: *Nature* 171 (apr. 1953), s. 740–741. DOI: 10.1038/171740a0. URL: https://www.nature.com/articles/171740a0 (cit. 13. 06. 2023).

[22] Wolfram Saenger. *Principles of Nucleic Acid Structure*. Springer Science & Business Media, dec. 2013. ISBN: 9781461251903.

[23] Alexander Rich a Shuguang Zhang. "Z-DNA: the long road to biological function". In: *Nature Reviews Genetics* 4 (júl 2003), s. 566–572. DOI: 10.1038/nrg1115. URL: https://www.nature.com/articles/nrg1115 (cit. 13. 06. 2023).

[24] P. Michael Ho et al. "A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences." In: *The EMBO Journal* 5 (okt. 1986), s. 2737–2744. DOI: 10.1002/j.1460-2075.1986.tb04558.x. URL: https://doi.org/10.1002/j.1460-2075.1986.tb04558.x (cit. 15. 06. 2023).

[25] P. Shing Ho. "Thermogenomics: Thermodynamic-based approaches to genomic analyses of DNA structure". In: *Methods* 47 (mar. 2009), s. 159–167. DOI: 10.1016/j.ymeth.2008.09.007. URL: https://doi.org/10.1016/j.ymeth.2008.09.007 (cit. 15. 06. 2023).

[26] A Rich, A Nordheim a A H J Wang. "The Chemistry and Biology of Left-Handed Z-DNA". In: *Annual Review of Biochemistry* 53 (jún 1984), s. 791–846. DOI: 10.1146/annurev.bi.53.070184.004043. URL: https://doi.org/10.1146/annurev.bi.53.070184.004043 (cit. 15. 06. 2023).

[27] Haiyuan Zhang et al. "Reversible B/Z-DNA Transition under the Low Salt Condition and Non-B-Form PolydApolydT Selectivity by a Cubane-Like Europium-L--Aspartic Acid Complex". In: *Biophysical Journal* 90 (máj 2006), s. 3203–3207. DOI: 10.1529/biophysj.105.078402. URL: https://doi.org/10.1529/biophysj.105.078402 (cit. 15. 06. 2023).

[28] G.P. Schroth, P.J. Chou a P.S. Ho. "Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes." In: *Journal of Biological Chemistry* 267 (jún 1992), s. 11846–11855. DOI: 10.1016/s0021-9258(19)49776-7. URL: https://pubmed.ncbi.nlm.nih.gov/1601856/ (cit. 15. 06. 2023).

[29] Madhabi M Bhanjadeo et al. "Biophysical interaction between lanthanum chloride and (CG)n or (GC)n repeats: A reversible B-to-Z DNA transition". In: *International Journal of Biological Macromolecules* 216 (sept. 2022), s. 698–709. DOI: 10.1016/j.ijbiomac.2022.07.020. URL: https://www.sciencedirect.com/science/article/pii/S0141813022014441 (cit. 15. 06. 2023).

[30] Mauroesguerroto. *English: A-DNA, B-DNA, and Z-DNA conformations of DNA*. Wikimedia Commons, okt. 2014. URL: https://commons.wikimedia.org/wiki/File:Dnaconformations.png (cit. 15. 06. 2023).

[31] Juyong Lee et al. "Transition between B-DNA and Z-DNA: Free Energy Landscape for the B-Z Junction Propagation". In: *The Journal of Physical Chemistry B* 114 (júl 2010), s. 9872–9881. DOI: 10.1021/jp103419t. URL: https://pubs.acs.org/doi/10.1021/jp103419t (cit. 15. 06. 2023).

[32] Sook Ho Kim et al. “Unveiling the pathway to Z-DNA in the protein-induced B–Z transition”. In: *Nucleic Acids Research* 46 (máj 2018), s. 4129–4137. DOI: 10.1093/nar/gky200. URL: https://academic.oup.com/nar/article-abstract/46/8/4129/4951845 (cit. 15. 06. 2023).

[33] Anitha Suram et al. “First evidence to show the topological change of DNA from B-dNA to Z-DNA conformation in the hippocampus of Alzheimer’s brain”. In: *Neuromolecular Medicine* 2 (2002), s. 289–297. DOI: 10.1385/nmm:2:3:289. URL: https://pubmed.ncbi.nlm.nih.gov/12622407/ (cit. 15. 06. 2023).

[34] Alan Herbert. “Mendelian disease caused by variants affecting recognition of Z-DNA and Z-RNA by the Zα domain of the double-stranded RNA editing enzyme ADAR”. In: *European Journal of Human Genetics* 28 (júl 2019), s. 114–117. DOI: 10.1038/s41431-019-0458-6. URL: https://doi.org/10.1038/s41431-019-0458-6 (cit. 15. 06. 2023).

[35] So-I. Shin et al. “Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome”. In: *DNA Research* 23 (júl 2016), s. 477–486. DOI: 10.1093/dnares/dsw031. (Cit. 18. 06. 2023).

[36] Regina Z Cer et al. “Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools”. In: *Nucleic Acids Research* 41 (nov. 2012), s. D94–D100. DOI: 10.1093/nar/gks955. (Cit. 18. 06. 2023).

[37] Subramaniyam Ravichandran, Vinod Kumar Subramani a Kyeong Kyu Kim. “Z-DNA in the genome: from structure to disease”. In: *Biophysical Reviews* 11 (máj 2019), s. 383–387. DOI: 10.1007/s12551-019-00534-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6557933/ (cit. 05. 03. 2024).

[38] Randy Johnson a Chris Wolcott. *abcsFrederick/non-B_gfa*. GitHub, aug. 2023. URL: https://github.com/abcsFrederick/non-B_gfa (cit. 05. 03. 2024).

[39] A. M. Deaton a A. Bird. “CpG islands and the regulation of transcription”. In: *Genes & Development* 25.10 (2011), s. 1010–1022. DOI: 10.1101/gad.2037511. (Cit. 18. 06. 2023).

[40] D. Takai a P. A. Jones. “Comprehensive analysis of CpG islands in human chromosomes 21 and 22”. In: *Proceedings of the National Academy of Sciences* 99.6 (2002), s. 3740–3745. DOI: 10.1073/pnas.052410099. (Cit. 18. 06. 2023).

[41] A. E. Morgan, T. J. Davies a M. T. Mc Auley. “The role of DNA methylation in ageing and cancer”. In: *Proceedings of the Nutrition Society* 77.4 (2018), s. 412–422. DOI: 10.1017/S0029665118000150. (Cit. 18. 06. 2023).

[42] H. Gowher a A. Jeltsch. “Mammalian DNA methyltransferases: new discoveries and open questions”. In: *Biochemical Society Transactions* 46 (2018), s. 1191–1202. DOI: 10.1042/BST20170574. (Cit. 18. 06. 2023).

[43] D. Ramasamy, A. K. Deva Magendhra Rao, T. Rajkumar et al. "Non-CpG methylation—a key epigenetic modification in cancer". In: *Briefings in Functional Genomics* 20 (2021), s. 304–311. DOI: 10.1093/bfgp/elab035. (Cit. 18. 06. 2023).

[44] A. Fuso a M. Lucarelli. "CpG and Non-CpG Methylation in the Diet-Epigenetics--Neurodegeneration Connection". In: *Current Nutrition Reports* 8 (2019), s. 74–82. DOI: 10.1007/s13668-019-0266-1. (Cit. 18. 06. 2023).

[45] H. S. Jang, W. J. Shin, J. E. Lee et al. "CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function". In: *Genes (Basel)* 8 (2017), s. 148. DOI: 10.3390/genes8060148. (Cit. 18. 06. 2023).

[46] Michal Urban. *antroit47/cpg_islands*. GitHub, jan. 2024. URL: https://github.com/antroit47/cpg_islands (cit. 04. 03. 2024).

[47] Lucas Nell. *lucasnell/TaJoCGI*. GitHub, dec. 2023. URL: https://github.com/lucasnell/TaJoCGI (cit. 04. 03. 2024).

[48] Long-Cheng Li a Rajvir Dahiya. "MethPrimer: designing primers for methylation PCRs". In: *Bioinformatics* 18 (nov. 2002), s. 1427–1431. DOI: 10.1093/bioinformatics/18.11.1427. (Cit. 18. 06. 2023).

[49] Fábio Madeira et al. "Search and sequence analysis tools services from EMBL-EBI in 2022". In: *Nucleic Acids Research* 50 (apr. 2022), s. 276–279. DOI: 10.1093/nar/gkac240. (Cit. 19. 06. 2023).

[50] Michal Urban. *cpg_islands/test.py at main · antroit47/cpg_islands*. GitHub, jan. 2024. URL: https://github.com/antroit47/cpg_islands/blob/main/test.py (cit. 05. 03. 2024).

[51] Václav Brázda et al. *DNA analyser*. bioinformatics.ibp.cz, 2016. URL: https://bioinformatics.ibp.cz/#/ (cit. 05. 03. 2024).

[52] Václav Brázda et al. "G4Hunter web application: a web server for G-quadruplex prediction". In: *Bioinformatics* 35 (feb. 2019). Ed. John Hancock, s. 3493–3495. DOI: 10.1093/bioinformatics/btz087.

[53] Václav Brázda et al. "R-Loop Tracker: Web Access-Based Tool for R-Loop Detection and Analysis in Genomic DNA Sequences". In: *International journal of molecular sciences* 22 (nov. 2021), s. 12857–12857. DOI: 10.3390/ijms222312857. (Cit. 05. 03. 2024).

[54] Amina Bedrat, Laurent Lacroix a Jean-Louis Mergny. "Re-evaluation of G-quadruplex propensity with G4Hunter". In: *Nucleic Acids Research* 44 (jan. 2016), s. 1746–1759. DOI: 10.1093/nar/gkw006. URL: https://academic.oup.com/nar/article/44/4/1746/1854457?login=true (cit. 05. 03. 2024).

[55] Vaclav Brazda et al. “G4Killer web application: a tool to design G-quadruplex mutations”. In: *Bioinformatics* 36 (jan. 2020), s. 3246–3247. DOI: 10.1093/bioinformatics/btaa057. (Cit. 05. 03. 2024).

[56] Václav Brázda et al. “Palindrome analyser – A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences”. In: *Biochemical and Biophysical Research Communications* 478 (sept. 2016), s. 1739–1745. DOI: 10.1016/j.bbrc.2016.09.015. (Cit. 05. 12. 2023).

[57] Dmitry B. Veprintsev a Alan R. Fersht. “Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA”. In: *Nucleic Acids Research* 36 (mar. 2008), s. 1589–1598. DOI: 10.1093/nar/gkm1040. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2275157/ (cit. 05. 03. 2024).

[58] Evan You. *Introduction — Vue.js*. v2.vuejs.org, jan. 2022. URL: https://v2.vuejs.org/v2/guide/ (cit. 05. 03. 2024).

[59] Evan You. *Vue.js*. v2.vuejs.org, dec. 2023. URL: https://v2.vuejs.org/eol/ (cit. 05. 03. 2024).

[60] Vue. *Introduction | Vue.js*. vuejs.org, 2022. URL: https://vuejs.org/guide/introduction.html (cit. 05. 03. 2024).

[61] Byung-Jae Kwak, Nah-Oak Song a L.E. Miller. “Performance analysis of exponential backoff”. In: *IEEE/ACM Transactions on Networking* 13 (apr. 2005), s. 343–355. DOI: 10.1109/tnet.2005.845533. (Cit. 05. 03. 2024).

[62] Kacper Madej a Torstein Hønsi. *Heatmap | Highcharts*. highcharts.com, sept. 2023. URL: https://www.highcharts.com/docs/chart-and-series-types/heatmap (cit. 06. 03. 2024).

[63] Oracle. *JDK 11 Documentation*. Oracle Help Center, 2024. URL: https://docs.oracle.com/en/java/javase/11/ (cit. 07. 03. 2024).

[64] Phillip Webb. *Spring Boot Reference Documentation*. Spring.io, 2012. URL: https://docs.spring.io/spring-boot/docs/current/reference/htmlsingle/ (cit. 07. 03. 2024).

[65] Rod Johnson et al. *Spring Framework Documentation :: Spring Framework*. docs.spring.io, jan. 2024. URL: https://docs.spring.io/spring-framework/reference/index.html (cit. 07. 03. 2024).

[66] The PostgreSQL Global Development Group. *PostgreSQL 12.18 Documentation*. PostgreSQL Documentation, feb. 2024. URL: https://www.postgresql.org/docs/12/index.html (cit. 07. 03. 2024).

[67] Andrei Tokar. *H2 Database Engine (redirect)*. h2database.com, apr. 2024. URL: https://h2database.com/ (cit. 07. 03. 2024).

[68] IBM. *What is a REST API? | IBM*. www.ibm.com, 2023. URL: https://www.ibm.com/topics/rest-apis (cit. 07. 03. 2024).

[69] Codecademy. *What Is REST?* Codecademy, 2024. URL: https://www.codecademy.com/article/what-is-rest (cit. 07. 03. 2024).

[70] SmartBear Software. *Swagger Documentation*. Swagger.io, 2021. URL: https://swagger.io/docs/ (cit. 07. 03. 2024).

[71] Redgate. *Flyway Documentation - Flyway - Product Documentation*. documentation.red-gate.com, máj 2024. URL: https://documentation.red-gate.com/flyway (cit. 07. 03. 2024).

[72] Apache. *The Apache Groovy programming language - Documentation*. groovy-lang.org, apr. 2024. URL: https://groovy-lang.org/documentation.html (cit. 07. 03. 2024).

[73] Allie Sadler. *Manuals*. Docker Documentation, dec. 2023. URL: https://docs.docker.com/manuals/ (cit. 08. 03. 2024).

[74] Len Bass, Paul Clements a Rick Kazman. *Software Architecture in Practice*. 3. vyd. Addison-Wesley, 2013. ISBN: 9780321815736.

[75] Sung Chul Ha et al. "The structures of non-CG-repeat Z-DNAs co-crystallized with the Z-DNA-binding domain, hZα ADAR1". In: *Nucleic acids research* 37 (dec. 2008), s. 629–637. DOI: 10.1093/nar/gkn976. (Cit. 09. 03. 2024).

[76] Serge Saxonov, Paul Berg a Douglas L. Brutlag. "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters". In: *Proceedings of the National Academy of Sciences of the United States of America* 103 (jan. 2006), s. 1412–1417. DOI: 10.1073/pnas.0510310103. URL: https://pubmed.ncbi.nlm.nih.gov/16432200/ (cit. 10. 03. 2024).

[77] Dmitriy Umerenkov et al. "Z-flipon variants reveal the many roles of Z-DNA and Z-RNA in health and disease." In: *Life Science Alliance* 6 (júl 2023). DOI: 10.26508/lsa.202301962. (Cit. 11. 03. 2024).

Prílohy

Zoznam obrázkov

Zoznam tabuliek

Zoznam DNA kódov

Zoznam kódov

Zoznam algoritmov