VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV AUTOMATIZACE A INFORMATIKY
INSTITUTE OF AUTOMATION AND COMPUTER SCIENCE

IMPLEMENTACE NÁSTROJE PRO ANALÝZU LOKÁLNÍCH STRUKTUR DNA

IMPLEMENTATION OF LOCAL DNA STRUCTURES ANALYSIS TOOL

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. Patrik Kaura
AUTHOR

VEDOUCÍ PRÁCE — prof. RNDr. Ing. Jiří Šťastný, CSc.
SUPERVISOR

BRNO 2019

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav automatizace a informatiky
Student:	**Bc. Patrik Kaura**
Studijní program:	Strojní inženýrství
Studijní obor:	Aplikovaná informatika a řízení
Vedoucí práce:	**prof. RNDr. Ing. Jiří Šťastný, CSc.**
Akademický rok:	2018/19

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Implementace nástroje pro analýzu lokálních struktur DNA

Stručná charakteristika problematiky úkolu:

Obsahem práce je implementace programového nástroje v prostředí Python pro automatizovanou analýzu lokálních struktur DNA a jeho ověření provedením ukázkové analýzy velkého množství DNA sekvencí.

Cíle diplomové práce:

Cílem práce je návrh specializované platformy pro analýzu lokálních struktur DNA.
Práce bude zahrnovat analýzu současného stavu automatizované analýzy lokálních struktur DNA.
Vytvořte automatizační nástroj pro komunikaci a řízení výpočetního jádra využívající stávající komunikační rozhraní.
Zaměřte se na přívětivé zpracování velkého množství sekvencí a pohodlnou práci s výsledky.
Pro implementaci využijte prostředí Python, nástroj publikujte jako knihovnu pro správce balíčků pip (package manager).
Navržený software bude ověřen provedením ukázkové analýzy velkého množství DNA sekvencí.

Seznam doporučené literatury:

BRÁZDA, V. a kol. Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. Biochemical and Biophysical Research Communications. 2016. sv. 478, č. 4, s. 1739--1745. ISSN 0006-291X. URL: http://www.sciencedirect.com/science/article/pii/S0006291X16314620.

BRÁZDA, V. a kol. Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. BioMed Research International. 2018. č. 24 July, ISSN 2314-6133. URL: https://doi.org/10.1155/2018/1097018.

DOWNEY, A. Think Python. 2nd edition, updated for Python 3. Sebastopol, CA: O'Reilly Media, 2016. ISBN 1491939362.

PILGRIM, M. Ponořme se do Python(u) 3: Dive into Python 3. Praha: CZ.NIC, c2010. CZ.NIC. ISBN 978-80-904248-2-1.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2018/19

V Brně, dne

L. S.

doc. Ing. Radomil Matoušek, Ph.D. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

ABSTRAKT

Tato diplomová práce je zaměřena na popis a realizaci aplikace API wrapper, která pracuje nad výpočetním jádrem ibp bioinformatics. První polovina práce se zaměřuje na shrnutí základních znalostí z oblasti výzkumu DNA, rovněž i specifikaci problému a popisu zvolených technologií. Druhá polovina se soustředí na samotnou realizaci, distribuci a vyhodnocení použitelnosti aplikace na konkrétních sekvencích DNA.

ABSTRACT

This diploma thesis is focused on the description and implementation of the API wrapper application, which works on top of computational core ibp bioinformatics. The first half of the thesis is focused on the summary of basic knowledge in the field of DNA research, as well as the specification of the problem and description of selected technologies. The other half deals with the actual implementation, distribution, and evaluation of application applicability on specific DNA sequences.

KLÍČOVÁ SLOVA

REST API, Python, Jupyter notebook, pypi, DNA, lokální struktury, guaninový kvadruplex

KEYWORDS

REST API, Python, Jupyter notebook, pypi, DNA, local structures, guanine quadruplex

BIBLIOGRAFICKÁ CITACE

KAURA, Patrik. *Implementace nástroje pro analýzu lokálních struktur DNA.* Brno, 2019. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/117299. Diplomová práce. Vysoké učení technické v Brně, Fakulta strojního inženýrství, Ústav automatizace a informatiky. Vedoucí práce Jiří Šťastný.

PODĚKOVÁNÍ

Na tomto místě bych chtěl poděkovat mému vedoucímu prof. RNDr. Ing. Jiřímu Šťastnému, CSc. za vedení diplomové práce. Dále bych chtěl poděkovat Ing. Janu Kolomazníkovi Ph.D. a Ing. Jiřímu Lýskovi Ph.D. z Mendelovy univerzity za odborné rady při zpracovávání technické části. V neposlední řadě bych chtěl poděkovat doc. Mgr. Václavu Brázdovi Ph.D. z ústavu Biofyzikální chemie a molekulární onkologie AV ČR za možnost podílení se na analýzách v oblasti lokálních struktur DNA. Nakonec bych chtěl poděkovat především mé rodině za projevenou trpělivost a podporu mého studia.

ČESTNÉ PROHLÁŠENÍ

Prohlašuji, že tato práce je mým původním dílem, zpracoval jsem ji samostatně pod vedením RNDr. Ing. Jiřího Šťastného, CSc. a s použitím literatury uvedené v seznamu literatury.

V Brně dne 1. 3. 2018 ……

Bc. Patrik Kaura

OBSAH

1 ÚVOD

Od objevu deoxyribonukleové kyseliny v roce 1869 švýcarským lékařem Friedrichem Miescherem se v oboru biochemie zrodila řada významných objevů usnadňujících lidstvu pochopení tohoto pozoruhodného polymeru. Trvalo však téměř sto let, než doktor Francis Crick s doktorem Jamesem D. Watsonem byli schopni popsat její dvoušroubovicouvou strukturu. Za tento objev v roce 1962 společné s anglickým molekulárním biologem Mauricem Wilkinsem obdrželi Nobelovu cenu za fyziologii a lékařství. S postupem dalšího vývoje byly objeveny i další možné prostorové konfigurace DNA. Za zmínku strojí například konfigurace Z-DNA, která je oproti běžnější pravotočivé B-DNA levotočivá.

Mimo již zmíněné geometrické uspořádání rozeznáváme i výrazně odlišné konfigurace, totiž lokální struktury DNA. Ty se formují za velice specifických podmínek. Na jejich vznik má vliv chemické složení prostředí, interakce s proteiny či struktura samotné DNA. Toto, ale i jiné vlivy mají velký dopad na formování struktur, jako je guaninový kvadruplex či invertované repetice (vlásenky, křížové struktury). Již zmíněné struktury mohou mít vliv na změny v evolučních a regulačních vlastnostech DNA. Existuje i možná souvislost s predispozicemi k různým vážným onemocněním. Ukazuje se využití kvadruplexů například v léčbě nádorových onemocnění. Vzhledem k výskytu kvadruplexů především v oblasti telomer lidského genomu je zde určitá možnost tyto telomerové kvadruplexy stabilizovat a tím omezit růst nádorových buněk.

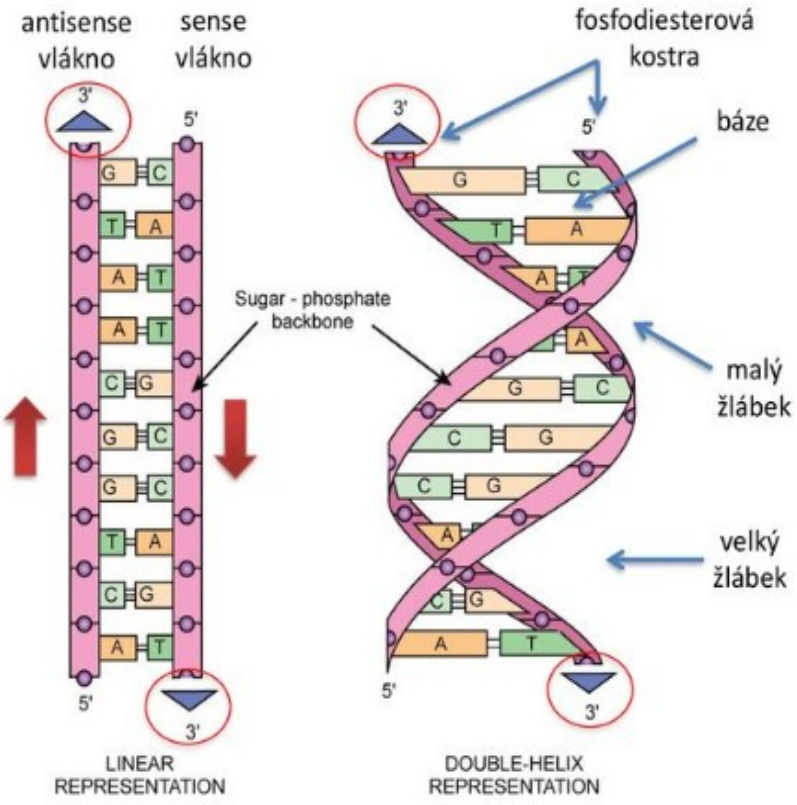
Cílem této práce není vyvozování závěru jednotlivých analýz z oblasti biochemie, ale tvorba nástroje umožňujícího jednoduchý přístup k rozborům zkoumaných vzorků. Nástroj musí být schopen hromadného zadávání testovacích sekvencí a provádění jednotlivých analýz na vzdáleném výpočetním serveru aplikace ibp bioinformatics. Vzhledem k okruhu lidí používajících tento nástroj musí být navržen tak aby byl uživatelsky přívětivý i pro pracovníky s minimální kvalifikací v oblasti programování. O dalších požadavcích pojednává kapitola 5 zabývající se projektovou přípravou. Kromě návrhu, realizace a publikace softwaru je součástí práce i série několika testů nad skupinami sekvencí genomů ověřujících funkčnost v různých případech, viz kapitoly 6 a 7.

2 LOKÁLNÍ STRUKTURY DNA

V této úvodní kapitole diplomové práce jsou shrnuty základní poznatky z oblasti biochemie, na kterých práce dále staví. Vzhledem k technickému zaměření této práce jsou tyto poznatky pouze krátkým pojednáním o daných tématech a nezacházejí příliš do detailních popisů. V první části kapitoly je popsána základní struktura deoxyribonukleové kyseliny. Druhá polovina se zabývá již konkrétními lokálními strukturami, které jsou v následujících kapitolách zmíněny.

2.1 DEOXYRIBONUKLEOVÁ KYSELINA (DNA)

DNA je nukleová kyselina, která je nositelkou genetické informace organismů s výjimkou nebuněčných, u nichž tuto funkci nahrazuje RNA. Je složena ze dvou polynukleotidových řetězců, které se otáčejí kolem společné osy a tvoří pravotočivou dvojitou šroubovici. Tyto dva řetězce rovněž tvoří antiparalelní uspořádání. Ve středu se nacházejí tzv. nukleové báze, cukry deoxyribózy a fosfátové skupiny jsou vně struktury dvoušroubovice. Struktura je obecně členěna do tří základních skupin, a to primární, sekundární a terciální. [1] V následujících několika podkapitolách jsou shrnuty základní popisy těchto struktur. Schéma struktury DNA viz obr. 1.

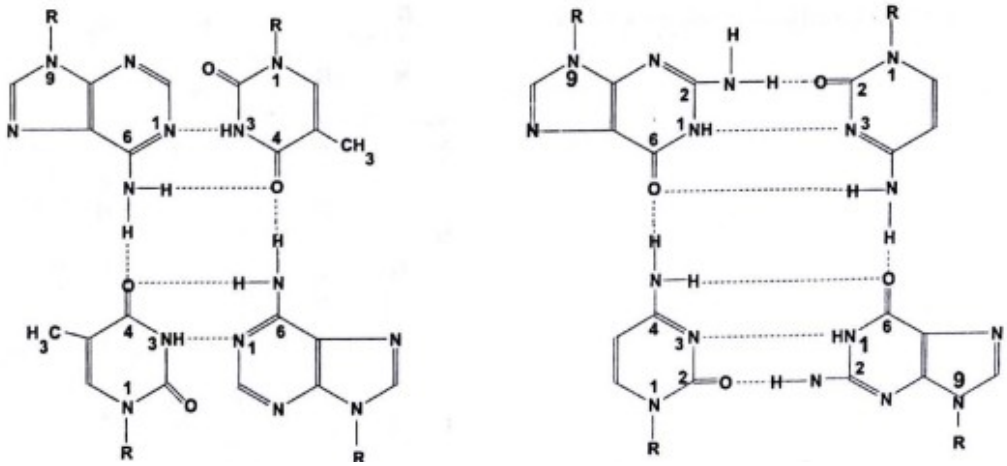


Obr 1: Schéma struktury deoxyribonukleové kyseliny [2]

2.1.1 PRIMÁRNÍ STRUKTURA

Primární struktura popisuje seřazení jednotlivých nukleotidů ve struktuře DNA. Složkami nukleotidů jsou 2-deoxy-D-ribóza ($C_5H_{10}O_4$), kyselina trihydrogenfosforečná (H_3PO_4), purinové a pyrimidinové báze. Mezi purinové báze patří adenin a guanin. Pyrimidinové báze tvoří především cytozin, tymin a uracil. Poslední zmiňovaný uracil se však vyskytuje pouze ve struktuře RNA. Jednotlivé báze jsou spojeny pomocí vodíkových vazeb. Tímto spojením vzniká dvou-řetězcová DNA. Dle Watsonova-Crickova pravidla o párování vazeb se pomocí vodíkových vazeb propojují vždy purinové a pyrimidinové báze. Konkrétně jde o vazbu adeninu s tyminem (AT pár) a guaninu s cytozine (GC pár). Schéma jednotlivých párů jsou vyobrazena na obrázku níže obr. 2.

Kromě bází je rozlišována tzv. direkcionalita řetězců. Jednotlivé nukleotidy jsou spojeny pomocí 3‘ nebo 5‘ konci DNA. Z tohoto důvodu se zavádí další rozlišení polynukleotidových řetězců, a to dle směru. Jeden z řetězců DNA jde od konce 3 k 5 a druhý opačným směrem. Čísla konců vyjadřují uhlíkové pozice v molekule ribózy či deoxyribózy. Schéma jednotlivých zakončení a struktury bází viz výše obr. 1. [1]



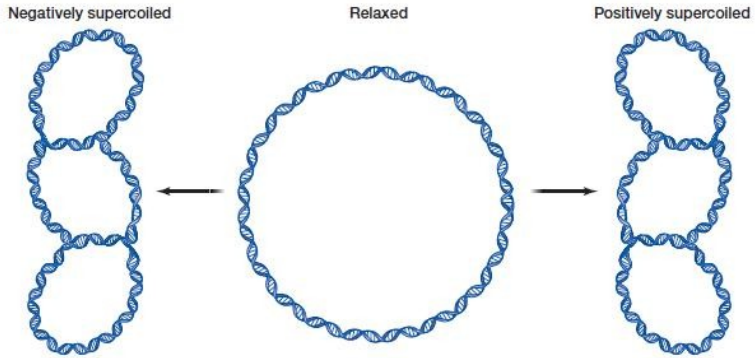
Obr 2: Párování adeninu s tyminem (vlevo) a guaninu s cytozinem (vpravo) [1]

2.1.2 SEKUNDÁRNÍ STRUKTURA

Sekundární struktura popisuje vzájemné geometrické uspořádání polynukleotidových řetězců v prostoru. Nejčastější forma sekundární struktury je již zmíněná dvojšroubovice. Pro tuto strukturu platí tzv. Chargaffovo pravidlo, dle kterého je poměr purinových a pyrimidinových bází roven jedné. Kromě lineární formy s volnými konci rozeznáváme i kruhovou variantu, která má jednotlivé konce spojené, a tedy tvoří uzavřenou smyčku. Existuje i varianta ohnuté DNA, kde v určitém místě dochází k ohybu původní struktury s volnými konci. K již zmíněné dvouřetězcové formě existuje varianta jednořetězová, třířetězcová (triplex) a čtyřřetězcová (kvadruplex) varianta uspořádání. [1]

2.1.3 TERCIÁLNÍ STRUKTURA

Pokud se hovoří o tzv. terciální struktuře, je tím myšleno uspořádání do tzv. nadšroubovice, neboli supercoil původní struktury DNA. Tato nadšroubovice se může tvořit ve formě kruhové i lineární. Rozlišujeme negativní a pozitivní nadšroubovice. Negativní se otáčejí proti směru hodinových ručiček, kladné ve směru. [1] Ukázka negativní a pozitivní nadšroubovice viz obr. 3.



Obr 3: Schéma terciální struktury tzv. nadšroubovice [3]

2.2 TYPY SEKVENCÍ

Obecně sekvence bází dělíme na dvě základních skupiny. Sekvence jedinečné jsou takové, které se v celém genomu vyskytují pouze jednou. Druhá skupina obsahuje tzv. repetice. [1] Na výzkum v této oblasti jsou použity nástroje zmíněné v kapitole 3.

2.2.1 REPETICE

Oproti sekvencím jedinečným se repetice mnohonásobně opakují ve zkoumaném genomu. Příkladem může být například výskyt repetice CAG v 5‘ konci genu u lidí trpících neurodegenerativní Huntingtonovou chorobou. [4] Repetice obecně lze dále dělit na tandemové, obrácené, přímé, dlouhé koncové, křížové a rozptýlené.

Tandemové repetice jsou takové, u kterých se vyskytuje repetitivní vzor bezprostředně za sebou. Průměrně jsou obsaženy v 5 až 15% DNA a jsou považované za často se opakující. Příkladem může být například sekvence CTG**ATTAATTA**CGTC s opakujícím se vzorem ATTA. Obrácená repetice je sekvence také známá pod názvem palindrom. Tato sekvence má v komplementárním řetězci shodnou sekvenci pouze s jinou direkcionalitou. Ukázkou může být například sekvence **GACTTC** ve směru od 5‘ k 3‘ konci. V komplementárním řetězci jsou báze uspořádány **CTTCAG** ve směru od 3‘ k 5‘ konci. Tyto sekvence mají tendenci tvořit z řetězců tzv. vlásenky nebo křížové struktury. Přímá repetice je velice podobná dříve zmíněné tandemové. Rozdíl spočívá ve výskytu mezery vyplněné jinými bázemi. Příkladem může být sekvence **ATGC**AATCA**ATGC.** Dlouhé koncové repetice jsou přímé repetice

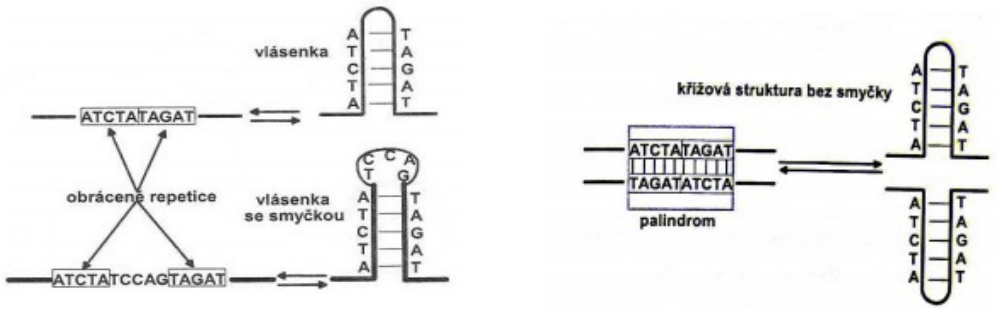
vyskytující se na obou koncích DNA řetězce. Příkladem je uspořádání **ATCTA** na obou koncích řetězce. Poslední zmíněný typ jsou tzv. rozptýlené repetice. Jedná se o repetice podobné přímým s tím rozdílem, že opakování vzoru je náhodně rozmístěno v sekvenci. Ty se dále dělí na krátké repetice do 300 bp a dlouhé, jejichž délka přesahuje hodnotu 300 bp. Zkratka bp pochází z anglického base pair a značí počet párovaných bází. [1]

2.3 LOKÁLNÍ STRUKTURY V DNA

Opakující se motivy jsou velice běžným jevem vyskytujícím se v řetězcích DNA. Mezi tyto motivy řadíme již zmíněné křížové struktury, vlásenky, triplexy, kvadruplexy, tetraplexy a Z-DNA. Lokální struktury mohou způsobovat mutace genomu u prokaryotických i eukaryotických organismů. Nynější poznatky z oblasti molekulární biologie naznačují nejen souvislost s predispozicemi k různým onemocněním, rovněž se ukazuje i vliv na rychlé evoluční změny v oblasti vývojových a regulačních vlastností DNA. [1] V následujících podkapitolách jsou popsány struktury křížové, vlásenky a kvadruplexy. Důvod tohoto omezení je spojen s oblastí použitelnosti vyvíjeného nastroje, který v době tvorby této práce je schopen nacházet tyto struktury v předložených sekvencích.

2.3.1 VLÁSENKOVÁ A KŘÍŽOVÁ STRUKTURA

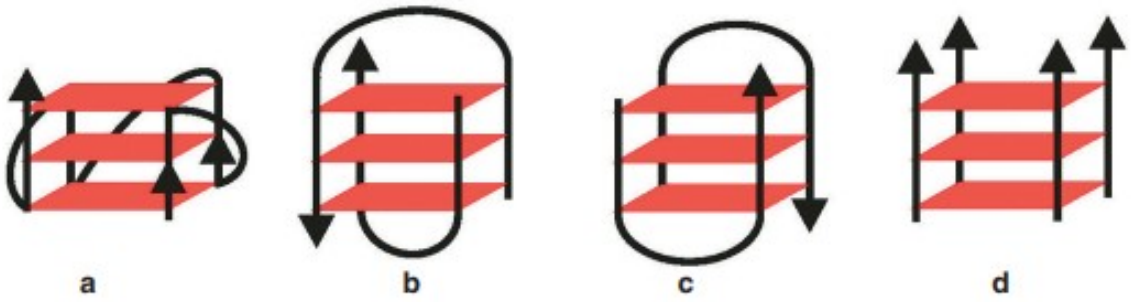
Vlásenka je obecně dvoušroubovicová struktura vznikající v místech obrácené repetice. Pokud jednotlivé repetice spolu bezprostředně sousedí, jedná se o běžnou vlásenku. V případě že mezi repeticemi vzniká určitý prostor, vzniká vlásenka se smyčkou viz obr. 3. V případě křížové struktury se hovoří o párování obrácených repetic na obou polynukleotidových řetězcích DNA. Obdobně s vlásenkovou strukturou vzniká i křížová bez i se smyčkou. Důvod tvorby smyčky je obdobný jako v předchozím případě. Křížové struktury zobrazuje rovněž obr. 4. [1]



Obr 4: Schéma struktury vlásenky (vlevo) a křížové repetice (vpravo) [1]

2.3.2 GUANINOVÝ KVADRUPLEX

Jedná se o místo s vysokým obsahem guaninu, proto se těmto útvarům častěji říká guaninový kvadruplex či g-kvartet. Základem pro vznik kvadruplexu je tedy, jak již bylo řečeno, guanin, formující tzv. guaninovou tetrádu. Jednotlivé guaniny jsou propojeny pomocí vodíkových vazeb a z geometrického hlediska tvoří planární útvar. Jedná se tedy o vrstvenou strukturu, jejíž volné prostory jsou obvykle vyplněny jednomocnými ionty. Tyto struktury lze dělit dle mnoha kriterií. Jednou z možných forem rozdělení je dělení dle počtu vláken, které strukturu formují. Rozlišujeme tedy kvadruplexy unimolekulární, bimolekulární a lineární. Unimolekulární a bimolekulární formují tzv. smyčku, naopak lineární jsou tvořeny čtyřmi samostatnými vlákny DNA. Další forma dělení spočívá na základě direkcionality řetězců tvořících kvadruplex. Hovoříme o paralelním nebo antiparalelním uspořádání. Dále je pak možné tyto struktury dělit dle množství tetrád a podobně. Ukázka možných konfigurací kvadruplexů je na obr. 5. [5]



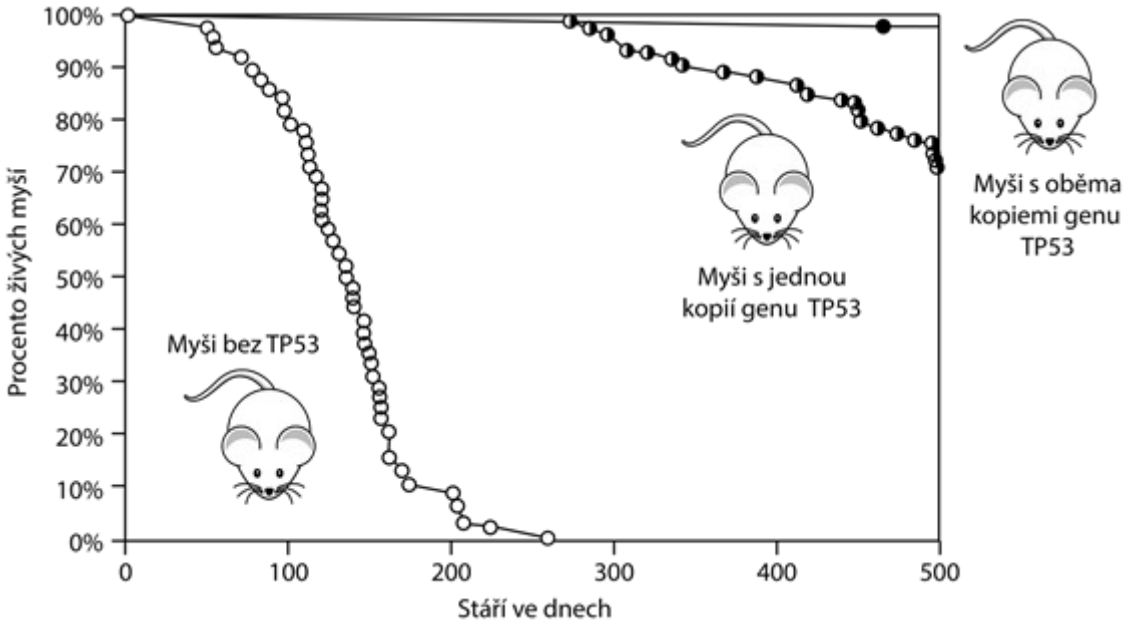
Obr 5: Schéma možných konfigurací guaninového kvadruplexu [5]

2.4 PROTEIN P53

V předchozích kapitolách byly shrnuty tři druhy lokálních struktur v DNA. Kromě hledání těchto struktur implementuje nástroj bioinformatics i analýzu vazeb proteinu p53 na zadanou část řetězce. Proto je v následující části shrnut popis tohoto proteinu a jeho důležitost pro výzkum. Produkt genu p53 je jaderný protein patřící do kategorie tumor supresorových genů. Obecně má v buňce funkci detektoru poškození DNA. Standardně je tento gen neaktivní, aktivuje se až při poškození genomu. Pokud proběhne oprava DNA, buňka naváže na běžný buněčný cyklus. V opačném případě buňka přejde do stavu programované buněčné smrti a poškozená část se dále nereplikuje. [7]

Tento protein, který je zván strážcem genomu, se stal objektem mnoha výzkumů již od doby objevení v roce 1979 doktorem Davidem Lanem. Výzkum na laboratorních myších dokazuje, že jedinci bez genomu TP53 hynou na rakovinu podstatně dříve než jedinci s oběma kopiemi genu TP53. Podobnému osudu čelili i jedinci s pouze jednou kopií tohoto genu viz obr. 6. Dále se ukazuje, že zvýšené množství hladiny proteinu p53 vede k zastavení

buněčného dělení, ale pouze do okamžiku, kdy je protein přítomen. Hledaný protinádorový účinek je pak možný pomocí více různě zacílených látek. Tento typ terapie je dnes podrobován klinickým zkouškám na vybraných pacientech. [6]



Obr 6: Graf zobrazující délku dožití myší dle genomu p53 [6]

3 ANALÝZA EXISTUJÍCÍ NÁSTROJŮ

Tato kapitola shrnuje několik analytických nástrojů používaných pro zpracování sekvencí, v nichž hledáme repetice či guaninové kvadruplexy. V první polovině této kapitoly jsou shrnuty základní popisy těchto webových aplikací. Druhá polovina se zabývá srovnáním vybraných vzorků s nástrojem bioinformacics.ibp, který je objektem zájmu této práce.

3.1 REPEATMASKER

Nástroj RepeatMasker je vyvíjen v Seattlu Institutem pro systémovou biologii. Aplikace se obsluhuje pomocí jednoduchého webového formuláře. Po zaslání textového řetězce či souboru ve formátu FASTA je sekvence podrobena analýze repetic. Výsledkem je pak jednoduchý textový soubor s výslednými částmi sekvence a nalezenými repeticemi. Aplikace obsahuje i verzi spustitelnou v lokálním prostředí. [8] Aplikace je dostupná na adrese http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.

3.2 EMBOS EXPLORER

Webová aplikace EMBOS explorer obsahuje nástroj palindrome a je k nalezení na adrese http://emboss.bioinformatics.nl/cgi-bin/emboss/palindrome. Jedná se o velice jednoduchý webový formulář s velice podobnou funkcí jako u předchozího nástroje. Oproti předchozímu nástroji má palindrome možnost nastavení minimální a maximální délky palindromu. Dále je zde možnost nastavení maximální mezery mezi oblastmi s repeticí. Vzhledem k možnému výskytu chyb v analyzovaných sekvencích umožňuje i nastavení počtu rozdílných bází např. **ATTA** a **ATCA** by byly palindromy v případě nastavení hodnoty odchylky na jedna. Zkratka EMBOS odkazuje na komunitu The European Molecular Biology Open Software Suita vyvíjející aplikace pro komunitu molekulárních biologů. [9]

3.3 PALINDROMIC SEQUENCE FINDER

Je nejjednodušší webový nástroj z již zmiňovaných a je k nalezení na webové adrese http://insilico.ehu.eus/palindromes/. Poskytuje pouze základní informace o nalezených palindromech ve zkoumaných sekvencích. Stejně jako nástroj EMBOS umožňuje volbu minimální a maximální délky sekvence palindromu. Vstup je možný pouze pomocí krátkého textového řetězce. Nástroj vyvinul doktor Joseba Bikandi na University of the Basque Country. [10]

3.4 PALINDROME SEARCH

Je poslední zmíněný nástroj na analýzu repetic na adrese http://bioinfo.cs.technion.ac.il/projects/Engel-Freund/new.html. Stejně jako předchozí zmíněné umožňuje nastavení minimální délky palindromu. Maximální délku mezery, maximální počet odlišných bází viz kap 3.2. Vstup vlastní sekvence je možný pouze ve formátu FASTA. Nástroj byl vytvořen Izraelským technologickým institutem Technion. [11]

3.5 QGRS MAPPER

Nástroj QGRS mapper je první zmíněný analytický nástroj sloužící k odhalování oblastí bohatých na guanin tzv. G-Rich sekvence. Stejně jako v předchozích případech jde o webový formulář umožňující vstup pomocí textového řetězce, FASTA souboru či importu z NCBI databáze. Výsledkem je pak přehledné tabulkové zobrazení jednotlivých predikovaných kvadruplexů. Nástroj zpracovala americká univerzita Ramapo College of New Jersey a je dostupný na http://bioinformatics.ramapo.edu/QGRS/analyze.php. [12]

3.6 QUADBASE2

QUADBASE2 na rozdíl od všech předchozích softwarů neumožňuje vložení vlastní sekvence ale poskytuje pouze analýzy nad sekvencemi typu Bacteria a Archaea. Stejně jako nástroj QGRS analyzuje výskyt kvadruplexů. Nástroj poskytuje nastavení vlastností použitých algoritmů, jako například volbu typu algoritmu, hledání i v komplementárním řetězci apod. Vývoj zabezpečuje doktor Parashar Dhapola z CSIR-Institute of Genomics and Integrative Biology, New Dehli a je dostupný na http://quadbase.igib.res.in/ProQuad. [13]

3.7 NON-B DB

Poslední zmíněný nástroj není přímo analytickým nástrojem, ale databází. Ta poskytuje několik předem definovaných genomů, nad kterými se provádí vybraná analýza. Databáze obsahuje kompletní genom člověka, psa, krávy a dalších. Nad zvoleným genomem je pak spuštěn vyhledávač jednotlivých struktur. Kromě samotných kvadruplexů může analyzovat i různé druhy repetic. Databázi spravuje National Cancer Institute Center for Cancer Research a je dostupná na https://nonb-abcc.ncifcrf.gov. [14]

3.8 IBP BIOINFORMATICS

Webová aplikace IBP bioinformatics je vyvíjena Ústavem informatiky Mendelovy univerzity ve spolupráci s Biofyzikálním ústavem Akademie věd České republiky. V době tvorby této práce je aplikace rozdělena na dvě samostatná prostředí, Palindrome analyser a G4hunter. Aplikace Palindrome analyser obsahuje analytické nástroje pro vyhledávání palindromů, a

výpočet afinity vazeb k proteinu p53 na zadanou sekvenci DNA. Druhá zmíněná část obsahuje algoritmus G4hunter vyhledávající guaninové kvadruplexy a funkci G4killer, která snižuje hodnotu tzv. Gscore v zadané sekvenci. Palindrome analyser je průběžně integrován do druhé zmíněné aplikace z důvodu jednodušší udržitelnosti. [odkaz na článek] Obě aplikace kromě výpočetního REST API obsahují i tzv. SPA (single page application) rozhraní. Ty umožňují zadávání jednoduchých analýz nad výpočetním jádrem. Aplikace je dostupná na adrese http://bioinformatics.ibp.cz. [15, 16]

3.9 SROVNÁNÍ APLIKACÍ

Vlastnosti všech aplikací shrnuje následující tabulka. Software ibp bioinformatics obsahuje dle tabulky nejvíce funkcí. Problém roztříštěnosti jednotlivých nástrojů do různých aplikací byl hlavní motivací při tvorbě tohoto nástroje. Následuje tabulka shrnující jednotlivé funkce viz tab. 1. V následujících několika kapitolách budou popsány implementované algoritmy sloužící k vyhledávání lokálních struktur, které nástroj ibp obsahuje.

Tab 1: Tabulka srovnávající jednotlivé aplikace

	VLOŽENÍ SEKVENCE	IMPORT FASTA SOUBOR	IMPORT TEXT. SOUBORU	IMPORT Z NCBI DATABÁZE	ANALÝZA PALINDROMŮ	MOŽNOST NASTAVENÍ	ANALÝZA KVADRUPLEXŮ	MOŽNOST NASTAVENÍ
REPEATMASKER	✓	✓	✓	X	✓	X	X	X
EMBOS EXPLORER	✓	✓	✓	X	✓	✓	X	X
PALINDROME SEQUENCE FINDER	✓	X	X	X	✓	✓	X	X
PALINDROME SEARCH	✓	✓	✓	X	✓	✓	X	X
QGRS MAPPER	✓	X	✓	X	X	X	✓	✓
QUADBASE2	X	X	X	X	X	X	✓	✓
NON-B DB	X	X	X	X	X	X	✓	✓
IBP BIOINFORMATICS	✓	✓	✓	✓	✓	✓	✓	✓

3.10 ALGORITMUS G4HUNTER

Nástroj bioinformatics pro hledání kvadruplexu využívá algoritmu původně zveřejněného v článku Re-evaluation of G-quadruplex propensity with G4Hunter. Z biologického hlediska se jedná o vyhledávání míst bohatých na guanin, kde by se mohly tvořit kvadruplexy.

Dle algoritmu je guaninovým bázím přiřazována kladná hodnota a cytosinovým bázím hodnota záporná. Jednotlivé báze před samotným ohodnocením jsou nejdříve umístěny do skupin bezprostředně sousedících bází. Pokud má skupina pouze jednu bázi, je jí přidělena hodnota +1 případně -1. Pro skupinu čítající dvě báze pak +2 nebo -2 až do počtu čtyř bází, kdy i pro větší skupiny je volena hodnoticí hodnota maximálně +4 případně -4. Pro takto ohodnocené báze v délkově definovaných oknech je pak vypočten aritmetický průměr a stanoveno výsledné skóre. Dle hodnoty skóre pak mohou být oblasti filtrovány pomocí zvoleného prahu. Výsledné vytříděné a ohodnocené oblasti pak jsou spojovány, pokud jsou sousední, a je jim spočteno nové skóre, ale již pro jinou délku okna. Ukázku výpočtu nad několika sekvencemi shrnuje tabulka níže viz tab. 2. [17]

Tab 2: Výpočet g4hunter skóre pro vybrané sekvence [17]

SEKVENCE	OHODNOCENÍ	PRŮMĚR	SKÓRE
CCC	-3-3-3	-9/3	-3
GGGGGG	4+4+4+4+4+4	24/6	4
ATGGATGGATGATGAT	0+0+2+2+0+0+2+2+0+0+1+0+0+1+0+0	10/16	0.625

3.11 NÁSTROJ G4KILLER

Nástroj g4killer má za úkol snížení skóre zadané sekvence. Pomocí prohledávání stavového prostoru zaměňuje v sekvencích báze guaninu za znak W s nulovou hodnotou. Pro každou pozměněnou sekvenci je poté spočtena hodnota skóre pomocí algoritmu g4hunter. Proces postupuje stavovým prostorem až do okamžiku dosažení požadované hodnoty. [17]

3.12 ALGORITMU P53 PREDICTOR

Tento algoritmus počítá rozdíl v hodnotě afinity k proteinu p53 vzhledem k referenční sekvenci. Jako ideální sekvence byla experimentálně dle článku Algorithm for prediction of tumour supressor p53 affinity zvolena sekvence GGACATGCCCGGGCATGTCC. Dle předlohové tabulky s vypočtenými koeficienty pro jednotlivé báze definuje afinitu vazby k proteinu p53 viz tab. 3. Tabulka však ilustruje pouze prvních pět bází z důvodu rozměrnosti tabulky.

Tab 3: Předlohová tabulka afinity k p53 proteinu [18]

NEJVĚTŠÍ AFINITA ($\Delta \log K_d$)	1	2	3	4	5
	G	G	A/G	C	A
A	0.05	0.03	0.00	0.59	0.00
T	0.07	0.10	0.15	0.31	0.21
G	0.00	0.00	0.00	0.55	0.28
C	0.11	0.18	0.40	0.00	0.47

Pro předloženou sekvenci je pak pomocí následujícího vzorce (1) spočtena hodnota predikce vazby. Hodnota $\Delta\log K_d$ značí hodnotu rozdílu od maximální hodnoty afinity. Hodnota referenční afinity je stanovena pro již zmíněnou sekvenci -7.51. [17, 18]

$$\log K_d(x) = \log K_d(ref) + \sum \Delta \log K_d(i, N) \quad (1)$$

3.13 ALGORITMUS PALINDROME ANALYSIS

Algoritmus, jak je již uvedeno v názvu, hledá tzv. palindromy. Funkce nejdříve provede výběr testované části sekvence, pro kterou se hledá inverzní repetici. U přímo navazujících částí řetězce se provede výběr porovnávané sekvence, u které je provedeno odvození komplementárních bází. Porovnávaná část je poté invertována a srovnána s testovanou. Pokud je testovaný a porovnávaný řetězec shodný, zařadí se do výstupní množiny palindromů. Pokud shoda není, zvětší se rozestup mezi testovanou sekvencí a novou porovnávanou sekvencí o +1. Volba velikosti maximálních rozestupů je možná pomocí volby parametru spacer. Algoritmus kromě možnosti volby parametru spacer umožňuje i volbu počtu odlišných bází v porovnávané sekvenci. Následující tabulka ilustruje ukázku nalezení palindromu nad sekvencí viz tab. 4. [16, 17]

Tab 4: Ukázka výsledku algoritmu palindrome analyser [17]

SEKVENCE	GGACATGCCCGGGCATGTCC
VÝBĚR TESTOVANÉ SEKVENCE	GGACATGCCC
VÝBĚR POROVNÁVANÉ SEKVENCE	GGGCATGTCC
REVERZE POROVNÁVANÉ SEKVENCE	GGGCATGTCC | **CCCGTACAGG**
INVERZE POROVNÁVANÉ SEKVENCE	GGGCATGTCC | **GGACATGCCC**
POROVNÁNÍ SEKVENCÍ	1: **GGACATGCCC** |||||||||| 20: **GGACATGCCC**

4 PŘEHLED POUŽITÝCH TECHNOLOGIÍ

V průběhu tvorby této práce bylo použito několik níže zmíněných technologií. V první polovině této kapitoly bude věnována pozornost především technologiím použitým na straně obsluhované aplikace. Druhá polovina pojednává především o prostředcích spjatých se samotným vývojem aplikace API wrapperu.

4.1 JAVA

Aplikace bioinformacics je vytvořena pomocí programovacího jazyka Java. Tento jazyk je objektově orientovaný vyvinutý firmou Sun Microsystems v roce 1995. Typicky je kompilován to bytecode, který je provozován na tzv. Java virtual machine (JVM) nezávislé na architektuře počítače.

Pro tvorbu REST API bylo použito framework Spring. Jedná se o open-source aplikační framework používaný na platformě Java EE (Enterprise Edition). Pro automatizaci byl použit jazyk Gradle založený na jazyce Groovy. Oba tyto zmíněné jazyky jsou navrženy pro platformu Java. [19]

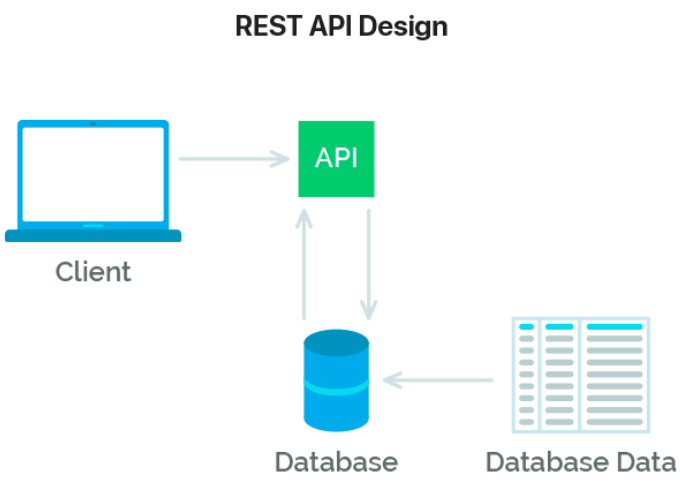
4.1.1 REST API

REST je zkratkou pro Representational State Transfer. Jedná se o architekturu rozhraní pracujících v distribuovaných prostředích. Toto rozhraní je pak použito jako standardizovaný přístupový bod pro jednotlivé aplikace, které pomocí volání HTTP metod získávají/předávají potřebná data pro svoji činnost viz obr. 7.

Rozhraní implementuje čtyři základní metody popsané zkratkou CRUD. Tedy Create pro vytváření, Retrieve pro získávání, Update pro aktualizaci a Delete pro mazání dat na serveru. Každá metoda má vlastní adresu a množinu požadavků s ní spojenou. Komunikační protokol je postaven na bázi klient server. Každý požadavek musí obsahovat všechny informace pro jeho vykonání protože server jako takový je bezstavový. [20, 21]

Jako hlavní formát pro komunikaci s Api je použit JSON (JavaScript Object Notation). JSON je formát používaný pro výměnu dat. Je jednoduše čitelný, snadno analyzovatelný a strojově generovaný. Od svého vzniku je založen na dvou základních strukturách:

- Kolekce párů název hodnota v programovacích jazycích realizovaných jako:
  - slovník (dictionary)
  - hash tabulka
  - struktura apod.
- Seřazený seznam hodnot realizovaný převážně jako
  - pole
  - vektor
  - list apod.



Obr 7: Schéma architektury REST API [21]

4.2 PYTHON

Programovací jazyk Python je vysokoúrovňový programovací jazyk, který byl navržen v roce 1991. Je dynamicky interpretovaný, tzn. že kód je překládán až při běhu programu. Samotné jádro jazyka Python je implementováno v jazyce C.

Python podporuje procedurální, funkcionální a objektové paradigmata. Dále je jazyk dynamicky typovaný s automatickou správou paměti. V době tvorby této práce byly dostupné dvě aktuální verze, a to 2.7.16 a 3.7.2. Ty jsou však vzájemně nekompatibilní a pro konverzi projektů je třeba značných změn v syntaxi. Pro tvorbu projektu byla použita verze 3.6 s možností použití i ve verzi 3.7. Python je hojně používán v oblasti datové analytiky, strojového učení, vývoje serverových aplikací a dalších. Z důvodu velké oblíbenosti a poměrně jednoduché syntaxe je velice oblíbený v akademické sféře, proto byl zvolen i pro tvorbu tohoto projektu. [22, 23]

4.2.1 JUPYTER NOTEBOOK

Projekt Jupyter je neziskový open-source projekt vzniklý v roce 2014 z původního IPython projektu. Nástroj Jupyter Notebook je webovou aplikací typu klient server a tedy rozhraním mezi konzolovou verzí jazyka Python a webovým editorem. V interaktivním prostředí uživatel tvoří kód, u kterého má možnost okamžité kontroly výstupu. Zdrojové kódy jsou během tvorby ukládány do samostatných souborů a mohou tak být znovu využity a dále sdíleny. Instalace může být nasazena v lokálním prostředí případně na vzdáleném serveru viz. Google Colaboratory a IBM Watson Studio. Projekt podporuje přes čtyřicet programovacích jazyků včetně velice oblíbeného jazyka R. [24]

5 PROJEKTOVÁ PŘÍPRAVA

V následující kapitole je shrnuta potřebná projektová příprava skládající se ze stanovení cílů projektu, objektového návrhu celé aplikace, studiua dokumentace REST API webové aplikace Palindrome analyser a G4 hunter. V neposlední řadě se autor věnuje návrhovým vzorům použitých při vývoji.

5.1 SPECIFIKACE PROJEKTU

Software musí být navržen vzhledem k uživatelům, kteří jsou převážně pracovníci bez patřičné programátorské kvalifikace. Proto musí být navrhnut tak, aby spouštění zvolených analýz probíhalo co nejvíce automaticky bez nutnosti vnějších zásahů.

5.1.1 FUNKČNÍ POŽADAVKY

- Přihlášení se na uživatelský účet
- Výpis všech nahraných sekvencí
- Možnost nahrání sekvencí z NCBI databáze nebo textového souboru
- Funkce hromadného zadání analýz nad skupinou sekvencí
- Výpis všech hotových analýz
- Zobrazení výsledků a jejich případný export

5.1.2 NEFUNKČNÍ POŽADAVKY

- Vysoký stupeň automatizace
- Rozšířitelnost o další analýzy
- Udržitelnost – vhodné rozdělení do komponent

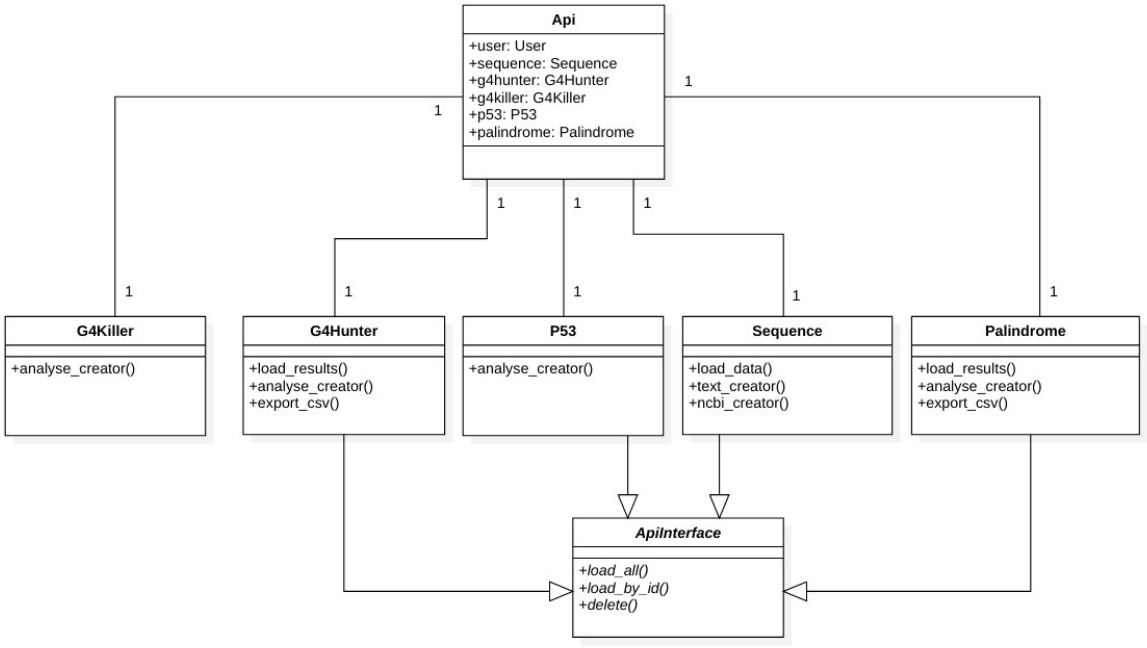
5.2 APLIKACE API WRAPPER

Cílem aplikace je, jak už bylo zmíněno v zadání, umožnit uživateli jistou míru automatizace při zadávání jednotlivých analýz lokálních struktur. Z tohoto důvodu byla vytvořena aplikace typu API wrapper. Komunikace a zpracování dat je prováděno automaticky a uživatel interaguje s aplikací pomocí volání metod centrálního objektu, který si sám vytvoří v programovacím prostředí. Předávání dat probíhá pomocí tabulkových objektů z knihovny pandas.

5.2.1 NÁVRHOVÝ VZOR FACADE PATTERN

Facade pattern je návrhový vzor používaný v oblasti objektového programování. Vzor obsahuje centrální objekt zprostředkovávající unifikované rozhraní pro celou skupinu tříd. Centralizací se sníží počet tříd jednotlivých subsystémů, se kterými musí uživatel interagovat. Zvolený návrhový vzor urychluje proces učení se daného softwaru, avšak omezuje jeho flexibilitu. Konkrétní použití je zobrazeno na obr. 8.

V projektu je komunikace s uživatelem zprostředkovávána pomocí centrálního objektu Api, který umožňuje volání jednotlivých metod a interakci s uživatelem pomocí knihovny pandas. Tato knihovna je součástí většiny standardních nástrojů v oblasti strojového učení a zpracování dat. [25]



Obr 8: UML diagram návrhového vzoru facade pattern

5.2.2 DOKUMENTACE API (SWAGGER)

Aplikace Palindrome analyser a G4hunter mají své REST API rozhraní dokumentované pomocí open source balíku Swagger. Tento balík obsahuje celou řadu nástrojů pro tvorbu standardizované dokumentace, testování a vizualizaci ve webovém rozhraní. Jedná se o hojně využívaný nástroj především v oblasti vývoje tzv. RESTful services. [20, 21] Dokumentace je provedena formou HTML stránky s možnosti přímé komunikace s danou aplikací. Uživatel je tak schopen testovat cílové endpointy a zároveň i uživatelsky testovat chování RESTového rozhraní.

V dokumentaci jsou pro jednotlivé endpointy specifikovány tyto parametry:

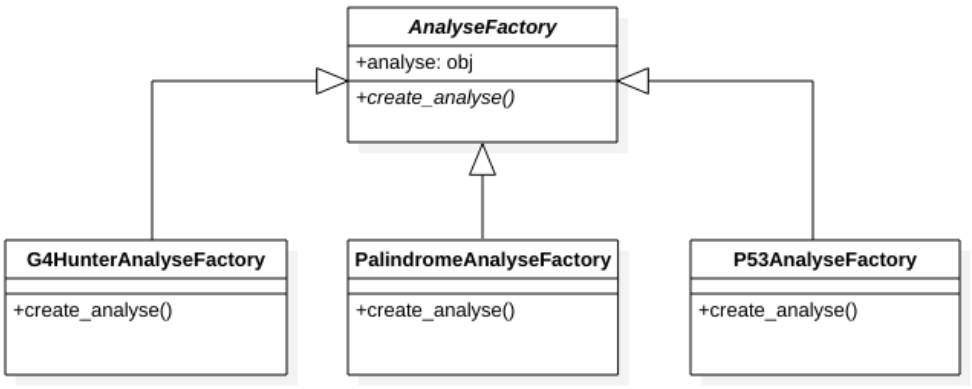
- HTTP metoda GET, POST, PUT apod.
- Adresa endpointu např. *api*/analyse/g4hunter
- Parametry endpointu, JSON formát zprávy
- Formát odpovědi, response content type, status code apod.

5.3 OBJEKTOVÝ NÁVRH PROJEKTU

Struktura programu je složena ze dvou hlavních komponent. Caller objektů přímo se dotazujících API a interface objektů, které zprostředkovávají rozhraní pro použití knihovny pandas. Celá tato struktura je uzavřena do třídy Api, jak již bylo zmíněno výše.

5.3.1 NÁVRHOVÝ VZOR FACTORY METHOD PATTERN

Z důvodu větší flexibility byl pro konstrukci tříd volajících jednotlivé přístupové body použit objektový návrhový vzor factory method pattern. Tento návrhový vzor je použit pro tvorbu objektů, ze kterých je v části interface vytvořen výsledný pandas dataframe. Princip spočívá v delegaci tvorby instance třídy konstruktorem na jinou metodu tzv. Factory method, která dle algoritmu rozhoduje o tom, jaký objekt třídy bude vytvořen. Na UML diagramu níže lze spatřit objektový návrh jednotlivých factory class pro analýzy. Stejný postup byl zvolen i při vytváření objektů sekvencí viz obr. 9. [25]



Obr 9: UML diagram návrhového vzoru factory method

5.4 DIAGRAM AKTIVIT – SPUŠTĚNÍ ANALÝZY

Pro spuštění nové analýzy musí být uživatel řádně přihlášen ke svému uživatelskému účtu, případně může využít hostitelský účet, kde však nemá dostupné dříve nahrané sekvence.

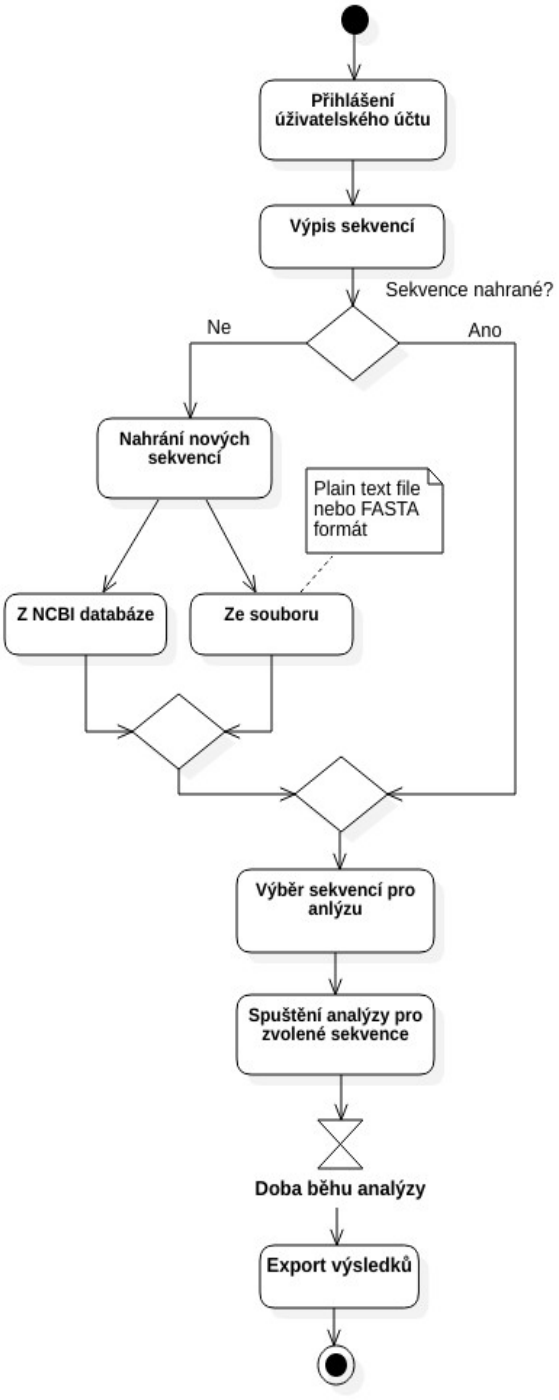
Základní tok:

1. Uživatel vytvoří instanci objektu Api a provede přihlášení
2. Ze seznamu nahraných sekvencí vybere vhodné pro analýzu
3. Zvolené sekvence vloží jako jeden z parametrů do metody sekvence
4. Uživatel spustí analýzu, průběžně je o stavu zpravován stavovým řádkem
5. Ze seznamu dokončených analýz vybere žádané analýzy
6. Zvolí možnost zobrazení dat případně export do csv souboru

Alternativní tok:

2.2. Pokud nejsou sekvence dostupné, zvolí jednu z nabízených možností importu nových sekvencí

Na následujícím diagramu aktivit je shrnut celý proces vypracování jednoduché analýzy nad vzorkem vybraných sekvencí viz obr.10.



Obr 10: Sekvenční diagram průběhu analýzy

5.5 ANALÝZA SOUČASNÉHO STAVU REST API

V této podkapitole jsou shrnuty jednotlivé endpointy RESTového rozhraní, které byly použity při tvorbě. Dokumentace je dostupná na adrese http://bioinformatics.ibp.cz:8888/swagger-ui.html#/, kde jsou uvedeny všechny požadavky pro provedení dotazů nad zvolenými funkcemi. Dokumentace je rozdělena do čtyř hlavních kategorií:

- /api/analyse
- /api/sequence
- /api/user
- /api/jwt-controller

Funkce pod záštitou adresy /api/user jsou obecně používané pro registraci a odstranění uživatele. Dále funkce obsahují možnost změny či resetování hesla. Tato funkcionalita však nemá být součástí aplikace api wrapper, a proto tyto endpointy nejsou v projekty využity. V prvním kroku analýzy je nutné umožnit přihlášení do uživatelského účtu. Tzn. je nutné obstarat JSON Web Token, který slouží pro autentizaci a autorizaci uživatele. Tyto funkce se nachází pod endpointem /api/jwt-controller. Dvě potřebná rozhraní zobrazuje tab. 5. [17]

Tab 5: Přístupové body pro obsluhu JSON web tokenů

ADRESA	METODA	FUNKCE
/api/jwt	PUT	Vytvoření tokenu pro registrovaného uživatele.
/api/jwt	POST	Vytvoření tokenu pro host uživatele.

V druhém následném kroku po autentizaci uživatele je nutné nahrát zvolené sekvence. Funkce jsou dosažitelné na adrese /api/sequence. Rozhraní obsahuje tři možné scénáře pro nahrání sekvence, dále funkcionalitu pro zobrazení všech či jedné zvolené sekvence. Možnosti upravování tzv. tagů nebudou implementovány, protože nejsou pro funkčnost důležité. Nahrání sekvence jako souboru rovněž nebude implementováno, protože se úzce překrývá s textovým vstupem. Nahrání souboru tak bude řešeno programově otevřením daného souboru. Poslední zmiňovaný endpoint slouží pro vypsání dat zvolené sekvence. Zpracovávaná rozhraní jsou zobrazena v tab. 6. [17]

Tab 6: Přístupové body pro obsluhu sekvencí (výběr)

ADRESA	METODA	FUNKCE
/api/sequence	GET	Kompletní výpis sekvencí.
/api/sequence/id	GET	Výpis sekvence dle id.
/api/sequence/id	DELETE	Smazání sekvence dle id.
/api/sequence/import/ncbi	POST	Import z NCBI databáze.
/api/sequence/import/text	POST	Import z textového řetězce.

V poslední etapě analýzy je nutné její spuštění nad zvolenou sekvencí. Tyto funkce jsou dostupné pod adresou /api/analyse/. Vzhledem k nastavené politice unifikace jednotlivých analýz je patrné, že přístupové body pro g4hunter i palindrome analyser budou identické. Proto v následující tabulce jsou uvedeny adresy bez specifikovaného druhu analýzy viz. tab. 7. [17]

Tab 7: Obecný vzor přístupových bodů pro obsluhu analýz

ADRESA	METODA	FUNKCE
/api/analyse/{funkce}	GET	Kompletní výpis sekvencí.
/api/analyse/{funkce}	POST	Vytvoření analýzy.
/api/analyse/{funkce}/id	DELETE	Smazání analýzy dle id.
/api/analyse/{funkce}/id	GET	Výpis analýz dle id.

Funkce jako je export do souboru csv jsou specifické pro konkrétní analýzy a nejsou zde dále specifikovány. Poslední dostupné funkce této aplikace jsou g4killer a p53 predictor. V případě g4killeru se jedná o funkci pro snižování tzv. gscore v předložené sekvenci. V případě funkce p53 predictor je počítána hodnota afinity k tomuto proteinu. Ve Swagger dokumentaci je specifikován pro tuto funkci pouze jediný endpoint vykonávající kompletní funkčnost viz. tab. 8. [17]

Tab 8: Přístupové body pro g4killer

ADRESA	METODA	FUNKCE
/api/analyse/g4killer	POST	Spuštění g4killer nad sekvencí.
/api/analyse/p53predictor	POST	Spuštění p53pred. nad sekvencí.

Dalším možným výstupem z aplikace jsou tzv. heatmapy. Jedná se o vizualizace výskytů jednotlivých lokálních struktur v celé sekvenci DNA. Heatmapy je možné získávat ve dvou variantách a s daty o počtech výskytů lokálních struktur v segmentu sekvence, případně s procentuálním vyjádřením vzhledem k délce segmentu. Data jsou v aktuální verzi přístupná pouze pro endpoint g4hunter viz. Tab 9. [17]

Tab 9: Přístupový bod pro získání heatmapy

ADRESA	METODA	FUNKCE
/api/analyse/{funkce}/{id}/heatmap	GET	Získání dat heatmap.

6 REALIZACE PROJEKTU

Tato kapitola se zabývá konečnou realizací projektu. V první části popisuje zvolené knihovny s krátkým popisem a případným zdůvodněním jejich výběru. V druhé polovině kapitoly je ilustrován příklad použití a popisuje se zde zvolené rozhraní aplikace.

6.1 KNIHOVNY V PRODUKČNÍ ČÁSTI A PIPENV

Při realizaci byly použity i některé externí knihovny. Ty byly voleny z důvodu zlepšení udržitelnosti kódu i větší bezpečnosti. Pro jednodušší správu knihoven byl vybrán nástroj pipenv. Tento nástroj automaticky s tvorbou projektu vytváří své vlastní izolované virtuální prostředí, spravuje kompletní strom závislostí jednotlivých knihoven a nabízí další nástroje pro obsluhu knihoven. Nástroj dále umožňuje i bezpečnostní kontrolu nainstalovaných balíků a v případě zjištěných známých bezpečnostních hrozeb nabízí konkrétní řešení.

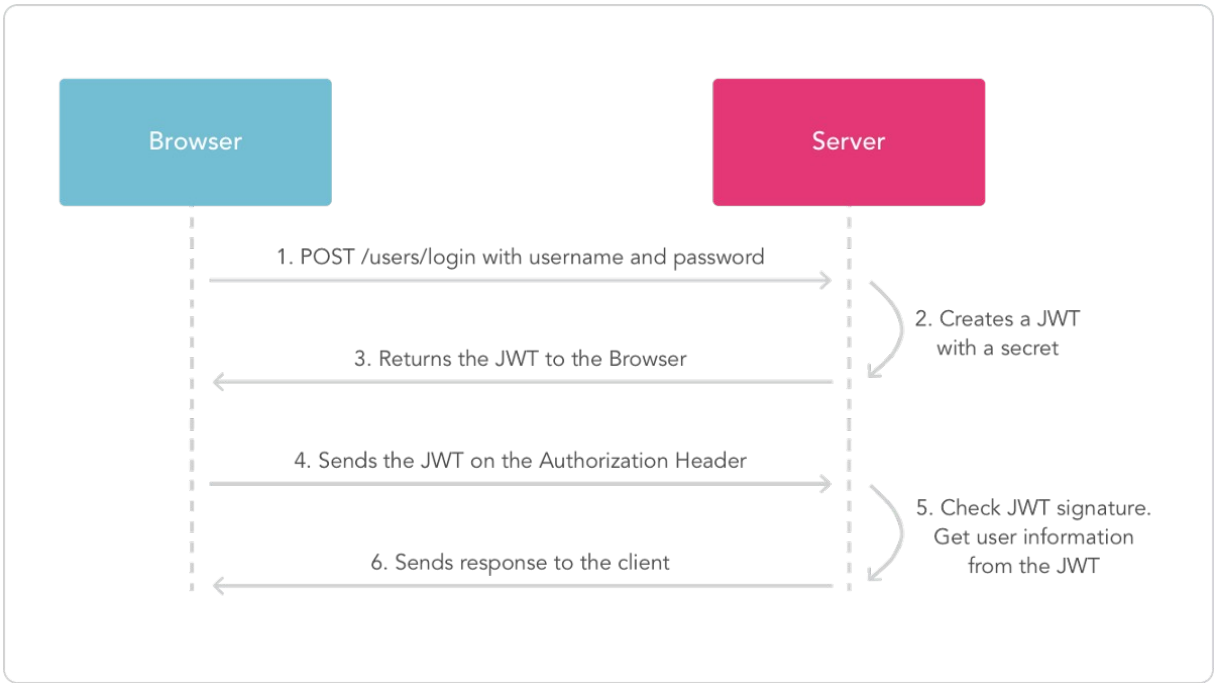
6.1.1 REQUESTS A REQUESTS TOOLBELT

V projektu se především operuje s funkcemi nad HTTP protokolem. Zejména v otázce komunikace s REST API aplikace. Z tohoto důvodu bylo nutno zvolit některou z možných knihoven realizujících tyto funkce. Jak již je uvedeno v názvu, byla vybrána knihovna requests. Requests byla zvolena z důvodu její jednoduché použitelnosti, dále z důvodu dobré dokumentace a snadného použití. Druhá zmíněná knihovna nabízí implementaci některých funkcí, které nejsou obsaženy v původní implementaci requests. V projektu byla použita pouze jediná funkce, a to MultipartEncoder pro odesílání souborů sekvencí na API.

6.1.2 PYJWT A JSON WEB TOKEN

JSON Web Token (JWT) je otevřený průmyslový standart zdokumentovaný v RFC7519. Informace jsou zakódovány ve formě JSON objektu, který je poté přenášen v textové formě v hlavičce HTTP operace. Po přihlášení uživatele na straně API se vytvoří tzv. token. Ten obsahuje informace o uživateli, čas vytvoření a čas expirace. Samotný objekt je pak podepsán dvojicí klíčů, veřejným a privátním. Z toho vyplývá, že pouze strana API může tento token vytvořit, a tedy i podepsat privátním klíčem. Strana API pak vykonává příkazy pouze pro uživatele, kteří předložili podepsané tokeny, které neprošly datem expirace. [26]

Z důvodu nutnosti autorizace a získávání základních dat o uživateli a obsluhu JSON web tokenů byla vybrána knihovna PYJWT. Tato knihovna podporuje obě strany komunikace, a to jak produkční, která vlastní privátní klíč, tak i konzumentskou, která se pomocí tokenu autorizuje. V projektu je využita pouze část konzumní, kdy se jednotlivé příkazy podepisují tímto objektem. Schéma komunikace pomocí JSON web tokenu je vyobrazeno na obr. 11.



Obr 11: Schéma principu funkce JSON web tokenu [26]

6.1.3 PANDAS

Pandas je, jak již bylo zmíněno v předchozí kapitole, hlavním komunikačním kanálem mezi uživatelem a API. Jedná se o open source nástroj poskytující vysokou výkonost, flexibilitu a jednoduché použití v aplikacích zaměřených na datovou analytiku. Jak již bylo zmíněno výše, knihovna pandas byla zvolena z důvodu jednoduchého použití a přehledné interpretace ve formě tabulkových hodnot.

6.1.4 MATPLOTLIB

Knihovna Matplotlib je zaintegrována do většiny dostupných nástrojů z oblasti datové analytiky. V případě použitých produkčních knihoven je úzká návaznost na předchozí zmíněné tabulkové objekty nástroje pandas. Matplotlib je v projektu použit za účelem vytváření dvourozměrných grafických interpretací, například tzv. heatmap při analýzách G4hunter. Knihovna podporuje práci v interaktivním prostředí Jupyter notebook a je dobře zdokumentována. Z těchto důvodů byla také zvolena při finální realizaci projektu.

6.1.5 TQDM

Knihovna tqdm nemá přímý vliv na funkčnost celého projektu. Z procesu zpracování sekvence vyplývá nutnost upozornit uživatele na probíhající proces výpočtu analýzy a na úroveň dokončení. Z tohoto důvodu volba padla na vizualizaci pomocí stavového řádku. Nástroj tqdm umožňuje v prostředí Jupyter notebook konfigurovat a pomocí programové logiky nastavovat chování tohoto indikačního prvku. Volba je zdůvodněna pouze z estetického hlediska a nemá na konečnou funkčnost velký vliv.

6.2 KNIHOVNY POUŽITÉ PŘI VÝVOJI

V několika následujících odstavcích se práce věnuje knihovnám, které byly použity pouze při vývoji. Tyto nástroje jsou dostupné pro instalaci pouze v repositáři aplikace a nejsou tak uvolňovány společně s produkční verzí. Všechny zmíněné nástroje slouží pro zlepšení kvality a spolehlivosti kódu aplikace.

6.2.1 PYTEST

Pro vytváření automatických unit testů padla volba na knihovnu pytest. Ta podporuje vytváření parametrizovaných testů a je dobře integrována do vývojového prostředí Pycharm, které bylo využito při tvorbě projektu. Mezi další výhody knihovny patří automatická lokalizace souborů s testy pro automatické spouštění, detailní výpis chyb právě spuštěných testů a dobře zpracovaná dokumentace.

6.2.2 VCR.py

Spuštěné automatické testy sledují hodnoty dat, která jsou navrácena po provedeném volání aplikace ibp analyser. Testy proto musí být napsány tak, aby nebyly příliš rychlé vzhledem k RESTovému rozhraní a neprovedly změny bez patřičného navrácení původního stavu. Z důvodu neměnnosti navrácených dat byla zvolena technika tzv. server mocking. Jedná se o techniku, kdy je pomocí objektu simulováno chování reálného objektu. V tomto případě se volaný objekt chová jako samotné API a na volané příkazy odpovídá stejným způsobem. To umožňuje rychlejší odbavení jednotlivých testů a zároveň neovlivní stav skutečné aplikace.

Z důvodu velkého množství volání byla zvolena knihovna VCR.py. Tato knihovna oproti konkurenčním dokáže automaticky zaznamenat síťový provoz do tzv. kazet a v případě dalšího spuštění namísto skutečného volání předložit výsledek uložený na kazetě. Kazety se skládají z jednoduchých JSON objektů, které jsou pak uloženy ve složce s testovacími skripty. Díky této technice se doba testování zkrátila z desítek sekund na sekundu jednu.

6.2.3 BLACK

Black je nástroj pro automatické formátování zdrojového kódu podle standartu PEP8. Dokument PEP8 a PEP257 stanovují konvence pro psaní zdrojového kódu aplikací a jejich dokumentace. Směrnice definuje použité styly psaní komentářů, definici odsazování, použité kódování zdrojových souborů a mnoho dalšího. Obecně v python komunitě existuje úzus na užívání těchto doporučení. Takto naformátovaný kód aplikace vypadá ve všech zdrojových souborech stejně a tím se i zlepšuje udržitelnost aplikace pro další vývoj. [22] Formátování nástrojem black bylo stejně jako u předchozího TQDM voleno pouze z estetických důvodů.

6.2.4 PYLINT A COALA BEARS

Tyto dva poslední zmiňované nástroje patří do kategorie tzv. linterů. Ty pomocí statické analýzy zdrojového kódu aplikace dokážou vyhodnotit úroveň komplexnosti kódu, porušení doporučovaných stylistických pravidel viz PEP8, vyhodnocení chyb, podezřelých konstruktů a doporučit řešení jednotlivých problémů. I přes částečné překrytí byly zvoleny tyto dva nástroje z důvodu vhodného vzájemného se doplňování.

6.3 DISTRIBUCE APLIKACE

Jak je uvedeno v zadání diplomové práce, aplikace má být dostupná pro instalaci pomocí instalátoru balíčků pip. Aby mohla být aplikace takto distribuována, je ji nutné uveřejnit v repositáři Python Package Index (Pypi). Jedná se o veřejně dostupný zdroj obsahující více než sto sedmdesát tisíc balíků k datu tvorby práce. Před nahráním projektu je nutné sestavit soubor setup.py. [22]

```
setup(
     name="dna_analyser_ibp",
     version="1.4",
     description="DNA analyser API wrapper tool for Jupiter notebooks.",
     long_description=long_description,
     long_description_content_type="text/markdown",
     author="Patrik Kaura",
     author_email="160702@vutbr.cz",
     keywords="dna, ibp, quadruplex, g4hunter, g4killer, palindrome, p53"
     license="GPLv3",
     url="https://gitlab.com/PatrikKaura/DNA_analyser_IBP/",
     packages=find_packages(),
     install_requires=["requests", "pandas", "tqdm",
                       "pyjwt", "matplotlib", "requests_toolbelt"],
     classifiers=[
       "Intended Audience :: Education",
       "Intended Audience :: Science/Research",
       "License :: OSI Approved :: GNU General Public License v3 (GPLv3)",
       "Operating System :: OS Independent",
       "Programming Language :: Python",
       "Programming Language :: Python :: 3",
       "Programming Language :: Python :: 3.6",
       "Programming Language :: Python :: 3.7",
       "Topic :: Scientific/Engineering :: Bio-Informatics",
     ],
     zip_safe=False
 )
```

Do souboru setup je nutno, jak je patrné z ukázky výše, vyplnit základní popis projektu jako je jméno, náhledový popis, dlouhý popis, tj. soubor README.MD, licenci a další parametry. Nutno je také z důvodu jednodušší vyhledatelnosti v seznamu repositáře vyplnit tzv. klasifikátory. V neposlední řadě je vyplněn údaj o přenášených závislostech viz. parametr install_requires. Ten lze nastavit ručně metodou zvolenou v tomto případě nebo pomocí requirement souboru. Soubor však nebyl využit z důvodu odlišné práce se závislostmi viz. podkapitola zabývající se nástrojem pipenv.

Pro nahrání projektu bylo využito nástroje twine. Tento jednoduchý nástroj umožňuje zaregistrovat, pomocí souboru setup zabalit a odeslat novou verzi souboru do repositáře. Celý balík je tak dostupný na adrese https://pypi.org/project/dna-analyser-ibp/ a to v aktuální verzi 1.4, která je poslední k datu tvorby dokumentu.

6.3.1 VOLBA LICENCE PRODUKTU

Společně s distribucí vyvstává otázka s volbou licence, kterou je projekt zaštítěn v open-source repositáři python balíků pypi. Volba byla usnadněna pomocí rozcestníku na webové stránce https://choosealicense.com/. Tento webový zdroj zprostředkovává formou jednoduchého rozcestníku informace o aktuálně používaných licencích. Z důvodu úzké návaznosti na vývoj s API byla zvolena forma licence, u které všechny odvozené projekty musejí zveřejňovat své zdrojové soubory. Zvolena tedy byla licence GNU General Public License v3.0 ve znění z roku 2007. Pro zjednodušení následující tab. 10 popisuje práva a limity v používání softwaru.

Tab 10: Tabulka obsahující parametry licence GPL-3.0 [27]

POVOLENO	Komerční užití
	Distribuce
	Modifikace
	Patentové užití
	Privátní užití
PODMÍNĚNO	Otevřený zdrojový kód
	Nutná zmínka původní licence
	Publikování pod stejnou licencí
	Dokumentace změn oproti originálu
NEGARANTOVÁNO	Spolehlivost
	Záruky

6.4 UKÁZKA JEDNODUCHÉ ANALÝZY G4HUNTER

V této podkapitole je popsán postup vytvoření jednoduché analýzy guaninového kvadruplexu nad nahranou sekvencí chromozomu Homo sapiens GRCh38.p12. V prvním kroku je nutné nainstalovat balíček dna-analyser-ibp z repositáře pypi, a to pomocí jednoho z následujících příkazů.

```
pipenv install dna-analyser-ibp
```

```
pip install dna-analyser-ibp
```

V druhém kroku je importována a vytvořena instance objektu Api. Při její inicializaci jsou vyplněny uživatelské údaje. Objekt umožňuje změnu serveru, na který se požadavky budou odesílat. V aktuální verzi je výchozí hodnota nastavena na adresu lokální instance serveru. Problematika bezpečného zadávání přihlašovacích údajů je shrnuta v kapitole 5.4.

```
from DNA_analyser_IBP.api import Api

API = Api()
```

```
from DNA_analyser_IBP.api import Api

API = Api(server='http://hostname:port/api')
```

Ve třetí etapě procesu je nutné nahrát požadovanou sekvenci. Jak již bylo zmíněno na začátku podkapitoly, bude nahrán chromozom Homo sapiens. Tento vzorek nahrajeme z databáze NCBI, a to z důvodu jednodušší manipulace.

```
API.sequence.ncbi_creator(circular=True,
           tags=['Homo','sapiens', 'chromosome'],
           name='Homo sapiens chromosome 12',
           ncbi_id='NC_000012.12')
```

```
Sequence Homo sapiens chromosome 12 uploading: 100 % uploaded
```

V předposledním kroku uložíme tabulkový objekt sekvence do proměnné, kterou následně poskytneme metodě vytvářející analýzu. V metodě dále vyplníme hodnoty nutné pro spuštění analýzy, a to threshold a window size.

```
sapiens_sequence = API.sequence.load_all(filter_tag='Homo')

API.g4hunter.analyse_creator(sequence=sapiens_sequence,
                 tags=['test','Homo','sapiens'],
```

```
                    threshold=1.4, window_size=30)
```

```
   Analyse Homo sapiens chromosome 12 processing: 100 % processed
```

Poslední krok exportu je obdobný se spuštěním g4 hunter analýzy. Stejně jako v tomto kroku uložíme výsledek analýzy (tabulkový objekt pandas) do proměnné a tu poté poskytneme metodě zobrazující výsledky. Z důvodu nutnosti omezení vstupu na jednu konkrétní analýzu je nutné pomocí příkazu iloc převést tabulku (dataframe) na objekt sloupce (series).

```
  sapiens_result = API.g4hunter.load_all(filter_tag=['Homo'])

  API.g4hunter.load_results(
                    g4hunter_analyse=sapiens_result.iloc[0])
```

Po zavolání příkazu load_results se uživateli zobrazí výsledky G4 analýzy viz tab. 11.

Tab 11: Tabulka s výstupem analýzy g4hunter

ID	POSITION	LENGTH	SCORE	ABS_SCORE	SEQUENCE	SUB_SCORE
1	57258	28	1.321	1.321	TTTGGGGAATGACATTTTATGGGGGAGA	1.44,1.44, 1.48,1.48
2	63308	28	1.321	1.321	GAGATGGGGTTGATCTTAGGGAGGGAAG	1.44,1.4,1 .4,1.4
3	63313	25	1.399	1.399	GGGGTTGATCTTAGGGAGGGAAGCG	1.4
4	63694	27	-1.296	1.296	GCCAATACCCCAGTATGTCCCCCAGTT	-1.4,- 1.48,-1.4
5	72097	25	-1.440	1.440	CCAAAACCCTAGTTAATATCCCCCC	-1.44
6	82858	25	-1.440	1.440	CCAAAACCCTAGCTGATATCCCCCC	-1.44
7	84108	29	1.275	1.275	CAAGGGTGTGTGGGGCCGTGGGAGGCATG	1.44,1.44, 1.44,1.44

Následný export výsledků pro analýzu překrytí s původní anotovanou sekvencí viz kapitola 6 je možné jednoduše provést příkazem csv_export. Uživatel však musí této funkci poskytnout cestu do cílového adresáře, kde mají být data uložena.

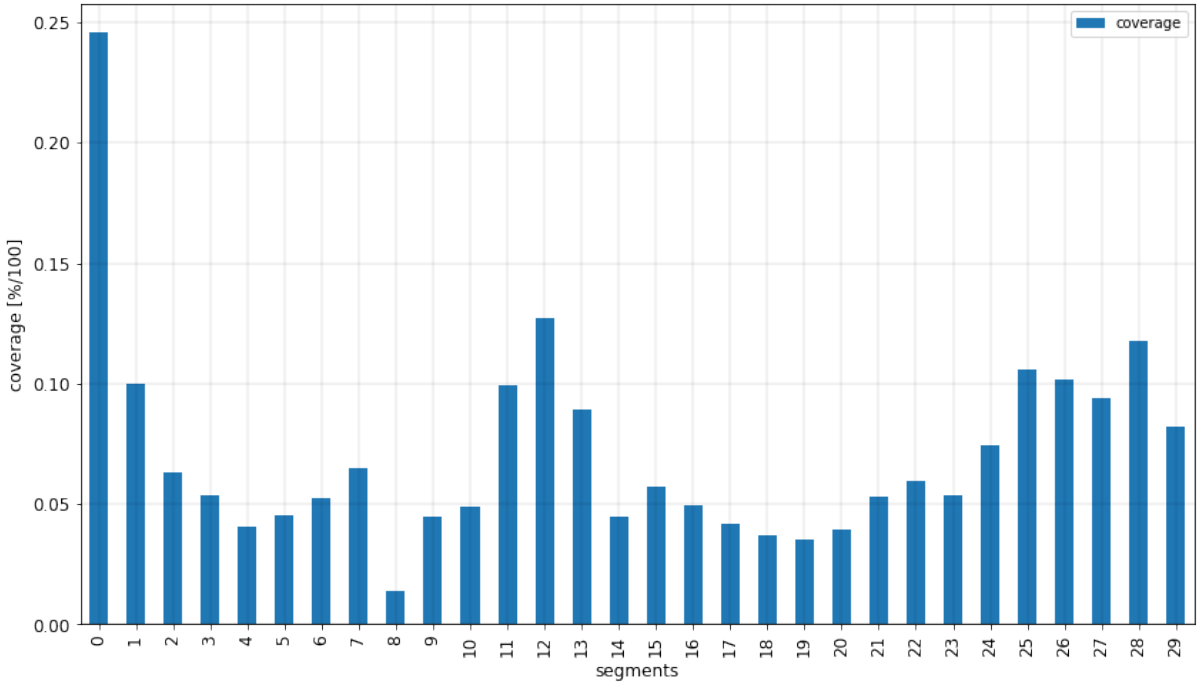
```
  API.g4hunter.export_csv(
       g4hunter_pandas=sapiens_result,
       out_path='/path/to/csv/result/dir')
```

```
File created → /path/to/csv/result/dir/Homo_sapiens_chromosome_12_{id}.csv
```

6.4.1 UKÁZKA G4HUNTER HEATMAP

Společně s výsledky analýzy vzniká i grafická interpretace výsledků tzv. heatmapa. Ta reprezentuje oblasti s výskytem guaninových kvadruplexů dvěma způsoby. Celkovým počtem kvadruplexů v daném segmentu a procentuálním vyjádřením vůči počtu bází v segmentu. Výsledek je graficky prezentován pomocí knihovny matplotlib, kterou pro vizualizace dat z tabulkových proměnných je nutné doinstalovat viz kapitola 5. Níže je uvedena ukázka volání s výslednými grafy viz obr. 12.

```
API.g4hunter.load_heatmap(
                    g4hunter_analyse=sapiens_result.iloc[0],
                    coverage=True, segment_count=31)
```



Obr 12: Graf zobrazující frekvenční rozložení výskytů kvadruplexů

6.5 PROBLEMATIKA AUTENTIZACE UŽIVATELE

V druhé etapě analýzy existuje požadavek na získání uživatelských přihlašovacích údajů. Problém nastává při jejich zpracování. Protože je stanoven předpoklad na přenositelnost jednotlivých zdrojových souborů mezi uživateli, bylo nutno vyřešit problém s neustálým mazáním těchto údajů před samotným sdílením. Kdyby tyto údaje byly součástí inicializační metody objektu Api, musel by uživatel před každým sdílením smazat část dat vkládaných do

této metody. Proto jsou tato data získávána jinak, a to pomocí vstupní metody input v případě uživatelského jména. A pomocí standardní knihovny getpass je získán údaj s heslem k uživatelskému účtu. Přičemž uživatelské jméno není citlivý údaj, proto je při psaní zobrazováno v otevřené formě. Naopak heslo je již při vkládání skryto a není jej možno poté zkopírovat ze sdíleného souboru analýzy. Údaje jsou obstarávány ve zdrojovém kódu viz ukázka níže.

```
from DNA_analyser_IBP.api import Api

API = Api(server='http://hostname:port/api')
```

```
Enter your email        user@example.cz
Enter your password     ........
User user@example.cz logged in: 2019-03-30 19:46:56.661376
```

6.6 UKÁZKA NÁSTROJE G4KILLER

Nástroj G4killer je volán pouze pomocí jediné funkce objektu Api. Vstupem do této funkce není sekvence jako v případě standardní analýzy, ale pouze textový řetězec, nad kterým je spočteno nové skóre. Dále je nutné zvolit cílenou hodnotu gscore, která bude pomocí nástroje dosažena. Stejně jako v předešlé ukázce je nutné přihlášení uživatele viz příklad výše. Ukázka volání metody a zobrazený výsledek lze yhlédnout níže viz tab. 12.

```
API.g4killer.analyse_creator(
            origin_sequence='AATTATTTGGAAAGGGGGGGTTTTCCGA',
            threshold=0.5)
```

Tab 12: Tabulka s výsledky nástroje g4killer

origin_score	1.03571
target_threshold	0.5
origin_sequence	AATTATTTGGAAAGGGGGGGTTTTCCGA
mutation_sequence	AATTATTTGGAAAGGGWGWGTTTTCCGA
mutation_score	0.428571

Tabulková hodnota mutation_sequence a mutation_score ukazují výslednou sekvenci s hodnotou skóre co nejbližší požadované hodnotě. Hodnota origin_score a origin_sequence pak zobrazují informaci o stavu před spuštěním výpočtu. Jednotlivé změny v sekvenci genomu jsou nahrazeny písmenem W.

6.7 UKÁZKA NÁSTROJE P53PREDICTOR

P53predictor je rovněž jako předchozí zmíněný nástroj volán pouze pomocí jednoho příkazu. Stejně jako G4killer pracuje s textovým řetězcem sekvence, avšak pouze o fixní délce 20 bází.

```
API.p53.analyse_creator(sequence='GGACATGCCCGGGAATGTCC')
```

Výsledkem je pak velice podobná tabulková proměnná interpretující výsledky analýzy uživateli. Výsledky spuštěného nástroje nad ilustrační sekvencí zobrazuje tabulka 13.

Tab 13: Tabulka s výsledky nástroje p53 predictor

position	0
length	20
difference	-6.99
predictor	0.62
affinity	0.91
sequence	GGACATGCCCGGGAATGTCC

Hodnota position a length značí začátek a konec hodnocené sekvence. Tato data jsou však nejvíce relevantní při zpracování většího množství vzorků, které není v době publikace k dispozici. Proměnné difference a predictor informuje o vypočtené afinitě a rozdílu od ideální hodnoty -7.61 viz kapitola 3. Afinita k vazbě s proteinem p53 je zde vyjádřena hodnotou affinity s maximální hodnotou 1 tzn. 100%.

7 ANALÝZY LOKÁLNÍCH STRUKTUR

Pro ověření funkčnosti aplikace byly zpracovány analýzy lokálních struktur DNA především v oblasti výzkumu guaninových kvadruplexů. V první části je kapitola věnována problematice obstarávání DNA sekvencí a anotací vybraných organismů. V druhé obsáhlejší části je na konkrétních příkladech uvedena analýza jednotlivých souborů sekvencí.

7.1 DATABÁZE NCBI A ANOTACE SEKVENCE

National Center for Biotechnology Information (NCBI) je organizace mající za úkol poskytnout informace z oblasti genetiky od roku 1992. Organizace sdružuje odborné publikace, databázi genomů, proteinů a nukleotidů. V rámci projektu byla převážně využita možnost importování genomů právě z databáze NCBI. Rovněž databáze poskytuje k velkému množství genetických sekvencí tzv. anotace. Anotovaná místa sekvence jsou taková, kde se vyskytuje určitá významná struktura. Anotace je opatřena jménem, indexem začátku anotovaného místa, indexem konce anotované sekvence a dalšími popisnými informacemi. V následující ukázce lze zhlédnout zdrojový text s anotovanými místy viz. anotace Papaver somniferum cultivar HN1 chromosome 1 s ID NC_039358.1. [28]

```
>Feature ref|NC_039358.1|
92760 94313 gene
                  gene  LOC113324569
                  db_xref     GeneID:113324569
92760 94313 mRNA
                  product     beta-glucosidase 12-like
                  transcript_id     ref|XM_026572883.1|
                  gene  LOC113324569
                  db_xref     GeneID:113324569
92760 94313 CDS
                  product     beta-glucosidase 12-like
                  protein_id  ref|XP_026428668.1|
                  gene  LOC113324569
                  db_xref     GeneID:113324569
96586 98713 gene
                  gene  LOC113276494
                  db_xref     GeneID:113276494
97822 96586 mRNA
98713 97968
                  product     polyphenol oxidase, chloroplastic-like
                  transcript_id     ref|XM_026526103.1|
                  gene  LOC113276494
                  db_xref     GeneID:113276494
```

Dle webové stránky https://www.ncbi.nlm.nih.gov/projects/Sequin/table.html byly následně všechny soubory s informacemi o anotovaných místech rozloženy pro následné programové zpracování. Důležitá jsou pouze místa zdrojového textu, která začínají čísly případně symboly < a >. Poslední dva zmíněné symboly značí částečně nekompletní anotace. Poslední odchylkou od normálu jsou otočené indexy začátku a konce. Viz úmyslně přehozené indexy u poslední anotované oblasti mRNA. V případě tohoto přehození pořadí anotace vyjadřuje výskyt v komplementárním řetězci anotovaného genomu. [28, 29]

7.2 DATOVÝ FORMÁT FASTA

FASTA je datový formát užívaný v bioinformatice pro přenos nukleotidových nebo peptidových sekvencí. Každá báze je pak reprezentována jedním znakem definované znakové sady. Na začátku každého souboru FASTA je pak vyžadováno krátké popsání dané sekvence začínající znakem „>“. Doporučená maximální délka popisu činí 80 znaků, v praxi však může být i větší. Následující úsek textu ilustruje formát na konkrétní sekvenci Candidatus Heimdallarchaeota archaeon. [28]

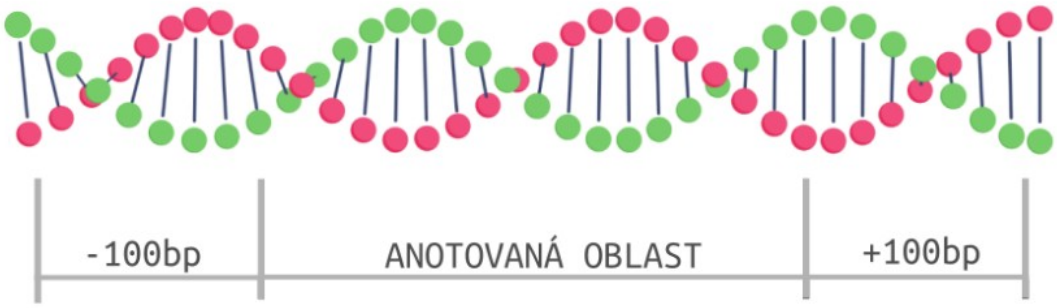
```
>MDVS01000001.1 Candidatus Heimdallarchaeota archaeon
TCTTAATTACAGCTTTACCTAATTCAAAAATAACTTTACTCATCCGCTCATAAACATAAGGGACTTTAAAAGCT
TTTTTACAATGATAACAATAATAACGACGGATTTCACCTGCTCGTGTCCAATACGAACCCCAAGTAGTAACAGG
```

7.3 POŽADAVKY ANALÝZY

Každá z níže uvedených analýz byla složena ze série několika neměnných kroků. V prvním nejdříve bylo nutné pomocí příkazu pro NCBI import nahrát sekvence z databáze NCBI. V kroku následujícím bylo nutné nad těmito analýzami spustit analytický nástroj G4hunter. Výsledná data bylo nutné exportovat do formátu csv a vůči staženým anotacím provést překrytí. Překrytím sekvence a anotace dostáváme informace o tom, ve kterých významných oblastech sekvencí se guaninové kvadruplexy vyskytují, a tato místa dále zkoumat.

7.4 POSTUP PŘEKRYTÍ

Postup překrytí je složen ze dvou samostatných částí. V první části je nutné rozložit soubor anotací a rozdělit ho do jednotlivých částí. V části překrývání je iterativně procházen seznam nalezených kvadruplexů, který se porovnává s iteračně procházeným seznamem anotací. Pro urychlení běhu programu se neprochází veškeré anotace, ale pouze ty, které jsou v určitém okolí. Pro každý kvadruplex je pak vyhodnocováno, jestli spadá do oblasti uvnitř anotace, případně blízkého okolí. Okolí je definováno jako -100bp před anotovanou oblastí a +100bp za anotovanou oblastí. Pokud algoritmus vyhodnotí, že kvadruplex spadá do jedné ze tří oblastí, zvedne inkrementální počítadlo o +1. Zároveň probíhá detekce typu anotace, tato informace je důležitá pro následný výzkum. Ukázka překrývané oblasti viz obr. 13. [16]



Obr 13: Schéma anotované oblasti s definovaným okolím ±100bp

7.5 CÍL TESTOVANÝCH ANALÝZ

Následující série podkapitol pojednává o zpracovávaných analýzách ve spolupráci s Biofyzikálním ústavem AV ČR. Výsledky jednotlivých analýz zde nejsou interpretovány, protože to není účelem této práce. Zpracované soubory sekvencí ilustrují praktickou použitelnost nad reálnými daty. Testovány byly čtyři nezávislé případy, které ilustrují případy uvedené v tab. 14.

Počet sekvencí je rozdělen do těchto kategorií:

- malý (desítky sekvencí)
- střední (stovky sekvencí)
- velký (tisíce sekvencí)

Datové náročnosti jsou v těchto třech kategoriích:

- malá (jednotky MB)
- střední (stovky MB)
- velká (jednotky GB)

Tab 14: Srovnání analyzovaných sekvencí

NÁZEV	POČET SEKVENCÍ	DATOVÁ NÁROČNOST
Mák setý	malý	velká
Kvasinka pivní	malý	malá
Prokariotyčtí / Bacteria	velký	velká
Prokariotyčtí / Archaea	střední	střední
Prok. / Archaea (rozšířeno)	velký	velká

7.5.1 ANALÝZA MÁKU SETÉHO (PAPAVER SOMNIFERUM)

Testovací analýza máku setého měla za úkol zpracovat kompletní genom této rostliny a vyhodnotit překrytí vůči anotovaným místům. Sekvence kompletního genomu je dostupná ke stažení z ncbi databáze na adrese https://www.ncbi.nlm.nih.gov/assembly/GCF_003573695.1. Základní informace o genomu viz. tab. 15

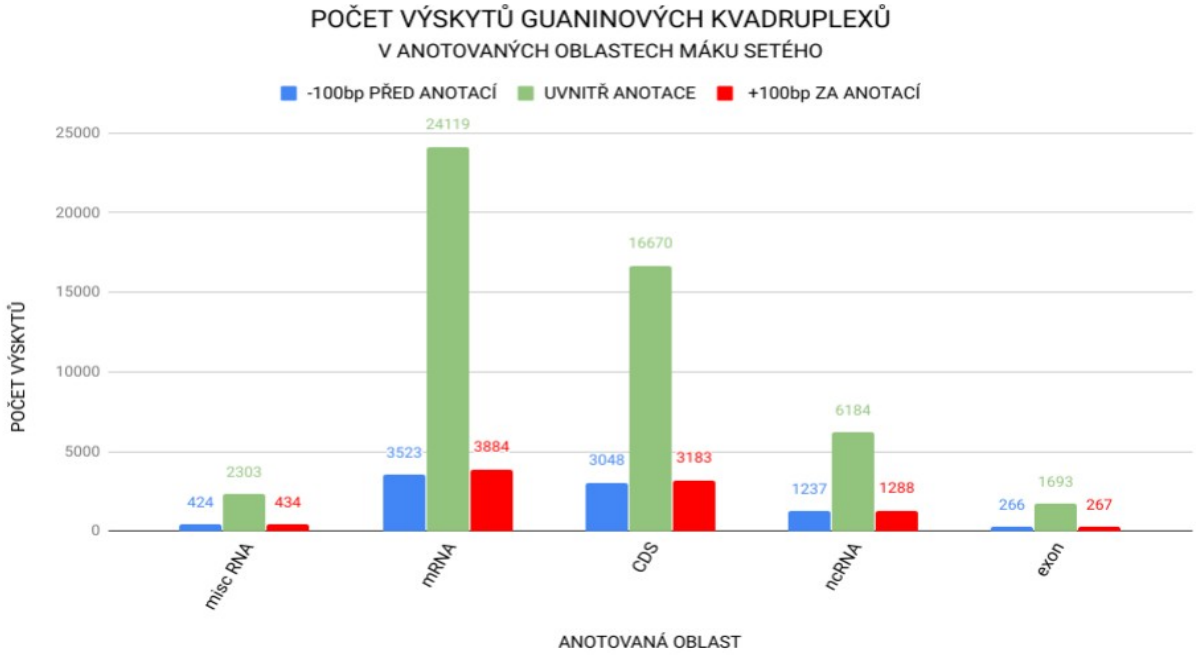
Tab 15: Základní informace o sekvenci máku setého

JMÉNO ORGANISMU	Papaver somniferum (opium poppy)
AUTOR SEKVENCE	Xi'an Jiaotong University
CELKOVÁ DÉLKA SEKVENCE	2,715,530,335
CELKOVÁ DATOVÁ VELIKOST [GB]	2,2

Pro již zmíněné sekvence byl proveden rozbor metodou G4hunter s nastavenou velikostí prohledávaného okna 25 a hodnotou prahu 1,4. Z analýzy vyplynuly následující výsledky viz. tab. 16. Výsledné překrytí vůči anotovaným oblastem se uvádí v poměru vůči frekvenci výskytu genomu v souboru anotace. Pro všechny vzniklé analýzy byl kromě tabulky překrytí sestaven graf, který naznačuje, kolem kterých významných oblastí sekvence se kvadruplexy vyskytují viz obr. 14.

Tab 16: Ukázka výsledků analýzy g4hunter

ID CHROMOZOMU	POČET KVADRUPLEXŮ	FREKVENCE VÝSKYTU	DÉLKA SEKVENCE
1	42869	0.172511	248,500,730
2	36806	0.167635	219,560,757
3	37566	0.163615	229,600,373
4	25701	0.156991	163,710,198
5	37344	0.172197	216,867,838
6	30218	0.167397	180,516,484
7	44577	0.164853	270,405,062
8	29035	0.164549	176,451,499
9	34387	0.168175	204,470,928
10	27993	0.168099	166,527,240
11	24071	0.171698	140,193,751



Obr 14: Graf zobrazující nejčastější výskyt kvadruplexů máku setého

Z výsledků zpracované sekvence máku setého vyplývá průměrná hodnota frekvence výskytu kvadruplexů na 1000 bází činící 0.167. Maximální počet predikovaných kvadruplexů obsahuje chromozom číslo 7. Tato hodnota však není anomálií vzhledem ke skutečnosti, že se jedná o nejdelší sekvenci v daném souboru. V případě máku setého jsou oblasti mRNA a CDS oblastmi s nejčastějším místem výskytu predikovaných kvadruplexů.

7.5.2 ANALÝZA KVASINKY PIVNÍ (SACCHAROMYCES CEREVISIAE)

Tato analýza vznikla za účelem publikace v článku Presence and localization of local DNA structures in S. cerevisiae genome ve spolupráci s AV ČR. Soubory sekvencí jsou dostupné z databáze NCBI na adrese https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2. Analýza proběhla s nastaveným parametrem prahu 1.2 a velikostí výpočetního okna 25. Výsledky analýzy shrnuje tab. 18.

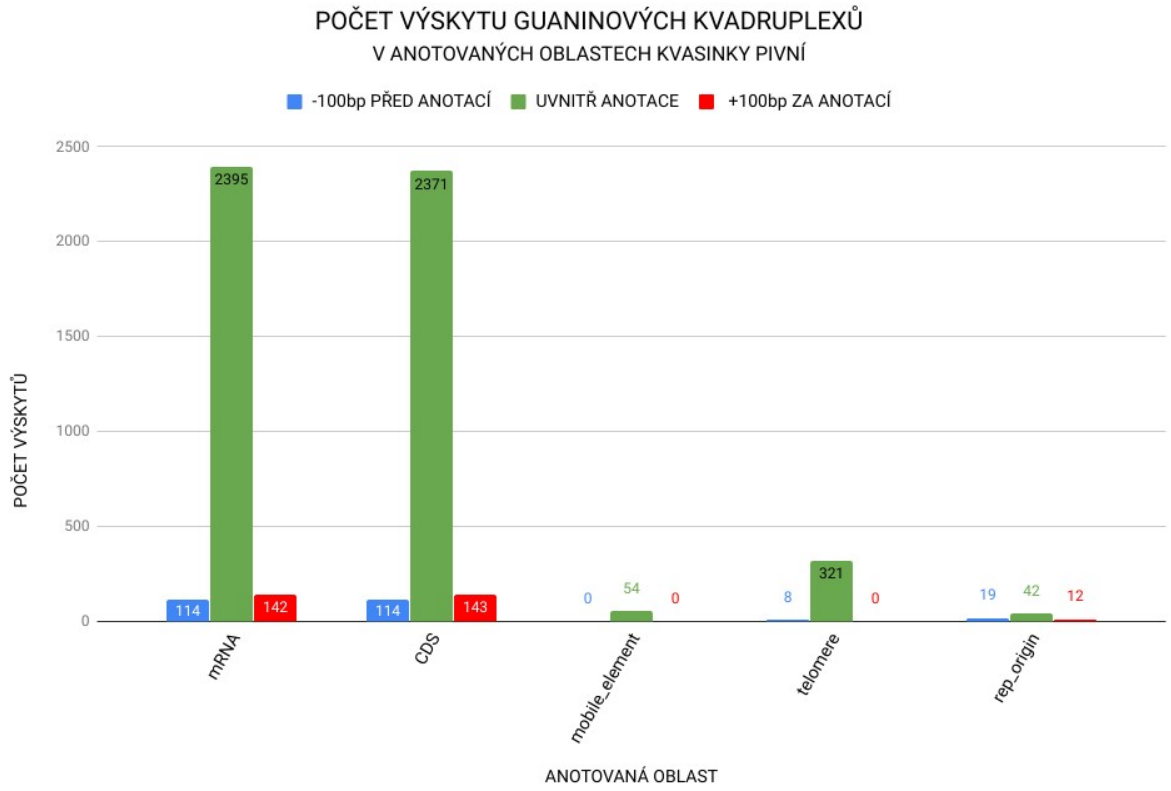
Tab 17: Základní informace o sekvenci kvasinky pivní

JMÉNO ORGANISMU	Saccharomyces cerevisiae S288C (baker's yeast)
AUTOR SEKVENCE	Saccharomyces Genome Database
CELKOVÁ DÉLKA SEKVENCE	12,175,105
CELKOVÁ DATOVÁ VELIKOST [MB]	5

Tab 18: Ukázka výsledku g4hunter analýzy

ID CHROMOZOMU	POČET KVADRUPLEXŮ	FREKVENCE VÝSKYTU	DÉLKA SEKVENCE
1	87	0.377903	230,218
2	248	0.304974	813,184
3	138	0.435854	316,620
4	429	0.280038	1,531,933
5	197	0.341496	576,874
6	86	0.318329	270,161
7	334	0.306158	1,090,940
8	157	0.279040	562,643
9	162	0.368276	439,888
10	192	0.257459	745,751

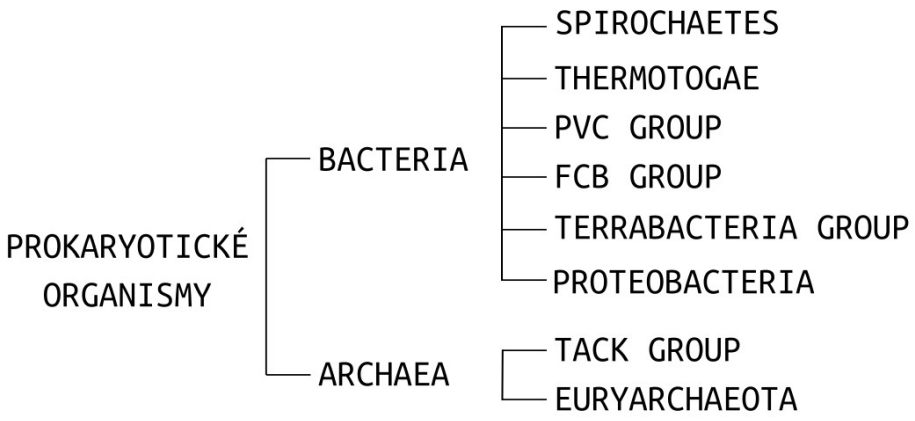
Průměrná frekvence výskytu kvadruplexů na 1000 bp je 0.326. Maximální počet kvadruplexů obsahuje chromozom číslo 12. Jako v případě máku setého je tato skutečnost zapříčiněna délkou dané sekvence. Nejvýznamnější oblastí výskytu guaninových kvadruplexů je v okolí genomů, mRNA a oblastí CDS. Ostatní oblasti výskytu predikovaných kvadruplexů jsou prakticky zanedbatelné a zobrazuje je graf na obr. 15.



Obr 15: Graf zobrazující nejčastější výskyt kvadruplexů kvasinky pivní

7.5.3 ANALÝZA PROKARYOTICKÝCH ŽIVOČICHŮ

Předposlední zpracovaná analýza se skládá ze dvou částí. Jedná se o analýzu 1620 prokaryotických jednobuněčných organismů typu bakterie a 140 organismů typu archea. Vzhledem k druhové rozmanitosti byly jednotlivé organismy rozděleny podle následujícího fylogenetického stromu viz obr. 16.



Obr 16: Fylogenetický strom zpracovávaného rodu Bacteria a Archaea

Shrnutí parametrů dvou hlavních skupin, do kterých je analýza rozdělena, shrnuje následující tabulka viz tab. 19. Jak bylo uvedeno na začátku kapitoly, jedná se o analýzy ověřující funkci aplikace ve dvou samostatných případech. V případě Archaea mluvíme o běžném rozsahu použitelnosti. Naopak Bacteria ilustruje použitelnost v extremnějších případech nízkých jednotek tisíců sekvencí.

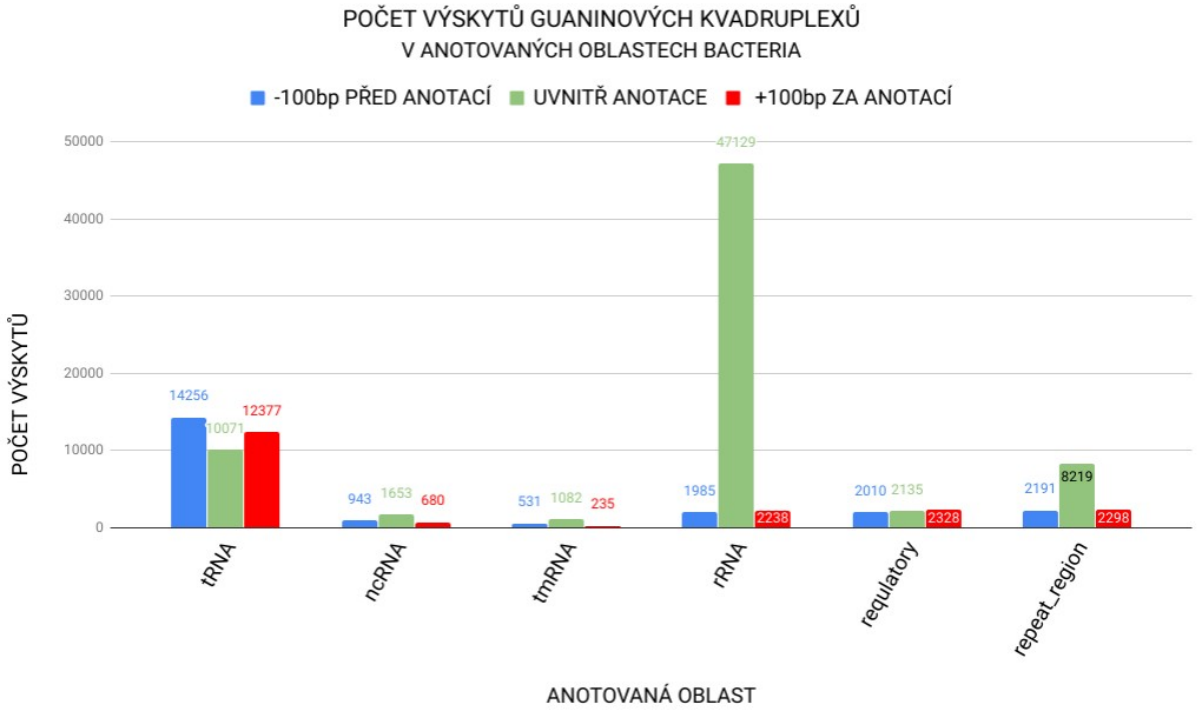
Tab 19: Základní informace o sekvencích z rodu Bacteria

JMÉNO SKUPINY ORGANISMŮ	Prokaryotičtí / Bacteria
AUTOR SEKVENCÍ	NCBI
CELKOVÁ DÉLKA SEKVENCE	5,747,603,551
CELKOVÁ DATOVÁ VELIKOST [GB]	1.2

Pro analýzu byly zvoleny parametry prahování 1.2 a hodnota velikosti prohledávaného okna 25 bází. Z důvodu velkého množství výsledných dat jsou zde uvedeny tabulky výsledků pouze z malého výběru. Výsledné překrytí s anotovanými oblastmi shrnují grafy na obr. 17 a obr. 18.

Tab 20: Ukázka výsledků analýzy g4hunter pro rod Bacteria

ID CHROMOZOMU	POČET KVADRUPLEXŮ	FREKVENCE VÝSKYTU	DÉLKA SEKVENCE
1	344	0.20126328	1709204
2	48	0.10788335	444925
3	357	0.21378705	1669886
4	370	0.21711373	1704176



Obr 17: Graf zobrazující nejčastější výskyt kvadruplexů rodu Bacteria

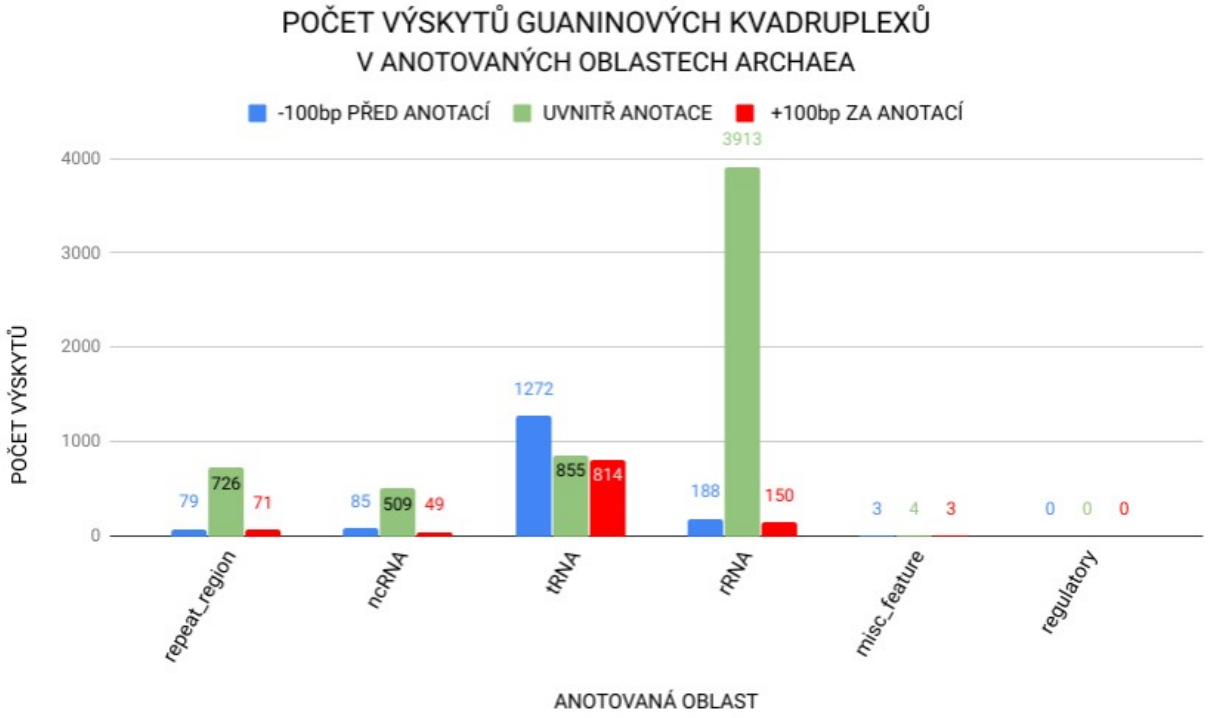
Tab 21: Základní informace o sekvencích z rodu Archaea

JMÉNO SKUPINY ORGANISMŮ	Prokaryotičtí / Archaea
AUTOR SEKVENCÍ	NCBI
CELKOVÁ DÉLKA SEKVENCE	337,153,312
CELKOVÁ DATOVÁ VELIKOST [MB]	125

Stejné parametry byly zvoleny jako u skupiny Bacteria. Tedy hodnota prahování 1.2 a velikost detekčního okna 25.

Tab 22: Ukázka výsledků analýzy g4hunter pro rod Archaea

ID CHROMOZOMU	POČET KVADRUPLEXŮ	FREKVENCE VÝSKYTU	DÉLKA SEKVENCE
1	145	0.15563384	931,674
2	268	0.29752824	900,755
3	200	0.22114426	704,387
4	5719	0.91394728	6,257,473

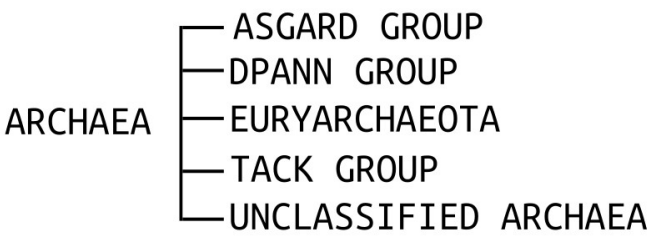


Obr 18: Graf zobrazující nejčastější výskyt kvadruplexů rodu Archaea

V obou analyzovaných skupinách existují shodné oblasti s nejčastějším výskytem predikovaných kvadruplexů, a to rRNA a tRNA. Průměrná frekvence výskytu kvadruplexů na 1000bp činí 1.341 pro rod bacteria a 1.282 pro rod Archaea. V rámci analýzy budou výsledky publikovány v článku The Presence and Localization of the G-quadruplex Forming Sequences in kingdom Bacteria. Výsledky zpracování rodu Archaea nebudou publikovány v tomto rozsahu, nýbrž v rozšířeném viz poslední podkapitola níže.

7.5.4 ANALÝZA PROKARYOTICKÝCH ARCHAEA (ROZŠÍŘENÁ)

Poslední zmíněná analýza vznikla za účelem publikace v článku The Presence and Localization of the G-quadruplex Forming ve spolupráci s AV ČR. V tomto případě se jedná však o rozdílný způsob zpracování sekvencí vzhledem k faktu, že neexistuje ucelený seznam s informacemi o identifikátorech v NCBI databázi pro jednotlivé organismy. Stejně jako v předchozím případě jsou organismy rozděleny dle fylogenetického stromu obr. 19.



Obr 19: Fylogenetický strom zpracovávaného rozšířeného rodu Archaea

Jednotlivé sekvence proto pomocí automatického skriptu musely být staženy přímo z FTP serverů NCBI databáze. Soubory však byly staženy ve formě tzv. Genomic assembly, které sdružují v jednom souboru všechny sekvence dostupné k organismu. Z důvodu dalšího zpracování bylo nutné provést sloučení jednotlivých sekvencí do jedné zastřešující celý organismus. Tyto výsledné soubory ve formátu FASTA poté byly nahrány do aplikace bioinformatics a postupně zpracovávány. Rozbor byl proveden s nastaveným parametrem prahu 1.2 a velikostí výpočetního okna 25. V Výsledky analýzy shrnuje tab. 22.

Tab 23: Základní informace o sekvencích z rozšířeného seznamu rodu Archaea

JMÉNO SKUPINY ORGANISMŮ	Prok. / Archaea (rozšířeno)
AUTOR SEKVENCÍ	NCBI
CELKOVÁ DÉLKA SEKVENCE	3,387,000,000
CELKOVÁ DATOVÁ VELIKOST [GB]	6,5

V době vzniku této práce poslední zmíněná analýza nebyla dosud dokončena, z tohoto důvodu se zde nevyskytují výsledky překrytí s anotovanými oblastmi. Průměrná frekvence výskytů na 1000 bází činí pro rozšířený set 1.206. Přičemž extrémní hodnoty frekvence výskytu 15.31 lze objevit u organismu Hadesarchaea archaeon.

8 ZÁVĚR

Cílem této diplomové práce bylo sestavit software schopný zpracovávat hromadné analýzy lokálních struktur DNA. V průběhu tvorby bylo nutné brát zřetel na nedostatečnou kvalifikaci v oblasti programování pracovníků, pro něž je tento nástroj vytvářen. Z tohoto důvodu již zadání počítá s použitím v prostředí jupyter notebook, které umožňuje rychlou kontrolu výsledků a jednoduchou kontrolu napsaných skriptů.

V úvodu se práce věnuje teoretickým znalostem z oblasti molekulární biochemie. Od základního rozdělení struktur DNA řetězců až po samotné lokální struktury, jimž se práce věnuje. Vzhledem k technickému zaměření tohoto textu jsou tyto znalosti však pouze základní. Tato skutečnost by ale neměla bránit v pochopení problematiky a rychlé orientaci v prezentovaném tématu.

V následující kapitole 3 je proveden základní rozbor nalezených nástrojů vhodných pro provádění zpracování sekvencí DNA za účelem vyhledání lokálních struktur. Jak je patrné ze srovnání, vyvíjený nástroj bioinformatics implementuje největší množství nástrojů pro vyhledávání těchto objektů. Druhá polovina kapitoly se dále zabývá základním popisem implementovaných algoritmů použitých při vývoji software bioinformatics.

Kapitoly 4 a 5 se zabývají jak projektovou přípravou. tak i navazujícím průzkumem použitých technologií. V přípravné části jsou shrnuty funkční a nefunkční požadavky spjaté s vývojem. Následná část popisuje zvolené návrhové vzory společně s jejich použitím na konkrétních případech tříd. Součástí technologického průzkumu je i popis technologie REST API. V kapitole 5 je pak proveden rozbor aktuálního stavu RESTového rozhraní aplikace bioinformatics s průvodním komentářem popisujícím jednotlivé přístupové body rozhraní.

Předposlední kapitola 6 je věnována konečné realizaci projektu. Kromě ukázek použití jednotlivých metod projektu se zde práce vyjadřuje i k použitým knihovnám nutným pro funkčnost projektu v době tvorby projektu. Vzhledem k pokračujícímu vývoji této aplikace není tento seznam knihoven garantován jako konečný resp. neměnný. V poslední části se pak kapitola věnuje problematice publikace na repositář python balíčků (Pypi) a volbě licence, pod kterou je software šířen.

Poslední 7. kapitola je určena pro konkrétní testovací případy. Ve fázi testování byla realizována řada analýz, jejichž výsledky jsou dále zpracovávány a publikovány v odborných časopisech. Kromě konkrétních výsledků z oblasti biochemie kapitola zmiňuje i parametry testovaných případů, jako je například datová náročnost či počet zpracovávaných sekvencí. V kapitole je rovněž popsán postup tzv. překrývání vůči anotovaným oblastem. Společně s výsledky algoritmu G4hunter pro hledání guaninových kvadruplexů jsou v této části i výstupní informace z překrývání sekvencí. V neposlední řadě je nutné zmínit i popis NCBI databáze, která byla při tvorbě těchto analýz klíčovým prvkem. V testovacích případech není testován algoritmus Palindrom analyser. Ten se v době tvorby práce postupně integruje do aplikace sdružující všechny algoritmy. Z tohoto důvodu jsou v projektu pouze navrženy struktury pro budoucí třídy, s jejichž implementací do projektu se v budoucnu počítá. Vzhledem k neustálému pokračování na vývoji je k datu odevzdání práce dostupná verze aplikace ve verzi 1.4 k volnému použití a šíření na adrese https://pypi.org/project/dna-

analyser-ibp/. Zdrojové soubory a následný další vývoj pak probíhá na repositáři Gitlab viz https://gitlab.com/PatrikKaura/DNA_analyser_IBP. Na přiloženém CD se nachází i zdrojové soubory k poslední zmiňované verzi, které se nemusejí shodovat s aktuální verzí v repositáři Pypi resp. Gitlab. Rovněž v elektronické příloze je ukázka analýzy Saccharomyces cerevisiae dostupné i na adrese https://gitlab.com/PatrikKaura/g4hunter_saccharomyces_cerevisiae.

Zadané cíle práce byly bez výjimek splněny a nástroj je plně provozuschopný. V blízké budoucnosti je však nutné doplnit příkazy pro volání funkcí spjatých s hledáním palindromů v sekvencích DNA. Další možné zlepšení spočívá ve vhodnějším odchytávání výjimek, především vyvolávaných z knihovny requests, z důvodu lepší informovanosti uživatele. Poslední zmíněné vylepšení v oblasti přístupnosti je plánované zavedení unifikovaného rozhraní například formou csv souboru s předem definovaným formátem seznamu sekvencí pro následné automatické zpracování.

9 SEZNAM POUŽITÉ LITERATURY

[1] ROSYPAL, Stanislav. *Úvod do molekulární biologie*. 4., (inovované) vyd. Brno: Stanislav Rosypal, 2005. ISBN 80-902-5625-2.

[2] Obr 1: Schéma struktury deoxyribonukleové kyseliny. In: *Slideplayer.cz* [online]. Praha: SlidePlayer.cz, 2019 [cit. 2019-04-19]. Dostupné z: https://slideplayer.cz/slide/13755847/

[3] Obr 3: Schéma terciální struktury tzv. nadšroubovice. In: *Https://biology.stackexchange.com* [online]. New York: Stack Exchange, 2019, 13. listopad 2016 [cit. 2019-04-19]. Dostupné z: https://biology.stackexchange.com/questions/52266/over-winding-vs-under-winding-of-a-dna

[4] Huntington disease. *Genetics Home Reference* [online]. Bethesda: National Library of Medicine of U.S., 2019 [cit. 2019-04-19]. Dostupné z: https://ghr.nlm.nih.gov/condition/huntington-disease#genes

[5] ŠPAČKOVÁ, Naděžda. Tři jsou málo, pět je moc aneb seznamte se s kvadruplexy. *Živa* [online]. Praha: Academia, 2009, **2009**(3), 98-100 [cit. 2019-04-19]. Dostupné z: http://ziva.avcr.cz/files/ziva/pdf/tri-jsou-malo-pet-je-moc-aneb-seznamte-se-s-kvadru.pdf

[6] ANDRYSÍK, Zdeněk. Hlavního strážce před rakovinou známe už 40 let. *Technet.cz* [online]. Praha: MAFRA, 2017, 13. listopad 2017 [cit. 2019-04-19]. Dostupné z: https://www.idnes.cz/technet/veda/pricina-vznik-rakovina-gen-tp53-david-lane.A171110_135316_veda_mla

[7] NOVOTNÁ, Božena a Jaroslav MAREŠ. *Vývojová biologie pro mediky*. Praha: Karolinum, 2005. ISBN 80-246-1023-X.

[8] SMIT, Arian a Robert HUBZLEY. RepeatMasker Web Server. *Institute for systems biology* [online]. Seattle: Institute for Systems Biology, 2018 [cit. 2019-04-19]. Dostupné z: http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

[9] RICE, P, A BLEASBY a I LONGDEN. Palindrome: Find inverted repeats in nucleotide sequence(s). *EMBOSS explorer* [online]. Wageningen: The European Molecular Biology Open Software Suite, 2009 [cit. 2019-04-19]. Dostupné z: http://emboss.bioinformatics.nl/cgi-bin/emboss/palindrome

[10] BIKANDI, J, R SAN MILLAN, A REMETERIA a Garaizar J. Palindromic sequences finder. *Http://insilico.ehu.eus* [online]. Leioa: Universidad del País Vasco, 2004 [cit. 2019-04-19]. Dostupné z: http://insilico.ehu.eus/palindromes/

[11] TLALIT, Freaund a Engel NOGA. Palindrome search. *The Laboratory of Computational Biology* [online]. Haifa: Technion Israel Institute of Technology, 2003 [cit. 2019-04-19]. Dostupné z: http://bioinfo.cs.technion.ac.il/projects/Engel-Freund/new.html

[12] OLEG, Kikin, D'Antonio LAWRENCE a Paramjeet S BAGGA. QGRS Mapper. *QGRS* [online]. New Jersey: Ramapo College of New Jersey, 2006 [cit. 2019-04-19]. Dostupné z: http://bioinformatics.ramapo.edu/QGRS/analyze.php

[13] DHAPOLA, Parashar. QUADBASE2. *QUADBASE* [online]. New Dehli: CSIR-Institute of Genomics and Integrative Biology, 2019 [cit. 2019-04-19]. Dostupné z: http://quadbase.igib.res.in/ProQuad

[14] REGINA, Z, E DUNCAN a Uma S. MUDUNURI. Non-B DB. *National Cancer Institute at Frederick* [online]. Frederic: The Advanced Biomedical Computing Center, 2013 [cit. 2019-04-19]. Dostupné z: https://nonb-abcc.ncifcrf.gov/apps/site/default

[15] BRÁZDA, V. a kol. Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochemical and Biophysical Research Communications*. http://www.sciencedirect.com/science/article/pii/S0006291X16314620., 2016, **2016**(4), 1739-1745. ISSN 0006-291X.

[16] BRÁZDA, V. a kol. Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *BioMed Research International*. 2018, **2018**(24), 10. ISSN 2314-6133.

[17] http://bioinformatics.ibp.cz [online]. Brno: Institute of Biophysics of the Czech Academy of Sciences, 2018 [cit. 2019-04-19]. Dostupné z: http://bioinformatics.ibp.cz

[18] VEPRINTSEV, Dmitry a Alan FERSHT. Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic acids research*. 2008, **2008**(36), 1589-1598 PMCID: PMC2275157.

[19] Java [online]. Redwood Shores: Oracle Corporation, 2019 [cit. 2019-04-19]. Dostupné z: https://go.java/index.html

[20] MALÝ, Martin. REST: architektura pro webové API. *Zdroják.cz* [online]. Praha: Devel.cz Lab, 2009 [cit. 2019-04-19]. Dostupné z: https://www.zdrojak.cz/clanky/rest-architektura-pro-webove-api/

[21] Obr 7: Schéma principu činnosti REST API. In: *MlsDev* [online]. Kyjev: MlsDev, 2019 [cit. 2019-04-19]. Dostupné z: https://mlsdev.com/blog/81-a-beginner-s-tutorial-for-understanding-restful-api

[22] PILGRIM, Mark. *Ponořme se do Python(u) 3: Dive into Python 3*. Praha: CZ.NIC, c2010. CZ.NIC. ISBN 978-80-904248-2-1.

[23] DOWNEY, Allen. *Think Python*. 2nd edition, updated for Python 3. Sebastopol, CA: O'Reilly Media, 2016. ISBN 14-919-3936-2.

[24] *Jupyter* [online]. Worldwide: Project Jupyter, 2019 [cit. 2019-04-19]. Dostupné z: https://jupyter.org/

[25] *Source making* [online]. Worldwide: sourcemaking.com, 2019 [cit. 2019-04-19]. Dostupné z: https://sourcemaking.com/

[26] MCLARTY, 2019 [cit. 2019-04-19]. Dostupné z: https://medium.com/sean3z/json-web-tokens-jwt-with-restify-bfe5c4907e3c

[27] *Choose an open source license* [online]. San Francisco: GitHub, 2019 [cit. 2019-04-19]. Dostupné z: https://choosealicense.com/

[28] *National Center for Biotechnology Information* [online]. Bethesda: National Library of Medicine U.S., 2019 [cit. 2019-04-19]. Dostupné z: https://www.ncbi.nlm.nih.gov/

[29] Submission of Annotation Using a Table. *NCIB* [online]. Bethesda: National Library of Medicine of U.S., 2019 [cit. 2019-04-19]. Dostupné z: https://www.ncbi.nlm.nih.gov/projects/Sequin/table.html

10 SEZNAM OBRÁZKŮ

11 SEZNAM TABULEK

A. OBSAH PŘILOŽENÉHO CD

NÁZEV	POPIS
DP_Kaura.pdf	Textová část diplomové práce.
DNA_analyser_IBP_v14.zip	Zdrojové kódy aplikace verze 1.4.
G4hunter_saccharomyces.zip	Ukázka analýzy Sacharomyces cerevisiae.