Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Teória hry chaosu a jej použitie na porovnanie postupností v bioinformatike

Bakalárska práca

2024
Sára Guziová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Teória hry chaosu a jej použitie na porovnanie postupností v bioinformatike

Bakalárska práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	prof. RNDr. Mária Lucká, PhD.

Bratislava, 2024
Sára Guziová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Sára Guziová
Študijný program: aplikovaná informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Teória hry chaosu a jej použitie na porovnanie postupností v bioinformatike
Chaos game theory and its application to sequence comparison in bioinformatics

Anotácia: Výpočet podobnosti medzi dvomi nukleotidovými sekvenciami je jedným zo základných problémov bioinformatiky. Súčasné metódy sú založené buď na výpočtovo náročnom zarovnaní sekvencií alebo na použití metód bez zarovnania. Nové možnosti v tomto smere prináša reprezentácia postupností pomocou hry chaosu, ktorá využíva grafické prostredie a transformuje postupnosti rôznej dĺžky na obrazy alebo matice rovnakej veľkosti. Takáto reprezentácia je vhodná aj na kódovanie čŕt v strojovom učení.
Cieľom bakalárskej práce je aplikovať teóriu hry chaosu na hľadanie podobnosti veľkých genomických postupností a porovnať ju z hľadiska presnosti s inými metódami bez zarovnania na vybraných dátových množinách.

Literatúra: Löchel, H. F., & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. Computational and Structural Biotechnology Journal, 19, 6263–6271. https://doi.org/10.1016/j.csbj.2021.11.008
Farkaš, T., Sitarčík, J., Brejová, B., & Lucká, M. (2019). SWSPM: A Novel Alignment-Free DNA Comparison Method Based on Signal Processing Approaches. Evolutionary Bioinformatics, 15. https://doi.org/10.1177/1176934319849071
Detlefsen, N. S., Hauberg, S., & Boomsma, W. (2022). Learning meaningful representations of protein sequences. Nature Communications, 13(1), 1–12. https://doi.org/10.1038/s41467-022-29443-w
Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biology, 18(1), 1–17. https://doi.org/10.1186/s13059-017-1319-7

Kľúčové slová: hra chaosu, porovnávanie postupností, DNA

Vedúci: prof. RNDr. Mária Lucká, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 10.10.2023

Dátum schválenia: 16.10.2023

doc. RNDr. Damas Gruska, PhD.
garant študijného programu

Abstrakt

Kľúčové slová: hra chaosu, porovnávanie postupností, DNA

Abstract

Abstract in the English language (translation of the abstract in the Slovak language).

Keywords: chaos game theory, sequence comparison, DNA

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

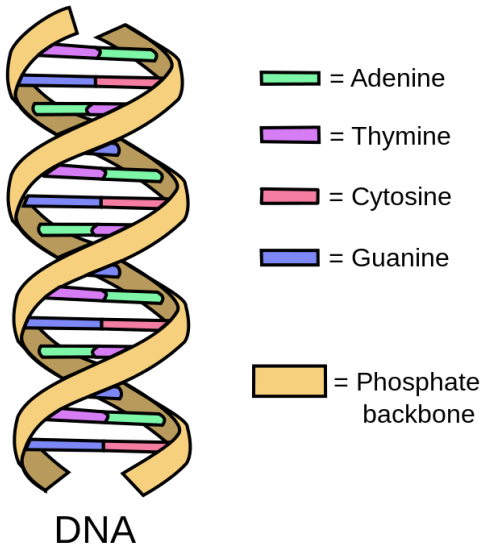
Kapitola 1

Východiská

Žijeme v ére, v ktorej je skúmanie genetických informácií jednoduchšie než kedykoľvek predtým, čo umožňuje postupné odkrývanie tajomstiev života. Bioinformatika je interdisciplinárna veda, ktorá analyzuje dáta získané biologickými metódami s pomocou počítačovej vedy a matematiky [29]. Vývoj efektívnych algoritmov v tejto oblasti prispel k dosiahnutiu nových vedeckých objavov aj identifikácii mnohých klinických aplikácií. Umožnil rozvoj farmakogenetiky, odboru, ktorý sa zaoberá návrhom liečby šitej na konkrétneho pacienta. HĺBková analýza DNA sekvencii zas viedla k určeniu korelácie medzi určitými génmi a vznikom ochorení [14]. Okrem uplatnenia v biomedicíne je skúmanie genetických informácií esenciálne pre fylogenetiku, ktorá mapuje evolučné vzťahy medzi živými organizmami. Evolučné vzťahy je možné vizualizovať konštrukciou fylogenetických stromov [9].

Nositeľom genetickej informácie organizmov sú nukleové kyseliny, ku ktorým zaraďujeme ribonukleovú kyselinu (RNA) a deoxyribonukleovú kyselinu (DNA). Monomérna jednotka DNA, deoxyribonukleotid, obsahuje fosfátovú zložku, sacharidovú zložku a dusíkatú bázu – adenín (A), guanín (G), cytozín (C) alebo tymín (T). DNA sa skladá z 2 polynukleotidových reťazcov, ktoré sú spojené vodíkovou väzbou na princípe komplementarity tak, že adenín sa vždy nachádza oproti tymínu a guanín oproti cytozínu, ako môžeme vidieť na obrázku 1.1. Pri zápise DNA je preto potrebné uviesť explicitne iba poradie nukleotidov nachádzajúcich sa v jednom z vlákien. RNA sa skladá iba z jedného reťazca a namiesto tymínu obsahuje uracil (U). Gén charakterizujeme ako určitý úsek na reťazci DNA, teda poradie nukleotidov, ktoré kóduje jednu vlastnosť organizmu.

Porovnávanie sekvencií je disciplína, ktorá sa od zrodu bioinformatiky enormne rozšírila. Kvôli neustálemu nárastu množstva genetických dát je potrebné hľadať stále nové a nové metódy na rýchlejšie a presnejšie porovnávanie. Pre dôkladné porozumenie témy porovnávania postupností v tejto kapitole uvádzame podstatné poznatky ohľadom genómov organizmov 1.1, sekvenovania DNA 1.2, formátov súborov používaných



Obr. 1.1: [15] Model DNA ilustrujúci princíp komplementarity. Adenín sa nachádza oproti tymínu a guanín oproti cytozínu.

v bioinformatike 1.3 a existujúcich metódach na porovnávanie postupností 1.4.

1.1 Charakteristika genómu

Genóm organizmu obsahuje jeho dedičné informácie, ktoré sú zakódované v DNA. Veľkosť genómu je špecifická pre druh, napríklad veľkosť genómu baktérie Escherichia coli, významného modelového organizmu, je 4 milióny párov znakov, zatiaľ čo ľudský genóm sa skladá z približne 3,2 miliardy párov znakov [10]. Genómy dvoch jedincov rovnakého druhu typicky obsahujú enormné množstvo rovnakých či podobných úsekov, pričom u ľudí je viac ako 99 percent znakov zhodných.

Približne 3 percentá bázových párov v ľudskom genóme kóduje proteíny, komplexné molekuly esenciálne pre reguláciu všetkých biologických funkcií. Triplet, poradie troch dusíkatých báz, je základná jednotka informácie potrebná pri ich syntéze. Zvyšných 97 percent sa nazýva nekódujúca DNA a je asociovaná s reguláciou génovej expresie [10].

Významná časť DNA sekvencií sa v genóme opakuje. Genóm môže obsahovať krátke unikátne sekvencie zopakované niekoľkokrát za sebou, viacero kópií jedného génu alebo duplicity veľkých génových zhlukov či dokonca celých chromozómov. Typický je aj frekventovaný výskyt palindrómov. Repetitívnosť a iné charakteristické vlastnosti sú často narušené náhodnými mutáciami, ako príklad môžeme uviesť deléciu (stratu jedného nukleotidu), inzerciu (vsunutie nového nukleotidu) a substitúciu (nahradenie nukleotidu) [10].

1.2 Sekvenovanie DNA

Sekvenovanie DNA je proces určenia poradia nukleotidov. Výsledné sekvencie sú zapísané do databáz a následne je možné ich skúmať a porovnávať. Vzhľadom na neustály progres v bioinformatických metódach a znižujúce sa finančné náklady na sekvenovanie objem dát v genomických databázach každý rok exponenciálne rastie [10], čo vyvoláva obavy súvisiace s ukladaním, prenosom a vyhľadávaním údajov [11].

Moderné technológie sekvenovania súhrnne nazývame sekvenovanie novej generácie (NGS). Hlavným rozdielom medzi staršími technológiami a NGS je prítomnosť kvality báz [23], teda informácie o pravdepodobnosti ich nesprávneho určenia. Veľmi rozšírené je momentálne nanopórové sekvenovanie, technológia tretej generácie. Zariadenia od Oxford Nanopore Technologies (ONT) sú založené na meraní iónového prúdu popri tom, ako časti DNA prechádzajú cez proteínový pór [22]. Dusíkaté bázy sú identifikované podľa toho, na akú dlhú dobu sa zastaví tok iónov pri prechode nukleotidu cez pór. Využitie ONT zariadení má viacero výhod, sú prenosné, umožňujú rýchle spracovanie vzoriek DNA a zobrazenie výsledkov v reálnom čase. Väčšina existujúcich algoritmov na spracovanie genomických dát bola vytvorená pre krátke sekvencie (v rozsahu stoviek báz) a za predpokladu, že dôjde k malej chybovosti pri sekvenovaní. Nanopórovým sekvenovaním sa dajú získať oveľa dlhšie sekvencie (až v rozsahu stoviek tisíc báz) na úkor pomerne častého výskytu chýb. Miera chybovosti však v posledných rokoch dramaticky klesla a pohybuje sa na úrovni okolo päť percent [19].

V živých bunkách sú genetické informácie zapísané v postupnosti monomérov spojených do dlhých lineárnych molekúl DNA. Aby sme s nimi mohli pracovať, musíme ich previesť do digitálnej podoby [7]. Výsledkom sekvenovania genómu je niekoľko sekvencií, nazývaných čítania (sequencing reads). Z informatického hľadiska sú to reťazce zložené zo znakov A, G, C a T [3], reprezentujúcich dusíkaté bázy. Nespracované čítania sú uložené v textových súboroch založených na kódovacom systéme znakov ASCII alebo EASCII (extended ASCII) [2]. Tento formát označujeme ako surové dáta a nie je vhodný na ich dlhodobé uloženie [7].

Keďže sú čítania výstupom biochemických a výpočtových metód, nedá sa zaručiť ich presnosť. DNA môže byť kontaminovaná aj pri manipulácií alebo obsahovať nukleové kyseliny patogénnych druhov. Reprezentovanie sekvencií preto vyžaduje schopnosť modelovať heterogénne, dynamické, neúplné a nedokonalé informácie. [24].

1.3 Formáty súborov

Ak chceme pracovať s celým genómom, musíme ho najskôr zostaviť z jednotlivých čítaní. Genóm sa postupne konštruuje na základe prekrývania čítaní a tento náročný proces vyžaduje aj spätnú korekciu chýb [32]. Zostavený genóm sa dá zobraziť pomo-

```
>NC_012920.1 Homo sapiens mitochondrion, complete genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
ATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATC
ATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCA
```

Obr. 1.2: [20] Časť súboru vo formáte FASTA. Hlavičku tvorí jeden riadok začítajúci symbolom > a zvyšné riadky obsahujú uloženú sekvenciu.

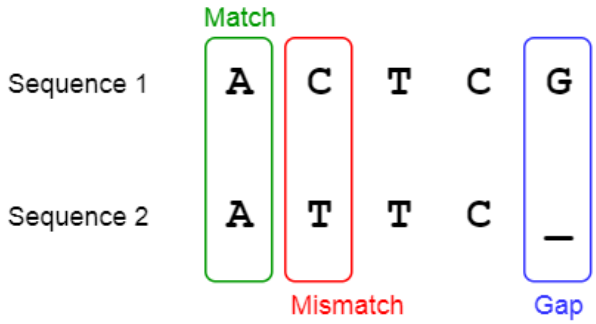
cou štandardných textových súborov. Existujú však špeciálne textové formáty, prispôsobené na genetické dáta. Spravidla je v nich informácia logicky rozdelená na hlavičku a telo, kde hlavička obsahuje identifikátor, popis alebo druhovú špecifikáciu a telo samotnú sekvenciu či jej kódovanie. V jednom súbore sa môže nachádzať aj viacero vhodne oddelených sekvencií. Medzi najznámejšie formáty patria [21, 22]:

- FASTA - jeden z najstarších a najjednoduchších formátov používaných v bioinformatike, využíva jednopísmenové kódy pre zápis nukleotidov alebo aminokyselín stanovené Medzinárodnou úniou pre čistú a aplikovanú chémiu (IUPAC), príklad sa nachádza na obrázku 1.2
- FASTQ - formát sekvenovania novej generácie, ktorý obsahuje rozšírenie formátu FASTA o kvality báz
- SAM - formát sekvenovania novej generácie, slúžiaci na ukladanie DNA sekvencií zarovnaných k referenčnej sekvencii
- BAM - binárná podoba formátu SAM podpodporujúca indexáciu pre rýchly náhodný prístup
- FAST5 - formát sekvenovania pomocou ONT zariadení

1.4 Porovnávanie postupností

Hľadanie podobnosti, respektíve rozdielnosti nukleotidových sekvencií je jedným zo základných problémov bioinformatiky. Pri skúmaní genetických dát môžeme zistiť, že sú identické, podobné alebo úplne rozdielne. Ak sa miera podobnosti stanoví na aspoň 30%, porovnávané sekvencie sa považujú za homologické. Homológia dvoch sekvencií indikuje, že organizmy, od ktorých boli sekvencie získané, majú spoločného predka v evolučnom vývoji [29].

Z informácií uvedených v podkapitole 1.1 vyplýva, že aplikácia všeobecných textových algoritmov porovnávania na genetické dáta bude vo väčšine prípadov neefektívna,



Obr. 1.3: [17] Príklad zarovnania dvoch sekvencií s rôznou dĺžkou tak, aby pod sebou boli čo najčastejšie rovnaké bázy. Na obrázku je vyznačená zhoda báz (match), nezhoda báz (mismatch) aj vložená medzera (gap).

pretože nevyužíva špecifické vlastnosti DNA sekvencií. Súčasné metódy porovnávania sa dajú rozdeliť do dvoch hlavných skupín na základe toho, či používajú alebo nepoužívajú zarovnanie porovnávaných sekvencii [9]. Oba prístupy majú svoje výhody a nevýhody, ktoré uvádzame v podčastiach 1.4.1 a 1.4.2.

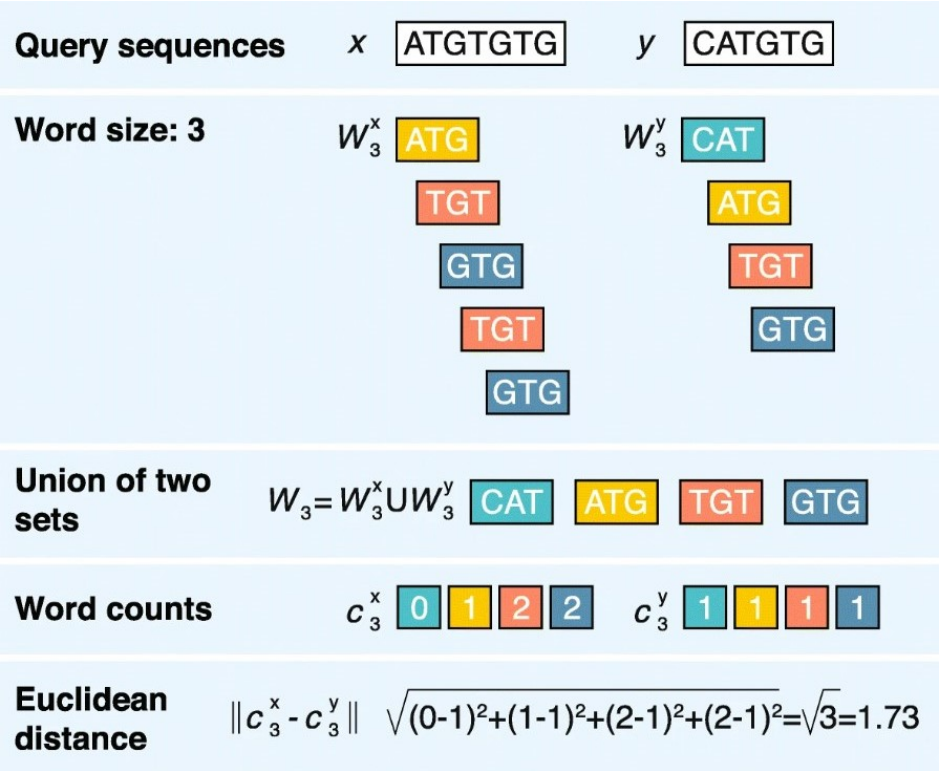
1.4.1 Metódy so zarovnaním

Programy založené na týchto metódach sekvencie najskôr zarovnajú a až následne hľadajú mieru zhody [32].

Prvým krokom je častokrát netriviálne zarovnanie sekvencií. Sekvencie sa zapíšu pod seba, na každom riadku sa teda nachádza práve jedna sekvencia a bázy na rovnakej pozícii ležia priamo pod sebou. V prípade, že sekvencie nie sú rovnako dlhé, medzi jednotlivé bázy sa vpíšu pomlčky. Skupiny za sebou idúcich pomlčiek nazývame medzery. Vkladanie medzier prebieha tak, aby pod sebou boli čo najčastejšie rovnaké bázy a zároveň sa minimalizoval počet medzier [3]. Druhým krokom je vypočítanie podobnosti na základe bodovacieho systému. Na každej pozícii určíme podobnosť báz a celková hodnota podobnosti sekvencií je potom súčet podobností na jednotlivých pozíciách [7].

Ak dochádza k zarovnaniu celých sekvencií, hovoríme o globálnom zarovnávaní. Lokálne zarovnávanie je naopak proces, pri ktorom sa najskôr vyhľadávajú podobné časti sekvencií a iba tie sa následne zarovnávajú a analyzujú [3]. Najjednoduchšie je zarovnávanie dvoch sekvencií. Príklad môžeme vidieť na obrázku 1.3. V tomto príklade došlo k zarovnaniu sekvencií rôznej dĺžky, preto je do druhej sekvencie vpísaná medzera. V prípade, že chceme porovnať viac ako dve sekvencie, musíme vykonať takzvané viacnásobné zarovnanie [3], ktoré je výpočtovo veľmi náročné.

Metódy so zarovnaním sú vo všeobecnosti presnejšie, majú však väčšiu časovú a



Obr. 1.4: [32] Proces porovnania 2 sekvencií pomocou rátania frekvencie k-mérov pre k = 3.

pamäťovú zložitosť. Počet možných zarovnaní sa rapídne zvyšuje s počtom zarovnávaných sekvencií a ich dĺžkou. Zarovnanie sekvencií v niektorých prípadoch ovplyvňujú aj predpoklady o vzťahoch medzi nimi a už malé zmeny vo vstupných parametroch môžu výrazne ovplyvniť výsledok [32]. Optimálne algoritmy zarovnávania sekvencií sú implementované najmä pomocou dynamického programovania, ktoré maximalizuje mieru podobnosti sekvencií [28]. Neexistuje konsenzus, ktorý by definoval univerzálny bodovací systém, zarovnanie aj samotný výsledok porovnávania sa teda môže výrazne líšiť na základe zvoleného bodovacieho systému pre danú metódu [32].

1.4.2 Metódy bez zarovnania

Metódy bez zarovnania sú pomerne rôznorodé a môžeme ich definovať ako akýkoľvek postup porovnávania sekvencií, pri ktorom nedochádza k vytvoreniu zarovnania medzi sekvenciami ani ich časťami počas celého behu programu. Vznikli ako prirodzená alternatíva k metódam so zarovnaním, ktoré sú v niektorých prípadoch tak výpočtovo náročné, že sa nedajú použiť.

Metódy bez zarovnania využívajú predovšetkým matematické koncepty, hlavne štatistiku a princípy z oblasti lineárnej algebry [32]. Ak sa sekvencie DNA namapujú do vektorových priestorov, ich analyzovanie je možné vykonávať efektívnejšie [27]. V sú-

časnosti sú najpoužívanejšie prístupy založené na rátaní frekvencie k-mérov [31]. K-mér je bioinformatické označenie pre podslovo dĺžky k nachádzajúce sa v sekvencii. Idea tejto metódy spočíva v tom, že podobné sekvencie zdieľajú rovnaké subsekvencie. Postup porovnania dvoch sekvencií pomocou rátania frekvencie k-mérov pre k = 3 je zobrazený na obrázku 1.4. Najskôr sa sekvencie rozdelia na podslová rovnakej dĺžky. Následne sa vytvorí množina podslov zo všetkých porovnávaných sekvencií a pre každú sekvenciu sa vytvorí vektor obsahujúci počty týchto subsekvencií v danej sekvencii. Nakoniec sa vypočíta miera odlišnosti vektorov pomocou zvolenej metriky, ako napríklad euklidovskej vzdialenosti. Samotné porovnanie vektorov tak priebieha v lineárnom čase. Vyššie číslo indikuje väčšiu mieru rozdielnosti. Pre identické sekvencie bude teda výsledná vzdialenosť rovná 0 [32].

K metódam bez zarovnania zaraďujeme aj princípy využívajúce teóriu informácii, Fourierovu transformáciu či systém iterovaných funkcií [31]. Inovatívny prístup k porovnávaniu sekvencií prinieslo použitie metód, ktoré vytvárajú grafické reprezentácie sekvencií. Existuje niekoľko spôsobov, ako vizualizovať genetické dáta. Jedným z nich je reprezentácia pomocou hry chaosu, ktorú podrobne opisujeme v nasledujúcej kapitole 2.

Najväčšou výhodou metód bez zarovnania je ich rýchlosť. Z toho dôvodu sú schopné porovnávať aj celé genómy organizmov. Navyše sú omnoho odolnejšie voči náhodným mutáciám či chybám vzniknutých počas sekvenovania [27]. Na základe týchto atribútov sú metódy bez zarovnania ideálne pre analyzovanie dlhých sekvencií získaných pomocou NGS [32]. Pri použití týchto metód sa však často stratí poradie symbolov alebo podslov sekvencie, čo má za následok menšiu presnosť oproti metódam so zarovnaním [27].

Kapitola 2

Teória hry chaosu

Výber reprezentácie genetických dát má zásadný vplyv na schopnosť extrahovať z nich potrebné informácie. V prípade spracovania DNA sekvencie sa snažíme vytvoriť abstrakciu, ktorá zachytáva dôležité vlastnosti pôvodného polynukleotidového reťazca [8]. Jednou z možností je reprezentácia sekvencií pomocou poznatkov z teórie hry chaosu. V tejto kapitole preto uvádzame základné informácie o hre chaosu 2.1, spôsoby reprezentácie sekvencií, menovite Chaos game representation 2.2 a Frequency matrix chaos game representation 2.3, a existujúce aplikácie teórie hry chaosu 2.4.

2.1 Hra chaosu

V nasledujúcich bodoch opisujeme hru chaosu [12] s využitím papiera a pera, postup sa dá jednoducho algoritmizovať a spustiť na počítači:

1. Na papieri vytvoríme 3 body tak, aby neležali v jednej línii. Označíme ich A, B a C a budeme ich nazývať vrcholy.

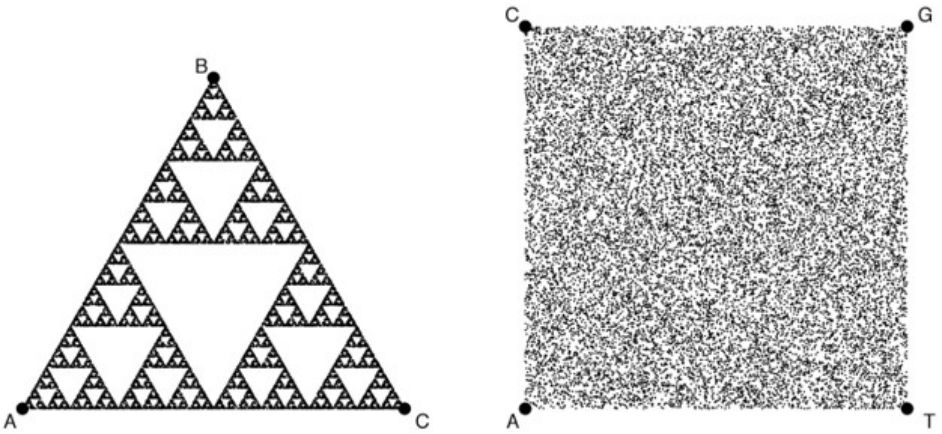
2. Zvolíme si nástroj na generovanie náhodných čísel, napríklad obyčajnú hraciu kocku. Vrcholu A priradíme čísla 1 a 2, vrcholu B čísla 3 a 4, vrcholu C čísla 5 a 6.

3. Vytvoríme jeden iniciálny bod niekde na ploche papiera.

4. Hodíme kockou. Ak padlo číslo 1 alebo 2, zvolený je vrchol A. Na papier zaznačíme nový bod do stredu medzi iniciálnym bodom a vrcholom A. Ak padlo iné číslo, postupujeme analogicky pre vrchol reprezentovaný daným číslom.

5. Pokračujeme v hádzaní kockou, po každom hode zaznačíme nový bod do stredu medzi posledným zaznačeným bodom a zvoleným vrcholom.

Intuitívne by sa dalo očakávať, že výsledný obrázok bude chaotický. Nový bod vzniká na náhodnom mieste, jeho pozícia je určená hodom kocky a posledným, tiež



Obr. 2.1: [1] Výsledok hry chaosu pri troch (vľavo) a štyroch (vpravo) vrcholoch. Fraktál vľavo sa nazýva Sierpinského trojuholník.

náhodne vzniknutým bodom. Opak je však pravdou. Po dostatočnom počte iterácií (niekoľko stoviek až tisícov) sa pred nami zjaví fraktálny útvar, nazývaný Sierpinského trojuholník podľa poľského matematika, ktorý ho v roku 1915 popísal. Body A, B a C predstavujú vrcholy tohto trojuholníka. Sierpinského trojuholník je zobrazený vľavo na obrázku 2.1. Fraktály môžeme charakterizovať ako na prvý pohľad komplikované geometrické objekty, ktoré vznikajú pomerne jednoduchou rekurzívnou funkciou a zároveň sú škálovo invariantné (sebapodobné) [16]. Formálne je objekt sebapodobný, ak jeho časť s príslušnou zmenou mierky má rovnaký tvar ako celkový objekt [12]. Neformálne to znamená, že pri škálovaní si zachováva svoju charakteristický vzor, pri priblížení či oddialení stále pozorujeme opakujúci sa motív a jednotlivé časti útvaru sú si navzájom podobné. V prípade, že sa iniciálny bod nachádza mimo trojuholníka ABC, niekoľko ďalších bodov sa môže taktiež nachádzať mimo tohto trojuholníka, nemá to ale vplyv na celkový výsledok.

Hra chaosu sa dá spustiť aj s iným počtom vrcholov, rôznym škálovacím faktorom a nie je limitovaná len na 2-dimenzionálny priestor. Napríklad pri 4 vrcholoch v 2D priestore však výsledkom nie je zaujímavý fraktál, ale štvorec rovnomerne vyplnený bodmi bez akejkoľvek štruktúry [12], ako môžeme vidieť na obrázku 2.1 vpravo.

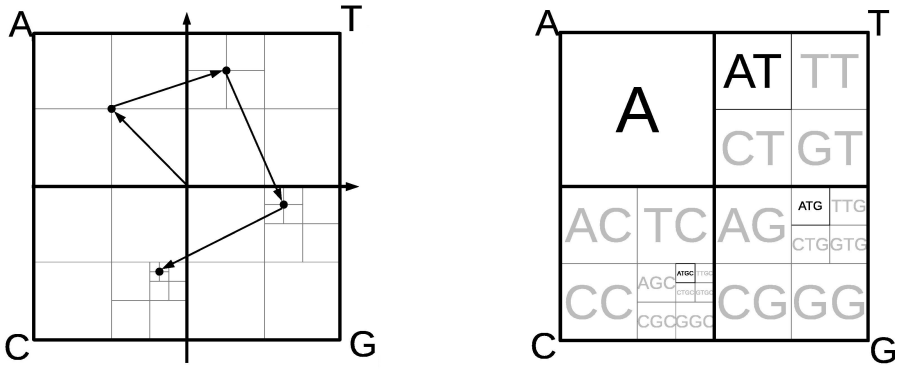
Matematicky môžeme hru chaosu opísať systémom iterovaných funkcií (IFS). IFS sa skladá z množiny lineárnych rovníc, na základe ktorých sa rátajú súradnice nového bodu. V súčasnosti sa používa najmä kompaktná notácia [1, 16] uvedená v rovnici (2.1)

$$P_i^j = P_{i-1}^j + sf(V_{i-1}^j - P_{i-1}^j), \tag{2.1}$$

kde

j je počet dimenzií priestoru

i je poradové číslo bodu



Obr. 2.2: [16] Ilustrácia CGR algoritmu pri štyroch vrcholoch označených A, C, G a T podľa dusíkatých báz nachádzajúcich sa v DNA sekvenciách a rozdelenie štvorca v dôsledku iteračného procesu.

P_0^j sú súradnice iniciálneho bodu

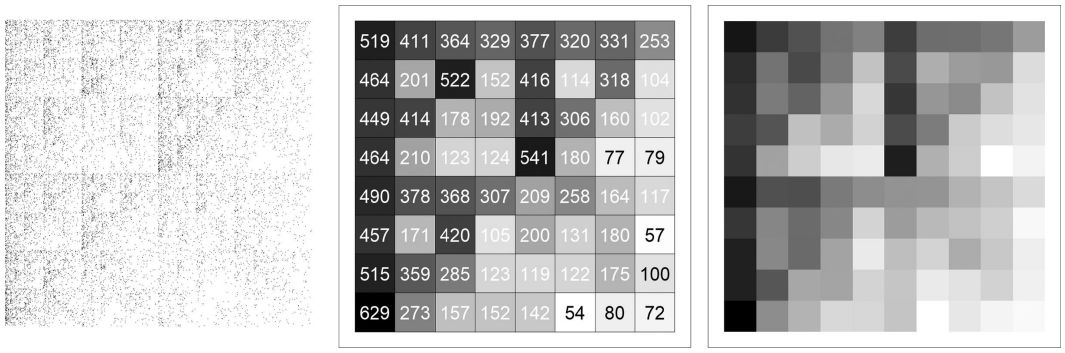
V_i^j sú súradnice i-tého zvoleného vrchola

sf je škálovací faktor.

2.2 Chaos game representation

Ako to súvisí s bioinformatikou? H. Joel Jeffrey sa vo svojej práci z roku 1990 "Chaos game representation of gene structure"[12] inšpiroval teóriou hry chaosu a rozhodol sa ju aplikovať na genetické dáta. Kedže nukleové kyseliny sa zapisujú ako reťazce 4 znakov dusíkatých báz, vrcholy jednotkového štvorca označil písmenami A, C, G a T (pre DNA sekvencie) alebo A, C, G a U (pre RNA sekvencie). Každý vrchol teda reprezentoval jeden možný nukleotid. Iniciálny bod umiestnil do stredu štvorca. Namiesto náhodného generovania čísel postupne prechádzal zvolenú sekvenciu a pre každú bázu vygeneroval nový bod medzi naposledy zaznačeným bodom a vrcholom príslušnej bázy. Vytvoril tak grafickú reprezentáciu danej sekvencie. Na rozdiel od náhodného generovania čísel sa pri prechádzaní sekvencií začali vytvárať fraktálne obrazce. Ilustrácia použitého algoritmu sa nachádza na obrázku 2.2 vľavo. Výsledok spracovania sekvencie pomocou tejto metódy dostal názov Chaos game representation (CGR).

Vo výskume Jeffrey aplikoval CGR na genetické informácie rôznych druhov organizmov, od najjednoduchších vírusov a baktérií až po komplexnejšie mnohobunkovce. Zatiaľ čo napríklad u bezstavovcov pozoroval takmer rovnomerné rozdelenie bodov bez zjavnej štruktúry, u stavovcov sa mu naskytol pohľad na opakujúci sa vzor, jasne viditeľný aj voľným okom. Iné typy vzorov sa zas vyskytovali v zobrazeniach nukleových kyselín rastlín, plesní, či niektorých vírusov. To ho priviedlo k myšlienke, že podobnosť sekvencií je možné určiť na základe skúmania týchto vzorov. Zaujímavé je, že algorit-



Obr. 2.3: [16] Ilustrácia FCGR algoritmu pre k = 3.

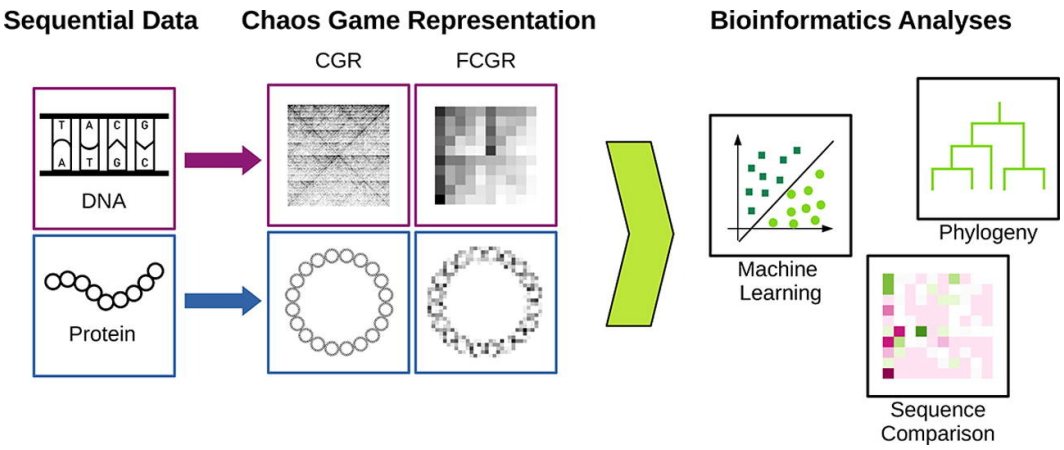
mus aplikoval aj na bežný anglický text, konkrétne na svoje práce. Algoritmus postupne prechádzal text a zobrazoval len písmená A, C, G a T (respektíve U), ostatné ignoroval. Na zobrazení takýchto textov bola taktiež vidieť určitá štruktúra, ktorá navyše nebola podobná žiadnym vzorom z genetických informácií organizmov.

Vizualizácia s využitím CGR odhalila dovtedy neznáme štruktúry nukleovým kyselín a umožnila skúmanie globálnych aj lokálnych vzorov v sekvenciách. Ako je však možné, že pri spracovaní sekvencií sa objavili fraktálne obrazce? Práve odchýlka od rovnomerného rozdelenia zvolených vrcholov spôsobí vznik určitej štruktúry [1] a potvrdzuje tým, že usporiadanie nukleotidov v genetických postupnostiach nie je náhodné.

2.3 Frequency matrix chaos game representation

Jednotkový štvorec *ACGT* môžeme rozdeliť na 4 kvadranty tak, že každý kvadrant prislúcha jednej z dusíkatých báz, ktoré tvoria štruktúru DNA. Ako algoritmus CGR postupne prechádza sekvenciou, pre každú bázu vygeneruje bod, ktorý vieme charakterizovať jeho x-ovou a y-ovou súradnicou. Tento bod sa určite nachádza v kvadrante príslušnej bázy vzhľadom na fakt, že je zakreslený v polovici cesty od nejakého bodu vrámci štvorca *ACGT* k jej vrcholu [12]. Iteratívne sa priestor každého kvadrantu dá opäť rozdeliť, ako môžeme vidieť na obrázku 2.2 vpravo. Napríklad kvadrant T sa ďalej delí na AT, CT, GT a TT.

Bod zodpovedajúci sekvencii dĺžky k sa tak v nachádza v štvorci sa s dĺžkou 2^{-k}. Túto vlastnosť vieme využiť pri skúmaní frekvencie k-mérov. Ak chceme zobraziť frekvencie subsekvencií dĺžky k, stačí rozdeliť priestor pôvodného CGR zobrazenia na sieť zloženú z 2^k x 2^k štvorcov a spočítať počet bodov v jednotlivých štvorcoch [13]. Výsledkom tohoto procesu je matica frekvencií všetkých k-mérov, ktorá sa dá ďalej vizualizovať pomocou grayscale. Tmavší odtieň priradený štvorcu indikuje väčšiu frekvenciu príslušnej subsekvencie. Takáto reprezentácia genetických dát dostala názov Frequency matrix chaos game representation (FCGR). Ilustrácia FCGR algoritmu sa nachádza na obrázku 2.3.



Obr. 2.4: [16] Zhrnutie procesu použitia teórie hry chaosu za účelom vytvorenia reprezentácií sekvencií, ktoré sa dajú použiť ako vstupné dáta pre bioinformatické analýzy.

2.4 Aplikácie teórie hry chaosu

Teória hry chaosu má v bioinformatike viacero možných aplikácií. Hlavným využitím je už spomínané porovnávanie postupností bez potreby zarovnania, nakoľko zobrazenie pomocou CGR a FCGR je možné vytvoriť zo sekvencie ľubovoľnej dĺžky. Pri použití týchto metód sa poradie symbolov úplne stratí [1] a nie je relevantné pre porovnanie sekvencií. Netriviálnou časťou využitia CGR a FCGR je zvolenie spôsobu, akým sa vizuálne alebo numerické reprezentácie sekvencií porovnajú. Spôsoby porovnania reprezentácií je možné rozdeliť na 2 základné kategórie: buď sa priamo porovnajú súradnice bodov v CGR alebo sa analyzujú frekvencie k-mérov (pre rôzne k) podobne ako pri iných metódach bez zarovnania [13]. Ak majú sekvencie grafickú podobu, vieme na ne aplikovať existujúce metódy spracovania obrázkov, nie je však ľahké interpretovať vizualizácie bez predchádzajúcich znalosti CGR a navyše môže dôjsť ku stratovej kompresii pri ich ukladaní a prenose [16]. Metódy založené na CGR a FCGR si vo všeobecnosti vedia poradiť aj s výpočtovo náročnými úlohami, ako porovnanie celých genómov, a teda našli uplatnenie najmä vo fylogenetickej analýze. Pri genómoch fylogeneticky blízkych organizmov sa pomocou CGR dajú získať aj cenné informácie o možných inzerciách, deléciách a substituciách [13] a pozorovať tak evolučný proces. FCGR sa používa aj na skúmanie často sa opakujúcich subsekvencií, respektíve vygenerovanie úplne absentujúcich subsekvencií [16], čo môže mať veľký význam pre identifikáciu génov spôsobujúcich niektoré ochorenia.

Tu však možnosti aplikácie teórie hry chaosu nekončia. Za posledné obdobie rapídne narástlo využitie umelej inteligencie, najmä strojového učenia, vo všetkých odvetviach a bioinformatika nie je výnimkou. Mnohé techniky, ako napríklad umelé neurónové siete, však vyžadujú vopred stanovenú veľkosť vstupu. CGR a FCGR je možné použiť na predspracovanie DNA sekvencií s rôznou dĺžkou tak, aby vytvorili vstupné dáta s

identickou veľkosťou. Metódy založené na CGR a FCGR sa dajú upraviť a použiť aj na iné genetické informácie, napríklad pri využití 20 vrcholov takto vieme vizualizovať proteíny na základe aminokyselín, z ktorých sa skladajú. S využitím Huffmanovho kódovania sa CGR dá uplatniť aj v kompresii genomických dát vzhľadom na pozoruhodnú možnosť rekonštrukcie sekvencie pri zapamätaní súradníc posledného bodu. Okrem toho existujú algoritmy na šifrovanie genetických informácií založené na CGR [16]. Zhrnutie procesu použitia teórie hry chaosu v bioinformatike, nájdeme na obrázku 2.4.

Napriek tomu, že tieto metódy pôvodne vznikli pre bioinformatiku, sú aplikovateľné aj na posúdenie kvality náhodného generátora čísel v počítačových systémoch. Pri náhodnom a rovnomernom rozdelení vygenerovaných čísel sa totižto nezobrazí žiaden viditeľný vzor. Uplatnenie teórie hry chaosu sa ďalej skúmalo pri analýze hudby [18] a zvukových signálov [4], identifikácii autorstva [26] či v ekonomike [6].

Kapitola 3

Vyhodnocovanie porovnávacích algoritmov

Dôležitou súčasťou vývoja porovnávacích algoritmov je aj testovanie a prezentovanie výsledkov. Parametre slúžiace na ich vyhodnocovanie sú najmä presnosť porovnania, časová a pamäťová zložitosť. Na posúdenie kvality navrhnutej metódy porovnávania postupností je nutné ju otestovať na reálnych dátach. V tejto kapitole preto opisujeme možnosti získania genetických dát 3.1. Okrem toho spomíname aj problémy spojené s porovnávaním metód navzájom 3.2 a webovú aplikáciu na porovnávanie metód AF-project 3.3.

3.1 Dostupnosť genetických dát

Ako sme už naznačili v kapitole 1, genetické dáta sa ukladajú do rozsiahlych databáz. DNA, RNA a proteínové sekvencie, ktoré v týchto databázach nájdeme, často pochádzajú z genómových projektov alebo vedeckých prác. Niektoré vedecké časopisy vyslovene vyžadujú, aby sekvencie spomínané v publikovaných prácach boli uložené vo verejne dostupnej databáze [14]. Genetické dáta od rôznych druhov organizmov je tak možné vyhľadať a stiahnuť priamo z webovej stránky niektorej z databáz.

Mnohé veľké databázové projekty fungujú vďaka medzinárodnej spolupráci. Ako príklad môžeme uviesť International Nucleotide Sequence Database Collaboration (INSDC) [20], ktorá združila GenBank spravovanú Národným centrom pre biotechnologické informácie (NCBI) v Spojených štátoch, DNA Data Bank Japonska (DDBJ) a Európský Nukleotidový Archív (ENA). Databázy pod INSDC sú synchronizované a prístup k dátam je bezplatný a neobmedzený. Nové genetické dáta z celého sveta sú tak k dispozícii. V databázach nájdeme široké spektrum nespracovaných čítaní, zostavených genómov, referenčných sekvencií, množstvo doplňujúcich informácií k jednotlivým vzorkám, ale aj nástroje na ich analýzu.

Pre výskumy v oblasti medicíny je veľmi dôležitá dostupnosť sekvencií pochádzajúcich od ľudí s doplňujúcimi informáciami ako životospráva daného človeka. Pri zverejňovaní takýchto dát je nutné myslieť aj na etický aspekt. UK Biobank je zdrojom až 500 000 ľudských genómov, skúmaných napríklad pri hľadaní príčin vzniku rakoviny [5]. Sú zverejnené tak, aby z nich nebolo možné identifikovať konkrétne osoby. Vzhľadom na povahu dát sa však na získanie prístupu treba registrovať a všetky informácie sa smú využiť len na medicínsky výskum vo verejnom záujme.

3.2 Problémy s porovnávaním metód

Momentálne existuje okolo 100 rôznych metód, ktoré nevyužívajú zarovnanie [32]. Pre výskumných pracovníkov môže byť náročné zorientovať sa v takej kvantite dostupných softvérov a vybrať si najvhodnejší pre konkrétny výskum. To vedie k potrebe objektívneho posúdenia ich kvality s cieľom zistiť, ktoré z nich dokážu rýchlo vygenerovať presné výsledky. Pri publikovaní metód sa väčšinou uvádza nejaké porovnanie s inými prácami, často sa však nejedná o komplexnejšie zhodnotenie všetkých doterajších prístupov. Navyše testovanie prebieha len na malej vzorke dát vyselektovanej priamo autorom metódy. Špecifický výber datasetov môže viesť ku skresleným výsledkom. Ak v článku prezentujúcom metódu nie je publikovaný odkaz na zdrojový kód, nie je taktiež možné nezávisle overiť výsledky.

Porovnávanie výkonnosti bioinformatických softvérov len z publikácií nie je uskutočniteľné, pretože autori používajú nekonzistentné hodnotiace stratégie [31]. Okrem toho sa práce testujú na rôznych genetických dátach, ktoré sa líšia napríklad veľkosťou datasetu, dĺžkou jednotlivých sekvencií či rôznorodosťou organizmov, od ktorých sekvencie pochádzajú. Nezanedbateľným faktorom je aj voľba zariadenia, na ktorom sa softvér spúšťa. Absencia nestranného porovnávania metód tak môže byť hlavným problémom, prečo napriek ich početnosti nie sú častokrát využívané v praxi.

3.3 AFproject

Možnosťou porovnávania metód sa bližšie zaoberal kolektív autorov na čele s Andrzejom Zielezinskim v práci "Benchmarking of alignment-free sequence comparison methods"[31]. V rozsiahlej štúdii vyhodnocovali 24 rôznych softvérov a dohromady vykonali 1 020 493 359 porovnaní sekvencií. Nakoniec zostavili katalóg otestovaných metód so zverejnenými zdrojovými kódmi, ktorý umožňuje jednoduchší výber a prístup k efektívnym bioinformatickým softvérom. Zaujímavým zistením bolo, že žiadna z metód nedosiahla najlepší výsledok pri všetkých datasetoch.

Výsledkom tejto iniciatívy je aj bezplatná webová aplikácia AFproject [30] pre

štandardizované hodnotenie metód, ktorá je dostupná pre verejnosť a môžeme ju nájsť na webovej adrese `https://afproject.org/app/`. Podľa autorov je vybavená interaktívnym rozhraním tak, aby bola jednoduchá pre používateľa, a poskytla rýchly a nestranný nástroj na porovnanie metód navzájom. Služba bola vytvorená primárne pre metódy bez použitia zarovnania (AF je iniciálovou skratkou pre alignment-free), no je všeobecne aplikovateľná aj na metódy využívajúce zarovnanie. Motiváciou pre vytvorenie tohto projektu bol nedostatok jasne definovaného konsenzu o porovnávaní metód. AFproject umožňuje developerom preskúmať výkonnosť vlastnej metódy pre zvolené typy údajov a oblasti výskumu a objektívne ju porovnať so súčasnými najmodernejšími nástrojmi, čím sa urýchľuje vývoj nových, presnejších riešení.

Otestovanie vlastnej metódy s pomocou AFproject-u prebieha v štyroch krokoch. Prvým krokom je stiahnutie testovacích dát, ktoré sú priamo dostupné vo webovej aplikácii. Dáta sú rozdelené do 12 datasetov a 5 kategórií na základe oblasti výskumu. Prehľad týchto datasetov uvádzame v tabuľke 3.1.

Druhým krokom je vypočítanie podobností sekvencií vlastnou metódou. Vstupom do programu sú postupnosti zo zvoleného datasetu vo formáte FASTA. Výstup z programu musí obsahovať porovnanie každej dvojice sekvencií. Najjednoduchším z akceptovaných formátov je Tap-Separated Value Format (TSV), teda súbor s troma stĺpcami oddelenými tabulátorom, ktoré obsahujú identifikátory dvoch porovnávaných postupností a výsledok porovnania. Ďalším je Phylip Distance Matrix, matica obsahujúca výsledky porovnania. Pre niektoré datasety je akceptovaný aj formát Newick Tree, kde sú identifikátory sekvencie usporiadané do fylogenetického stromu.

Tretím krokom je nahranie výsledkov v niektorom z akceptovaných formátov cez formulár, kde treba uviesť aj základné informácie o vlastnej metóde, konkrétne názov metódy a jej parametre. Nepovinne môže autor uviesť aj opis metódy či link na webovú stránku svojho projektu.

V štvrtom kroku prebehne automatické vyhodnotenie metódy. Nahrané výsledky z vlastnej metódy sa porovnajú s referenčnými dátami pomocou metriky špecifickej pre príslušnú kategóriu datasetu. Ďalej prebehne porovnanie s inými uverejnenými metódami. Autor dostane zhodnocujúcu správu a taktiež vizualizáciu dosiahnutých výsledkov, pri väčšine datasetov vo forme fylogenetického stromu. Nakoniec sa môže rozhodnúť, či správu zverejní alebo ju ponechá ako súkromnú.

Tabuľka 3.1: [30] Prehľad všetkých datasetov dostupných na stiahnutie vo webovej aplikácii AFproject

Názov datasetu	**Oblasť výskumu**	**Typ sekvencií**
Cis-regulatory modules (CRM)	Regulačné sekvencie	nekódujúca DNA
Low sequence identity ($< 40\%$)	Klasifikácia proteínových sekvencií	proteín
High sequence indentiy ($\geq 40\%$)	Klasifikácia proteínových sekvencií	proteín
SwissTree	Inferovanie génových stromov	proteín
29 E.coli/Shigella strains	Fylogenetika	nezostavené čítania
29 E.coli/Shigella strains	Fylogenetika	celý genóm
25 fish mitochondrial genomes	Fylogenetika	mitochondriálny genóm
14 plant species	Fylogenetika	nezostavené čítania
14 plant species	Fylogenetika	celý genóm
27 E.coil/Shigella strains	Horizontálny prenos genetickej informácie	celý genóm
8 Yersinia species	Horizontálny prenos genetickej informácie	celý genóm
33 simulated genomes	Horizontálny prenos genetickej informácie	celý umelo vytvorený genóm

Kapitola 4

Ciele práce

Hlavným cieľom našej práce bolo aplikovať teóriu hry chaosu na hľadanie podobnosti veľkých genomických postupností a porovnať ju z hľadiska presnosti s inými metódami bez zarovnania. Na dosiahnutie tohto cieľu bolo potrebné vykonať nasledovné kroky:

- Preštudovať teóriu hry chaosu a dôležité poznatky z oblasti bioinformatiky
- Navrhnúť spôsob spracovania formátov súborov, v ktorých sú sekvencie uložené
- Vytvoriť vizuálne aj numerické reprezentácie sekvencií na základe teórie hry chaosu
- Vybrať reprezentáciu sekvencií vhodnú na presné a rýchle porovnávanie
- Navrhnúť, pomocou ktorých metrík bude možné reprezentácie sekvencií porovnávať
- Vybrať si dátové množiny
- Vyhodnotiť presnosť našej metódy na vybraných dátových množinách
- Vyskúšať, ako sa líši porovnanie celého genómu alebo len jeho časti

V podkapitole 4.1 uvádzame funkcie, ktoré by výsledný systém na spracovanie, analýzu a porovnávanie postupností mal obsahovať.

4.1 Funkcie systému

Aby sa dalo s genetickými dátami pracovať, systém bude vedieť spracovať formát súboru FASTA, rozdeliť uložené informácie na hlavičku a telo obsahujúce sekvenciu. Systém bude taktiež vedieť spracovať formát súboru multi-FASTA obsahujúci niekoľko nezostavených čítaní.

Používateľ si pomocou systému bude vedieť vytvoriť numerickú aj vizuálnu verziu chaos game representation sekvencie DNA. Vizuálnu CGR si bude môcť zobraziť aj priblížiť, aby videl detailnejšiu štruktúru sekvencie. Ďalej si používateľ bude vedieť vytvoriť numerickú aj vizuálnu verziu frequency matrix chaos game representation sekvencie DNA pre zadané číslo k (dĺžka k-mérov). Vizuálnu FCGR si bude opäť môcť zobraziť a pre všetky k-méry zmapovať, v akom množstve sa vyskytli v sekvencií. Používateľ bude môcť jednoducho zistiť, ktorý k-mér sa v sekvencii vyskytol najmenej a ktorý najviac často. Pre bližšiu analýzu je dôležitá aj schopnosť systému vygenerovať zoznam k-mérov, ktoré sa v sekvencii vôbec nevyskytli. Všetky reprezentácie sekvencie DNA bude možné uložiť vo vhodnom formáte.

Podstatnou funkciou systému bude vykonanie porovnania dvoch sekvencií v podobe frekvenčných matíc. Používateľ si bude môcť vybrať z piatich metrík, ktorými sa frekvenčné matice dajú porovnať. Na základe zadanej cesty do niektorého priečinka bude systém vedieť automaticky vykonať porovnanie všetkých párov sekvencií z daného priečinka. Používateľ si bude môcť zistiť, ktoré dve sekvencie z priečinka sa podobajú najviac a ktoré najmenej. Výsledky porovnania sa budú dať uložiť vo vhodnom formáte.

Kapitola 5

Návrh riešenia

5.1 Spracovanie vstupných súborov

5.2 Reprezentácia sekvencií

5.3 Porovnanie sekvencií

Počet porovnaní, ktoré je potrebné vykonať pre vyhodnotenie datasetu s n súbormi, sa dá vypočítať pomocou rovnice (5.1)

$$P = \frac{n(n-1)}{2}. \tag{5.1}$$

5.4 Výber metrík

Euklidovská vzdialenosť

$$d_{euc}\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(p_i - q_i\right)^2} \tag{5.2}$$

Manhattanská vzdialenosť

$$d_{man}\left(p,q\right) = \sum_{i=1}^{n}\left|p_i - q_i\right| \tag{5.3}$$

Canberrská vzdialenosť

$$d_{can}\left(p,q\right) = \sum_{i=1}^{n}\frac{\left|p_i - q_i\right|}{\left|p_i\right| + \left|q_i\right|} \tag{5.4}$$

Kosínusová vzdialenosť

$$d_{cos}(p,q) = 1 - \frac{p \cdot q}{\|p\|_2 \|q\|_2} = 1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}} \tag{5.5}$$

Jensen-Shannon vzdialenosť

$$d_{js}(p,q) = \sqrt{\frac{D(p \parallel m) + D(q \parallel m)}{2}} \tag{5.6}$$

$$m = \frac{1}{2}(p+q) \tag{5.7}$$

D je Kullbackova–Leiblerova divergencia.

$$D(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{5.8}$$

5.5 Výber dátových množín

Datasety sme prevzali z webovej aplikácie AFproject [30]. Ich výber prebiehal tak, aby sa líšili:

- ríšou organizmov, od ktorých sekvencie pochádzajú: baktérie, rastliny aj živočíchy
- typom sekvencií: nezostavené čítania, mitochondriálny genóm, celý genóm, simulovaný genóm
- počtom súborov: od 8 do 203
- priemerným počtom znakov na jednu sekvenciu v datasete: od nezostavených čítaní dĺžky 150 po genómy so stovkami miliónov nukleotidov

Keďže sekvencie v jednotlivých datasetoch pochádzajú od evolučne blízkych organizmov, ich dĺžky sú obdobné. Vybrané datasety sú určené pre dve oblasti výskumu, Fylogenetika a Horizontálny prenos genetickej informácie. Opisujeme ich v podčastiach 5.5.1 a 5.5.2. Na označovanie datasetov v práci budeme používať dvojicu názov datasetu a typ sekvencií.

5.5.1 Fylogenetika

Fylogenetika mapuje evolučné vzťahy medzi živými organizmami [9]. Skúma teda, ako sa genetické informácie dedia. Takýto prenos génov nazývame aj vertikálny. Na skúmanie fylogenetických vzťahov je potrebné porovnávať veľké genomické postupnosti, táto oblasť teda zodpovedá hlavnému cieľu našej práce.

Tabuľka 5.1: [30] Podrobný prehľad datasetov z oblasti výskumu Fylogenetika

Názov datasetu	**Typ sekvencií**	**Počet súborov**	**Priemerný počet znakov**
29 E.coli/Shigella strains	nezostavené čítania	203	150 na 1 čítanie
29 E.coli/Shigella strains	celý genóm	29	5 miliónov
25 fish mitochondrial genomes	mitochondriálny genóm	25	16 tisíc
14 plant species	nezostavené čítania	98	150 na 1 čítanie
14 plant species	celý genóm	29	350 miliónov

Podrobný prehľad datasetov z oblasti výskumu Fylogenetika uvádzame v tabuľke 5.1. Dva datasety obsahujú aj nezostavené čítania vo formáte multi-FASTA. Súbory obsahujúce nezostavené čítania sú rozdelené do 7 skupín podľa hĺbky sekvenčného pokrytia, preto je pri nahrávaní výsledkov potrebné odovzdať 7 súborov.

5.5.2 Horizontálny prenos genetickej informácie

Horizontálny prenos genetickej informácie znamená, že došlo k odovzdaniu genetického materiálu inak ako z rodičov na potomkov. Tento proces pozorujeme najmä u baktérií, ktoré sa tak adaptujú na nové podmienky a stávajú sa napríklad rezistentnými voči antibiotikám [25]. Určenie fylogenetických vzťahov môže byť v takomto prípade náročnejšie, preto sme sa rozhodli otestovať našu metódu aj na datasetoch pre túto oblasť.

Podrobný prehľad datasetov z oblasti výskumu Horizontálny prenos genetickej informácie uvádzame v tabuľke 5.2. Zaujímavým je najmä dataset “33 simulated genomes“ (celý umelo vytvorený genóm) obsahujúci umelo vytvorené genomické postupnosti pomocou simulátora. Simulácie sú rozdelené do 5 skupín podľa rozsahu horizontálneho prenosu, preto je pri nahrávaní výsledkov potrebné odovzdať 5 súborov.

5.6 Vyhodnotenie metódy

Aby sme našu metódu vedeli vyhodnoť čo najpresnejšie a najobjektívnejšie, rozhodli sme sa využiť webovú aplikáciu AFproject [30]. Vďaka tomu sme získali aj zhodnocujúce správy porovnávajúce našu prácu s inými a taktiež vizualizácie dosiahnutých výsledkov vo forme fylogenetických stromov. Na zápis výsledkov porovnania všetkých párov sekvencií z datasetu sme zvolili formáty súborov Tap-Separated Value (TSV) a

Tabuľka 5.2: [30] Podrobný prehľad datasetov z oblasti výskumu Horizontálny prenos genetickej informácie

Názov datasetu	**Typ sekvencií**	**Počet súborov**	**Priemerný počet znakov**
27 E.coli/Shigella strains	celý genóm	27	5 miliónov
8 Yersinia species	celý genóm	8	5 miliónov
33 simulated genomes	celý umelo vytvorený genóm	165	2 milióny

Phylip distance matrix, aby boli kompatibilné s rozhraním AFprojectu. Tieto formáty sú navyše prehľadné a ľahko pochopiteľné. Výsledné hodnoty porovnaní sa normalizujú a vo formáte Phylip distace matrix sa zaokrúhlujú na 3 desatinné čísla.

5.7 Experiment

Kapitola 6

Implementácia

Kapitola 7

Dosiahnuté výsledky

Záver

Literatúra

[1] Jonas S Almeida. Sequence analysis by iterated maps, a review. *Briefings in bioinformatics*, 15(3):369–375, 2014.

[2] Gaëtan Benoit, Claire Lemaitre, Dominique Lavenier, Erwan Drezen, Thibault Dayris, Raluca Uricaru, and Guillaume Rizk. Reference-free compression of high throughput sequencing data with a probabilistic de bruijn graph. *BMC bioinformatics*, 16(1):1–14, 2015.

[3] Broňa Brejová and Tomáš Vinař. Metódy v bioinformatike. *Fakulta matematiky, fyziky a informatiky Univerzita Komenského v Bratislave*, 2011.

[4] Madison Cohen-McFarlane, Kevin Dick, James R Green, and Rafik Goubran. Chaos game representation of audio signals. In *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2021.

[5] Megan C. Conroy et al. Uk biobank: a globally important resource for cancer research. *British Journal of Cancer*, 128(4):519–527, 2023.

[6] Constantin P Cristescu, Cristina Stan, and Eugen I Scarlat. Modeling with the chaos game (i). simulating some features of real time series. *UPB Sci Bull Ser A*, 71:95–100, 2009.

[7] Fatima Cvrčková. *Úvod do praktické bioinformatiky*. Academia, 2006.

[8] Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

[9] Tomáš Farkaš, Jozef Sitarčík, Broňa Brejová, and Mária Lucká. Swspm: A novel alignment-free dna comparison method based on signal processing approaches. *Evolutionary Bioinformatics*, 15:1176934319849071, 2019.

[10] Umesh Ghoshdastider and Banani Saha. GenomeCompress: a novel algorithm for DNA compression, 2005.

[11] Rosario Gilmary, Akila Venkatesan, and Govindasamy Vaiyapuri. Compression techniques for DNA sequences: A thematic review. *J. Comput. Sci. Eng.*, 15(2):59–71, 2021.

[12] H Joel Jeffrey. Chaos game representation of gene structure. *Nucleic acids research*, 18(8):2163–2170, 1990.

[13] Jijoy Joseph and Roschen Sasikumar. Chaos game representation for comparison of whole genomes. *BMC bioinformatics*, 7(1):1–10, 2006.

[14] Arthur M. Lesk. bioinformatics. [Citované 2024-01-10] Dostupné na `https://www.britannica.com/science/bioinformatics`.

[15] LibreTexts. Storing genetic information. [Citované 2023-05-20] Dostupné na `https://bio.libretexts.org/Courses/Lumen_Learning/Biology_for_Non-Majors_I_%28Lumen%29/08%3A_DNA_Structure_and_Replication/8.02%3A_Storing_Genetic_Information`.

[16] Hannah Franziska Löchel and Dominik Heider. Chaos game representation and its applications in bioinformatics. *Computational and structural biotechnology journal*, 19:6263–6271, 2021.

[17] Vijini Mallawaarachchi. Pairwise sequence alignment using biopython. [Citované 2024-01-25] Dostupné na `https://towardsdatascience.com/pairwise-sequence-alignment-using-biopython-d1a9d0ba861f`.

[18] Brian Meloon and Julien C Sprott. Quantification of determinism in music using iterated function systems. *Empirical Studies of the Arts*, 15(1):3–13, 1997.

[19] Qingxi Meng, Shubham Chandak, Yifan Zhu, and Tsachy Weissman. Reference-free lossless compression of nanopore sequencing reads using an approximate assembly approach. *Scientific Reports*, 13(1):2082, 2023.

[20] National Institutes of Health et al. International nucleotide sequence database collaboration. [Citované 2024-04-22] Dostupné na `https://www.insdc.org/`.

[21] Jill Roughan. Your essential guide to different file formats in bioinformatics. [Citované 2023-05-27] Dostupné na `https://www.formbio.com/blog/your-essential-guide-different-file-formats-bioinformatics`.

[22] Hiruna Samarakoon, Hasindu Gamaarachchi, et al. Flexible and efficient handling of nanopore sequencing signal data with slow5tools. *Genome Biology*, 24(1):69, 2023.

[23] Muhammad Sardaraz and Muhammad Tahir. SCA-NGS: Secure compression algorithm for next generation sequencing data using genetic operators and block sorting. *Science Progress*, 104(2):00368504211023276, 2021.

[24] Milton Silva, Diogo Pratas, and Armando J Pinho. Efficient DNA sequence compression with neural networks. *GigaScience*, 9(11):giaa119, 2020.

[25] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015.

[26] Catalin Stoean and Daniel Lichtblau. Author identification using chaos game representation and deep learning. *Mathematics*, 8(11):1933, 2020.

[27] Susana Vinga. Alignment-free methods in computational biology, 2014.

[28] Susana Vinga and Jonas Almeida. Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513–523, 2003.

[29] Aimin Yang, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang. Review on the application of machine learning algorithms in the sequence data mining of dna. *Frontiers in Bioengineering and Biotechnology*, 8:1032, 2020.

[30] Andrzej Zielezinski et al. Afproject. [Citované 2024-03-19] Dostupné na `https://afproject.org/app/`.

[31] Andrzej Zielezinski et al. Benchmarking of alignment-free sequence comparison methods. *Genome biology*, 20(1):1–18, 2019.

[32] Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18:1–17, 2017.