Politechnika Śląska

PROJEKT INŻYNIERSKI

Narzędzie do analiz filogenetycznych.

Filip WSPANIAŁY
Nr albumu: ⟨**306982**⟩

Kierunek: ⟨Inżynieria biomedyczna⟩
Specjalność: ⟨Informatyka i aparatura medyczna⟩

PROWADZĄCY PRACĘ
⟨**dr inż. Anna Tamulewicz**⟩
KATEDRA ⟨**Katedra Informatyki Medycznej i Sztucznej Inteligencji**⟩
Wydział Inżynierii Biomedycznej

Zabrze 2026

Tytuł pracy

Narzędzie do analiz filogenetycznych.

Streszczenie

Celem niniejszej pracy inżynierskiej było zaprojektowanie i zaimplementowanie systemu informatycznego do analiz filogenetycznych sekwencji białkowych. W ramach pracy pogłębiono wiedzę z zakresu filogenetyki oraz zintegrowano uznane algorytmy i narzędzia bioinformatyczne (MAFFT, IQ-TREE, Biopython), tworząc kompleksowe rozwiązanie do przetwarzania danych. Zaprezentowano architekturę aplikacji umożliwiającą przygotowanie danych wejściowych, wyrównanie sekwencji, analizę filogenetyczną z automatycznym wyborem modelu ewolucyjnego oraz wizualizację drzew filogenetycznych. W pracy opisano szczegółowo strukturę programu, uzasadnienie wyboru narzędzi oraz wyniki testów na rzeczywistych danych z NCBI.

Słowa kluczowe

filogenetyka, bioinformatyka, wyrównanie sekwencji, analiza filogenetyczna, model ewolucyjny, drzewo filogenetyczne

Thesis title

Phylogenetic analysis tool.

Abstract

The objective of this engineering thesis was to design and implement a computational system for phylogenetic analysis of protein sequences. The work deepened knowledge in phylogenetics and integrated established bioinformatics algorithms and tools (MAFFT, IQ-TREE, Biopython), creating a comprehensive solution for data processing. The application architecture is presented, enabling input data preparation, sequence alignment, phylogenetic analysis with automatic evolutionary model selection, and visualization of phylogenetic trees. The thesis provides a detailed description of the program structure, rationale for tool selection, and test results on real datasets from NCBI.

Key words

phylogenetics, bioinformatics, sequence alignment, phylogenetic analysis, evolutionary model, phylogenetic tree

Spis treści

Rozdział 1

Wstęp

1.1 Wprowadzenie do tematu

Analiza filogenetyczna wykorzystuje dane z dziedziny filogenetyki do rekonstrukcji ewolucyjnych zależności i podobieństw międzygatunkowych. Efektem procesu jest drzewo filogenetyczne, generowane w oparciu o wybrany model substytucji, które za pomocą rozgałęzień ilustruje odległości ewolucyjne między analizowanymi taksonami. Kluczowym przełomem w rozwoju tej dyscypliny okazał się postęp informatyki i programowania, umożliwiający automatyzację metod obliczeniowych oraz znaczące przyspieszenie analiz sekwencji genetycznych. Rozwój ten przyczynił się bezpośrednio do powstania nowych algorytmów wyrównywania sekwencji oraz metod inferencji drzew filogenetycznych. Wraz ze wzrostem złożoności algorytmów oraz liczby dostępnych narzędzi, proces analizy filogenetycznej przestał być jednorazowym obliczeniem, a stał się wieloetapowym zadaniem wymagającym doboru metod, parametrów oraz interpretacji wyników. W praktyce badawczej prowadzi to do konieczności łączenia wielu narzędzi programistycznych oraz zarządzania złożonymi procesami obliczeniowymi.

1.2 Osadzenie problemu w dziedzinie

Współczesna analiza filogenetyczna osiągnęła wysoki poziom zaawansowania dzięki ciągłemu doskonaleniu algorytmów i opracowywaniu nowych metod rekonstrukcji drzew ewolucyjnych. Natomiast brak pewności co do optymalności najlepszych algorytmów powoduje, że nawet systemy oceny i porównywania metod opierają się głównie na statystyce. Kluczowym wyzwaniem w analizach filogenetycznych pozostaje pytanie o prawdopodobieństwo, że uzyskane rozwiązanie jest rzeczywiście najlepsze dla danego zbioru danych. Istniejące systemy analityczne mogą wskazać optymalny algorytm, metodę lub model substytucji dla konkretnej sekwencji, jednak ostateczny wybór wymaga integracji wielu narzędzi w spójny workflow. W praktyce badawczej analizy filogenetyczne wymagają

wielokrotnego testowania algorytmów, modeli substytucji oraz parametrów wejściowych, co prowadzi do powstawania złożonych, trudnych do odtworzenia, porównania oraz powtarzalnego uruchamiania workflowów analitycznych. Brakuje systemów, które w sposób zintegrowany umożliwiałyby porównywanie wyników różnych metod, zarządzanie eksperymentami analitycznymi oraz wspomaganie decyzji o wyborze końcowego drzewa filogenetycznego.

1.3 Cel pracy

Celem pracy było zaprojektowanie oraz zaimplementowanie systemu wspomagającego analizy filogenetyczne, umożliwiającego integrację poszczególnych etapów procesu analitycznego w spójny workflow. System ma na celu wsparcie użytkownika w rekonstrukcji drzew filogenetycznych poprzez automatyzację kluczowych kroków analizy oraz uporządkowaną prezentację wyników.

1.4 Zakres pracy

Zakres pracy obejmuje analizę podstaw teoretycznych filogenetyki oraz przegląd wybranych metod i narzędzi wykorzystywanych w analizach filogenetycznych. W ramach pracy dokonano porównania dostępnych podejść do wyrównywania sekwencji, inferencji drzew filogenetycznych oraz doboru modeli ewolucyjnych. Praca obejmuje zaprojektowanie i implementację systemu integrującego wybrane narzędzia analityczne w spójny workflow, umożliwiającego przeprowadzenie analizy filogenetycznej oraz wizualizację uzyskanych wyników. Zakres pracy nie obejmuje opracowywania nowych algorytmów filogenetycznych ani formalnej oceny biologicznej poprawności uzyskanych drzew.

1.5 Zwięzła charakterystyka rozdziałów

1.6 Określenie wkładu autora

Autor odpowiadał za zaprojektowanie architektury oraz implementację systemu wspomagającego analizy filogenetyczne. W ramach pracy autor dokonał integracji wybranych narzędzi do wyrównywania sekwencji, inferencji drzew filogenetycznych oraz doboru modeli ewolucyjnych w spójny workflow analityczny. Autor był również odpowiedzialny za przygotowanie mechanizmów wizualizacji wyników analizy oraz obsługę danych wejściowych, w tym pozyskiwanie sekwencji genetycznych z publicznie dostępnej bazy NCBI.

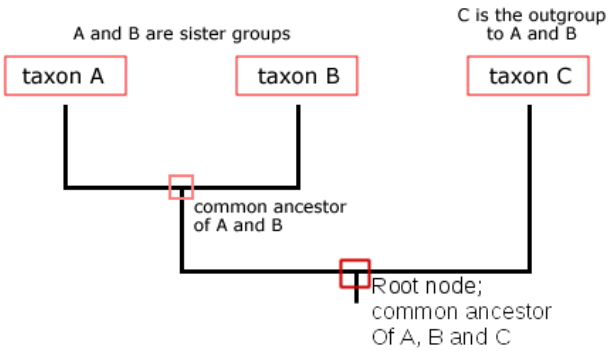
Rozdział 2

Podstawy teoretyczne analizy filogenetycznej

2.1 Filogenetyka i drzewa filogenetyczne

Filogenetyka jest dziedziną biologii zajmującą się badaniem filogenezy, czyli historii rozwoju rodowego organizmów oraz relacji pokrewieństwa pomiędzy taksonami. Obejmuje ona analizę przebiegu procesów ewolucyjnych prowadzących do różnicowania organizmów i powstawania nowych linii rozwojowych. Analiza filogenetyczna umożliwia określanie zależności ewolucyjnych między gatunkami i taksonami na podstawie różnych źródeł danych, takich jak zapisy paleontologiczne, anatomia porównawcza oraz dane molekularne.

W niniejszej pracy wykorzystywane są metody filogenetyki molekularnej, które opierają się na analizie sekwencji DNA lub białek w celu rekonstrukcji relacji ewolucyjnych. Wynikiem takiej analizy jest przedstawione na rysunku 2.1 drzewo filogenetyczne — struktura graficzna przedstawiająca hipotezę pokrewieństwa pomiędzy badanymi taksonami. W zależności od zastosowanej metody rekonstrukcji oraz modelu ewolucyjnego, długości gałęzi drzewa mogą odzwierciedlać miarę zmian genetycznych lub mieć charakter wyłącznie topologiczny. [1][2]



Rysunek 2.1: Drzewo filogenetyczne; źródło: [1]

2.2 Sekwencje genetyczne jako dane wejściowe

Sekwencje genetyczne stanowią podstawowe dane wejściowe wykorzystywane w analizach filogenetycznych realizowanych przez system. Są to uporządkowane ciągi symboli reprezentujących nukleotydy DNA lub aminokwasy budujące białka. W zależności od rodzaju analizy, system może operować na sekwencjach nukleotydowych lub sekwencjach aminokwasowych.

W praktyce analitycznej sekwencje genetyczne pozyskiwane są z badań własnych lub publicznych baz danych, takich jak NCBI, DDBJ, czy ENA i najczęściej zapisywane w formacie FASTA pokaznym na rysunku 2.2. Format ten umożliwia jednoznaczną identyfikację sekwencji oraz jej dalsze przetwarzanie przez narzędzia bioinformatyczne. Sekwencje mogą różnić się długością, stopniem kompletności oraz jakością danych, co wpływa na przebieg kolejnych etapów analizy. [1]

```
>NP_999866.1 cytochrome c oxidase subunit 4 isoform 1, mitochondrial [Danio rerio]
MLATTAFRLVGKRALSTSICLRGAHGVAKVEDYSLPAYFDRRESPLPEIKFVQQLSADQKSLKEKEKGSW
AALSKEEKIALYRISFKESFAEMNQGSGEWKSVVAGIFFFVGLTGLVVLWQRKYVYGDVPNTFDPEYKQK
EIQRMLDMRINPVQGFAAKWDYENNAWKK
```

Rysunek 2.2: Sekwencja białkowa; źródło: [2]

Ze względu na występowanie różnic długości sekwencji oraz obecność insercji i delecji, bezpośrednie porównywanie sekwencji nie jest możliwe. Z tego powodu przed rekonstrukcją drzewa filogenetycznego konieczne jest przeprowadzenie etapu wyrównywania sekwencji, który umożliwia ich porównywanie w ujednoliconej postaci.

2.3 Wyrównywanie sekwencji

Wyrównanie sekwencji jest kluczowym etapem w procesie analizy filogenetycznej, umożliwiającym porównywanie sekwencji genetycznych pochodzących od różnych organizmów. W wyniku procesów ewolucyjnych, takich jak insercje i delecje (indele), sekwencje mogą różnić się długością oraz zawierać przesunięcia pozycji homologicznych, co uniemożliwia ich bezpośrednie porównanie.

Celem wyrównania sekwencji jest identyfikacja pozycji homologicznych pomiędzy sekwencjami poprzez wprowadzenie przerw (ang. gaps), tak aby możliwe było ich dalsze przetwarzanie w kolejnych etapach analizy (rysunek 2.3). Wyrównanie pozwala na ujednolicenie długości sekwencji oraz określenie podobieństw i różnic wynikających z przebiegu ewolucji.

```
>seq1
M---LASRAF-SLIGRR--ALSTSICLR--A--------HGHAGVVKAEDFSLPAYVDRR
DVPLPEAAFVKQLSAQQKALKEKEKASWTALSVDEKVELYRIKFNETYAEMNKGTNEWKT
VLGGVLFFLGLTGVILIWQKIYMYGPIPHTFSDEWVSMQTKRMLDMRINPVEGISSQWDF
EKNEWKK
>seq2
M---LATRVF-NLIGRR--AISTSVCVR--A--------HG--SVVKSEDYALPVYVDRR
DYPLPDVAHVKNLSASQKALKEKEKASWSSLSMDEKVELYRLKFNESFAEMNRSTNEWKT
IVGTALFFIGFTALLLIWEKHYVYGPIPHTFEEEWVAKQTKRMLDMKVAPIQGFSAKWDY
DKNEWKK
```

Rysunek 2.3: Wyrównane sekwencje białkowe; źródło: [2]

W zależności od zastosowanego algorytmu możliwe jest wykonywanie wyrównań globalnych, obejmujących całe sekwencje, lub lokalnych, koncentrujących się na ich fragmentach. Przykładami klasycznych algorytmów wykorzystywanych w tym celu są: algorytm Needlemana–Wunscha dla wyrównań globalnych oraz algorytm Smitha–Watermana dla wyrównań lokalnych. Dobór odpowiedniej metody wyrównania ma istotny wpływ na jakość dalszej analizy filogenetycznej. Wyrównanie może wymagać także poprawek manualnych, zwłaszcza w przypadku sekwencji o niskim stopniu podobieństwa lub zawierających liczne indels. W praktyce badawczej często stosuje się podejście iteracyjne, polegające na wielokrotnym wyrównywaniu sekwencji z różnymi parametrami oraz ręcznej korekcie wyników w celu uzyskania optymalnego wyrównania.[1][3]

2.4 Metody rekonstrukcji drzew filogenetycznych

Rekonstrukcja drzewa filogenetycznego stanowi złożone zagadnienie statystyczne i algorytmiczne, a jej wynik zależy od przyjętych założeń biologicznych oraz zastosowanej metody analizy. W praktyce badawczej dostępnych jest wiele metod rekonstrukcji filogenezy, które mogą prowadzić do odmiennych wyników nawet dla tego samego zestawu danych. Z tego względu często stosuje się podejście polegające na porównywaniu rezultatów uzyskanych z wykorzystaniem różnych metod.

Wśród podstawowych metod rekonstrukcji drzew filogenetycznych wyróżnia się metody:

- metoda największej parsymonii,
- metody odległościowe,
- metody największej wiarygodności,
- metody bayesowskie.

Metody te różnią się sposobem modelowania procesu ewolucyjnego oraz podejściem do oceny najlepszego drzewa filogenetycznego. W zależności od zastosowanej metody, długości gałęzi drzewa mogą reprezentować liczbę zmian ewolucyjnych, estymowaną odległość genetyczną lub mieć charakter wyłącznie topologiczny.

Metody rekonstrukcji drzew filogenetycznych różnią się zakresem przyjmowanych założeń oraz stopniem złożoności obliczeniowej. Metoda największej parsymonii opiera się na minimalizacji liczby zmian ewolucyjnych i nie wykorzystuje jawnych modeli probabilistycznych. Metody odległościowe bazują na macierzy odległości genetycznych pomiędzy sekwencjami, upraszczając analizę kosztem utraty części informacji. Metody największej wiarygodności oraz metody bayesowskie wykorzystują modele probabilistyczne opisujące proces substytucji, co pozwala na bardziej realistyczne modelowanie ewolucji, jednak wiąże się z większym kosztem obliczeniowym. [3]

Rozdział 3

Przegląd istniejących narzędzi i rozwiązań

3.1 Narzędzia do wyrównywania sekwencji

Szeroki wachlarz dostępnych programów do wyrównywania sekwencji umożliwia elastyczne dopasowanie narzędzia do rodzaju oraz rozmiaru analizowanych danych oraz potrzeb badawczych. Programy te wykorzystują różne algorytmy i strategie dopasowania, od klasycznych metod opartych na dynamicznym programowaniu, po bardziej zaawansowane heurystyki optymalizacyjne. Poniżej przedstawiono przegląd wybranych narzędzi do wielosekwencyjnego wyrównywania sekwencji[1]:

- **Clustal W i Clustal Omega**
- **MAFFT**
- **MUSCLE**
- **T-Coffee**
- **PRANK**
- **ProbCons**

Do analizy i wprowadzania poprawek ręcznie mozna wykorzystać edytory wyrównań takie jak[1]:

- **Jalview**
- **AliView**
- **MacVim**

3.2 Narzędzia do inferencji filogenetycznej

Inferencja filogenetyczna polega na rekonstrukcji drzewa filogenetycznego w oparciu o dane sekwencyjne. Dostępne są różne programy i metody, które realizują tę funkcję, wykorzystując odmienne podejścia. Wybór odpowiedniego narzędzia zależy od rodzaju danych, liczby sekwencji oraz wymagań dotyczących dokładności i czasu obliczeń[1].

- **IQ-TREE 2**
- **RAxML**
- **MrBayes**
- **PhyML**
- **PHYLIP**
- **PAUP***
- **MEGA**

Po zakończeniu inferencji filogenetycznej często konieczne jest ocenienie wiarygodności uzyskanego drzewa. Do tego celu stosuje się metody takie jak bootstrap czy analiza bayesowska, które pozwalają na oszacowanie pewności poszczególnych gałęzi drzewa. Wiele z wymienionych narzędzi oferuje wbudowane funkcje do przeprowadzania takich analiz.

Po zakończeniu analizy filogenetycznej otrzymuje się w postaci pliku tekstowego zawierającego opis drzewa w formacie Newick lub Nexus, który może być dalej przetwarzany i wizualizowany za pomocą dedykowanych narzędzi.[1]

3.3 Narzędzia do wizualizacji drzew filogenetycznych

Do wizualizacji drzew filogenetycznych dostępne są różne narzędzia, które umożliwiają graficzne przedstawienie wyników analizy filogenetycznej. Poniżej przedstawiono wybrane programy do wizualizacji drzew:

- **FigTree**
- **Dendroscope**
- **iTOL**
- **ETE Toolkit**
- **Phylo.io**

3.4 Ograniczenia istniejących rozwiązań

Chociaż istnieją systemy i frameworki do automatyzacji analiz bioinformatycznych (np. Galaxy, Snakemake, Nextflow), w praktyce często wymaga się ręcznego ustawiania parametrów poszczególnych narzędzi oraz integracji wyników. Zaprojektowany system automatyzuje pełny łańcuch analizy — od wczytania sekwencji, przez wyrównanie i rekonstrukcję drzewa filogenetycznego, aż po wizualizację wyników — co ułatwia prowadzenie badań, zwiększa powtarzalność wyników i umożliwia prostsze korzystanie z analizy filogenetycznej, także osobom dopiero rozpoczynającym pracę z tymi metodami.

3.5 Opis zastosowanych technologii i narzędzi

3.5.1 Python

Python jest wysokopoziomowym językiem programowania, powszechnie wykorzystywanym w bioinformatyce oraz do automatyzacji analiz danych. W projekcie Python został użyty jako główny język implementacji systemu, odpowiadający za sterowanie przebiegiem analizy filogenetycznej, obsługę interfejsu użytkownika oraz integrację z zewnętrznymi narzędziami bioinformatycznymi. Implementacja została wykonana w środowisku programistycznym Visual Studio Code.

Biblioteki

W projekcie wykorzystano następujące biblioteki Pythona:

- **Biopython** – analiza danych biologicznych
- **subprocess** – integracja z zewnętrznymi narzędziami (MAFFT, IQ-TREE)
- **Matplotlib** – wizualizacja drzew.
- **tkinter** – interfejs użytkownika do tworzenia aplikacji okienkowych.

3.5.2 Ubuntu - Linux

System wykorzystuje środowisko systemu operacyjnego Ubuntu (Linux). Wybór systemu Linux podyktowany był wysoką kompatybilnością z narzędziami bioinformatycznymi, takimi jak MAFFT oraz IQ-TREE 2, które są natywnie rozwijane i testowane w tym środowisku.

3.5.3 MAFFT

Narzędzie do wyrównywania sekwencji DNA i białek. MAFFT oferuje różne algorytmy, które można dostosować do rozmiaru i charakterystyki danych wejściowych. W projekcie MAFFT został wykorzystany do przeprowadzenia etapu wyrównywania sekwencji przed rekonstrukcją drzewa filogenetycznego.

3.5.4 IQ-TREE 2

Narzędzie do rekonstrukcji drzew filogenetycznych metodą największej wiarygodności (Maximum Likelihood). Program umożliwia automatyczny dobór modelu substytucji oraz oferuje wydajne algorytmy optymalizacji drzewa, co czyni go jednym z najczęściej wykorzystywanych narzędzi do inferencji filogenetycznej.

3.5.5 System kontroli wersji: Git

Git pozwala na monitorowanie zmian oraz zarządzanie historią w kodzie źródłowym.

3.5.6 Repozytorium zdalne: Github

Github to zdalne repozytorium w pełni zintegrowane z Git. Zostało użyte do bezpiecznego przechowywania projektu oraz umożliwienia nadzoru nad postępem prac.

Rozdział 4

Specyfikacja zewnętrzna systemu

4.1 Wymagania sprzętowe i programowe

4.2 Instalacja

4.3 Instrukcja obsługi

Rozdział 5

Specyfikacja wewnętrzna systemu

5.1 Idea systemu

5.2 Wymagania funkcjonalne

5.3 Architektura ogólna systemu

5.4 Opis najważniejszych modułów systemu

Rozdział 6

Testy i analiza działania systemu

6.1 Testy dla sekwencji białek z bazy NCBI

6.2 Testy dla sekwencji nukleotydowych z bazy NCBI

6.3 Analiza wyników

6.4 Analiza działania systemu

6.5 Potencjalny rozwój systemu

Rozdział 7

Podsumowanie i wnioski

Bibliografia

[1] Grzegorz Góralski *Wstęp do filogenetyki molekularnej i tworzenia drzew filogenetycznych* [online] `https://ggoralski.github.io/proba_mdbook/print.html#wst%C4%99p-do-filogenetyki-molekularnej-i-tworzenia-drzew-filogenetycznych`, data dostępu: 09.01.2026.
[2] PWN Encyklopedia [online] `https://encyklopedia.pwn.pl/haslo/filogeneza;3900970.html`, data dostępu: 09.01.2026.
[3] Krzysztof Spalik, Marcin Piwczyński *Rekonstrukcja filogenezy i wnioskowanie filogenetyczne w badaniach ewolucyjnych* [artykuł] KOSMOS, Problemy Nauk Biologicznych, t. 58, nr 3–4, 2009, s. 485–498.

Dodatki

Źródła

[1] Georgia Tech Biological Sciences *Phylogenetic Trees* [online] https://organismalbio.biosci.gatech.edu/biodiversity/phylogenetic-trees/, data dostępu: 09.01.2026.

[2] National Center for Biotechnology Information [online] https://www.ncbi.nlm.nih.gov/guide/data-software/, data dostępu: 09.01.2026.

Załączniki

Lista dodatkowych plików uzupełniających tekst pracy

W systemie do pracy dołączono dodatkowe pliki zawierające:

Spis rysunków