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 wybrane algorytmy i narzędzia bioinformatyczne (MAFFT [1], IQ-TREE [2], Biopython [3]), 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 [4].

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

Wprowadzenie

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 rekonstrukcji 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 [5][6].

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 przepływ pracy analitycznej. 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 procesów analitycznych. Brakuje systemów, które w sposób zintegrowany umożliwiałyby porównywanie wyników różnych metod, zarządzanie eksperymentami analitycznymi oraz wspomaganie decyzji o wyborze końcowego drzewa filogenetycznego [6].

1.3 Cel pracy

Celem pracy było zaprojektowanie oraz zaimplementowanie systemu wspomagającego analizy filogenetyczne, umożliwiającego integrację poszczególnych etapów procesu analitycznego w spójny przepływ pracy analitycznej. 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, rekonstrukcji drzew filogenetycznych oraz doboru modeli ewolucyjnych. Praca obejmuje zaprojektowanie i implementację systemu integrującego wybrane narzędzia analityczne w spójny przepływ pracy analitycznej, 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

W ramach niniejszej pracy zaprojektowano architekturę oraz zaimplementowano system wspomagający analizę filogenetyczną. Zrealizowano zarówno integrację wybranych narzędzi do wyrównywania sekwencji, rekonstrukcji drzew filogenetycznych oraz doboru modeli ewolucyjnych w spójny proces analityczny. Ponadto przygotowano mechanizmy wizualizacji wyników analizy oraz obsługę danych wejściowych, w tym pozyskiwanie sekwencji genetycznych z publicznie dostępnej bazy NCBI.

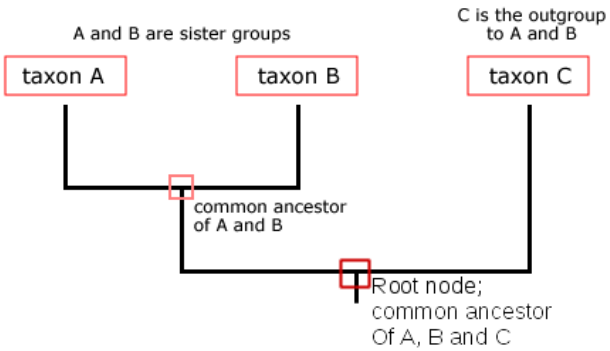
Rozdział 2

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. Przykładem wyniku takiej analizy jest drzewo filogenetyczne (rys. 2.1) — 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 [5][7].



Rysunek 2.1: Drzewo filogenetyczne; źródło: [8]

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 (rys. 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 [4].

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

Rysunek 2.2: Sekwencja białkowa; źródło: [4]

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 nukleotydowych lub aminokwasowych 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. Wyrównanie pozwala na ujednolicenie długości sekwencji oraz określenie podobieństw i różnic wynikających z przebiegu ewolucji.

W zależności od zastosowanego algorytmu możliwe jest wykonywanie wyrównań globalnych, obejmujących całe sekwencje lub lokalnych, koncentrujących się na ich fragmentach. Klasycznymi algorytmami wykorzystywanymi w wyrównaniach par sekwencji są algorytm Needleman–Wunscha dla wyrównań globalnych oraz algorytm Smitha–Watermana

dla wyrównań lokalnych. Algorytmy te nie znajdują jednak bezpośredniego zastosowania w analizach filogenetycznych, które wymagają jednoczesnego wyrównania większej liczby sekwencji. W tym celu stosuje się algorytmy przyrównania wielu sekwencji (MSA), takie jak metody progresywne FFT-NS-1 i FFT-NS-2 oraz metody iteracyjnego udoskonalania wyrównania, m.in. FFT-NS-i, L-INS-i, E-INS-i oraz G-INS-i, zaimplementowane w narzędziu MAFFT. Dobór metody wyrównania oraz jej parametrów ma istotny wpływ na jakość dalszej analizy filogenetycznej, w szczególności na poprawność rekonstrukcji drzewa ewolucyjnego.

Wyrównanie sekwencji może wymagać dodatkowych poprawek manualnych, zwłaszcza w przypadku sekwencji o niskim stopniu podobieństwa lub zawierających liczne insercje i delecje. W praktyce badawczej często stosuje się podejście iteracyjne, polegające na wielokrotnym wykonywaniu wyrównań z różnymi parametrami oraz ręcznej korekcie wyników w celu uzyskania biologicznie uzasadnionego wyrównania [1][5][9].

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 ewolucyjnych 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 [9].

Rozdział 3

Przegląd istniejących narzędzi

Współczesne analizy filogenetyczne realizowane są z wykorzystaniem wyspecjalizowanych narzędzi informatycznych, które wspomagają poszczególne etapy przetwarzania danych sekwencyjnych. Ze względu na zróżnicowanie dostępnych metod oraz narzędzi, istotnym elementem projektowania analizy jest dobór odpowiednich programów do realizacji poszczególnych etapów procesu rekonstrukcji. W niniejszym rozdziale przedstawiono przegląd wybranych narzędzi stosowanych w analizach filogenetycznych, ze szczególnym uwzględnieniem programów do wyrównywania sekwencji oraz narzędzi do inferencji filogenetycznej.

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 [5][10]:

- **Clustal W i Clustal Omega** - klasyczne narzędzia do wyrównywania sekwencji, stosujące metodę progresywną [10].
- **MAFFT** - narzędzie do wyrównania wielu sekwencji implementujące zestaw algorytmów progresywnych i iteracyjnych, umożliwiających wybór kompromisu pomiędzy dokładnością wyrównania a czasem obliczeń [1].
- **MUSCLE** - narzędzie, które stosuje iteracyjne udoskonalanie wyrównania, co pozwala na osiągnięcie wysokiej dokładności przy umiarkowanym czasie obliczeń [10].
- **T-Coffee** - narzędzie umożliwiające łączenie wyników różnych algorytmów wyrównywania w celu uzyskania bardziej wiarygodnego wyrównania konsensusowego, kosz-

tem zwiększonego czasu obliczeń [11].

- **PRANK** - uwzględniają procesy ewolucyjne, takie jak insercje i delecje, co pozwala na uzyskanie wyrównań lepiej odzwierciedlających historię ewolucyjną sekwencji [12].

Do analizy i wprowadzania poprawek ręcznie można wykorzystać edytory wyrównań takie jak [5]:

- **Jalview** - rozbudowane narzędzie do wizualizacji i edycji wyrównań sekwencji, umożliwiające ich ręczną korektę, analizę konserwacji pozycji, kolorowanie sekwencji oraz generowanie prostych drzew filogenetycznych [13].
- **AliView** – lekki i wydajny edytor wyrównań, zaprojektowany do pracy z bardzo dużymi zestawami sekwencji, oferujący podstawowe funkcje edycji oraz przeglądania wyrównań w wielu popularnych formatach [14].

3.2 Narzędzia do rekonstrukcji filogenetycznej

Dostępne są różne programy i metody, które wykorzystując odmienne podejścia realizują rekonstrukcję drzewa filogenetycznego w oparciu o dane sekwencyjne. Wybór odpowiedniego narzędzia zależy od rodzaju danych, liczby sekwencji oraz wymagań dotyczących dokładności i czasu obliczeń. Poniżej zostały przedstawione wybrane narzędzia [4].

- **IQ-TREE 2** – nowoczesne narzędzie do rekonstrukcji drzew filogenetycznych oparte na metodzie największej wiarygodności, oferujące szeroki wybór modeli substytucji oraz wydajne algorytmy obliczeniowe [2].
- **RAxML** – program wykorzystujący metodę największej wiarygodności, zoptymalizowany do analizy dużych zbiorów danych i powszechnie stosowany w badaniach filogenetycznych [15].
- **MrBayes** – narzędzie realizujące inferencję filogenetyczną z wykorzystaniem metod bayesowskich, umożliwiające estymację rozkładu prawdopodobieństwa drzew oraz parametrów modeli ewolucyjnych [16].
- **MEGA** – środowisko do analizy danych molekularnych, oferujące funkcje rekonstrukcji drzew filogenetycznych z wykorzystaniem różnych metod oraz przyjazny interfejs użytkownika [17].

Po zakończeniu rekonstrukcji drzewa filogenetycznego konieczna jest ocena wiarygodności uzyskanych wyników. W tym celu stosuje się metody, takie jak bootstrap, czy metody bayesowskie, które umożliwiają oszacowanie pewności poszczególnych węzłów drzewa

Wykorzystywana w niniejszym projekcie metoda bootstrap polega na wielokrotnym tworzeniu nowych zestawów danych poprzez losowe próbkowanie pozycji z wyrównania sekwencji oraz ponowną rekonstrukcję drzewa filogenetycznego. Częstość występowania danego węzła w kolejnych drzewach bootstrapowych traktowana jest jako miara jego wiarygodności. Im wyższa wartość bootstrap, tym większa pewność, że dana relacja filogenetyczna jest stabilna i dobrze wsparta przez dane [7].

Po zakończeniu analizy filogenetycznej wynikiem jest drzewo filogenetyczne zapisane w postaci pliku tekstowego, zawierającego opis struktury drzewa w formacie Newick lub w pliku formatu NEXUS, które może być dalej przetwarzane i wizualizowane za pomocą dedykowanych narzędzi [5].

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** - program desktopowy umożliwiający podstawową wizualizację oraz edycję drzew filogenetycznych, w tym modyfikację etykiet, długości gałęzi i wartości wsparcia [18].
- **Dendroscope** – zaawansowane narzędzie do wizualizacji dużych drzew filogenetycznych oraz sieci filogenetycznych, przystosowane do pracy z rozbudowanymi zbiorami danych [19].
- **iTOL** – aplikacja webowa oferująca interaktywną wizualizację drzew oraz możliwość dodawania warstw adnotacji, takich jak dane taksonomiczne, cechy fenotypowe czy metadane [20].
- **ETE Toolkit** – biblioteka programistyczna umożliwiająca automatyczną wizualizację, analizę i modyfikację drzew filogenetycznych [21].

3.4 Ograniczenia istniejących rozwiązań

Chociaż istnieją systemy i narzędzia programistyczne do automatyzacji analiz bioinformatycznych (np. Galaxy, Snakemake, Nextflow), w praktyce często wymaga się ręcznego ustawiania parametrów poszczególnych narzędzi oraz integracji wyników. W ramach niniejszego projektu inzynierskiego zaprojektowany został system automatyzujący pełny łańcuch analizy, od wczytania sekwencji, przez wyrównanie i rekonstrukcję drzewa filogenetycznego, aż po wizualizację wyników, co ułatwia prowadzenie badań, zwiększa powta-

rzalność wyników i umożliwia prostsze korzystanie z analizy filogenetycznej, także osobom dopiero rozpoczynającym pracę z tymi metodami.

3.5 Opis zastosowanych technologii i narzędzi

W celu realizacji systemu wspomagającego analizę filogenetyczną wykorzystano zestaw nowoczesnych technologii programistycznych oraz specjalistycznych narzędzi bioinformatycznych. Dobór zastosowanych rozwiązań podyktowany był koniecznością zapewnienia automatyzacji wieloetapowego procesu analitycznego, wysokiej wydajności obliczeniowej oraz możliwości integracji ze sobą nawzajem. W niniejszym podrozdziale przedstawiono opis użytych technologii, języków programowania oraz narzędzi, wraz z uzasadnieniem ich zastosowania w kontekście realizowanego projektu.

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

MAFFT (Multiple Alignment using Fast Fourier Transform) jest narzędziem do wielosekwencyjnego wyrównywania sekwencji DNA i białek, opartym na algorytmach progresywnych oraz iteracyjnych. Wykorzystuje ono transformację Fouriera (FFT) do szybkiego szacowania podobieństwa pomiędzy sekwencjami, co pozwala na znaczące przyspieszenie obliczeń w porównaniu do klasycznych metod opartych na pełnym dynamicznym programowaniu. W zależności od charakterystyki danych wejściowych MAFFT umożliwia zastosowanie różnych strategii wyrównywania, takich jak metody progresywne (np. FFT-NS-1, FFT-NS-2) oraz metody iteracyjnego udoskonalania wyrównania (np. L-INS-i, E-INS-i, G-INS-i), stanowiące kompromis pomiędzy dokładnością a czasem obliczeń. W projekcie MAFFT został wykorzystany do przeprowadzenia etapu wyrównywania sekwencji poprzedzającego rekonstrukcję drzewa filogenetycznego [1].

3.5.4 IQ-TREE 2

IQ-TREE 2 jest narzędziem do rekonstrukcji drzew filogenetycznych opartym na metodzie największej wiarygodności (Maximum Likelihood). Algorytm ten polega na poszukiwaniu takiej topologii drzewa oraz parametrów modelu substytucji, dla których prawdopodobieństwo obserwowanych danych sekwencyjnych jest maksymalne. Program wykorzystuje zaawansowane heurystyki przeszukiwania przestrzeni drzew, umożliwiające efektywną optymalizację topologii nawet dla dużych zbiorów danych. Dodatkowo IQ-TREE 2 implementuje algorytm ModelFinder[22] do automatycznego doboru najlepszego modelu substytucji oraz ultraszybką metodę bootstrap (UFBoot2)[23] do oceny wiarygodności węzłów drzewa filogenetycznego. Dzięki połączeniu wysokiej wydajności obliczeniowej oraz dokładnych metod statystycznych, narzędzie to jest powszechnie stosowane w analizach filogenetycznych [2].

3.5.5 Jalview

Jalview to rozbudowane narzędzie do wizualizacji i edycji wyrównań sekwencji, umożliwiające ręczną korektę wyrównań, analizę konserwacji pozycji oraz generowanie prostych drzew filogenetycznych. W projekcie Jalview został wykorzystany testowo do wizualizacji wyrównanych sekwencji, aby sprawdzić poprawność działania MAFFT [13].

3.5.6 System kontroli wersji: Git

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

3.5.7 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 systemowe

Minimalne wymagania sprzętowe:

- Procesor: dwurdzeniowy, 2 GHz lub szybszy
- Pamięć RAM: minimum 4 GB
- Przestrzeń dyskowa: co najmniej 5 GB wolnego miejsca

Minimalne wymagania systemowe:

- System operacyjny: Windows 10 (64-bit) lub nowszy
- Windows Subsystem for Linux (WSL) z zainstalowaną dystrybucją Ubuntu 20.04 LTS lub nowszą
- MAFFT w wersji 7.475 lub nowszej, zainstalowany w środowisku WSL
- IQ-TREE 2 w wersji 2.1.3 lub nowszej, zainstalowany w środowisku WSL

4.2 Instalacja

Aby uruchomic aplikację, należy wykonać następujące kroki:

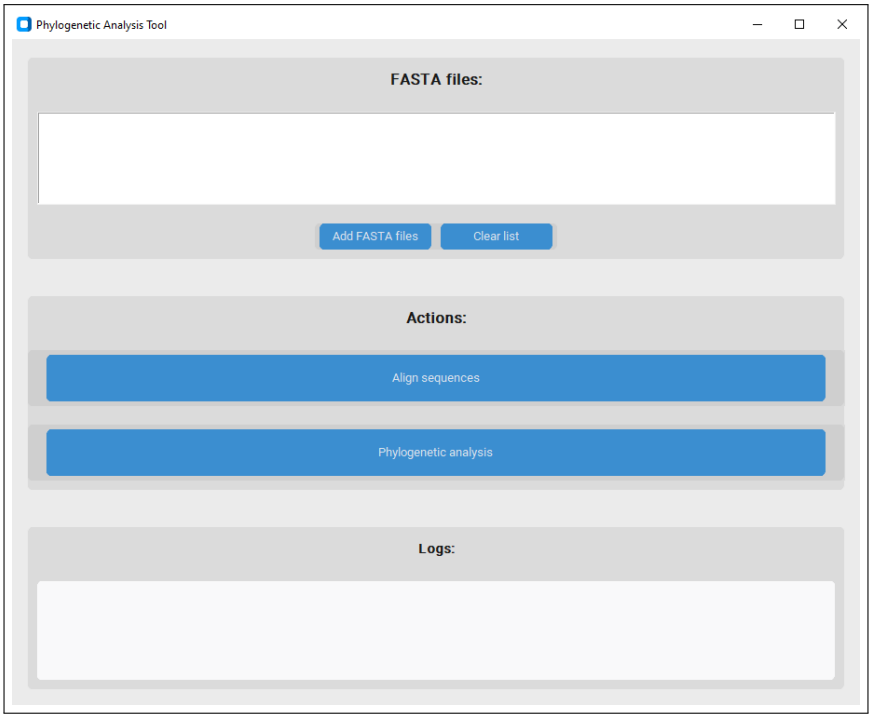
1. Zainstalować Windows Subsystem for Linux (WSL) oraz dystrybucję Ubuntu 20.04 LTS lub nowszą.
2. W środowisku WSL zainstalować narzędzia MAFFT [1] oraz IQ-TREE 2 [2].
3. Upewnić się, że narzędzia są dostępne w ścieżce systemowej (PATH) w WSL.
4. Pobrać kod źródłowy aplikacji z repozytorium GitHub github.com/FilipWspanialy/PhylogeneticAnalyses i uruchomić plik `main.exe` w systemie Windows.

4.3 Instrukcja obsługi

4.3.1 Interfejs użytkownika

Graficzny interfejs użytkownika (rys. 4.1) umożliwia użytkownikowi:

- wczytanie sekwencji w formacie FASTA,
- Wyrównanie sekwencji przy użyciu MAFFT,
- Przeprowadzenie analizy filogenetycznej z IQ-TREE 2,
- Wizualizację uzyskanego drzewa filogenetycznego.
- Podgląd logów z poszczególnych etapów analizy.

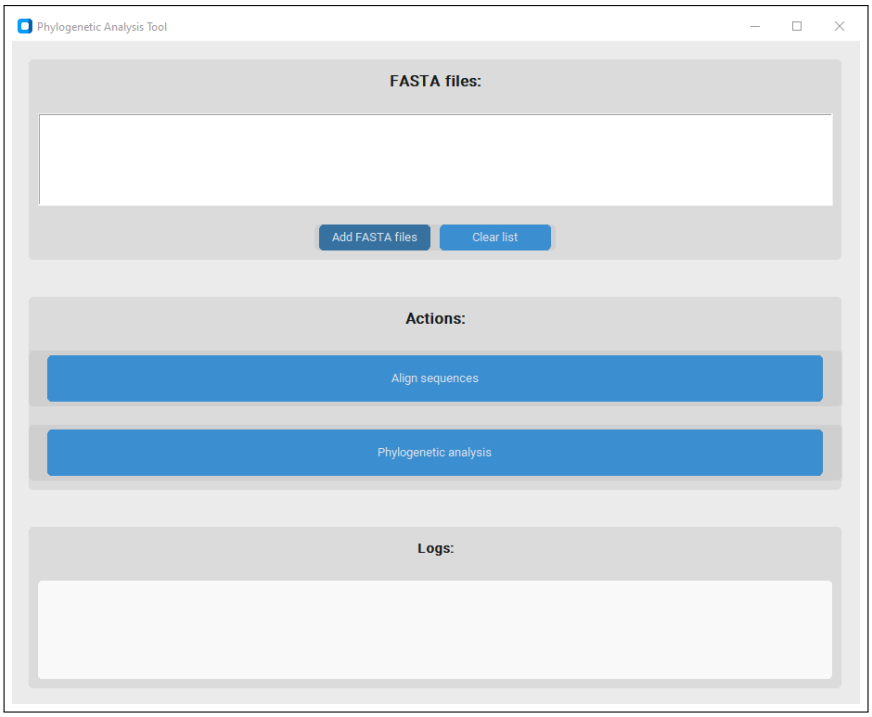


Rysunek 4.1: Interfejs użytkownika aplikacji

4.3.2 Przykład działania

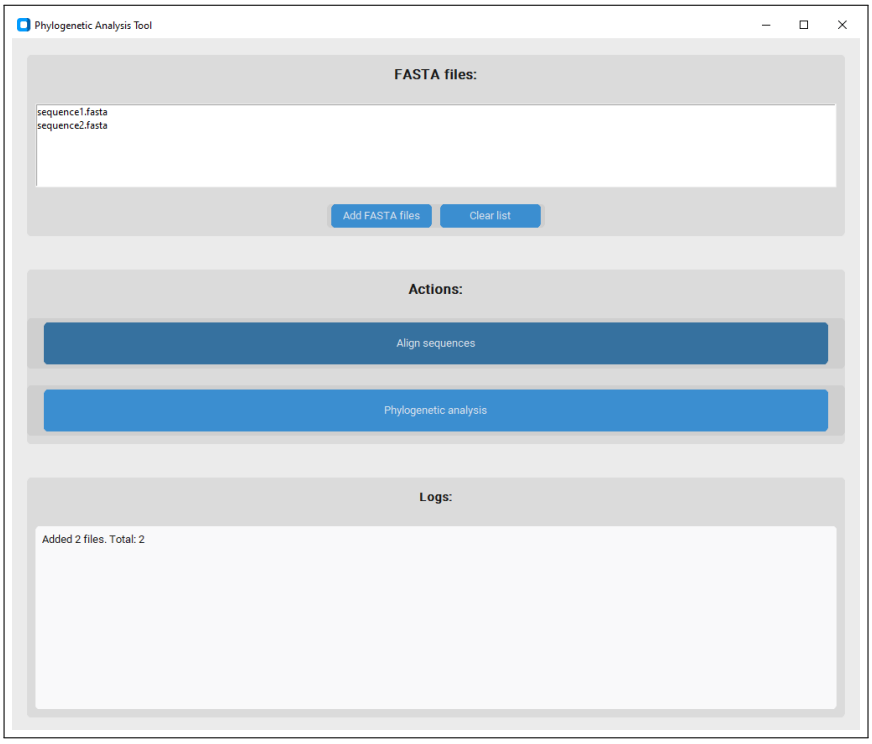
Poniżej przedstawiono przykładowy przebieg analizy filogenetycznej z wykorzystaniem aplikacji:

1. Uruchomienie aplikacji
2. Wczytanie plików FASTA z sekwencjami białkowymi (rys. 4.2).



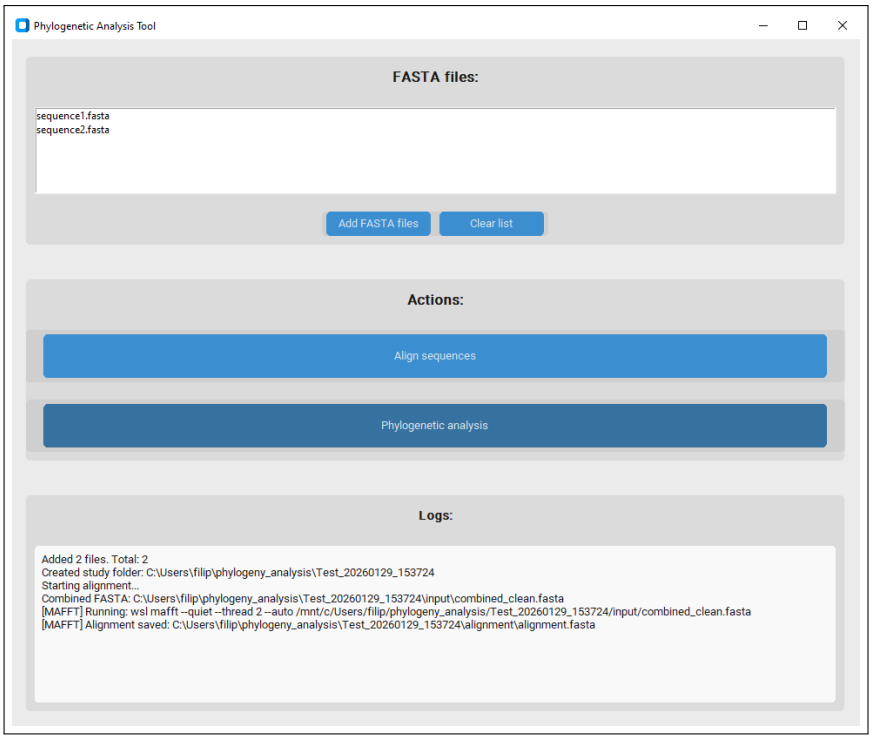
Rysunek 4.2: Wybór plików wejściowych

3. Uruchomienie wyrównania sekwencji za pomocą MAFFT, którego produkt zostaje zapisany w utworzonym przez aplikację folderze (rys. 4.3).



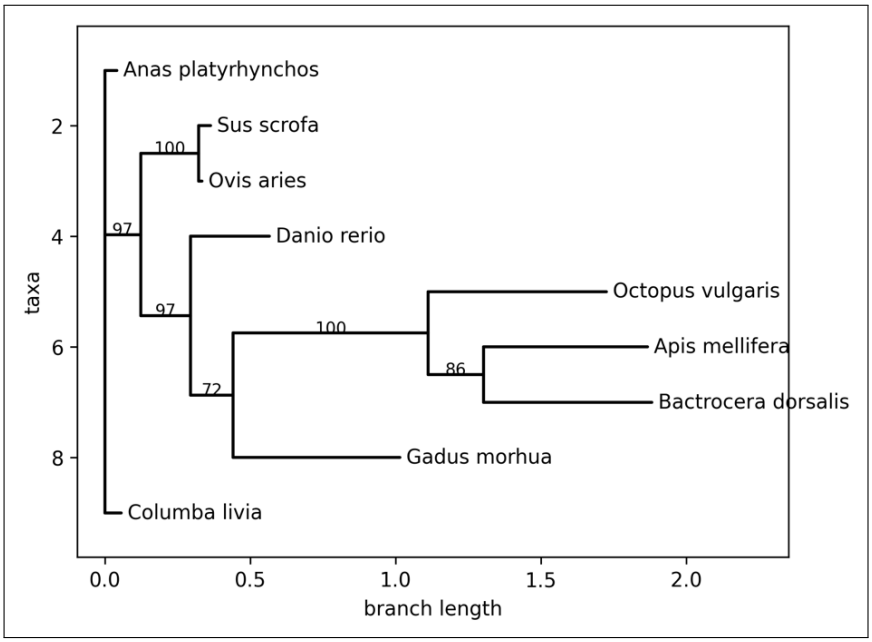
Rysunek 4.3: Wyrównanie sekwencji

4. Po zakończeniu wyrównania, uruchomienie analizy filogenetycznej z IQ-TREE 2 (rys. 4.4).



Rysunek 4.4: Analiza filogenetyczna

5. Po zakończeniu analizy następuje wizualizacja uzyskanego drzewa filogenetycznego (rys. 4.5).



Rysunek 4.5: Wizualizacja drzewa filogenetycznego

Wszystkie czynności są rejestrowane w logach dostępnych w interfejsie użytkownika, a pliki wynikowe z każdego etapu zapisywane są w katalogu, którego ścieżka zostaje podana w trakcie działania aplikacji.

4.3.3 Bezpieczeństwo

Aplikacja nie przechowuje żadnych danych osobowych użytkownika.

Rozdział 5

Specyfikacja wewnętrzna systemu

5.1 Idea systemu

Celem aplikacji jest integracja istniejących, sprawdzonych narzędzi do analiz filogenetycznych w jeden spójny proces umożliwiający przeprowadzenie analizy bez konieczności posiadania wiedzy eksperckiej z zakresu filogenetyki. Użytkownik ma możliwość intuicyjnej pracy z plikami wejściowymi, natomiast dobór parametrów analizy oraz modeli ewolucyjnych odbywa się w sposób automatyczny.

System realizuje kompletny proces analizy filogenetycznej, obejmujący przygotowanie danych wejściowych, wyrównywanie sekwencji, budowę drzewa filogenetycznego oraz wizualizację wyników. Całość została zaprojektowana z myślą o środowisku Windows, przy jednoczesnym wykorzystaniu narzędzi Linuksowych dostępnych poprzez Windows Subsystem for Linux.

5.2 Wymagania funkcjonalne

System spełnia następujące wymagania funkcjonalne:

- import plików sekwencji w formacie FASTA,
- łączenie oraz wstępne czyszczenie danych wejściowych,
- uruchamianie narzędzi do wyrównywania sekwencji,
- budowę drzewa filogenetycznego z automatycznym doborem modelu,
- automatyczne tworzenie struktury katalogów wynikowych,
- wizualizację drzewa filogenetycznego,
- logowanie przebiegu analizy w interfejsie użytkownika.

5.3 Architektura

Architektura systemu została zaprojektowana w sposób modułowy, co zapewnia czytelny podział odpowiedzialności pomiędzy poszczególne komponenty aplikacji oraz umożliwia jej dalszy rozwój. System składa się z logicznych warstw odpowiadających za interakcję z użytkownikiem, realizację analizy, zarządzanie danymi oraz prezentację wyników.

5.3.1 Warstwy systemu

1. **Warstwa interfejsu użytkownika** – odpowiada za interakcję z użytkownikiem, obsługę plików wejściowych oraz prezentację wyników i logów.
2. **Warstwa logiki aplikacji** – realizuje proces analizy filogenetycznej oraz steruje uruchamianiem narzędzi analitycznych.
3. **Warstwa zarządzania danymi** – odpowiada za przygotowanie, organizację i spójność danych wykorzystywanych w analizie.
4. **Warstwa wizualizacji** – generuje oraz udostępnia użytkownikowi wizualną reprezentację drzewa filogenetycznego.

5.3.2 Opis modułów systemu

Wszystkie funkcjonalności systemu zostały zaimplementowane w klasie `PhylogenyApp`. Ze względu na skalę projektu architektura systemu została zrealizowana w postaci jednej klasy, w obrębie której wydzielono logiczne moduły funkcjonalne grupujące powiązane metody. Takie podejście zapewnia czytelny podział odpowiedzialności przy jednoczesnym zachowaniu prostoty implementacyjnej.

Moduł interfejsu użytkownika

Moduł interfejsu użytkownika został zaimplementowany z wykorzystaniem biblioteki `tkinter`. Struktura GUI tworzona jest w metodzie `setup_ui` i obejmuje elementy umożliwiające wybór plików FASTA, uruchamianie poszczególnych etapów analizy oraz podgląd logów i wyników. Moduł obsługuje zdarzenia generowane przez użytkownika i pełni rolę warstwy pośredniczącej pomiędzy użytkownikiem a logiką aplikacji.

Moduł logiki aplikacji

Moduł logiki aplikacji realizuje właściwy proces analizy filogenetycznej, obejmujący wyrównywanie sekwencji oraz budowę drzewa filogenetycznego. Proces wyrównywania sekwencji realizowany jest w metodzie `align_sequence`, natomiast analiza filogenetyczna i konstrukcja drzewa wykonywane są w metodzie `phylogenetic_analysis`. W ramach tych

metod wykorzystywane są funkcje `run_mafft` oraz `run_iqtree`, które za pomocą biblioteki `subprocess` uruchamiają zewnętrzne narzędzia analityczne w środowisku Windows Subsystem for Linux.

Parametry wykorzystywane w MAFFT [1]

```
cmd = [
    "wsl", "mafft",
    "--quiet", "--thread",
    "2", "--auto",
    wsl_input
]
```

- `--quiet` – wyciszenie komunikatów informacyjnych,
- `--thread 2` – wykorzystanie dwóch wątków procesora,
- `--auto` – automatyczny dobór strategii wyrównania,
- `wsl_input` – ścieżka do pliku FASTA w formacie zgodnym z WSL.

Parametry wykorzystywane w IQ-TREE 2 [24]

```
cmd = [
    "wsl", "iqtree2",
    "-s", wsl_alignment,
    "-m", "MFP",
    "-bb", "1000",
    "-alrt", "1000",
    "-nt", "AUTO",
    "-pre", wsl_prefix,
    "-redo"
]
```

- `-s wsl_alignment` – plik wyrównanych sekwencji,
- `-m MFP` – automatyczny wybór modelu ewolucyjnego,
- `-bb 1000` – ultrafast bootstrap,
- `-alrt 1000` – test ALRT,
- `-nt AUTO` – automatyczne wykorzystanie wątków CPU,
- `-pre wsl_prefix` – prefiks plików wynikowych,
- `-redo` – nadpisanie istniejących wyników.

Moduł zarządzania danymi

Moduł zarządzania danymi odpowiada za przygotowanie i organizację danych wejściowych oraz wynikowych. W jego ramach wykorzystywane są metody `combine_fasta`, `sanitize_fasta_headers` oraz `build_label_map`, które umożliwiają łączenie plików FASTA, ujednolicanie nagłówków sekwencji oraz zachowanie mapowania etykiet wykorzystywanych na etapie wizualizacji drzewa. Dodatkowo metoda `create_study_dirs` odpowiada za automatyczne tworzenie struktury katalogów wynikowych. Moduł ten zapewnia spójność danych pomiędzy kolejnymi etapami analizy oraz niezależność pozostałych komponentów od szczegółów operacji plikowych.

Moduł wizualizacji danych

Moduł wizualizacji danych odpowiada za graficzną prezentację wyników analizy filogenetycznej. Do wizualizacji drzewa filogenetycznego wykorzystano bibliotekę `Matplotlib`. Na podstawie pliku wynikowego w formacie Newick generowana jest statyczna reprezentacja drzewa, uwzględniająca przemapowane etykiety sekwencji zgodnie z wcześniej utworzoną mapą nagłówków. Proces ten realizowany jest w metodzie `plot_tree`, która odpowiada za odczyt struktury drzewa, jego renderowanie oraz zapis wizualizacji do pliku graficznego.

Rozdział 6

Testy i analiza działania systemu

6.1 Cel testów

Celem przeprowadzonych testów było zweryfikowanie poprawności działania systemu na różnych zestawach danych wejściowych oraz ocena jakości uzyskiwanych wyników filogenetycznych. Testy miały na celu sprawdzenie, czy aplikacja prawidłowo integruje narzędzia MAFFT oraz IQ-TREE 2, a także czy uzyskane drzewa filogenetyczne są zgodne z oczekiwaniami biologicznymi.

6.2 Testy dla sekwencji białek z bazy NCBI

Testy dla sekwencji białkowych przeprowadzono w celu oceny działania systemu na danych o różnym stopniu konserwatywności. Analiza obejmowała zarówno białka silnie konserwatywne, jak i sekwencje bardziej zmienne, co pozwoliło na kompleksową ocenę poprawności wyrównywania sekwencji oraz rekonstrukcji drzew filogenetycznych.

6.2.1 Test 1: Cytochrom c

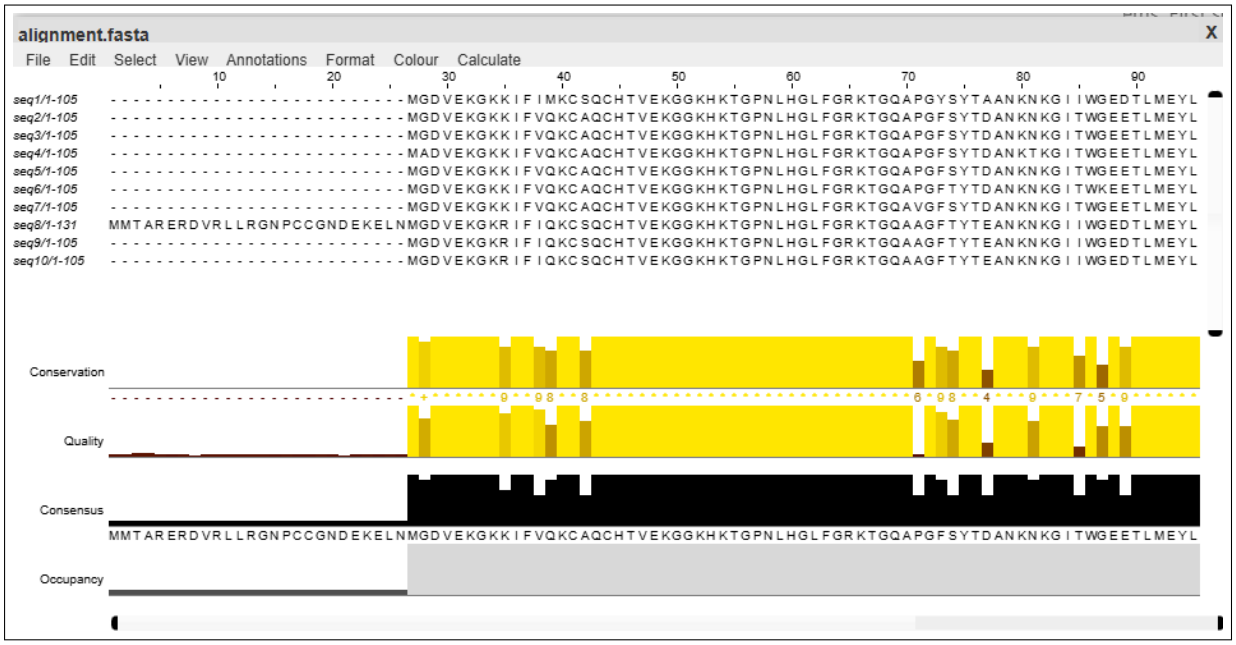
Wykorzystane w teście numer 1 sekwencje (tab. 6.1) kodują cytochrom c, który jest białkiem silnie konserwatywnym, powszechnie wykorzystywanym w analizach filogenetycznych jako marker referencyjny.

Cel testu:

- Weryfikacja poprawności działania stworzonej aplikacji pod względem integracji narzędzi MAFFT oraz IQ-TREE 2.
- Sprawdzenie działania systemu na niewielkim zbiorze sekwencji białkowych o dobrze poznanej filogenezie w celu oceny poprawności uzyskiwanych wyników.

Tabela 6.1: Charakterystyka sekwencji cytochromu c wykorzystanych w teście

Gatunek (łac.)	**Nazwa polska**	**Długość sekwencji (aa)**
Homo sapiens	Człowiek rozumny	105
Bos taurus	byk	105
Sus scrofa	świnia	105
Tenrec ecaudatus	Tenrek zwyczajny	105
Canis lupus familiaris	Pies domowy	105
Equus caballus	Koń	105
Camelus bactrianus	Wielbłąd baktriański	105
Saimiri boliviensis	Sajmiri czarnołbista	131
Saimiri boliviensis	Sajmiri czarnołbista	105
Saimiri boliviensis	Sajmiri czarnołbista	105

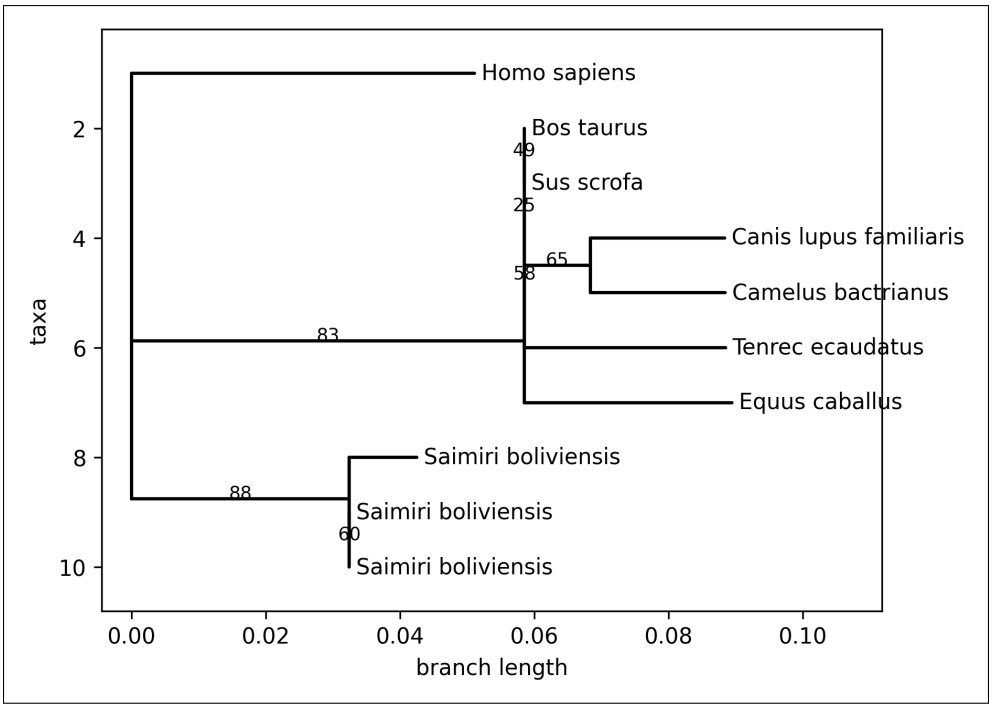


Rysunek 6.1: Fragment wyrównanych sekwencji białkowe - test 1

Wyniki testu:

Test przeprowadzony na krótkich sekwencjach białkowych cytochromu c zakończył się powodzeniem. Aplikacja poprawnie zrealizowała pełny proces analizy filogenetycznej, integrując narzędzia MAFFT oraz IQ-TREE 2, od etapu wyrównania sekwencji (rys. 6.1) po rekonstrukcję drzewa filogenetycznego (rys. 6.2).

Uzyskane drzewo filogenetyczne jest zgodne z ogólnymi oczekiwaniami biologicznymi i odzwierciedla znane relacje filogenetyczne pomiędzy analizowanymi gatunkami ssaków. Celowe uwzględnienie trzech sekwencji pochodzących od tego samego gatunku skutkowało ich bliskim sąsiedztwem na drzewie wynikowym, co potwierdza poprawność działania



Rysunek 6.2: Drzewo filogenetyczne dla cytochromu c - test 1

algorytmów wyrównywania oraz rekonstrukcji drzewa.

Wysokie wartości bootstrap (>70) obserwowane są dla węzłów odpowiadających głównym podziałom taksonomicznym, natomiast dla rozgałęzień pomiędzy blisko spokrewnionymi gatunkami wartości bootstrap są niższe (25–65). Wynika to z wysokiej konserwatywności białka cytochromu c, które dostarcza ograniczonej ilości sygnału filogenetycznego.

Test potwierdza poprawność działania zaprojektowanego systemu na prostych, kontrolowanych danych wejściowych oraz poprawną integrację wykorzystanych narzędzi bioinformatycznych.

6.2.2 Test 2: Podjednostka IV oksydazy cytochromu c

Tabela 6.2: Charakterystyka sekwencji białka COX4 wykorzystanych w analizie filogenetycznej

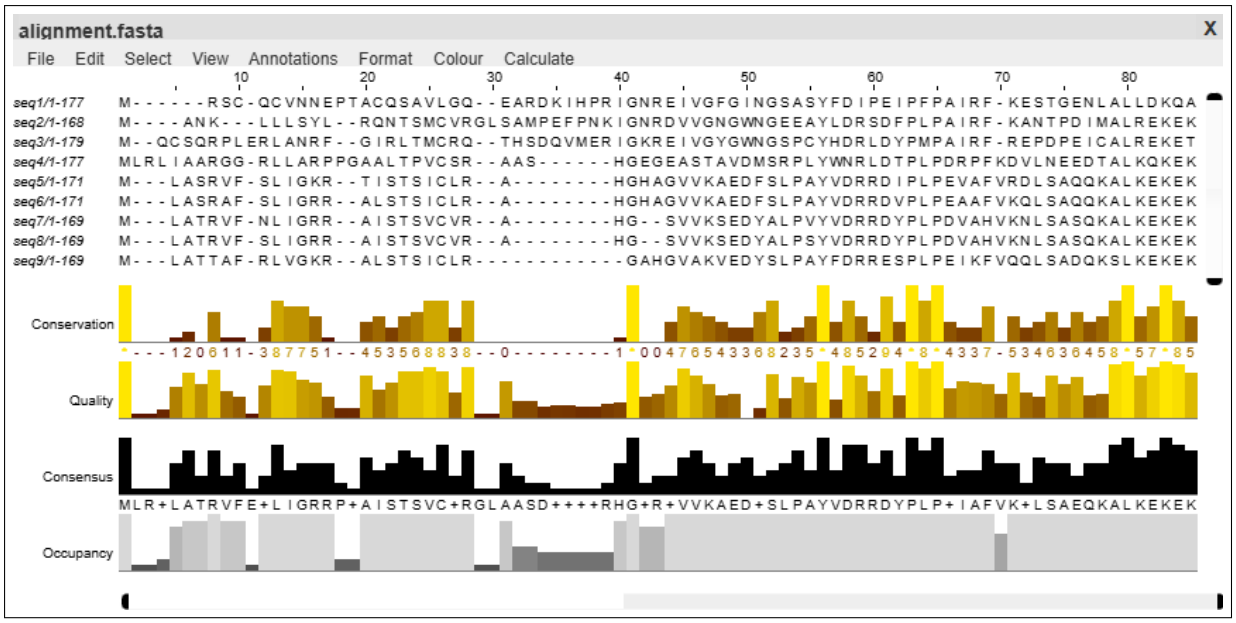
Gatunek (łac.)	**Nazwa polska**	**Długość sekwencji (aa)**
Octopus vulgaris	Ośmiornica zwyczajna	174
Apis mellifera	Pszczoła miodna	169
Bactrocera dorsalis	Muszka owocowa	170
Gadus morhua	Dorsz	176
Columba livia	Gołąb skalny	168
Anas platyrhynchos	Kaczka krzyżówka	168
Sus scrofa	Świnia	169
Ovis aries	Owca	169
Danio rerio	Danio pręgowany	171

Wykorzystane dane zawierają sekwencje białkowe COX4 dla różnych gatunków (tab. 6.2). COX4 jest integralną podjednostką oksydazy enzymu cytochromu c (kompleks IV), uczestniczącego w końcowym etapie łańcucha oddechowego. W przeciwieństwie do cytochromu c, sekwencja COX4 jest mniej konserwatywna ewolucyjnie i wykazuje większą zmienność pomiędzy odlegle spokrewnionymi gatunkami.

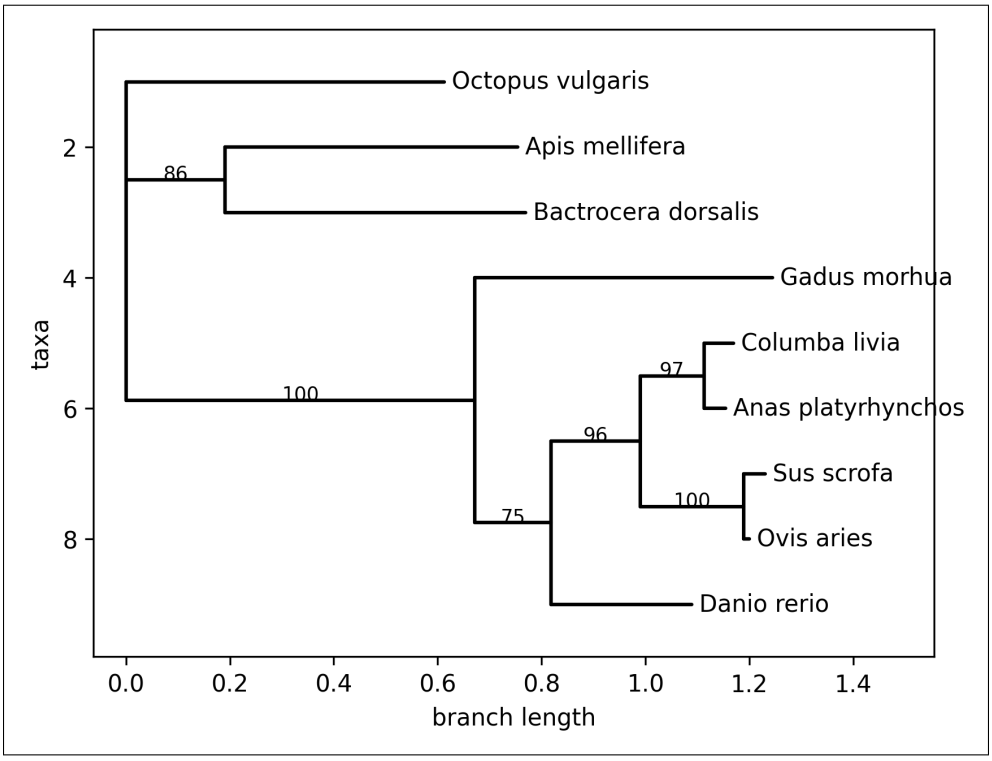
Cel testu:

- Weryfikacja działania aplikacji na białkach integralnych błony mitochondrialnej o większej zmienności sekwencji niż cytochrom c.

- Sprawdzenie jakości wyrównania i stabilności drzewa filogenetycznego dla białek o różnych długościach i izoformach.



Rysunek 6.3: Fragment wyrównanych sekwencji białkowe - test 2



Rysunek 6.4: Drzewo filogenetyczne dla COX4 - test 2

Wyniki testu:

Test przeprowadzono na sekwencjach białka COX4 pochodzących z różnych grup taksonomicznych, w tym ssaków, ryb, owadów i mięczaków. Aplikacja poprawnie zintegrowała narzędzia MAFFT oraz IQ-TREE 2, generując wyrównanie sekwencji (rys. 6.3) oraz drzewo filogenetyczne (rys. 6.4).

Uzyskane drzewo wykazuje logiczną strukturę filogenetyczną. Octopus vulgaris został wyraźnie odseparowany jako najbardziej odległy takson w zbiorze. Apis mellifera i Bactrocera dorsalis tworzą klad owadów z umiarkowanym wsparciem bootstrap (86). Gatunki ryb i ptaków grupują się w stabilne klady, np. Columba livia i Anas platyrhynchos (bootstrap 97) oraz Sus scrofa i Ovis aries (bootstrap 100).

Wysokie wartości bootstrap dla głównych kladów (75–100) wskazują na stabilność rekonstrukcji, natomiast niższe wartości dla owadów odzwierciedlają większą zmienność sekwencji w tej grupie. Test potwierdza poprawne działanie pipeline'u na białkach mitochondrialnych o zróżnicowanej długości i zmienności sekwencji.

6.3 Testy dla sekwencji nukleotydowych z bazy NCBI

6.3.1 Test 3: COX2–COX1

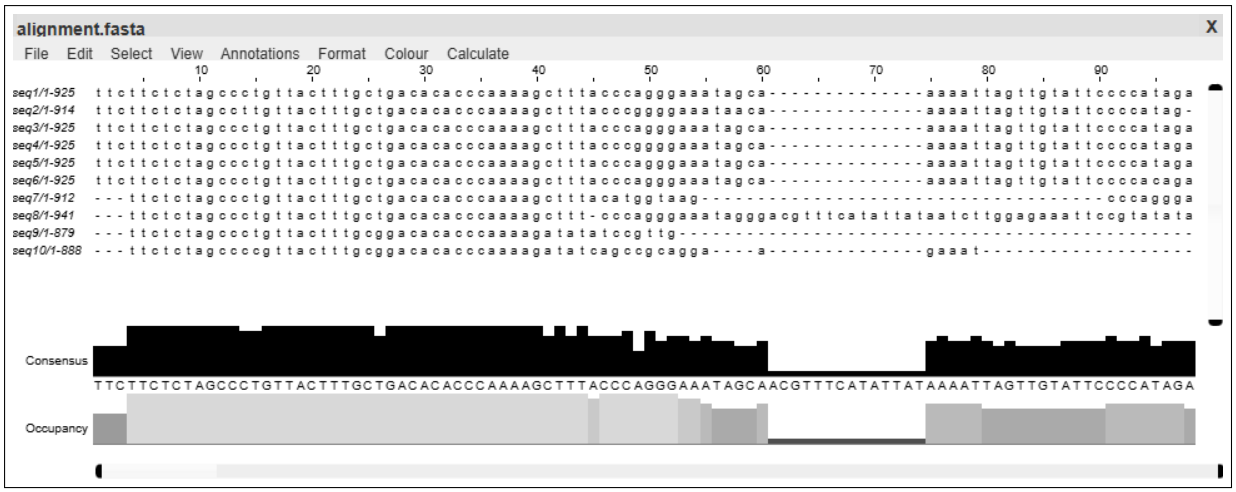
Tabela 6.3: Charakterystyka sekwencji mitochondrialnych wykorzystanych w analizie filogenetycznej

Gatunek (łac.)	**Nazwa polska**	**Długość sekwencji (bp)**
Calyptrophora clinata	Koralowiec klinowaty	925
Thouarella grasshoffi	Koralowiec grasshoffi	914
Primnoa sp.	Koralowiec primnoa	925
Narella versluysi	Koralowiec versluysi	925
Candidella imbricata	Koralowiec imbricata	925
Narella bellissima	Koralowiec bellissima	925
Chrysogorgia sp.	Koralowiec chrysogorgia	918
Chelidonisis aurantiaca	Koralowiec aurantiaca	920
Keratoisididae sp. F1	Koralowiec keratoisididae F1	922
Acanella sp.	Koralowiec acanella	923

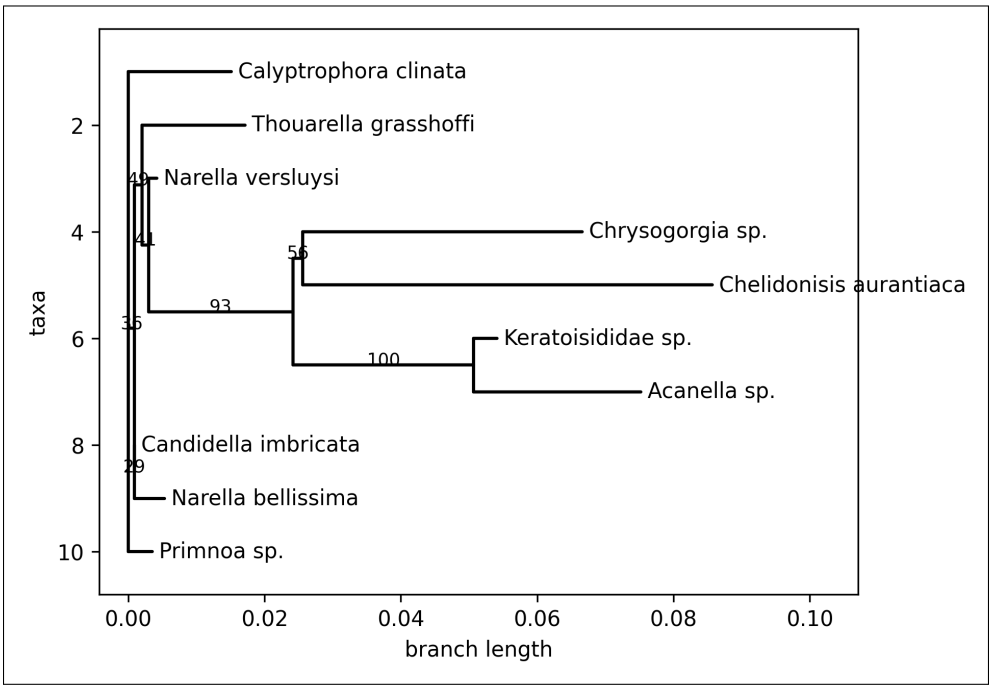
Sekwencje nukleotydowe wykorzystano do rekonstrukcji filogenezy na poziomie gatunków i bliskich rodzin koralowców (tab. 6.3). Ze względu na większą zmienność w porównaniu z białkami mitochondrialnymi, sekwencje te dostarczają bardziej szczegółowego sygnału filogenetycznego między taksonami blisko spokrewnionymi

Cel testu:

- Weryfikacja działania pipeline'u analizy filogenetycznej na sekwencjach nukleotydowych.
- Sprawdzenie, czy narzędzia MAFFT i IQ-TREE 2 radzą sobie z fragmentami genów mitochondrialnych o różnej długości i zmienności.



Rysunek 6.5: Fragment wyrównanych sekwencji nukleotydowych - test 3



Rysunek 6.6: Drzewo filogenetyczne dla COX2–COX1 - test 3

Wyniki testu:

Test przeprowadzono na sekwencjach nukleotydowych COX2–COX1 pochodzących od gatunków głębinowych koralowców (tab. 6.3). Aplikacja poprawnie zrealizowała pełen proces analizy filogenetycznej, obejmujący wyrównanie sekwencji (rys. 6.5) oraz rekonstrukcję drzewa filogenetycznego (rys. 6.6).

Uzyskane drzewo wykazuje logiczną organizację taksonomiczną. Calyptrophora clinata została odseparowana jako samodzielny takson, natomiast blisko spokrewnione gatunki, takie jak Thouarella grasshoffi, Narella versluysi, Candidella imbricata oraz Narella bellissima, grupują się w zagnieżdżone klady. Wartości bootstrap dla tych rozgałęzień są

zróżnicowane (29–99), co odzwierciedla zmienność sekwencji nukleotydowych.

Najwyższe wsparcie bootstrap obserwuje się dla kladów dobrze określonych genetycznie, natomiast niskie wartości dla części rozgałęzień wskazują na ograniczoną pewność relacji w obszarach o niewielkiej różnorodności lub fragmentarycznych danych. Test potwierdza poprawność działania pipeline'u na sekwencjach nukleotydowych oraz zdolność systemu do rozróżniania taksonów o różnym stopniu pokrewieństwa.

6.3.2 Test 4: Mitochondrialne sekwencje nukleotydowe

Charakterytyka danych:

Tabela 6.4: Charakterystyka pełnych genomów mitochondrialnych wykorzystanych w analizie

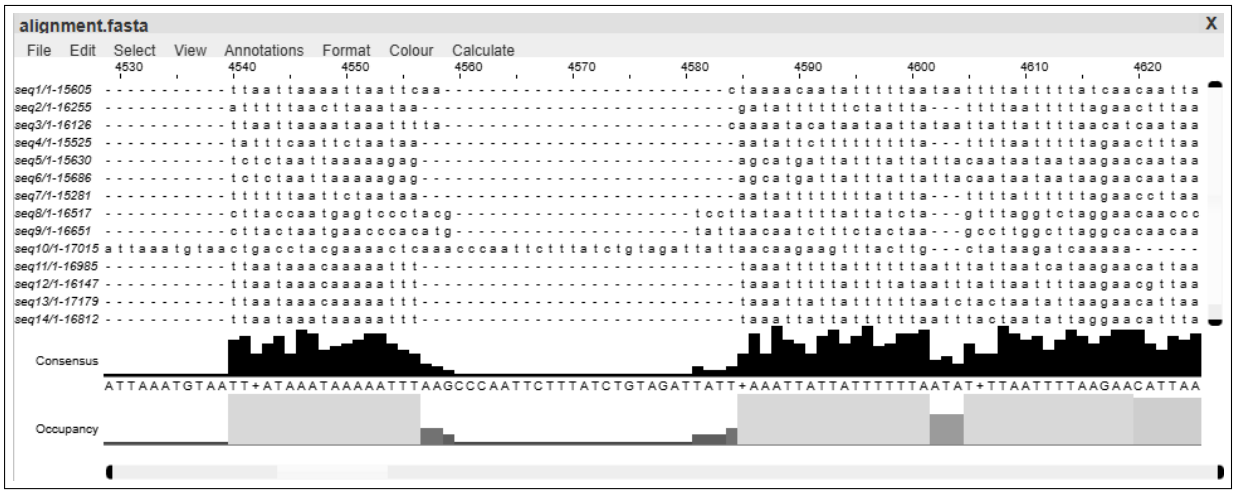
Gatunek (łac.)	**Nazwa polska**	**Długość sekwencji (bp)**
Loxocephala sichuanensis	Koralowiec sichuanensis	15605
Synanthedon bicingulata	Koralowiec synanthedon bicingulata	16255
Salurnis marginella	Koralowiec salurnis marginella	16126
Pintara bowringi	Koralowiec pintara bowringi	15525
Dolycoris penicillatus	Koralowiec dolycoris penicillatus	15630
Dolycoris indicu	Koralowiec dolycoris indicu	15686
Pyrausta phoenicealis	Koralowiec pyrausta phoenicealis	15281
Hemibagrus punctatus	Koralowiec hemibagrus punctatus	16517
Scleromystax barbatus	Koralowiec scleromystax barbatus	16651
Ortmanniana ligamentina	Koralowiec ortmanniana ligamentina	17015
Prothemus sanguinosus	Koralowiec themus sanguinosus	16985
Fissocantharis imparicornis	Koralowiec fissocantharis imparicornis	16147
Cyrebion gracilicornis	Koralowiec cyrebion gracilicornis	17179
Themus luteipes	Koralowiec themus luteipes	16812
Themus stigmaticus	Koralowiec themus stigmaticus	16970

Analizowany zbiór (tab. 6.4) obejmuje kompletne sekwencje genomu mitochondrialnego poszczególnych gatunków, co zapewnia bogaty sygnał filogenetyczny i umożliwia analizę relacji ewolucyjnych na poziomie całych genomów.

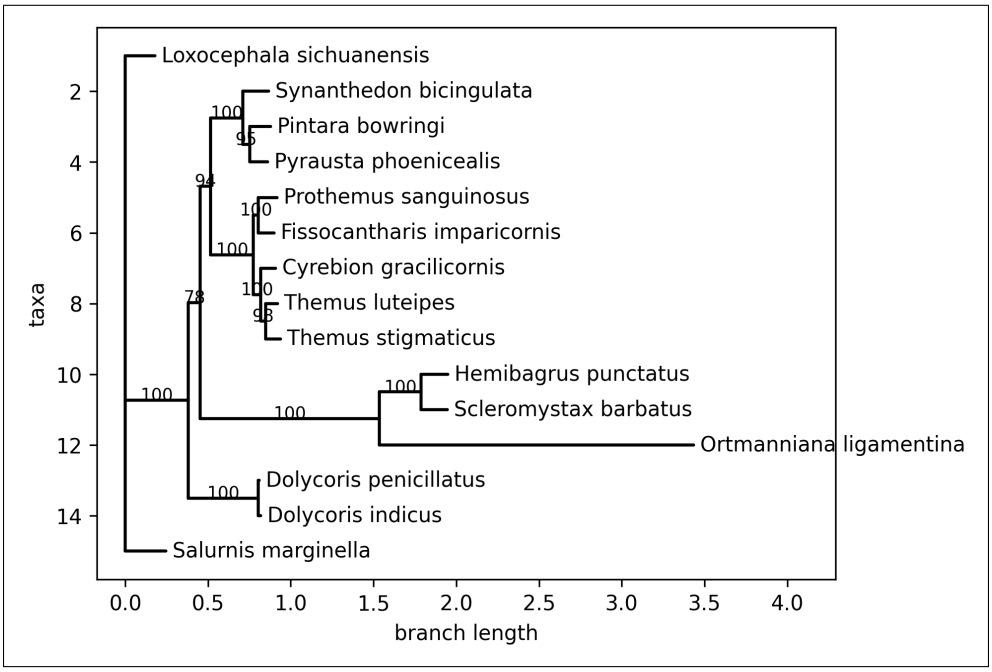
Cel testu:

- Weryfikacja poprawności działania aplikacji na dużych sekwencjach nukleotydowych.
- Ocena wydajności i stabilności pipeline'u przy pracy z pełnymi genomami mitochondrialnymi.

- Sprawdzenie jakości wyrównania oraz stabilności rekonstrukcji drzewa filogenetycznego przy sekwencjach o dużej długości.



Rysunek 6.7: Fragment wyrównanych sekwencji nukleotydowych - test 4



Rysunek 6.8: Drzewo filogenetyczne dla pełnych genomów mitochondrialnych - test 4

Wyniki testu:

Aplikacja poprawnie wykonała pełen proces analizy filogenetycznej dla pełnych genomów mitochondrialnych, obejmujący wyrównanie sekwencji (rys. 6.7) oraz rekonstrukcję drzewa filogenetycznego (rys. 6.8).

W porównaniu z testami na krótszych fragmentach genów, czas obliczeń był znacząco dłuższy, co bezpośrednio wynika z dużej długości analizowanych sekwencji. Jednocześnie

pipeline zachował stabilność działania, potwierdzając możliwość analizy dużych zestawów danych.

Uzyskane drzewo filogenetyczne wykazuje wysoką spójność z oczekiwanymi relacjami taksonomicznymi. Węzły odpowiadające głównym grupom charakteryzują się wysokimi wartościami bootstrap (>95), natomiast niższe wsparcie (78–85) obserwowane dla części rozgałęzień pomiędzy blisko spokrewnionymi gatunkami stanowi naturalny efekt ograniczonej różnorodności mitochondrialnej.

Test potwierdza, że system radzi sobie poprawnie z dużymi sekwencjami nukleotydowymi, zapewniając stabilne i wiarygodne wyniki analizy filogenetycznej.

6.4 Podsumowanie wyników

Wyniki analiz filogenetycznych uzyskane w trakcie testów wskazują, że aplikacja poprawnie integruje narzędzia MAFFT oraz IQ-TREE 2, umożliwiając pełen proces analizy filogenetycznej zarówno dla danych białkowych, jak i nukleotydowych.

Celem testów była weryfikacja poprawności działania systemu, a nie uzyskanie wyników naukowych. Uzyskane drzewa filogenetyczne wykazały spójność z oczekiwanymi relacjami taksonomicznymi, co potwierdza skuteczność zastosowanych algorytmów wyrównywania sekwencji i rekonstrukcji drzewa.

Zaobserwowano również przewidywaną zależność pomiędzy długością sekwencji a czasem analizy – dłuższe sekwencje wymagają więcej czasu obliczeniowego, a wraz ze wzrostem liczby analizowanych sekwencji konieczne jest zwiększenie mocy obliczeniowej.

6.5 Analiza działania systemu

W trakcie testów aplikacja wykazała stabilność działania na różnych zestawach danych wejściowych, pod warunkiem, że kod był odpowiednio dostosowany do specyfiki sekwencji. Konieczne były drobne modyfikacje, ponieważ różnorodność sposobów zapisu danych wejściowych czasami utrudniała automatyczne wyodrębnienie nazw sekwencji, co ograniczało uniwersalność systemu.

Przy próbach przetworzenia bardzo dużych zestawów pełnych genomów mitochondrialnych system napotkał ograniczenia sprzętowe związane z niewystarczającą ilością pamięci RAM. Pokazuje to, że pipeline jest stabilny i funkcjonalny dla większości typowych danych, jednak jego wydajność i skalowalność zależą od dostępnych zasobów sprzętowych oraz przygotowania danych wejściowych.

6.6 Potencjalny rozwój systemu

1. Rozszerzenie obsługi różnych formatów plików wejściowych (np. Clustal, Phylip), aby zwiększyć uniwersalność aplikacji.

2. Implementacja dodatkowych narzędzi do analizy filogenetycznej, takich jak RAxML czy MrBayes, umożliwiająca użytkownikowi wybór preferowanej metody rekonstrukcji drzewa.

3. Optymalizacja wydajności poprzez równoległe przetwarzanie sekwencji lub wykorzystanie chmury obliczeniowej do obsługi większych zestawów danych.

4. Dodanie funkcji automatycznego generowania raportów analizy w formacie PDF lub HTML, zawierających podsumowanie statystyk oraz wizualizacje.

5. Udoskonalenie interfejsu użytkownika poprzez możliwość konfiguracji parametrów analizy i lepszą wizualizację postępu pracy.

Rozdział 7

Podsumowanie i wnioski

Analizy filogenetyczne opierają się na modelach ewolucyjnych oraz założeniach statystycznych, które mają bezpośredni wpływ na uzyskane wyniki. Metody te bazują na prawdopodobieństwie, które ocenia najbardziej prawdopodobny scenariusz ewolucyjny w oparciu o dostępne dane i matematyczne modele zmian sekwencji.

W pracy przeanalizowano relatywnie niewielkie zbiory danych, jednak w przypadku większych zestawów sekwencji zapotrzebowanie na moc obliczeniową rośnie znacząco. Wyniki mogą się różnić w zależności od zastosowanej metody i parametrów analizy, co wymaga ostrożnej interpretacji.

Pomimo wiedzy na temat podobieństw międzygatunkowych i możliwości porównywania ich z cechami morfologicznymi, należy pamiętać, że proces ewolucji zachodzi przez bardzo długi czas i może obejmować zmiany trudne do uchwycenia przy użyciu dostępnych modeli. W związku z tym, nawet jeśli dwa organizmy wydają się podobne, analiza filogenetyczna opiera się na przyjętym wzorcu zmian sekwencji, który najlepiej odzwierciedla ewolucyjny sygnał, ale nie zawsze oddaje pełną złożoność historii ewolucyjnej.

Zaleca się, aby wyniki analiz traktować w kontekście biologicznym, uzupełniając je wiedzą morfologiczną i ekologiczną oraz stosując różne metody rekonstrukcji drzewa filogenetycznego w celu oceny stabilności wniosków.

Na podstawie przeprowadzonych testów można stwierdzić, że opracowany system poprawnie integruje narzędzia bioinformatyczne wykorzystywane w analizach filogenetycznych i umożliwia rekonstrukcję drzew filogenetycznych dla różnych typów danych sekwencyjnych. Uzyskane wyniki są zgodne z oczekiwaniami biologicznymi oraz literaturą przedmiotu, co potwierdza poprawność działania zaproponowanego rozwiązania. Jednocześnie należy podkreślić, że jakość wyników zależy od doboru danych wejściowych, modeli ewolucyjnych oraz parametrów analizy, co wymaga świadomej interpretacji rezultatów. Opracowane narzędzie może stanowić użyteczne wsparcie w analizach filogenetycznych, szczególnie w przypadku niewielkich i średnich zbiorów danych.

Bibliografia

[1] MAFFT Manual [online] https://mafft.cbrc.jp/alignment/software/manual/manual.html, data dostępu: 29.01.2026.

[2] Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear (2020)
*IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.**, in press. https://doi.org/10.1093/molbev/msaa015, data dostępu: 29.01.2026.

[3] Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L. (2009) *Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics**, 25(11), 1422–1423.

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

[5] 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.

[6] Płoński, P., Radomski, J. (2011) *Translacja drzew filogenetycznych.* Zeszyty Naukowe Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej, nr 9, seria: ICT Young. [online] https://www.ire.pw.edu.pl/~pplonski/papers/ICT_YOUNG_2011_pplonski_referat.pdf, data dostępu: 31.01.2026.

[7] PWN Encyklopedia [online] https://encyklopedia.pwn.pl/haslo/filogeneza;3900970.html, data dostępu: 09.01.2026.

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

[9] 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.

[10] Sievers, F. Higgins, D. G. (2017) *Clustal Omega for making accurate alignments of many protein sequences.Protein Science* 27(1), 135–145, [online] https://pmc.ncbi.nlm

.nih.gov/articles/PMC5734385/, data dostępu 30.01.2026
[11] Notredame, C. Higgins, D. G. Heringa, J. (2000) *T-Coffee: A novel method for multiple sequence alignments*, *Journal of Molecular Biology*, 302(1), 205–217, [online] https://tcoffee.org/Publications/Pdf/tcoffee.pdf, data dostępu 30.01.2026.
[12] Ari Löytynoja, Nick Goldman (2010) *PRANK: A probabilistic multiple sequence alignment program for DNA, codon and amino-acid sequences.* [online] https://ariloytynoja.github.io/prank-msa/, data dostępu: 30.01.2026.
[13] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., Barton, G. J. (2009) *Jalview Version 2—a multiple sequence alignment editor and analysis workbench.Bioinformatics*, 25(9), 1189–1191. [online] https://www.jalview.org/, data dostępu: 30.01.2026.
[14] Larsson, A. (2014) *AliView: a fast and lightweight alignment viewer and editor for large datasets.* [online]http://www.ormbunkar.se/aliview/, data dostępu: 30.01.2026.
[15] Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A. (2019) *RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics*, 35(21), 4453–4455. [online] https://github.com/amkozlov/raxml-ng, data dostępu: 30.01.2026.
[16] Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., Huelsenbeck, J. P. (2012) *MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology*, 61(3), 539–542. [online] https://nbisweden.github.io/MrBayes/, data dostępu: 30.01.2026.
[17] Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. (2018) *MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution*, 35(6), 1547–1549. [online] https://www.megasoftware.net/, data dostępu: 30.01.2026.
[18] Rambaut, A. (2018) *FigTree v1.4.4: Tree figure drawing tool.* [online] https://tree.bio.ed.ac.uk/software/figtree/, data dostępu: 30.01.2026.
[19] Huson, D. H., Scornavacca, C. (2012) *Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Systematic Biology*, 61(6), 1061–1067. [online] https://uni-tuebingen.de/en/faculties/mathematics-and-natural-sciences/subjects/computer-science/research/algorithms-in-bioinformatics/software/dendroscope/, data dostępu: 30.01.2026.
[20] Letunic, I., Bork, P. (2021) *Interactive Tree Of Life (iTOL) v6: an online tool for visualization and annotation of phylogenetic trees. Nucleic Acids Research*, 49(W1), W293–W296. [online] https://itol.embl.de/, data dostępu: 30.01.2026.
[21] Huerta-Cepas, J., Serra, F., Bork, P. (2016) *ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Molecular Biology and Evolution*, 33(6), 1635–1638. [online] https://etetoolkit.org/, data dostępu: 30.01.2026.
[22] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt von Haese-

ler, Lars S. Jermiin *ModelFinder: Fast model selection for accurate phylogenetic estimates* [artykuł] Nature Methods, 14:587–589, 2017. https://doi.org/10.1038/nmeth.4285, data dostępu: 31.01.2026.

[23] Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, Le Sy Vinh *UFBoot2: Improving the ultrafast bootstrap approximation* [artykuł] Mol. Biol. Evol., 35:518–522, 2018. https://doi.org/10.1093/molbev/msx281, data dostępu: 31.01.2026.

[24] IQ-TREE 2 Command Reference [online] https://iqtree.github.io/doc/Command-Reference, data dostępu: 29.01.2026.

Lista dodatkowych plików uzupełniających tekst pracy

Wszytskie dodatkowe pliki uzupełniające tekst pracy znajdują się na zdalnym repozytorium GitHub pod adresem github.com/FilipWspanialy/PhylogeneticAnalyses.

Spis rysunków

Spis tabel