Spis treści

Spis rysunków

1. Wstęp

1.1 Cel pracy

- wprowadzenie w obszar filogenetyki,
- zapoznanie z podstawowymi pojęciami dotyczącymi budowy drzew filogenetycznych,
- wyjaśnienie sposobu tworzenia drzew przy pomocy metod UPGMA i WPGMA,
- implementacja aplikacji do konstrukcji drzew filogenetycznych na podstawie analizy sekwencji nukleotydowych,
- graficzne przedstawienie wybranych drzew filogenetycznych.

1.2 Rozwiązania alternatywne

Obecnie istnieją różne programy filogenetyczne o wielu możliwościach, ale też ograniczeniach. Jako przykład można wymienić PAUP, TREE-PUZZLE czy PHYML.

1.3 Struktura DNA

Informacja o strukturze białek organizmów przechowywana jest w tzw. kwasach nukleinowych. Jednym z nich jest kwas deoksyrybonukleinowy (DNA). W jego skład wchodzą, m. in. nukleotydy nazywane też zasadami. Oznacza się je literami: A (adenina), C (cytozyna), T (tymina), G (guanina). Najbardziej stabilne są pary adenina-tymina i guanina-cytozyna. Para G-C jest związana trzema wiązaniami wodorowymi, a para A-T jest wiązana przez dwa wiązania wodorowe. Kolejność, w której nukleotydy są połączone razem tworzy kod, który ostatecznie określa kolejność aminokwasów w określonym białku [1].

1.4 Czym jest ewolucja?

Z biologicznego punktu widzenia, ewolucja to rozwój formy biologicznej z innych wcześniej istniejących form lub jej powstanie w postaci obecnie istniejącej na skutek działania doboru naturalnego i występowania modyfikacji. Przyczyną jej występowania

są zmiany warunków środowiskowych, skutkiem których formy muszą zostać do nich odpowiednio dostosowane. W każdej populacji istnieje więc pewna różnorodność biologiczna, zapewniana przez zmiany materiału genetycznego [7].

1.5 Zadania filogenetyki

Filogenetyka zajmuje się badaniem historii ewolucyjnej żyjących organizmów i przedstawia ich ewolucyjną dywergencję przy pomocy „drzew" - diagramów. Drzewa te mogą rozgałęziać się według różnych schematów. Proces ich powstawania nazywa się filogenezą. W jednym ze sposobów jej badania wykorzystywane są materiały kopalne, w których zawarte są informacje o przodkach obecnych form oraz czasie wystąpienia dywergencji. Są one jednak trudno dostępne, a opis cech morfologicznych często nie jest jednoznaczny. Inną metodą zdobycia danych molekularnych jest ich zapis w sekwencji DNA lub białek, gdzie nośnikami informacji o ewolucji i mutacjach są geny. W przeciwieństwie do poprzedniej metody, nie istnieje tu problem błędu systematycznego, dane łatwiej jest uzyskać oraz występują one w większej ilości. Ponadto, drzewa filogenetyczne skonstruowane na ich podstawie są bardziej wiarygodne i jednoznaczne. Dane molekularne są z tego powodu preferowanym, a czasem jedynym źródłem informacji, natomiast filogenetyka molekularna - podstawą badań powiązań ewolucyjnych pomiędzy genami (sekwencjami), a co za tym idzie - również między gatunkami. Jej podstawowym celem jest prawidłowa rekonstrukcja historii ewolucji organizmów na podstawie zmian między sekwencjami [7].

1.6 Drzewa filogenetyczne

1.6.1 Podstawowe założenia

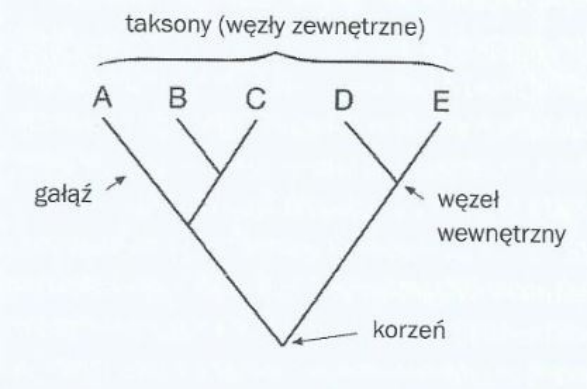
Podczas konstruowania drzew filogenetycznych konieczne jest przyjęcie pewnych założeń:

- sekwencje są homologiczne, co oznacza ich wspólne pochodzenie oraz, że ulegały dywergencji stopniowej,
- dywergencja jest dychotomiczna, to znaczy, że w każdym przypadku gałąź rodzicielska rozszczepia się na dokładnie dwie gałęzie potomne,
- każda pozycja w sekwencji ewoluuje niezależnie,
- analizowane sekwencje są różnorodne i dostarczają ilość informacji odpowiednią do konstrukcji jednoznacznych drzew [7].

1.6.2 Podstawowe pojęcia

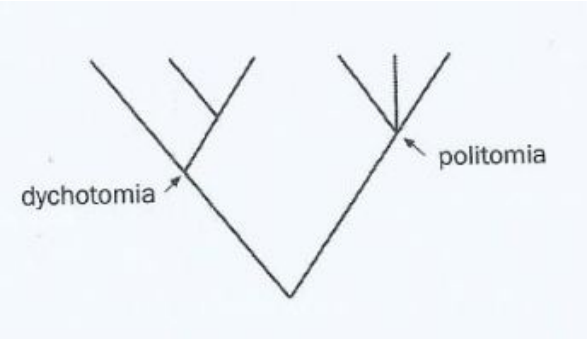
Aby móc zrozumieć metody powstawania drzew filogenetycznych, konieczne jest wcześniejsze zapoznanie się z pojęciami opisującymi ich elementy i budowę. Na ry-

sunku 1.1 przedstawiono przykład typowego drzewa filogenetycznego. Gałęzie to linie, które je tworzą. Zakończone są one tak zwanymi taksonami, z których każdy odpowiada jednemu gatunkowi (jednej sekwencji). Miejsce połączenia sąsiednich gałęzi ma nazwę węzła i określa domniemanego przodka danych dwóch gatunków. Wspólny przodek wszystkich taksonów należących do drzewa ma swój odpowiednik w punkcie nazywanym korzeniem znajdującym się na samym dole drzewa. Grupa monofiletyczna lub inaczej klad to grupa taksonów pochodzących od wspólnego przodka, który poza tym nie jest przodkiem żadnego innego taksonu. Jeśli grupa taksonów ma jednego wspólnego przodka, ale nie zawiera wszystkich jego potomków, nie może już być nazwana kladem, są to natomiast taksony parafiletyczne.



Rys. 1.1: Typowe drzewo filogenetyczne [7]

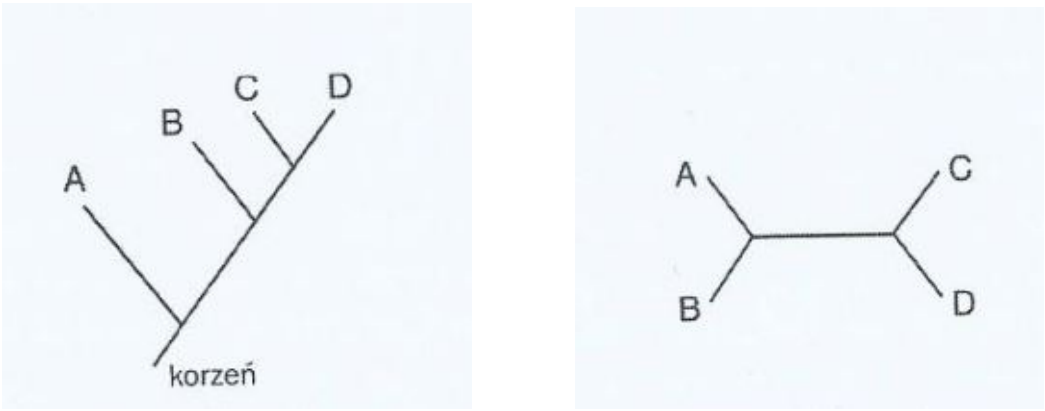
Gałęzie drzewa są ułożone według jego topologii. Ich podział na dwie gałęzie potomne określa się jako dychotomia (rys. 1.2). Czasem występuje także politomia, czyli sytuacja, gdy z punktu rozgałęzienia odchodzą więcej niż dwie gałęzie pochodne. Może ona być spowodowana tym, że przodek dał jednocześnie początek więcej niż dwóm potomkom (tzw. proces radiacji) albo brakiem możliwości precyzyjnego określenia kolejności podziałów - niepełnego rozwiązania filogenezy.



Rys. 1.2: Przykład wystąpienia dychotomii i politomii [7]

Istnieje możliwość konstruowania drzew ukorzenionych (rys. 1.3a) - zakładających znajomość wspólnego przodka - a także nieukorzenionych (rys. 1.3b), które wyłącznie

porządkują taksony zgodnie z ich wzajemnymi powiązaniami. W celu ustalenia kierunku drogi ewolucyjnej, konieczne jest ukorzenienie drzewa [7].

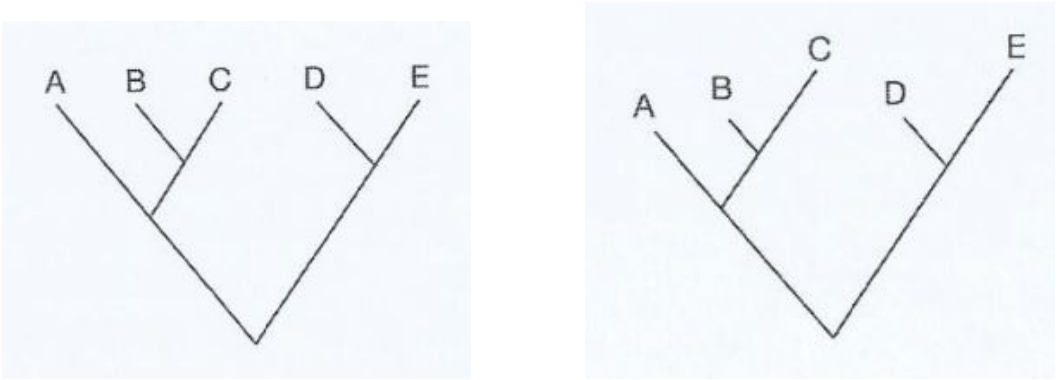


(a) Drzewo ukorzenione **(b)** Drzewo nieukorzenione

Rys. 1.3: Porównanie drzewa ukorzenionego i nieukorzenionego [7]

1.6.3 Formy reprezentacji drzew

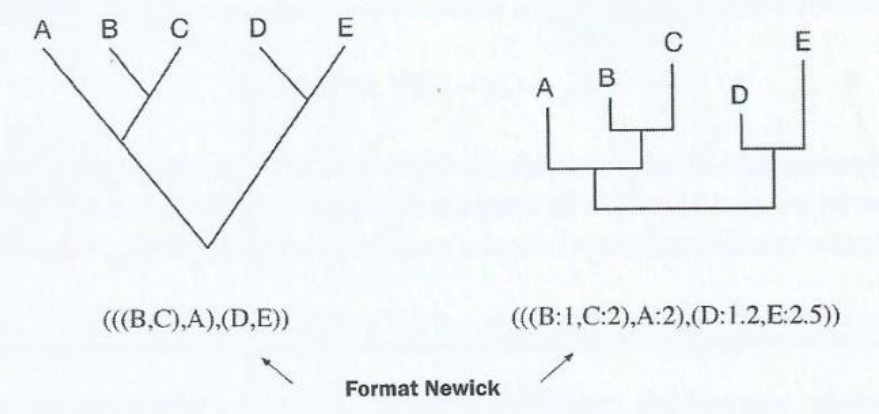
Jednym ze sposobów reprezentacji graficznej drzewa są filogramy (1.4b), gdzie długość gałęzi zależna jest od stopnia dywergencji ewolucyjnej. Są to drzewa wyskalowane. Prezentują one informacje nie tylko na temat występujących zależności, ale też o względnym czasie dywergencji poszczególnych gałęzi. W przypadku drzew niewyskalowanych, wszystkie gałęzie są jednakowej długości, co powoduje utratę części informacji. Takie drzewa nazywa się kladogramami [6, 7] (rys. 1.4a).



(a) Kladogram **(b)** Filogram

Rys. 1.4: Przykładowy kladogram i filogram [7]

W celu przekazania opisu topologii drzewa do programów komputerowych stosuje się specjalny format tekstowy - Newick (rys. 1.5).



Rys. 1.5: Przykład zapisu drzew w formacie Newick [7]

1.6.4 Problemy w szukaniu poprawnego drzewa

Do poprawnej konstrukcji drzewa filogenetycznego niezbędne jest znalezienie jego topologii i długości gałęzi, co nierzadko jest zadaniem trudnym i złożonym obliczeniowo. Liczba topologii drzewa może być bardzo duża już przy niewielkiej ilości taksonów. Rośnie ona wykładniczo zgodnie z zależnościami:

- drzewa ukorzenione:

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}, \tag{1.1}$$

- drzewa nieukorzenione:

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}, \tag{1.2}$$

gdzie N_R oznacza liczbę topologii drzew ukorzenionych, N_U liczbę topologii drzew nieukorzenionych, a n liczbę taksonów.

1.7 Etapy konstrukcji drzew filogenetycznych

Podczas konstruowania drzew odwzorowujących historię ewolucji gatunków, należy uwzględnić następujące etapy:

- wybór markerów molekularnych - dane z sekwencji nukleotydowych lub białkowych,
- przyrównanie sekwencji,
- wybór modelu ewolucji,

- wybór metody budowy drzewa,
- ocena wiarygodności uzyskanego drzewa.

Głównym założeniem niniejszej pracy jest przeprowadzenie etapu czwartego - wyboru metody konstrukcji drzewa oraz zaprezentowanie jej przy pomocy aplikacji.

1.7.1 Mutacje i przerwy w sekwencjach

Dobrym sposobem identyfikacji mutacji, które wystąpiły podczas ewolucji i spowodowały rozbieżności jest dopasowanie sekwencji parami. Najczęściej występujące mutacje to substytucje, insercje i delecje. Substytucje występują, gdy mutacja skutkuje zamianą jednego nukleotydu lub aminokwasu na inny. Powoduje to dopasowanie dwóch nieidentycznych aminokwasów. Insercje i delecje występują, gdy reszty aminokwasowe są dodawane lub usuwane. Insercje lub delecje (nawet te o długości tylko jednego znaku) są nazywane przerwami w przyrównaniu. Zwykle są one reprezentowane przez znak „-" dodawany do jednej lub drugiej sekwencji. Mogą wystąpić na końcu sekwencji lub w środku. Jednym ze skutków dodawania przerw jest spowodowanie, że całkowita długość każdego przyrównania jest taka sama. Dodanie przerw może pomóc w stworzeniu dopasowania odwzorowującego ewolucyjne zmiany, które miały miejsce [6].

1.7.2 Przyrównanie wielu sekwencji

Przyrównanie wielu sekwencji jest zbiorem trzech lub większej liczby sekwencji, które są częściowo lub całkowicie dopasowane. Homologiczne reszty aminokwasowe lub nukleotydy są przyrównane w kolumnach na całej długości sekwencji. Są one homologiczne w sensie ewolucyjnym - prawdopodobnie pochodzą od wspólnego przodka - oraz strukturalnym - w przypadku sekwencji białkowych, dopasowane reszty mają tendencję do zajmowania odpowiednich pozycji w trójwymiarowej strukturze każdego przyrównanego białka. Dopasowania wielu sekwencji nie są trudne do wygenerowania dla grupy bardzo blisko spokrewnionych sekwencji. Gdy tylko wykazują one pewną rozbieżność, problem wielokrotnego przyrównania staje się niezwykle trudny do rozwiązania. W szczególności trudno jest ocenić liczbę i lokalizację przerw [6].

1.7.3 Modele substytucji

Modele ewolucji kwasów nukleinowych i białek są wykorzystywane w metodach filogenetycznych jako podstawa do określenia odległości ewolucyjnych [4]. Istnieje kilka modeli substytucji, które są ograniczone pod tym względem, że w sytuacji, gdy na danej pozycji wystąpi zbyt wiele substytucji, dana pozycja zostaje wysycona. Oznacza to, że dywergencja ewolucyjna przekracza możliwości modeli do jej korekty, a rzeczywiste odległości ewolucyjne nie są możliwe do ustalenia [7]. Najprostszym modelem substytucji jest model Jukesa-Cantora (JC). Model opisuje jedną pozycję w dopasowaniu

sekwencji DNA. Zakłada on, że wszystkie cztery nukleotydy mają jednakową częstotliwość i że istnieje szybkość podstawienia α z dowolnego z czterech nukleotydów do dowolnego innego nukleotydu. Liczba miejsc, które różnią się między dwiema sekwencjami jest bezpośrednio obserwowalna przez porównanie dwóch sekwencji, ale nie uwzględnia ona wszystkich zmian, które zaszły, ponieważ mogło być więcej niż jedno podstawienie w danym miejscu. Dlatego należy obliczyć ewolucyjną odległość d, zdefiniowaną jako szacowana liczba substytucji, które wystąpiły na danej pozycji [4]. Zgodnie z modelem Jukesa-Cantora odległość tę określa wzór:

$$d_{AB} = -\frac{3}{4} ln[1 - \frac{4}{3} p_{AB}], \tag{1.3}$$

gdzie d_{AB} to odległość ewolucyjna pomiędzy sekwencjami A i B, a p_{AB} to obserwowana odległość pomiędzy tymi sekwencjami określona na podstawie odsetka substytucji na całej długości przyrównania [7].

1.7.4 Metody budowy drzew filogenetycznych

Drzewa można tworzyć przy pomocy metod należących do dwóch kategorii:

- metody oparte na odległościach:
  - klasteryzacja, np. metoda grupowania nieważonych par z arytmetycznymi średnimi (UPGMA), metoda grupowania ważonych par z arytmetycznymi średnimi (WPGMA), metoda łączenia sąsiadów (metoda najbliższego sąsiedztwa - NJ), uogólniona metoda NJ,
  - kryterium optymalności, np. metoda Fitcha-Margoliasha (FM), metoda minimalnej ewolucji (ME);
- metody oparte na znakach sekwencji taksonów:
  - metoda maksymalnej parsymonii (MP), ważona parsymonia, metoda kwarterów, metoda największej wiarygodności (ML), metoda NJML (połączenie metod NJ i ML), algorytm genetyczny (GA).

1.7.5 Porównanie metod UPGMA i WPGMA

Zarówno metoda UPGMA, jak i WPGMA to metody klastrowania. W metodzie UPGMA odległość między dwoma klastrami jest średnią odległością między wszystkimi obiektami każdego klastra, natomiast w metodzie WPGMA odległość między dwoma klastrami jest średnią arytmetyczną odległości między obiektami każdego klastra ważoną przez liczbę obiektów w każdym klastrze [3]. Algorytm obydwu metod:

- znajdź wartość minimalną (dwie najmniej różniące się/najbliższe sekwencje) macierzy odległości ewolucyjnych (utworzonej na podstawie modelu substytucji (1.7.3), np. Jukesa-Cantora),

- połącz najbliższe sekwencje tworząc klaster,
- oblicz nową macierz odległości ewolucyjnych:
  - UPGMA - oblicz średnią arytmetyczną biorąc pod uwagę wszystkie sekwencje należące do tworzonego klastera,
  - WPGMA - oblicz średnią arytmetyczną biorąc pod uwagę dwa klastery tworzące nowy klaster.
- powtarzaj, dopóki nie zostaną połączone wszystkie sekwencje.

Na rysunku 1.6 porównano wygląd przykładowej macierzy odległości ewolucyjnych w kolejnych krokach dla metody UPGMA i WPGMA [2, 5].

	A	B	C	D
A		0,40	0,35	0,6
B			0,45	0,7
C				0,55
D				

(a) Metoda UPGMA etap I

	A	B	C	D
A		0,40	0,35	0,6
B			0,45	0,7
C				0,55
D				

(b) Metoda WPGMA etap I

	A - C	B	D
A - C		(0,4 + 0,45)/2 = 0,425	(0,55 + 0,6)/2 = 0,575
B			0,7
D			

(c) Metoda UPGMA etap II

	A - C	B	D
A - C		(0,4 + 0,45)/2 = 0,425	(0,55 + 0,6)/2 = 0,575
B			0,7
D			

(d) Metoda WPGMA etap II

	A – C – B	D
A – C – B		(0,7 + 0,6 + 0,55)/3 = 0,617
D		

(e) Metoda WPGMA etap III

	A – C – B	D
A – C – B		(0,7 + 0,575)/2 = 0,6375
D		

(f) Metoda WPGMA etap III

Rys. 1.6: Porównanie kolejnych etapów w metodzie UPGMA i WPGMA

1.8 Metodologia

Problem przedstawiony w projekcie to konstrukcja drzew filogenetycznych. Jako jego rozwiązanie zaproponowano program, przyjmujący na wejściu fragmenty sekwencji nukleotydowych, które zostały przyrównane (1.7.2). Efektem jest macierz odległości ewolucyjnych zawierająca wyniki (odległości ewolucyjne) dla każdej pary sekwencji.

Uzyskuje się je przy pomocy modelu substytucji Jukesa - Cantora (1.7.3). Następnie, na podstawie wyżej wymienionej macierzy i zgodnie z metodą WPGMA (1.7.5) tworzone są drzewa filogenetyczne dla danych sekwencji. Graficznie przedstawiane są jako kladogramy (1.6.3), jednak informacja o długości poszczególnych gałęzi drzewa nie jest zatracana, lecz przedstawiona w formie kolejnej macierzy.

2. Specyfikacja wewnętrzna

Aplikację zaimplementowano w środowisku Matlab, z powodu względnie prostego sposobu operowania na macierzach i graficznej prezentacji drzew binarnych. Nazwa programu, podobnie jak jej głównej funkcji to *FilogeneticTrees*. W celu ułatwienia korzystania z aplikacji, utworzono graficzny interfejs użytkownika, również o tej samej nazwie. Jak wspomniano (1.8), parametrem przyjmowanym na wejściu są fragmenty sekwencji nukleotydowych jako łańcuchy znaków. Możliwe jest zatem wpisanie liter: „A", „G", „C", „T" oraz oznaczającego przerwę w sekwencji, „-". Program nie przyjmuje innych znaków, co zostało zapewnione funkcją *checkIfThereIsNoIllegalSign*. Kolejnym krokiem jest sprawdzenie zawartości wpisanych łańcuchów znaków. Odpowiada za nie funkcja *compareSequences*. Porównywane sekwencje powinny być takiej samej długości, nie powinny być identyczne, ani różnić się bardziej niż w 75%. Aby zabezpieczyć program przed przerwaniem jego działania w wymienionych przypadkach, utworzono funkcje, odpowiednio *checkIfLengthIsEqual* i *checkTheDifferencesBetweenSequences*. Jeśli wszystkie warunki zostały spełnione, przy pomocy funkcji *makeMatrixOfSequences* tworzona jest macierz, w której znajdują się wszystkie podane sekwencje (pominięte zostają pola tekstowe, które użytkownik pozostawił puste). Funkcja zwraca także długość sekwencji potrzebną w dalszych obliczeniach. Następuje też wyznaczenie macierzy odległości ewolucyjnych. Umieszczone w niej wartości to odległości (wyznaczone na podstawie ilości znaków różniących się) pomiędzy sekwencjami po zastosowaniu modelu substytucji Jukesa-Cantora (1.7.3) przy użyciu funkcji *jukesCantorSubstituteModel*. Podczas konstrukcji drzewa filogenetycznego, w programie głównym zostaje wywołana funkcja *createTreeByWpgmaMethod* jako argument przyjmująca utworzoną wcześniej macierz odległości ewolucyjnych. Wewnątrz niej został zawarty kod umożliwiający otrzymanie końcowej postaci drzewa filogenetycznego metodą WPGMA, a także parametrów koniecznych do jego graficznego zaprezentowania. Użyto tu takich funkcji, jak:

- *findFirstMinimumPosition* - odpowiedzialną za wyszukanie minimalnej wartości macierzy odległości ewolucyjnych, czyli zdefiniowanie, które sekwencje lub klastery są sobie najbliższe (ewolucyjnie), a co za tym idzie - należy połączyć w danej iteracji,

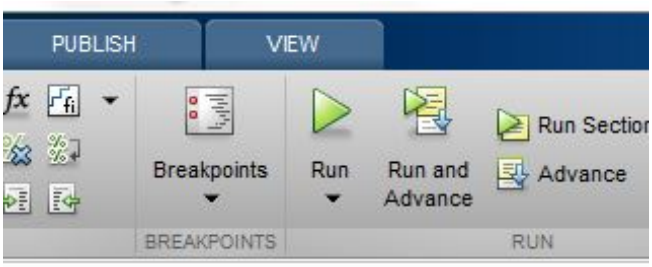
- *calculateBranchLength* - zwracającą długość gałęzi drzewa w danej iteracji (jest to połowa odnalezionej minimalnej wartości odległości); zwracana wartość zostaje dopisana do wektora długości gałęzi *branchLengthVector*, by posłużyć do obliczenia ostatecznych odległości pomiędzy sekwencjami,

- *makeClusterGroups* - przy jej pomocy tworzone są odpowiednie klastery lub do już istniejących dołączane są nowe sekwencje; jest to możliwe, między innymi dzięki funkcji *mergeRows*, która służy do łączenia dwóch istniejących klasterów, czyli kiedy nie jest dołączana żadna nowa sekwencja; w programie zrealizowano to jako połączenie dwóch wierszy macierzy klastrów *clusterGroupsArray*,
- *makeHelperClusterVectors* - za pomocą dwóch wektorów określającą, które sekwencje nie zostały jeszcze dołączone do tworzonego drzewa,
- *calculateNewDistanceMatrix* - tworzącą, na podstawie obecnych danych i pozycji wartości minimalnej, nową postać macierzy odległości ewolucyjnych, zgodnie z metodą WPGMA dla kolejnych iteracji,
- *calculateParametersToDrawTree* - definiującą takie parametry, jak liczba węzłów drzewa *nodesNumber*, wektor węzłów *nodes* określający ich położenie względem siebie czy wskazującą, które z węzłów powinny zostać oznaczone jako „liście" drzewa, czyli sekwencje.

Aplikacja umożliwia użytkownikowi wgląd w wartości macierzy odległości ewolucyjnych oraz przedstawia aktualny wygląd tworzonego drzewa filogenetycznego dla każdej iteracji. Zrealizowano to za pomocą wyżej wymienionego rozwiązania (kolejne postaci macierzy) oraz funkcji *displayTree* przyjmującej obliczone parametry i wyświetlającej obecną część drzewa. Ostatnią z wykorzystanych funkcji jest *signLeafsAsSequencesNumbers*, która pozwala na właściwe podpisanie węzłów drzewa jako sekwencji.

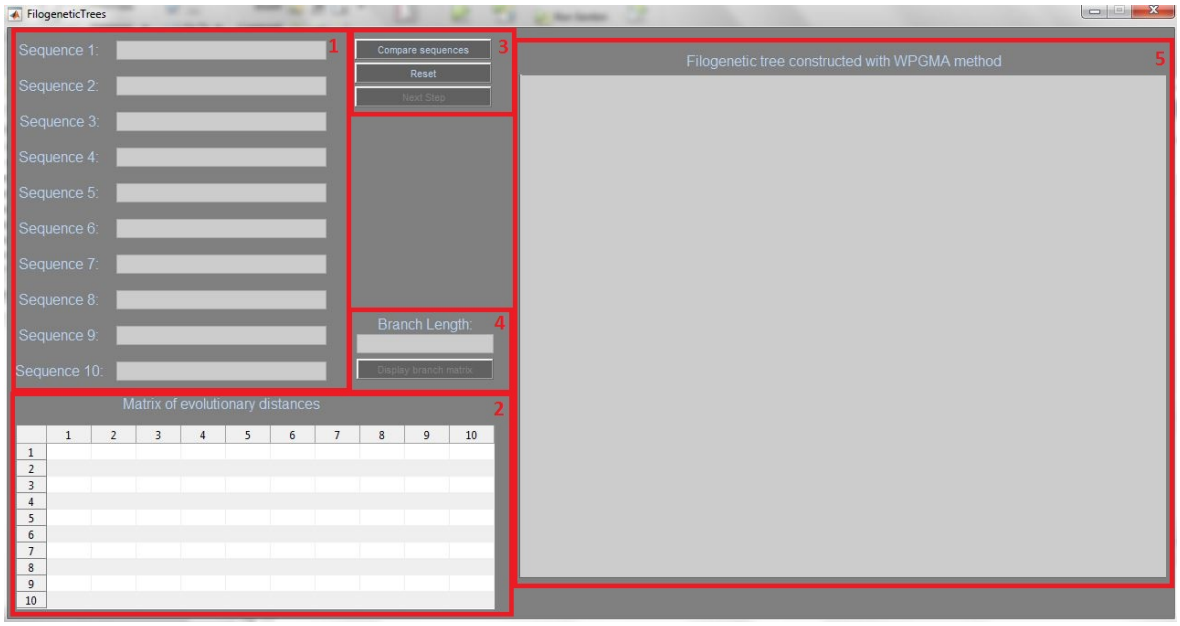
3. Specyfikacja zewnętrzna

Program został zaimplementowany w wersji Matlab R2015a. W celu korzystania z aplikacji, należy uruchomić plik *FilogeneticTrees.m* przyciskając przycisk **„Run”** (rys. 3.1).



Rys. 3.1: Uruchamianie aplikacji

Spowoduje to pojawienie się głównego okna aplikacji (rys. 3.2), gdzie następnie należy wpisać w odpowiednie pola fragmenty kolejnych, przyrównywanych sekwencji nukleotydowych (nr 1 na rys. 3.2), np. *„AGCTGTGA”*. Dopuszczalne są znaki „A”, „G”, „C” i „T” oraz przerwy (oznaczane symbolem „-”), np *„AGCTG-GA”*.



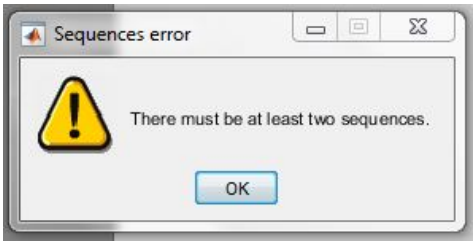
Rys. 3.2: Główne okno aplikacji

Pozostałe elementy głównego widoku aplikacji to:

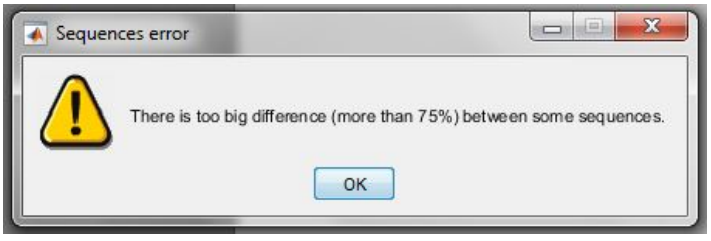
- macierz odległości ewolucyjnych (nr 2 na rys. 3.2) - w pierwszym kroku działania aplikacji, wiersze (oraz kolumny) odpowiadają przyrównywanym sekwencjom (np. wartość macierzy na pozycji (2,5) to odległość ewolucyjna między sekwencją drugą i piątą),
- przyciski **„Compare sequences”**, **„Reset”** oraz **„Next step”** (nr 3 na rys. 3.2) opisane w dalszej części pracy,
- okno służące wyświetlaniu długości aktualnie obliczanej gałęzi drzewa i przycisk **„Display branch matrix”** (nr 4 na rys. 3.2),
- miejsce na wyświetlenie dotychczas utworzonych gałęzi drzewa (nr 5 na rys. 3.2).

Sekwencje nie mogą się powtarzać, ani różnić od siebie w ponad 75% (jest to związane z wykorzystaniem modelu substytucji Jukesa-Cantora i wiarygodnością wykonywanych obliczeń). Jeśli te warunki nie będą spełnione, aplikacja wyświetli komunikat o błędzie (rys. 3.3a i 3.3c). Konieczne jest wpisanie co najmniej dwóch sekwencji, w przeciwnym razie program nie wykona żadnych obliczeń i wyświetli ostrzeżenie (rys. 3.3b).

(a) Błąd powtórzonej sekwencji



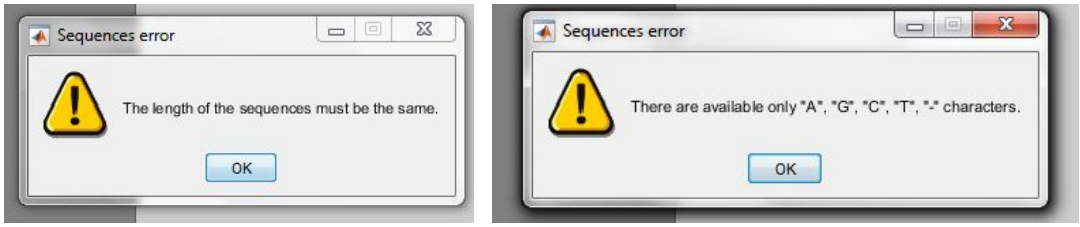
(b) Ostrzeżenie o zbyt małej liczbie wpisanych sekwencji



(c) Błąd zbyt dużej różnicy między sekwencjami

Rys. 3.3: Komunikaty o błędach dotyczących zawartości wpisanych sekwencji

Również w przypadku różnych długości sekwencji lub wpisania niedozwolonego znaku (innego niż „A”, „G”, „C”, „T” lub „-”) użytkownik zostaje o tym poinformowany (rys. 3.4).

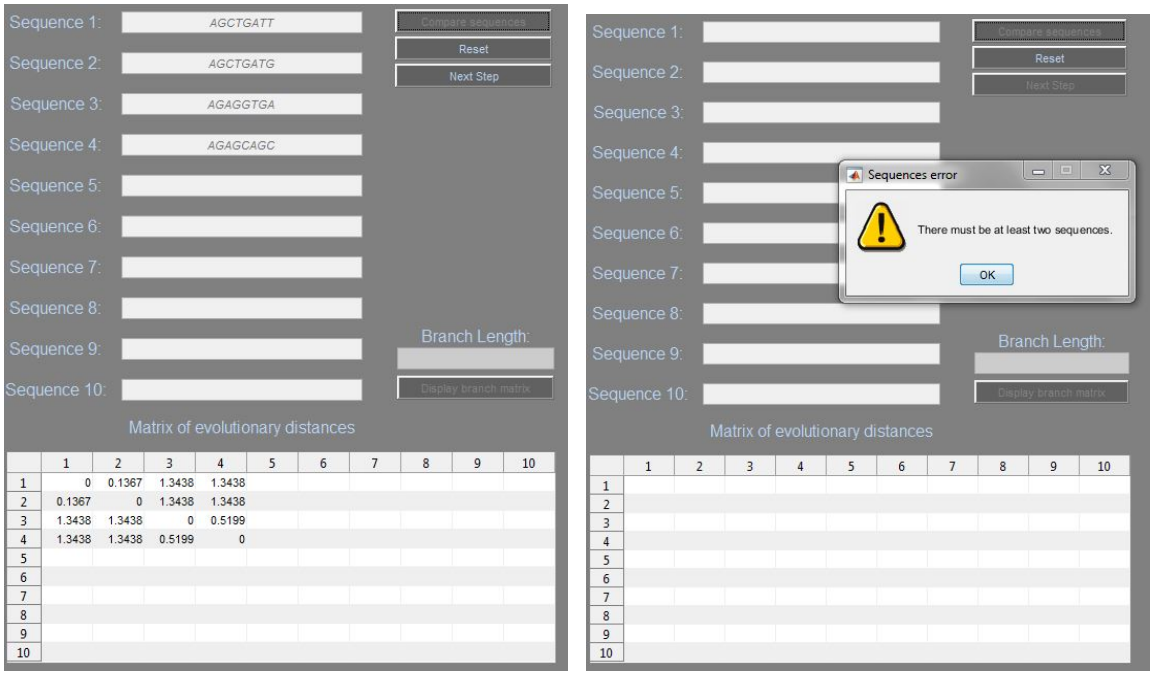


(a) Błąd różnych długości sekwencji **(b)** Błąd niedozwolonych znaków w sekwencji

Rys. 3.4: Komunikaty o błędach dotyczących długości i znaków w sekwencjach

Po upewnieniu się, że podane sekwencje mają właściwy format należy wcisnąć przycisk **„Compare sequences”**. Skutkuje to:

- zablokowaniem wyżej wymienionego przycisku,
- dezaktywacją pól tekstowych,



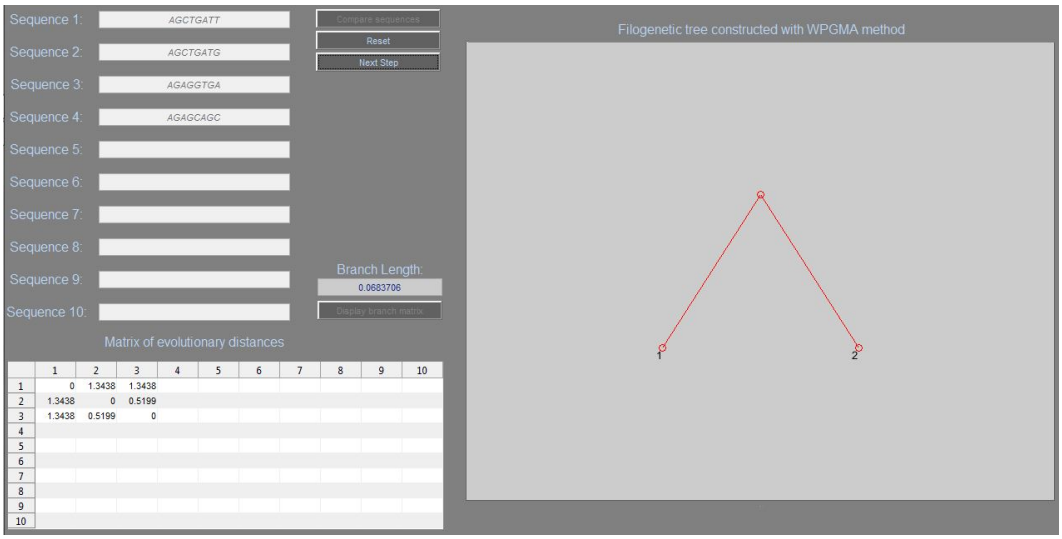
(a) Efekt wciśnięcia przycisku **„Compare sequenses”** w przypadku prawidłowego działania

(b) Efekt wciśnięcia przycisku **„Compare sequenses”** w przypadku wystąpienia błędu

Rys. 3.5: Widok aplikacji po wciśnięciu przycisku **„Compare sequences”**

- odblokowaniem przycisku **„Next step”** oraz wyświetleniem obliczonych wartości początkowej macierzy odległości ewolucyjnych lub - w przypadku wystąpienia błędu - pojawieniem się komunikatu (rys. 3.5).

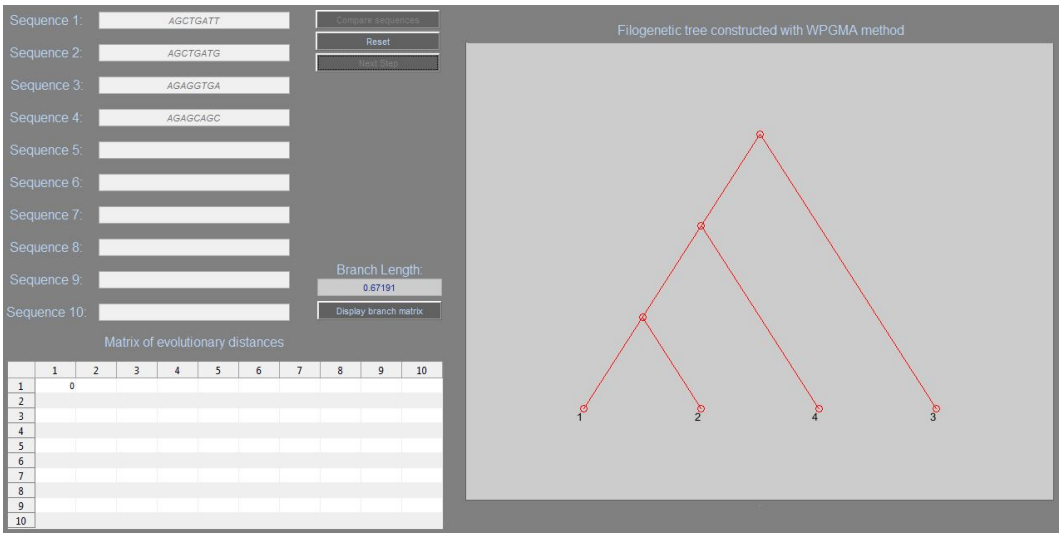
Kolejną czynnością jest wywołanie następnych kroków w tworzeniu drzewa filogenetycznego przy użyciu przycisku **„Next step”**. W efekcie obliczona zostaje nowa macierz odległości ewolucyjnych (o rozmiarze o 1 mniejszym w stosunku do poprzedniej iteracji) i wpisana na miejsce starej. Wyświetlone zostaje także poddrzewo końcowego drzewa filogenetycznego, które tworzone jest zgodnie z bieżącymi obliczeniami.



Rys. 3.6: Przykład efektu działania przycisku **„Next step”**

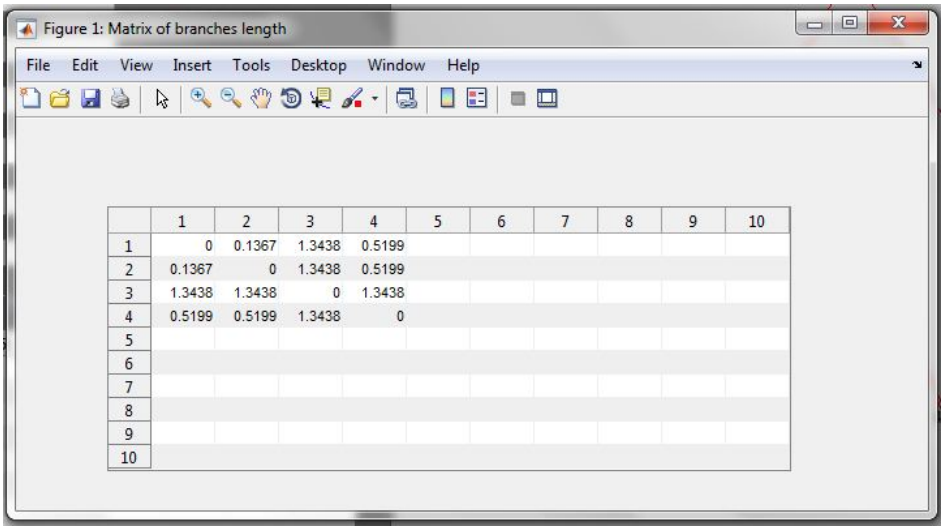
Dzięki temu użytkownik może obserwować rozrost drzewa krok po kroku (rys. 3.6) aż do uzyskania końcowego efektu. Dodatkowo, w polu tekstowym (nr 4 na rys. 3.2) wyświetlana jest długość bieżącej gałęzi drzewa.

Po przejściu do ostatniego kroku, zablokowany zostaje przycisk **„Next step”**, natomiast możliwe jest użycie przycisku **„Display branch matrix”** (rys. 3.7).



Rys. 3.7: Wygląd aplikacji po przejściu wszystkich kroków tworzenia drzewa

Pozwala on na wyświetlenie w osobnym oknie (rys. 3.8) macierzy zawierającej odległości pomiędzy wszystkimi sekwencjami. Wartości te obliczane są na podstawie wcześniej uzyskanych długości poszczególnych gałęzi.



	1	2	3	4	5	6	7	8	9	10
1	0	0.1367	1.3438	0.5199						
2	0.1367	0	1.3438	0.5199						
3	1.3438	1.3438	0	1.3438						
4	0.5199	0.5199	1.3438	0						
5										
6										
7										
8										
9										
10										

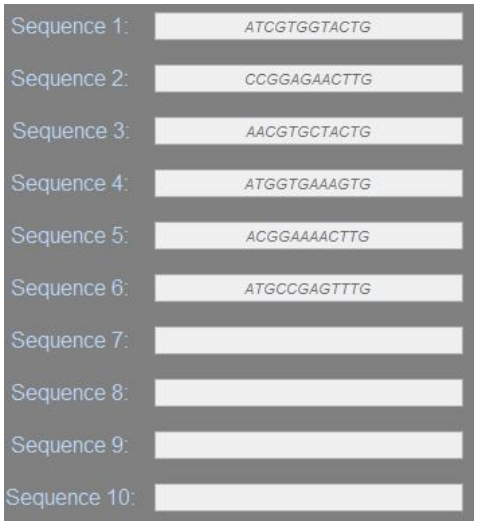
Rys. 3.8: Przykładowa macierz odległości pomiędzy sekwencjami obliczona na podstawie długości gałęzi drzewa

W aplikacji uwzględniono także jej przywrócenie do momentu wpisywania sekwencji. Aby to uczynić, należy użyć przycisku **„Reset”** dostępnego przez cały czas działania programu.

4. Symulacja działania programu

4.1 Wprowadzenie danych

Do symulacji użyto fragmentów sześciu losowych sekwencji nukleotydowych (rys. 4.1).



Rys. 4.1: Przykładowe sekwencje nukleotydowe

Po użyciu przycisku **„Compare sequences”** zostaje obliczona pierwotna postać macierzy odległości ewolucyjnych (rys. 4.2).

Matrix of evolutionary distances

	1	2	3	4	5	6	7	8	9	10
1	0	1.6479	0.1885	0.4408	1.6479	1.1281				
2	1.6479	0	1.6479	0.6082	0.1885	0.8240				
3	0.1885	1.6479	0	0.6082	1.6479	1.6479				
4	0.4408	0.6082	0.6082	0	0.6082	0.6082				
5	1.6479	0.1885	1.6479	0.6082	0	0.8240				
6	1.1281	0.8240	1.6479	0.6082	0.8240	0				
7										
8										
9										
10										

Rys. 4.2: Pierwotna postać macierzy odległości ewolucyjnych

4.2 Wyznaczanie kolejnych gałęzi drzewa

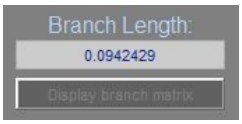
Wykorzystując przycisk **„Next step”** uzyskano nową macierz odległości ewolucyjnych, długość gałęzi oraz graficzną prezentację powstającego drzewa (rys. 4.3 - 4.17).

4.2.1 Krok pierwszy

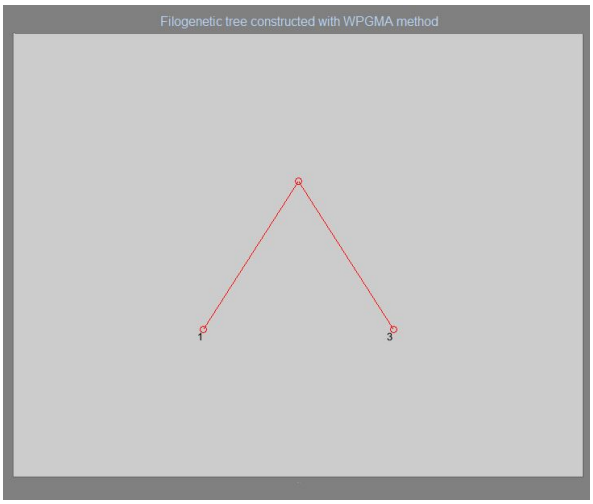
Matrix of evolutionary distances

	1	2	3	4	5	6	7	8	9	10
1	0	1.6479	0.5245	1.6479	1.3880					
2	1.6479	0	0.6082	0.1885	0.8240					
3	0.5245	0.6082	0	0.6082	0.6082					
4	1.6479	0.1885	0.6082	0	0.8240					
5	1.3880	0.8240	0.6082	0.8240	0					
6										
7										
8										
9										
10										

Rys. 4.3: Macierz odległości ewolucyjnych - krok pierwszy



Rys. 4.4: Długość dwóch pierwszych gałęzi drzewa



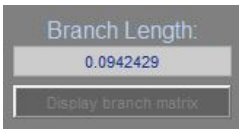
Rys. 4.5: Graficzne przedstawienie dwóch pierwszych gałęzi drzewa

4.2.2 Krok drugi

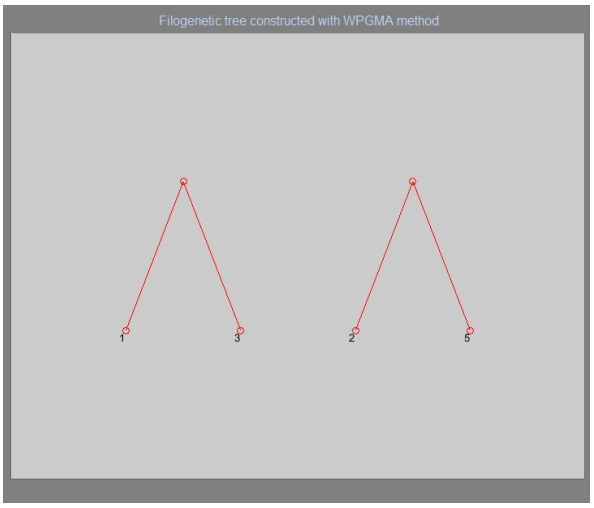
Matrix of evolutionary distances

	1	2	3	4	5	6	7	8	9	10
1	0	1.6479	0.5245	1.3880						
2	1.6479	0	0.6082	0.8240						
3	0.5245	0.6082	0	0.6082						
4	1.3880	0.8240	0.6082	0						
5										
6										
7										
8										
9										
10										

Rys. 4.6: Macierz odległości ewolucyjnych - krok drugi



Rys. 4.7: Długość kolejnych dwóch gałęzi drzewa - krok drugi



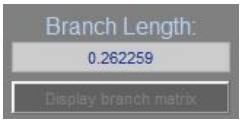
Rys. 4.8: Graficzne przedstawienie kolejnych dwóch gałęzi drzewa - krok drugi

4.2.3 Krok trzeci

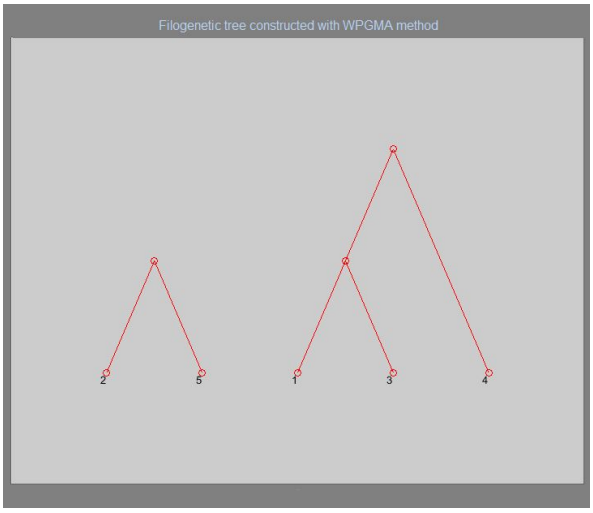
Matrix of evolutionary distances

	1	2	3	4	5	6	7	8	9	10
1	0	1.1281	0.9981							
2	1.1281	0	0.8240							
3	0.9981	0.8240	0							
4										
5										
6										
7										
8										
9										
10										

Rys. 4.9: Macierz odległości ewolucyjnych - krok trzeci

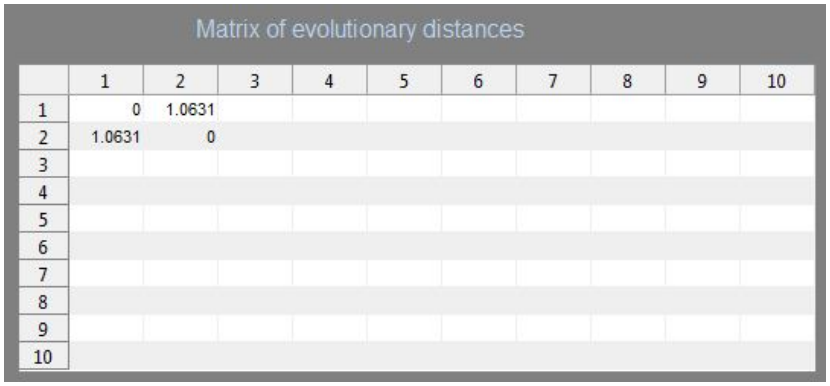


Rys. 4.10: Długość kolejnych dwóch gałęzi drzewa - krok trzeci



Rys. 4.11: Graficzne przedstawienie kolejnych dwóch gałęzi drzewa - krok trzeci

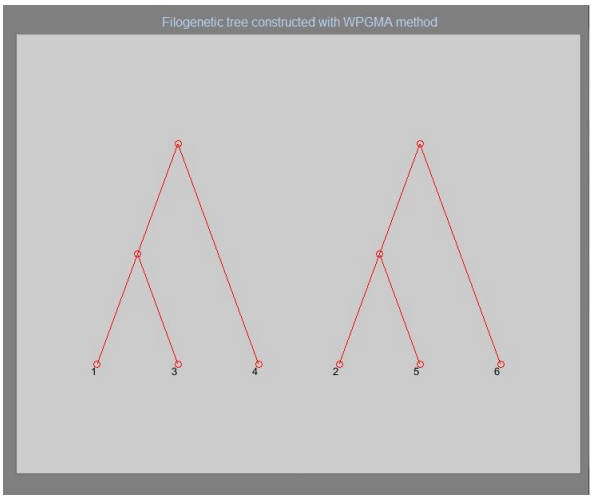
4.2.4 Krok czwarty



	1	2	3	4	5	6	7	8	9	10
1	0	1.0631								
2	1.0631	0								
3										
4										
5										
6										
7										
8										
9										
10										

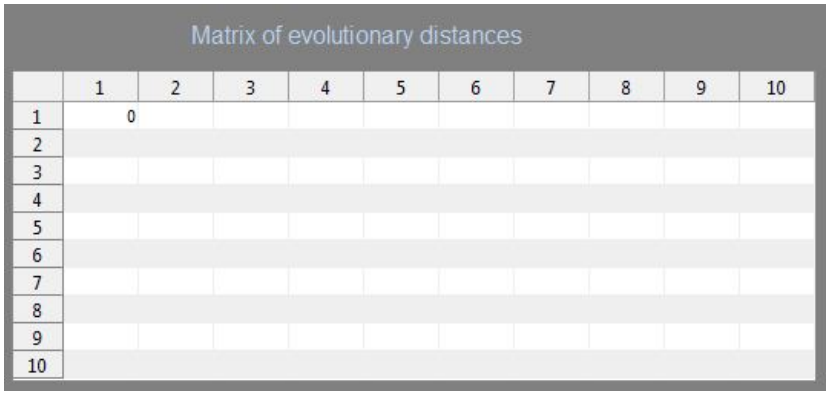
Rys. 4.12: Macierz odległości ewolucyjnych - krok czwarty

Rys. 4.13: Długość kolejnych dwóch gałęzi drzewa - krok czwarty

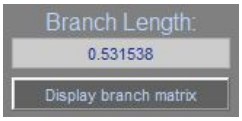


Rys. 4.14: Graficzne przedstawienie kolejnych dwóch gałęzi drzewa - krok czwarty

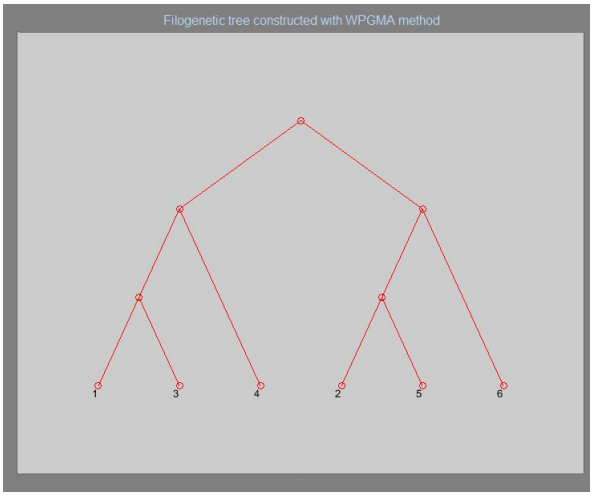
4.2.5 Krok piąty



Rys. 4.15: Macierz odległości ewolucyjnych - krok piąty

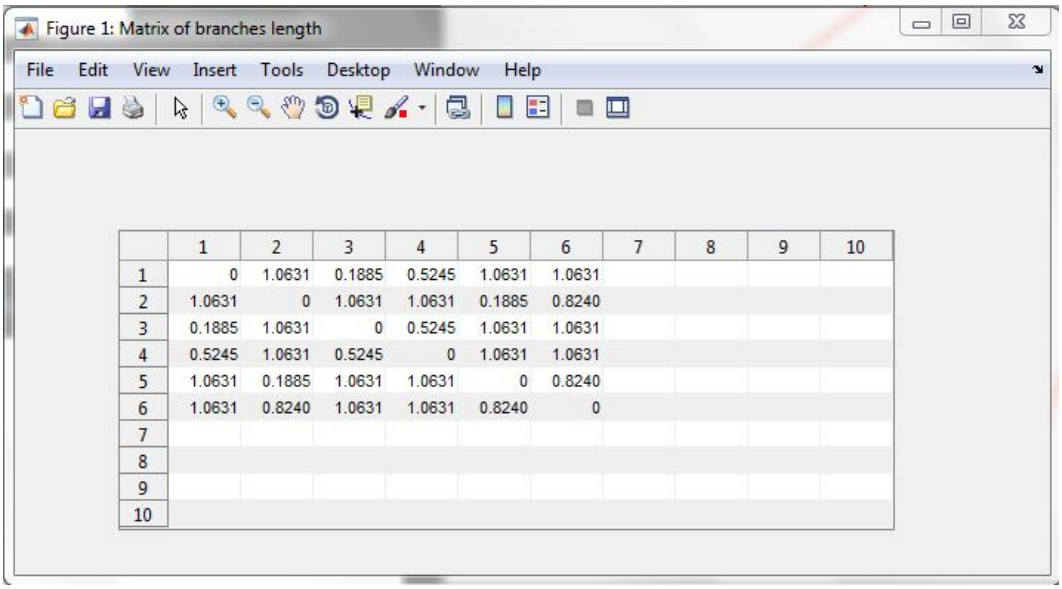


Rys. 4.16: Długość kolejnych dwóch gałęzi drzewa - krok piąty



Rys. 4.17: Graficzne przedstawienie kolejnych dwóch gałęzi drzewa - krok piąty

Po zakończeniu obliczeń, przycisk **„Next step”** zostaje zablokowany, natomiast możliwe staje się użycie przycisku **„Display branch matrix”**. Jego naciśnięcie powoduje pojawienie się nowego okna (rys. 4.18), w którym przedstawiono macierz odległości pomiędzy sekwencjami, obliczaną na podstawie wcześniej uzyskanych długości poszczególnych gałęzi.

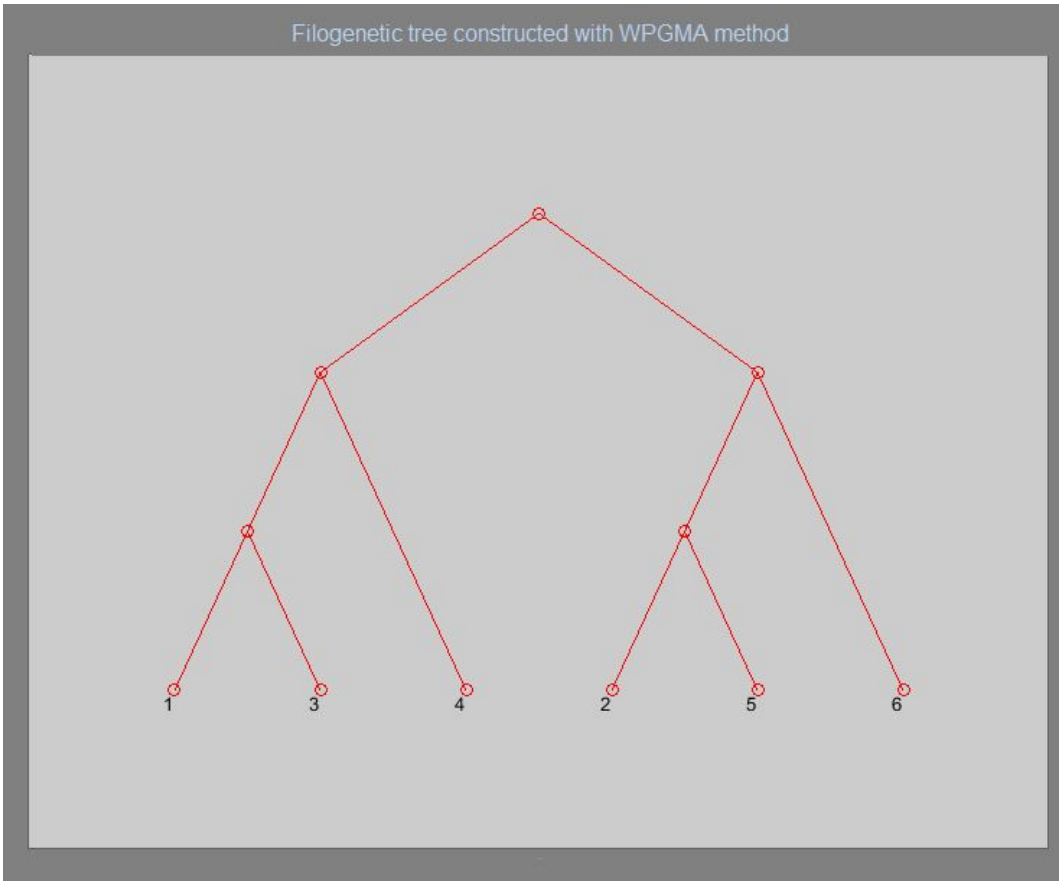


	1	2	3	4	5	6	7	8	9	10
1	0	1.0631	0.1885	0.5245	1.0631	1.0631				
2	1.0631	0	1.0631	1.0631	0.1885	0.8240				
3	0.1885	1.0631	0	0.5245	1.0631	1.0631				
4	0.5245	1.0631	0.5245	0	1.0631	1.0631				
5	1.0631	0.1885	1.0631	1.0631	0	0.8240				
6	1.0631	0.8240	1.0631	1.0631	0.8240	0				
7										
8										
9										
10										

Rys. 4.18: Macierz odległości pomiędzy sekwencjami utworzona na podstawie długości obliczonych gałęzi

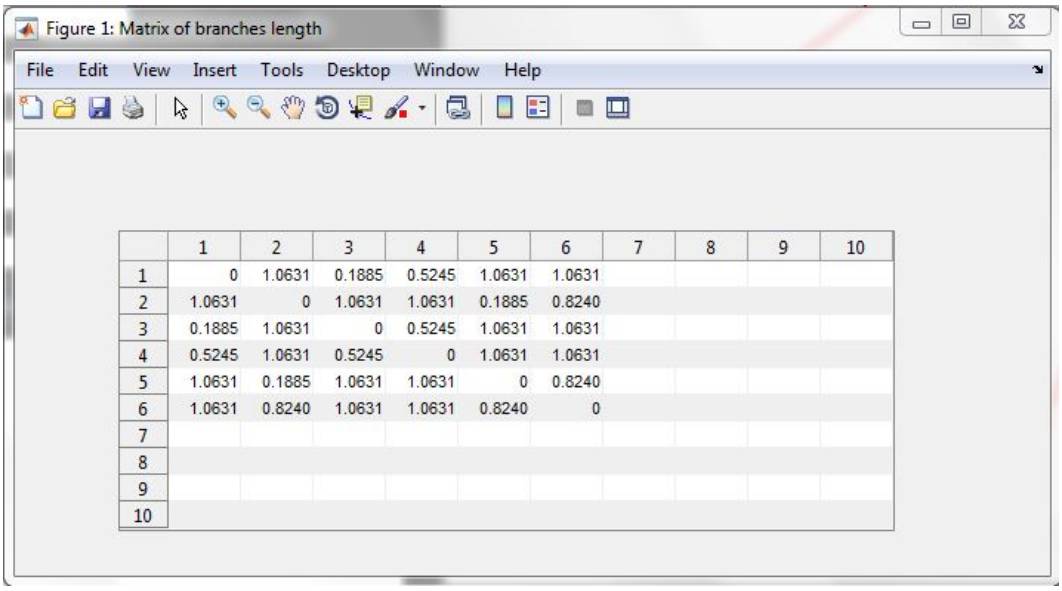
5. Podsumowanie

W projekcie przedstawiono i wyjaśniono podstawowe pojęcia z zakresu filogenetyki. W szczególności dotyczące drzew filogenetycznych oraz sposobów ich konstrukcji. Dwie z metod - UPGMA i WPGMA - opisano dokładniej porównując algorytmy, na których się opierają. Głównym celem pracy była implementacja aplikacji do konstrukcji drzew filogenetycznych. Osiągnięto go przy pomocy wspomnianej metody WPGMA. Program bazuje na przyrównaniu wielu (maksymalnie dziesięciu) fragmentów różnych sekwencji nukleotydowych. Macierz odległości ewolucyjnych pomiędzy badanymi sekwencjami jest wyznaczana w oparciu o model substytucji Jukesa-Cantora.



Rys. 5.1: Przykładowe drzewo filogenetyczne skonstruowane przy pomocy aplikacji utworzonej w projekcie

Efektem działania programu jest graficzne przedstawienie utworzonego drzewa filogenetycznego (rys. 5.1) oraz macierz odległości ewolucyjnych wyznaczana na podstawie długości poszczególnych gałęzi drzewa (rys. 5.2).



	1	2	3	4	5	6	7	8	9	10
1	0	1.0631	0.1885	0.5245	1.0631	1.0631				
2	1.0631	0	1.0631	1.0631	0.1885	0.8240				
3	0.1885	1.0631	0	0.5245	1.0631	1.0631				
4	0.5245	1.0631	0.5245	0	1.0631	1.0631				
5	1.0631	0.1885	1.0631	1.0631	0	0.8240				
6	1.0631	0.8240	1.0631	1.0631	0.8240	0				
7										
8										
9										
10										

Rys. 5.2: Przykładowa macierz odległości ewolucyjnych wyznaczona na podstawie długości gałęzi drzewa

Macierz ta różni się od pierwotnej postaci macierzy odległości ewolucyjnych opartej na różnicy pomiędzy znakami porównywanych sekwencji(rys. 5.3).

Matrix of evolutionary distances

	1	2	3	4	5	6	7	8	9	10
1	0	1.6479	0.1885	0.4408	1.6479	1.1281				
2	1.6479	0	1.6479	0.6082	0.1885	0.8240				
3	0.1885	1.6479	0	0.6082	1.6479	1.6479				
4	0.4408	0.6082	0.6082	0	0.6082	0.6082				
5	1.6479	0.1885	1.6479	0.6082	0	0.8240				
6	1.1281	0.8240	1.6479	0.6082	0.8240	0				
7										
8										
9										
10										

Rys. 5.3: Przykładowa macierz odległości ewolucyjnych wyznaczona na podstawie różnicy znaków pomiędzy sekwencjami

Można więc wnioskować, że wykorzystana metoda zyskuje na prostocie implementacji oraz niskiej złożoności obliczeniowej kosztem dokładności.

Bibliografia

[1] Alters S., „Biology: Understanding Life", Jones and Bartlett Publishers, Londyn, 2000

[2] Carr S. M., „UPGMA vs WPGMA", Text material, 2007

[3] Crow T. M., Albeke S. E., Buerkle C. A., Hufford K. M., „Provisional methods to guide species-specific seedtransfer in ecological restoration", Ecosphere, esa article, 2018, USA

[4] Higgs P. G. and Attwood T. K., „Bioinfrmatics and molecular evolution", Blackwell Publishing, United Kingdom

[5] Meneely P., Hoang R. D., Okeke I. N., Heston K., „Genetics. Genes, genomes and evolution.", Oxford, 2017

[6] Pevsner J., „Bioinformatics and functional genomics", third edition, Wiley Blackwell, 2015, Singapur

[7] Xiong J., „Podstawy bioinformatyki", Wydawnictwa Uniwersytetu Warszawskiego, 2009, Warszawa