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,
- graficzne przedstawienie wybranych drzew filogenetycznych.

1.2 Rozwiązania alternatywne

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

1.3 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 [1].

1.4 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 [1].

1.5 Drzewa filogenetyczne

1.5.1 Podstawowe założenia

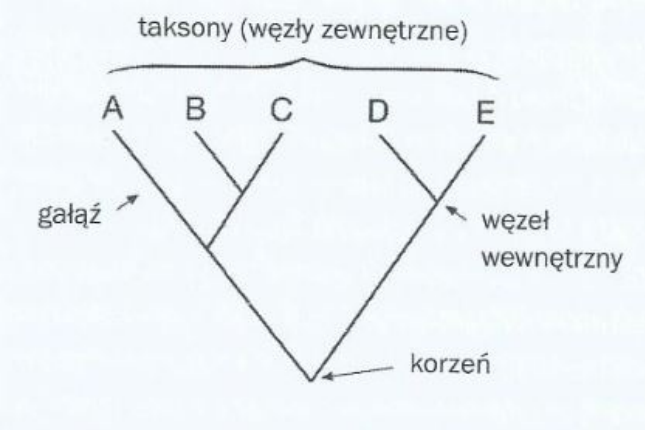
Podczas konstruowania drzew filogenetycznych konieczne jest przyjęcie pewnych założeń:

- 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 [1].

1.5.2 Podstawowe pojęcia

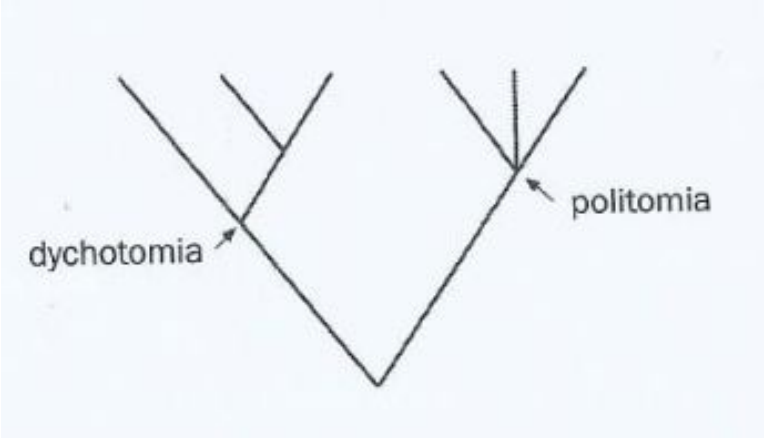
Aby móc zrozumieć metody powstawania drzew filogenetycznych, konieczne jest wcześniejsze zapoznanie się z pojęciami opisującymi ich elementy i budowę. Na rysunku 1.1 przedstawiono przykład typowego drzewa filogenetycznego. Gałęzie to linie, które je tworzą. Zakończone są one tak zwanymi taksonami, z 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.



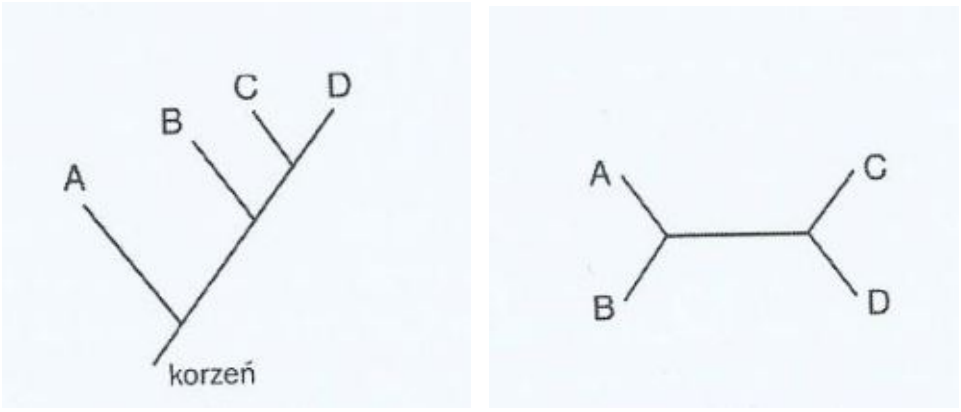
Rys. 1.1: Typowe drzewo filogenetyczne [1]

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 [1]

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 [1].

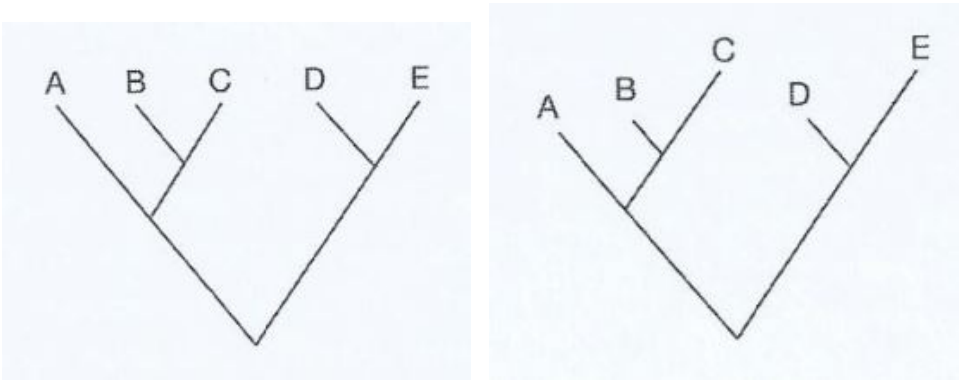


(a) Drzewo ukorzenione **(b)** Drzewo nieukorzenione

Rys. 1.3: Porównanie drzewa ukorzenionego i nieukorzenionego [1]

1.5.3 Formy reprezentacji drzew

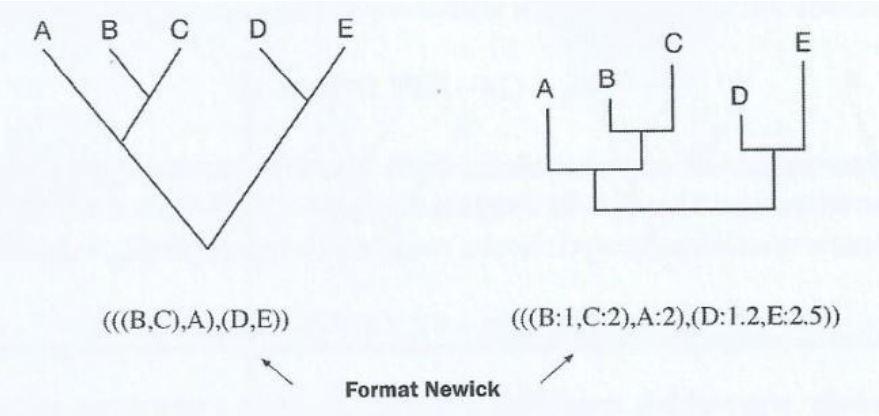
Jednym ze sposobów reprezentacji graficznej drzewa są filogramy (1.4b), gdzie długość gałęzi zależna jest od stopnia dywergencji ewolucyjnej. Są to drzewa wyskalowane. Prezentują one informacje, nie tylko na temat występujących zależności, ale też o względnym czasie dywergencji poszczególnych gałęzi. W przypadku drzew niewyskalowanych, wszystkie gałęzie są jednakowej długości, co powoduje utratę części informacji. Takie drzewa nazywa się kladogramami [1, 2] (rys. 1.4a).



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

Rys. 1.4: Przykładowy kladogram i filogram [1]

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 [1]

1.5.4 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:
$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}. \tag{1.2}$$

1.6 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.

1.6.1 Mutacje i luki w sekwencjach

Dobrym sposobem identyfikacji mutacji, które wystąpiły podczas ewolucji i spowodowały rozbieżność sekwencji dwóch badanych białek jest dopasowanie sekwencji parami. Najczęściej występujące mutacje to substytucje, insercje i delecje. W sekwencjach białkowych substytucje występują, gdy mutacja skutkuje zamianą kodonu dla jednego aminokwasu na inny. Powoduje to dopasowanie dwóch nieidentycznych aminokwasów. Insercje i delecje występują, gdy reszty są dodawane lub usuwane. Zwykle są one reprezentowane przez znak „-" dodawany do jednej lub drugiej sekwencji. Insercje lub delecje (nawet te o długości tylko jednego znaku) są nazywane lukami w wyrównaniu. Luki te mogą wystąpić na końcu białka lub w środku. Jednym ze skutków dodawania przerw jest spowodowanie, że całkowita długość każdego wyrównania jest taka sama. Dodanie luk może pomóc w stworzeniu dopasowania odwzorowującego ewolucyjne zmiany, które miały miejsce [2].

1.6.2 Przyrównanie wielu sekwencji

Przyrównanie wielu sekwencji jest zbiorem trzech lub większej liczby sekwencji białka (lub kwasu nukleinowego), które są częściowo lub całkowicie wyrównane. Homologiczne reszty są wyrównane w kolumnach na całej długości sekwencji. Są one homologiczne w sensie ewolucyjnym - prawdopodobnie pochodzą od wspólnego przodka - oraz strukturalnym - wyrównane reszty mają tendencję do zajmowania odpowiednich pozycji w trójwymiarowej strukturze każdego wyrównanego białka. [2] Dopasowania wielu sekwencji nie są trudne do wygenerowania dla grupy bardzo blisko spokrewnionych sekwencji białkowych (lub DNA). Gdy tylko sekwencje wykazują pewną rozbieżność, problem wielokrotnego wyrównania staje się niezwykle trudny do rozwiązania. W szczególności trudno jest ocenić liczbę i lokalizację luk. [2]

1.6.3 Modele substytucji

Modele ewolucji kwasów nukleinowych i białek są wykorzystywane w metodach filogenetycznych jako podstawa do określenia odległości ewolucyjnych [3]. 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 korekty homoplazji, a rzeczywiste odległości ewolucyjne nie są możliwe do ustalenia [1]. Najprostszym modelem substytucji jest model Jukesa-Cantora (JC). Model opisuje jedno miejsce w dopasowaniu sekwencji DNA. Podstawą w tym miejscu może być A, C, G lub T. Model zakłada, że wszystkie cztery zasady mają jednakową częstotliwość i że istnieje szybkość podstawienia α z dowolnej z czterech zasad DNA do dowolnej innej zasady. Liczba miejsc, które różnią się między dwiema sekwencjami D jest bezpośrednio obserwowalna przez

porównanie dwóch sekwencji, ale nie uwzględnia ona wszystkich zmian, które zaszły, ponieważ mogło być więcej niż jedno podstawienie na miejsce. Dlatego należy obliczyć ewolucyjną odległość d, zdefiniowaną jako szacowana liczba substytucji, które wystąpiły na miejscu. [3] Zgodnie z modelem Jukesa-Cantora odległość tą określa wzór:

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

1.6.4 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 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).

1.6.5 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 [4]. Algorytm obydwu metod:

- znajdź wartość minimalną (dwie najmniej różniące się/najbliższe sekwencje) macierzy odległości ewolucyjnych,
- 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 [5, 6].

	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

1.7 Metodologia

Problem przedstawiony w projekcie to konstrukcja drzew filogenetycznych. Jako jego rozwiązanie zaproponowano program, przyjmujący na wejściu fragmenty sekwencji nukleotydowych, które zostają przyrównane (1.6.2). Efektem jest macierz odległości ewolucyjnych zawierająca wyniki (odległości ewolucyjne) dla każdej pary sekwencji. Uzyskuje się je przy pomocy modelu substytucji Jukesa - Cantora (1.6.3). Następnie, na podstawie wyżej wymienionej macierzy i zgodnie z metodą WPGMA (1.6.5) tworzone są drzewa filogenetyczne dla danych sekwencji. Graficznie przedstawiane są jako kladogramy (1.5.3), 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 (1.7), parametrem przyjmowanym na wejściu są fragmenty sekwencji nukleotydowych jako łańcuchy znaków. Możliwe jest zatem wpisanie liter: „A", „G", „C", „T" oraz oznaczającego lukę w sekwencji, „-". Program nie przyjmuje innych znaków, co zostało zapewnione funkcją *checkIfThereIsNoIllegalSign*.
Kolejnym krokiem jest przyrównanie wielu sekwencji (1.6.2). 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 właściwe porównywanie sekwencji, którego efektem jest powstanie macierzy odległości ewolucyjnych. Umieszczone w niej wartości to odległości (ilości znaków różniących się) pomiędzy sekwencjami po zastosowaniu modelu substytucji Jukesa-Cantora (1.6.3) przy użyciu funkcji *jukesCantorSubstituteModel*.
Podczas przyrównywania sekwencji, 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 określający ich położenie względem siebie *nodes* 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.

Bibliografia

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

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

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

[4] 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

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

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