Bioinformatyka - zadanie 2

Badanie filogenetyki roślin nagonasiennych na podstawie reprezentantów rodzin

Konrad Komisarczyk

17 maja 2021

Wstęp

Rośliny nagonasienne (*Gymnospermae*) to grupa roślin charakteryzująca się nie wytwarzaniem otaczających nasiona owoców. Wraz z siostrzanym kladem roślin okrytonasiennych razem tworzą klad roślin nasiennych. *Gymnospermae* współcześnie reprezentowana jest przez ponad 1000 gatunków. Kiedyś tych gatunków było więcej, jednak podejrzewa się, że konkurencja ze strony roślin okrytonasiennych prowadzi do wymierania gatunków nagonasiennych i spadku różnorodności w tej grupie.

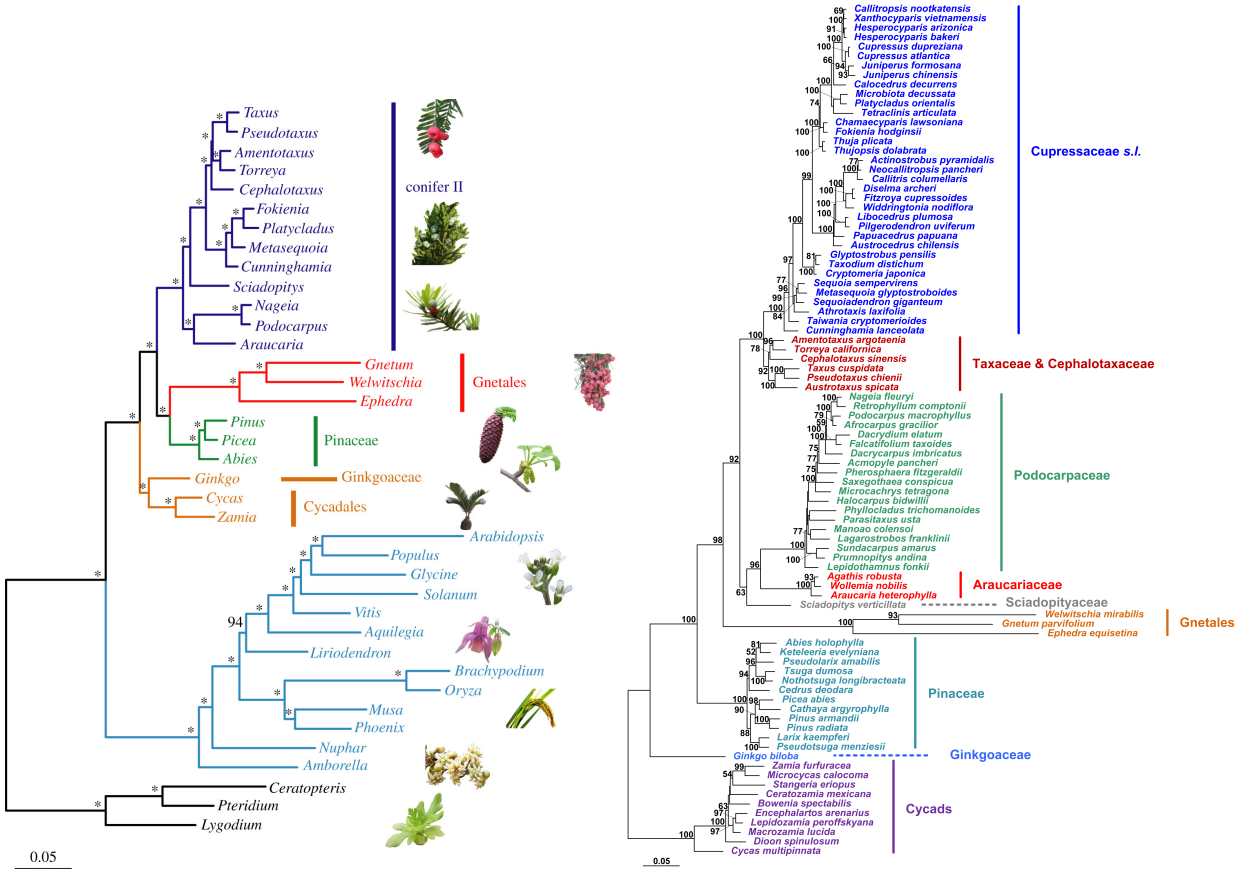
Problemy filogenetyki roślin nagonasiennych

Filogenetyka nagonasiennych jest ciekawym zagadnieniem ze względu na wciąż trwającą dyskusję nad systematyką wewnątrz grupy. Wciąż pojawiają się inne drzewa filogenetyczne i klasyfikacje w obrębie grupy (Ran et al. 2018) (Lu et al. 2014) (Christenhusz et al. 2010) (Simpson 2010), pomiędzy którymi istnieją znaczne różnice.

Podobieństwo między współczesnymi wynikami stanowi odłączanie miłorzębowatych (*Ginkgoaceae*) i sagowcowatych (*Cycadaceae*) jako wcześniejszych grup rozwojowych i wskazywanie sosnowatych (*Pinaceae*) jako bazowej grupy względem gniotowców (*Gnetales*). Na głębszych poziomach drzew ponadto prace drugiego dziesięciolecia XXI wieku zgadzają się co do wielu rzeczy, m.in. zawsze pary araukariowate (*Araucariaceae*) - zastrzalinowate (*Podocarpaceae*), cisowate (*Taxaceae*) - cyprysowate (*Cupressaceae*) oraz welwiczjowate (*Welwitschiaceae*) - gniotowate (*Gnetaceae*) są uważane jako pary siostrzane, a przęślowate (*Ephedraceae*) jako grupa bazalna względem ostatniej z nich (tworzy razem z tą parą gniotowce (*Gnetales*)).

Różnice we współczesnych drzewach filogenetycznych dotyczą umieszczenia grup sosnowatych (*Pinaceae*) (czy są one bazowe względem araukariowców i cyprysowców, czy tylko gniotowców) i sośnicowatych (*Sciadopityaceae*) (czy stanowią one klad bazowy względem araukariowatych i zastrzalinowatych, czy cisowatych i cyprysowatych).

W przeszłości, przed erą molekularnych badań filogenetycznych, problematyczna była grupa gniotowców (*Gnetales*), która ze względu na swoje cechy fenotypowe dawniej nie była klasyfikowana jako rośliny nagonasienne, a grupa siostrzana do roślin okrytonasiennych.



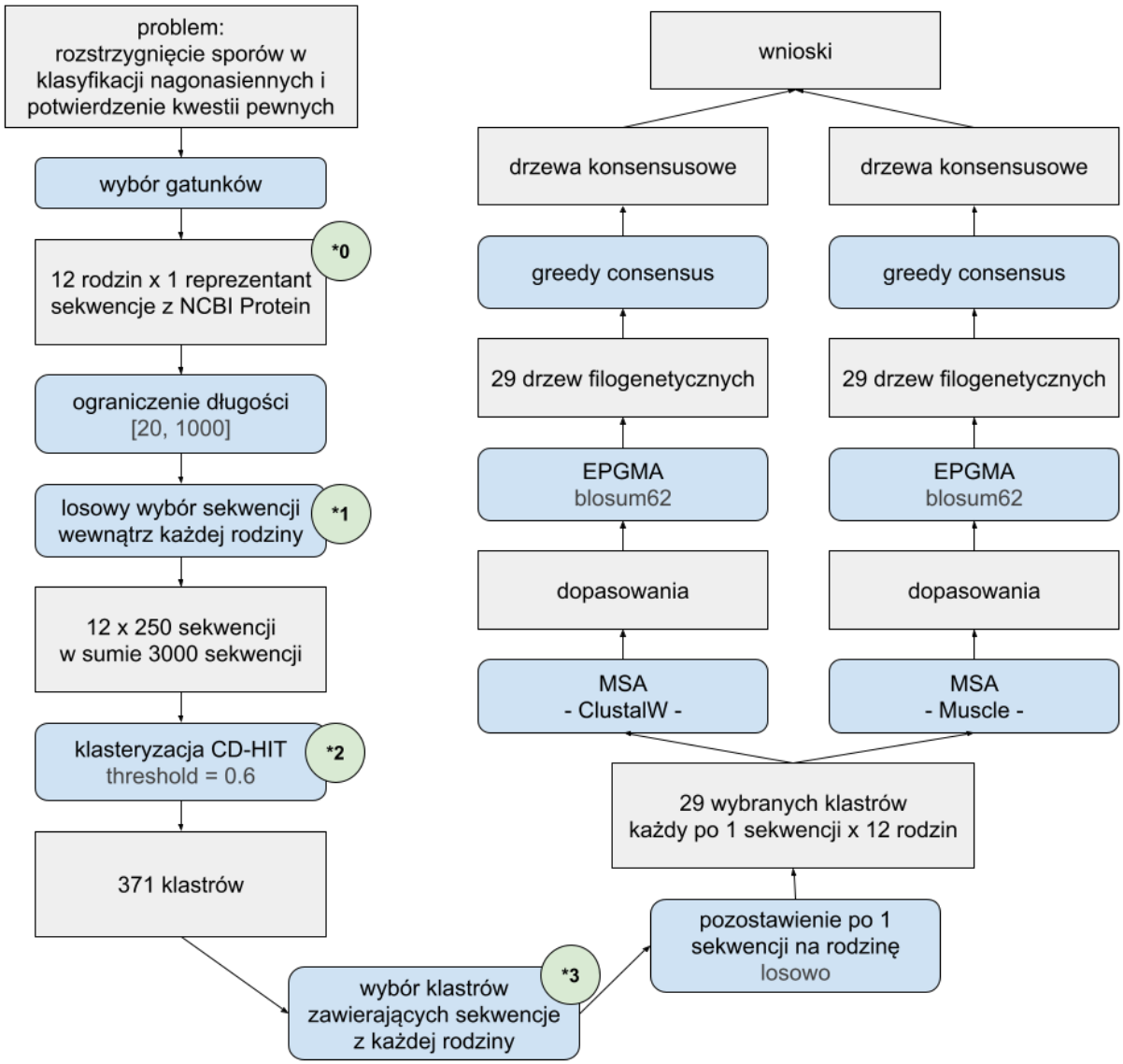
Rysunek 1: Drzewa filogenetyczne z artykułów (Ran et al. 2018) (po lewej) i (Lu et al. 2014) (po prawej)

Cel pracy

Celem pracy jest analiza filogenetyczna roślin nagonasiennych na podstawie sekwencji białkowych 12 gatunków z grupy - po jednym przedstawicielu każdej rodziny. Analiza ma potwierdzić ugruntowane elementy systematyki wewnątrz grupy oraz przedstawić dowody po jednej ze stron w opisanych powyżej kwestiach spornych.

Metody

Poniższy diagram (Rysunek 2) przedstawia pipeline którym przeprowadzone było badanie. Cały pipeline, jak i jego poszczególne elementy uruchamiane były kilka razy z różnymi wartościami wskazanych parametrów. Z kolejnych uruchomień zebrano wiedzę pozwalającą udoskonalić przebieg obliczeń. Dokładny opis obliczeń opisany jest w podsekcjach tej sekcji.



Rysunek 2: Diagram blokowy przedstawiający przebieg badania. Oznaczone na diagramie wartości parametrów to wartości użyte do otrzymania ostatecznych wyników. Zielone kółka stanowią odnośniki, które będą wspominane w dalszej części pracy.

Wybrane gatunki

Wybrano po 1 przedstawicielu z 12 omawianych rodzin. Zakładając przyjęty podział roślin na rodziny, każdy gatunek jest tu traktowany jako reprezentant swojej rodziny.

Sagowiec ...

Miłorząb dwuklapowy

Modrzew europejski

Ephedra sinica

Gnetum ...

welwiczjowate coś ? (9)

Sośnica japońska

jalowiec pospolity

cis pospolity

Araukaria ...

zastrzalinowate coś? (11)

to ostatnie

TODO zdjęcia gatunków

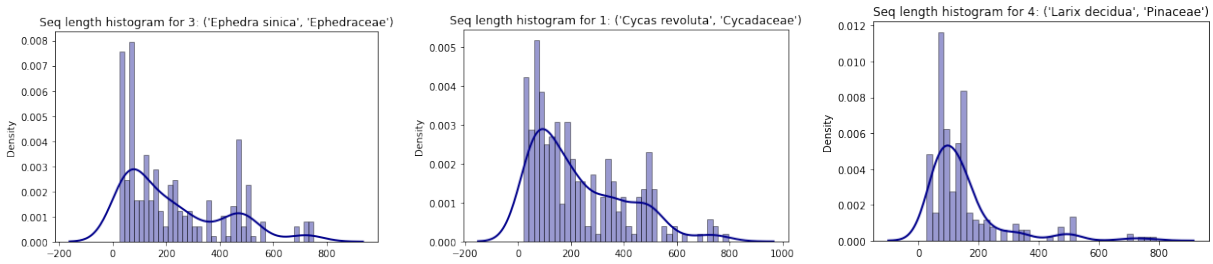
Wybór sekwencji

Z bazy danych NCBI Protein pobrano sekwencje białkowe w formacie FASTA.

Dla każdego gatunków ograniczono liczbę sekwencji użytych do analizy. Ograniczenie przebiegło w dwóch etapach. W pierwszym odfiltorwano sekwencje spoza zakresu długości `[20, 1000]`. Przeprowadzone wcześniej obliczenia wykazały, że na etapie `"*3"` nie powstają klastry zawierające sekwencje o długościach spoza tego przedziału. W drugim wybrano losowo po 250 sekwencji każdego gatunku.

W poniższej tabeli znajduje się liczba pobranych sekwencji dla poszczególnych gatunków:

Lp.	Rodzina	Gatunek	Pobranych sekwencji	Wybranych sekwencji
1	*Cycadaceae*	*Cycas revoluta*	416	250
2	*Ginkgoaceae*	*Ginkgo biloba*	8384	250
3	*Ephedraceae*	*Ephedra sinica*	329	250
4	*Pinaceae*	*Larix decidua*	515	250
5	*Gnetaceae*	*Gnetum gnemon*	517	250
6	*Welwitschiaceae*	*Welwitschia mirabilis*	457	250
7	*Sciadopityaceae*	*Sciadopitys verticillata*	472	250
8	*Cupressaceae*	*Juniperus communis*	381	250
9	*Araucariaceae*	*Araucaria angustifolia*	286	250
10	*Podocarpaceae*	*Nageia nagi*	340	250
11	*Zamiaceae*	*Zamia furfuracea*	383	250
12	*Taxaceae*	*Taxus baccata*	803	250



Rysunek 3: Histogramy i krzywe gęstości długości wybranych sekwencji wewnątrz poszczególnych gatunków (od lewej): Ephedra sinica, Cycas revoluta, Larix decidua. Histogramy pozostałych gatunków bardzo przypominają ten pierwszy.

Klastrowanie sekwencji i wybór klastrów

W kolejnym etapie przeprowadzono klasteryzację zbioru wszystkich wybranych sekwencji z użyciem algorytmu CD-HIT (Li and Godzik 2006). Ustalono następujące wartości parametrów `threshold = 0.6`, `wordsize = 4`. Wybór wartości parametru `threshold` zostanie wyjaśniony w kolejnym paragrafie. Parametr `wordsize` dobierany jest odpowiednio do wartości `threshold`. W efekcie działania algorytmu otrzymano 371 klastrów.

Spośród 371 klastrów, wybrano 29, które zawierały co najmniej jedną sekwencję z każdego gatunku. Żaden z klastrów nie zawierał dokładnie jednej sekwencji z każdego gatunku. Sekwencje wewnątrz klastrów nie zawsze były tej samej długości. Parametr `threshold = 0.6` pozwolił otrzymać większą liczbę wybranych klastrów, niż jego większe wartości, które także były testowane `(0.65, 0.7, 0.75, 0.8, 0.85)`.

Wybrane klastry oczyszczono pozostawiając po 1 sekwencji z każdego gatunku. Pozostawiono pierwszą sekwencję z danego gatunku według kolejności w wynikowych plikach programu CD-HIT.

Dopasowanie sekwencji

Wewnątrz każdego wybranego klastra przeprowadzono dopasowanie sekwencji. Uliniowienie przeprowadzono równolegle z użyciem dwóch metod: Muscle (Edgar 2004) i ClustalW (Thompson, Higgins, and Gibson 1994). Oba algorytmy uruchomiono z domyślnym zestawem parametrów.

Drzewa filogenetyczne

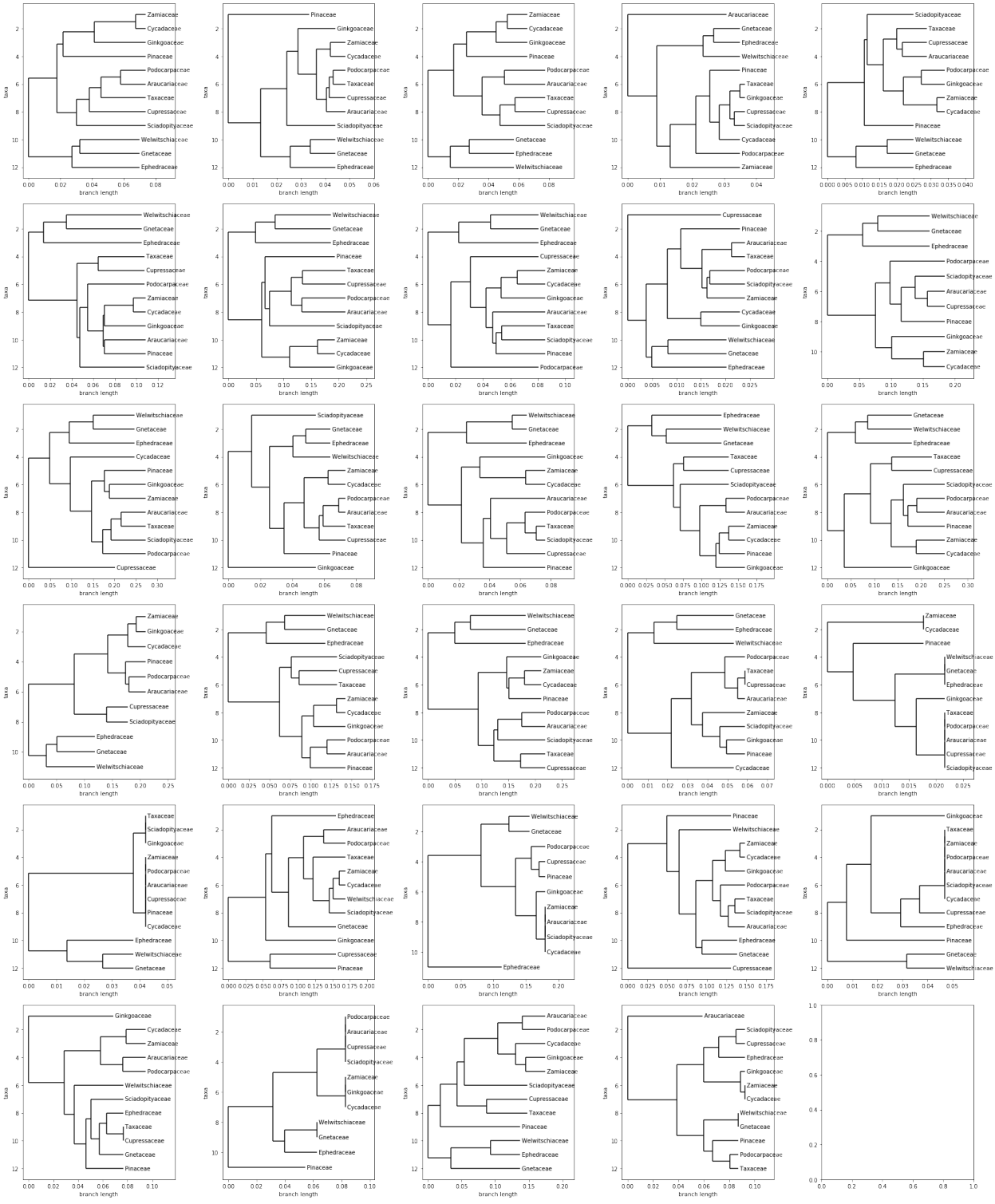
Na podstawie otrzymanych dopasowań, dla każdej kombinacji klastra i algorytmu dopasowania, zbudowano drzewa filogenetyczne. Drzewa zbudowano metodą UPGMA, zatem otrzymane drzewa były ukorzenione. Jako źródło odległości między aminokwasami wykorzystano macierz `BLOSUM62`. Macierz `BLOSUM62` jest domyślną macierzą używaną przez Blast do sekwencji białkowych (Hakeem et al. 2017), była już w przeszłości używana do badań filogenetycznych w obrębie królestwa roślin (Richardt et al. 2007) (Turnaev et al. 2020).

Drzewo konsensusowe

Dla obu algorytmów dopasowania zbudowano z drzew filogenetycznych drzewa konsensusowe. Drzewa konsensusowe zbudowano wykorzytując metody decyzyjne greedy consensus (`cutoff < 0.5`) i majority consensus (`cutoff >= 0.5`). Porównano drzewa dla wartości parametru `cutoff` z zakresu `{0, 0.2, 0.4, 0.6, 0.8}`. Dla większych niż `0.2` parametrów `cutoff` grupy siostrzane które powstawały składały się z dużej liczby członków, co oznaczało, że takie drzewa dostarczały mało informacji. Ostatecznie zdecydowano się zaprezentować drzewa dla `cutoff = {0.2, 0}`.

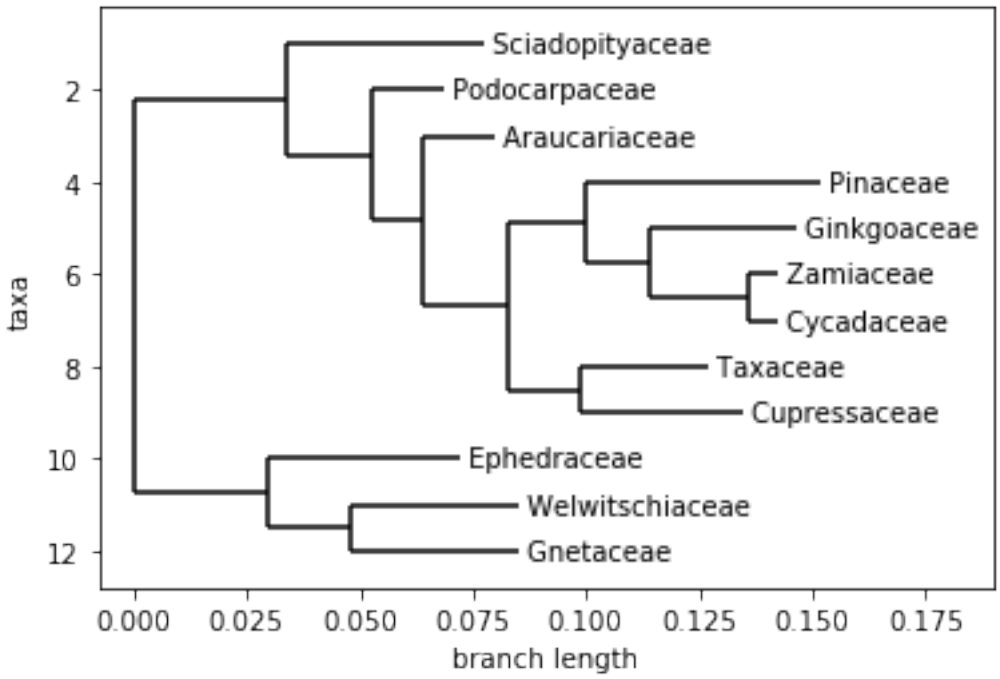
Wyniki

W tej sekcji zostaną zaprezentowane i omówione drzewa drzewa konsensusowe otrzymane przy użyciu obu wybranych metod MSA.



Rysunek 4: Drzewo konsensusowe (cutoff = 0.2) dla gałęzi pipelinu używającej algorytmu MSA Muscle.

ClustalW



Rysunek 5: Drzewo konsensusowe (cutoff = 0.2 i 0 - są identyczne) dla gałęzi pipelinu używającej algorytmu MSA ClustalW.

Muscle

Wnioski i dyskusja

Propozycje usprawnienia zastosowanej metodologii

Kod źródłowy

Kod źródłowy użyty do przeprowadzenia obliczeń wraz z wynikami cześciowymi i wykorzystanymi sekwencjami znajduje się na platformie Github pod adresem: https://github.com/konrad-komisarczyk/Gymnospermae-phylogenetics

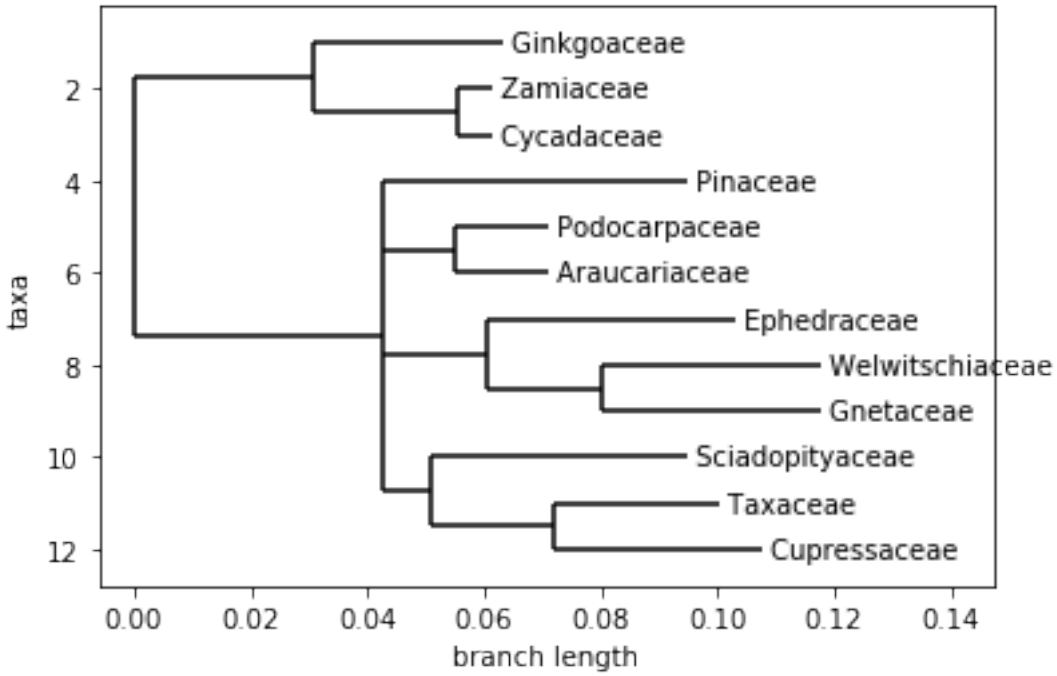
Bibliografia

Christenhusz, Maarten, James Reveal, Aljos Farjon, Martin Gardner, Robert Mill, and Mark Chase. 2010. "A New Classification and Linear Sequence of Extant Gymnosperms." *Nov. Magnolia Press Phytotaxa* 19 (November): 55–70. https://doi.org/10.11646/phytotaxa.19.1.3.

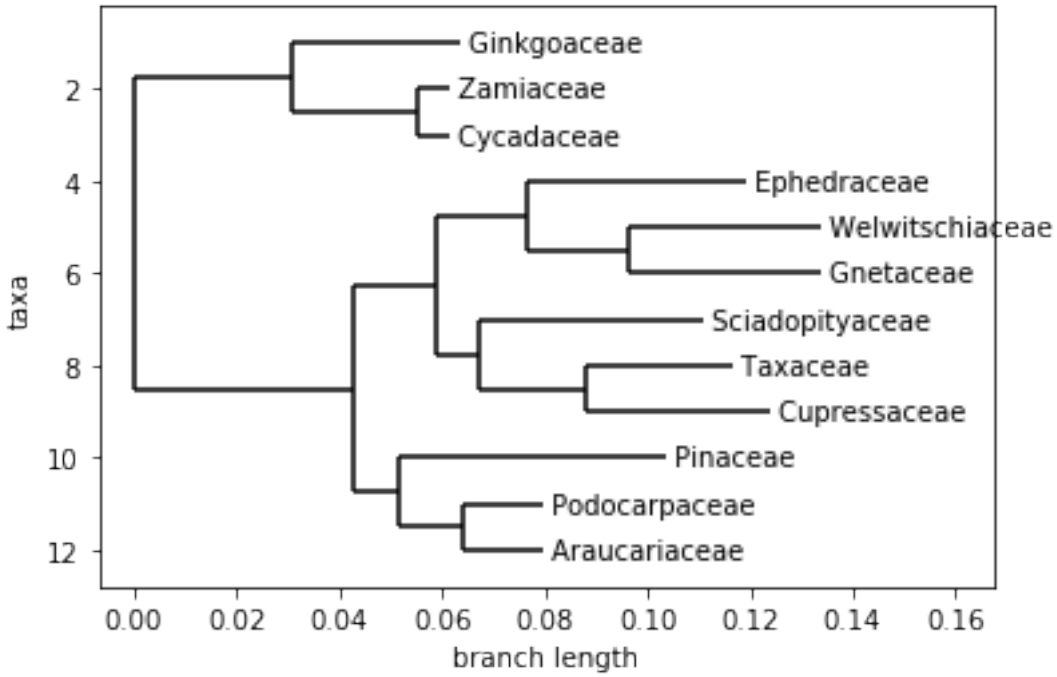
Edgar, Robert C. 2004. "MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity." *BMC Bioinformatics* 5 (1): 113. https://doi.org/10.1186/1471-2105-5-113.

Hakeem, Khalid Rehman, Adeel Malik, Fazilet Vardar-Sukan, and Munir Ozturk. 2017. *Plant Bioinformatics: Decoding the Phyta.* Springer.

Li, W., and A. Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13): 1658–9. https://doi.org/10.1093/bioinformatics/btl158.



Rysunek 6: Drzewo konsensusowe (cutoff = 0.2) dla gałęzi pipelinu używającej algorytmu MSA Muscle.



Rysunek 7: Drzewo konsensusowe (cutoff = 0) dla gałęzi pipelinu używającej algorytmu MSA Muscle.

Lu, Ying, Jin-Hua Ran, Dong-Mei Guo, Zu-Yu Yang, and Xiao-Quan Wang. 2014. "Phylogeny and Divergence Times of Gymnosperms Inferred from Single-Copy Nuclear Genes." Edited by Sven Buerki. *PLoS ONE* 9 (9): e107679. https://doi.org/10.1371/journal.pone.0107679.

Ran, Jin-Hua, Ting-Ting Shen, Ming-Ming Wang, and Xiao-Quan Wang. 2018. "Phylogenomics Resolves the Deep Phylogeny of Seed Plants and Indicates Partial Convergent or Homoplastic Evolution Between Gnetales and Angiosperms." *Proceedings of the Royal Society B: Biological Sciences* 285 (1881): 20181012. https://doi.org/10.1098/rspb.2018.1012.

Richardt, Sandra, Daniel Lang, Ralf Reski, Wolfgang Frank, and Stefan A. Rensing. 2007. "PlanTAPDB, a Phylogeny-Based Resource of Plant Transcription-Associated Proteins." *Plant Physiology* 143 (4): 1452–66. https://doi.org/10.1104/pp.107.095760.

Simpson, Michael G. 2010. *Plant Systematics*. 2nd ed. Burlington, MA: Academic Press.

Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. 1994. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." *Nucleic Acids Research* 22 (22): 4673–80. https://doi.org/10.1093/nar/22.22.4673.

Turnaev, Igor I., Konstantin V. Gunbin, Valentin V. Suslov, Ilya R. Akberdin, Nikolay A. Kolchanov, and Dmitry A. Afonnikov. 2020. "The Phylogeny of Class B Flavoprotein Monooxygenases and the Origin of the Yucca Protein Family." *Plants* 9 (9): 1092. https://doi.org/10.3390/plants9091092.