Wstęp do bioinformatyki - Drzewa filogenetyczne

Michał Stawikowski

10.05.2021

Spis treści

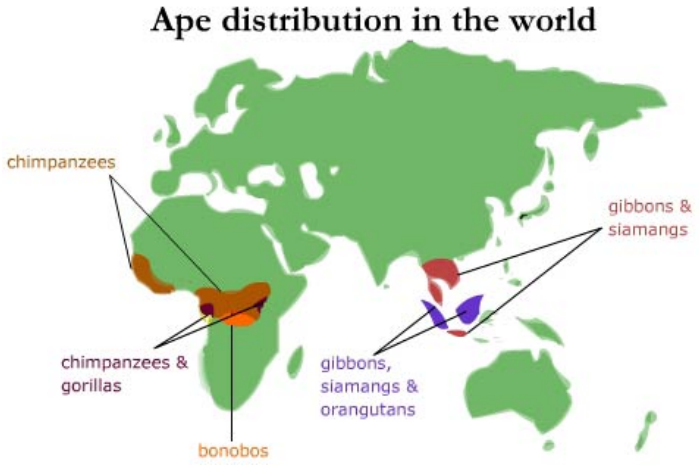
1 Wstęp i opis problemu

1.1 Opis problemu

Celem projektu jest odtworzenie drzewa filogenetycznego dla wybranych, jednych z najpopularniejszych wśród badaczy przedstawicieli naczelnych na podstawie analizy arbitralnie, losowo wybranych sekwencji aminokwasów białek o zbliżonej długości. Wśród badanej grupy organizmów znalazły się:

- Człowiek [*Homo sapiens*]
- Szympans zwyczajny [*Pan troglodytes*]
- Bonobo - szymapns karłowaty [*Pan paniscus*]
- Goryl nizinny [*Gorilla gorilla gorilla*]
- Orangutan sumatrzański [*Pongo abelii*]
- Gibon srebrzysty [*Hylobates moloch*]
- Makak królewski [*Macaca mulatta*]
- Lemurek myszaty [*Microcebus murinus*]

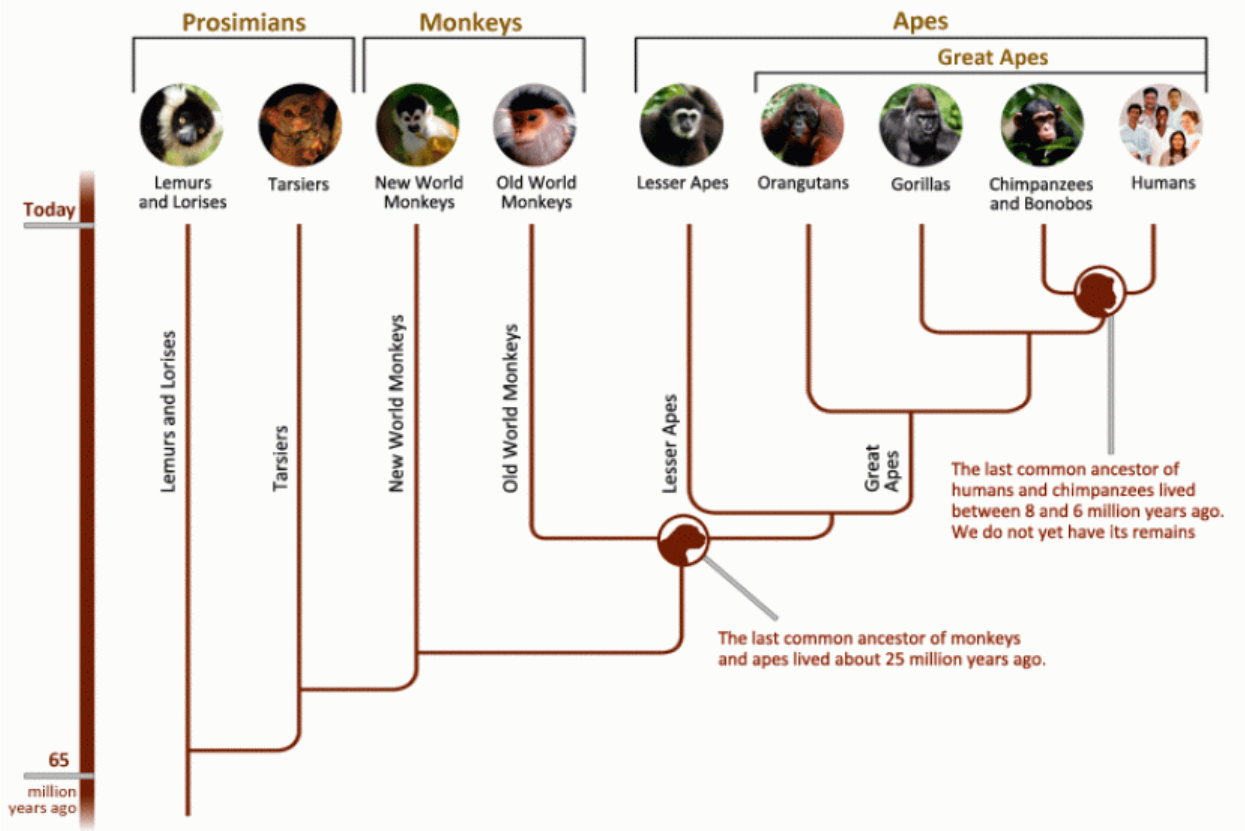
Na wybór powyższych organizmów wpłynęło to, że są one jednymi z najbardziej podobnych zwierząt do człowieka, dzięki czemu porównywanie ich sekwencji białek z ludzkimi mogą dać interesujące i rzeczowe wyniki. Dodatkowo naczelne są jedną z dokładniej zbadanych grup zwierząt, co może zapewnić nam dużą liczbę wysokiej jakości sekwencji białkowych do analizy i łatwą weryfikację wyników. Dodatkowo otrzymane wyniki będzie można porównać z faktycznym rozmieszczeniem organizmów na świecie i zbadać zbieżność tych dwóch analiz - czy to, że zwierzęta żyją w podobnych rejonach geograficznych znaczy, że mają podobne sekwencje białek - mapa 1 (Lemurek myszaty występuje na Madagaskarze, a Makak królewski od Afganistanu po Indie i Chiny, co nie jest zaznaczone na poniższej mapie). Oczywiście do takich badań potrzeba dużej ilości danych i metod bardziej zaawansowanych niż te użyte w tym projekcie, ale zawsze możemy porównać wyniki z aktualnym stanem wiedzy.



Rysunek 1: Rozmieszczenie wybranych gatunków na świecie. Źródło: https://www.sheppardsoftware.com/content/animals/animals/mammals/apevsmonkey.htm

1.2 Oczekiwane wyniki

Dzięki temu, że analizujemy sekwencje białek organizmów pochodzących z tego samego rzędu oraz dobrze zbadanych, nasze końcowe wyniki możemy porównać z wieloma drzewami filogenetycznymi naczelnych, co pozwoli nam zweryfikować poprawność wyników badania. Przykładowe drzewo filogenetyczne zostało zaprezentowane na diagramie 2 poniżej.



Rysunek 2: Przykładowe drzewo filogenetyczne naczelnych. Źródło: https://humanorigins.si.edu/evidence/genetics

Zgodnie z przedstawionym wykresem 2, dwoma najbliższymi człowiekowi gatunkami są Szympans zwyczajny oraz Bonobo. Kolejnymi **człowiekowatymi** gatunkami biorąc pod uwagę bliskość genetyczną są Goryl nizinny oraz Orangutan Sumatrzański. Następnymi w kolejce są **gibbonowate** - Gibon srebrzysty oraz **koczkodanowate**, czyli w naszym przypadku Makak królewski. Najbardziej oddalonymi naczelnymi w naszej analizie jest rodzina **lemurowatych** i nasz Lemurek myszaty. Uwzględniając powyższe informacje, po wynikach, będziemy spodziewać się podobnego podziału i hierarchii podobieństwa. Szczególnie będziemy zwracać uwagę na to, do którego gatunku naczelnych najbliżej będzie miał człowiek. Dodatkowo, możemy przeanalizować to, czy orangutany, którym geograficznie bliżej do np. gibbonów lub Makaka królewskiego znajdą się w naszej analizie sekwencji białek bliżej gorylów. Obserwować będziemy też czy nastąpi znaczący podział na **człekokształtne**, **małpy** (apes and monkeys) oraz **małpiatki**.

2 Wybór oraz analiza danych

2.1 Wybór danych

Zgodnie z wytycznymi rozważyliśmy 8 gatunków zwierząt i dla każdego z nich pobraliśmy zbliżone liczby sekwencji białek - tabela 3. W celu próby znalezienia jak największej liczby **homologów** lub po prostu podobnych białek ograniczyliśmy naszą analizę do sekwencji o długości pomiędzy 1500 a 1600 aminokwasów białkowych. Taki zakres zapewnił też wystarczającą liczbę sekwencji do analizy dla każdego z rozważanych naczelnych. Maksymalną liczbę sekwencji dla gatunku ograniczyliśmy do 1000, aby nie stworzyć dużej dysproporcji - rozważone zostały także 2000 i 3000 sekwencji dla gatunku, jednak prowadziły one do mniej zrównoważonych klastrów i gorszych wyników. W sumie analizie poddane zostało prawie 7000 sekwencji. Dane pobrane zostały przy użyciu białkowych baz danych udostępnionych przez National Center for Biotechnology Information **NCBI**. Wszystkie operacje na danych od pobierania i filtrowanie sekwencji po końcowe uliniowienie, klastrowanie i tworzenie drzew wykonane zostały przy użyciu pakietu dla języka Python - **Biopython** i jego wybranych modułów oraz później wymienionych narzędzi i algorytmów, także z poziomu języka Python.

```
1 download_seqs(species, labels)
```

```
Found 1000 sequences for Human.
Found 1000 sequences for Chimpanzee.
Found 756 sequences for Bonobo.
Found 414 sequences for Gorilla.
Found 751 sequences for Orangutan.
Found 1000 sequences for Gibbon.
Found 1000 sequences for Rhesus.
Found 815 sequences for Lemur.
Total 6736 records.
```

Rysunek 3: Liczba pobranych sekwencji z uproszczonymi nazwami gatunków

2.2 Klastrowanie i wybór klastrów

Używając danych opisanych wcześniej, w celu zmniejszenia złożoności obliczeniowej późniejszych kroków przeprowadziliśmy klastrowanie przy użyciu **UCLUST**, czyli metody klastrowania, która wykorzystuje USEARCH do przypisywania sekwencji do klastrów. UCLUST oferuje kilka zalet w porównaniu z szeroko stosowanym programem CD-HIT, w tym większą szybkość, mniejsze zużycie pamięci, lepszą czułość, klastrowanie przy niższych tożsamościach i klasyfikację znacznie większych zbiorów danych. Konkretnie użyta została komenda *cluster_fast* używana dla sekwencji w formacie FASTA lub FASTQ przy użyciu wariantu algorytmu UCLUST zaprojektowanego w celu maksymalizacji szybkości. Parametr progu tożsamości (identity treshold) id został ustawiony na 0.9 [1]. Pełna komenda:

```
usearch -cluster_fast {path_in}/All.fasta -id {iden} -clusters {path_out}/Clustering/c_
```

Powstałe klastry zapisane zostały do późniejszej analizy i edycji - tabela 4 - w celu wybrania klastrów, które posłużą do stworzenia drzew filogoenetycznych po wcześniejszym ich uliniowieniu (alignment).

Z tak uzyskanych klastrów następnie wybrane zostały te, które zawierały po jednym przedstawicielu dla każdego badanego gatunku, a następnie zawartość klastrów przefiltrowano tak, aby każdy zawierał tylko pierwszą znalezioną sekwencje dla każdego naczelnego:

```
[ ]: 1 cluster(iden=0.9)
Found 853 clusters.

[ ]: 1 choose_clusters()
Found 54 proper clusters.
```

Rysunek 4: Utworzone oraz wybrane klastry

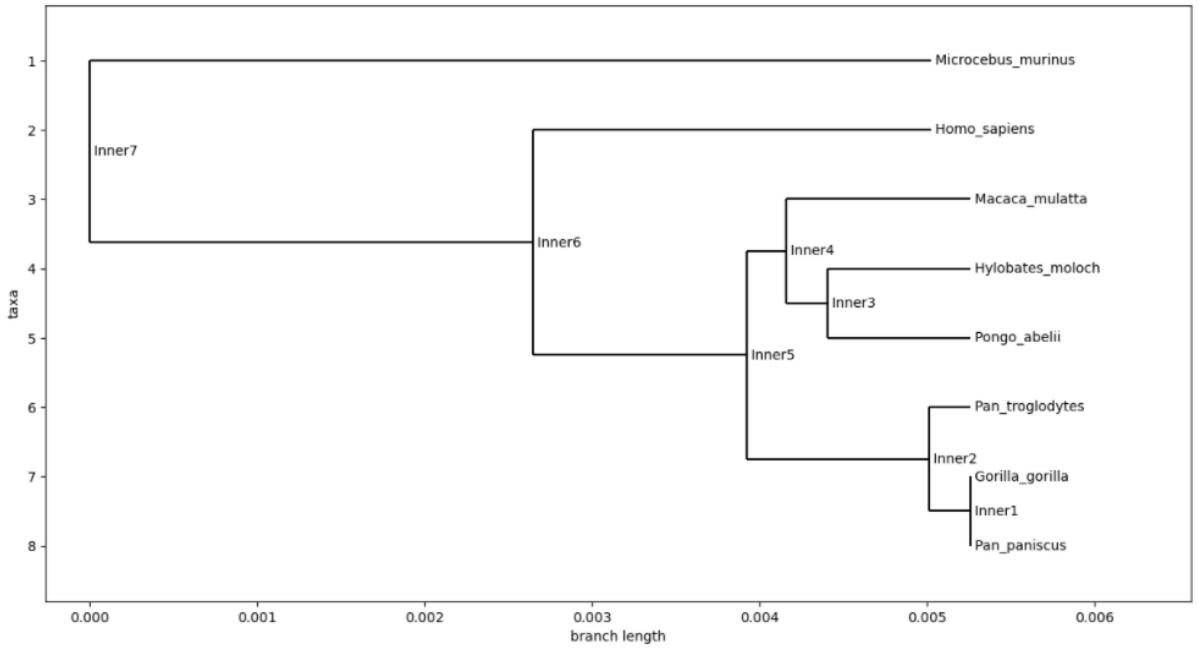
2.3 Uliniowienie sekwencji oraz budowa pojedynczych drzew

W celu uliniowienia znalezionych sekwencji użyliśmy wierszowego narzędzia poleceń dla programu do wielokrotnego uliniowienia **MUSCLE**. MUSCLE to skrót od MUltiple Sequence Comparison by Log- Expectation. Uważa się, że MUSCLE osiąga zarówno lepszą średnią skuteczność, jak i lepszą szybkość działania niż ClustalW2 lub T-Coffee, w zależności od wybranych opcji [2]. Dla wybranych sekwencji, dla wszystkich klastrów proces trwał około 90 sekund.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Rysunek 5: Przykład macierzy BLOSUM

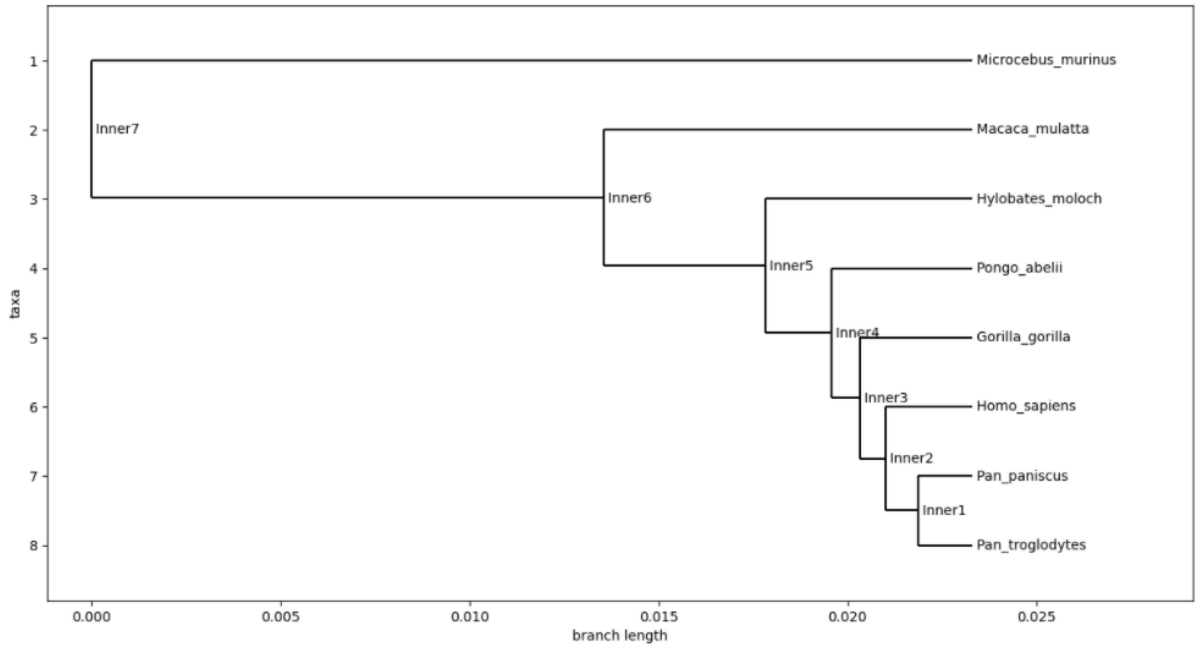
Po wykonaniu uliniowień następnym krokiem było zbudowanie drzew filogenetycznych dla każdego pozostałego klastra. Drzewa budowane były przy użyciu macierzy punktowania **blosum62**. Macierz BLOSUM (BLOcks SUbstitution Matrix) to macierz substytucyjna stosowana do dopasowania sekwencji białek. Macierze BLOSUM - figura 5 - są używane do oceny dopasowań między ewolucyjnie rozbieżnymi sekwencjami białek. Opierają się na lokalnych uliniowieniach. Macierze BLOSUM powstają poprzez analizę bazy danych BLOCKS pod kątem bardzo konserwatywnych regionów rodzin białek (które nie mają luk w dopasowaniu sekwencji), a następnie liczone są częstości aminokwasów i prawdopodobieństwa ich substytucji [3]. Następnie na podstawie tak powstałych macierzy zbudowane zostały drzewa przy użyciu algorytmu **UPGMA** (unweighted pair group method with arithmetic mean), czyli prostej metody hierarchicznego grupowania aglomeracyjnego (oddolnego). Algorytm UPGMA konstruuje zakorzenione drzewo (dendrogram), które odzwierciedla strukturę obecną w macierzy podobieństwa par (lub macierzy niepodobieństwa). Na każdym kroku dwa najbliższe klastry są łączone w klaster wyższego poziomu [4]. Cały proces przebiegał około 35 sekund. Tak powstałe drzewa prezentowały różne podziały, niektóre były bardzo rozbieżne z oczekiwanymi, a niektóre bardzo do nich zbliżone: diagramy 6 i 7 (Przedstawiono skrócone nazwy w etykietach gatunków). W trakcie analizy przetestowaliśmy różne rodzaje macierzy punktowania (macierze z rodziny PAM) i metod budowy drzewa (np. NJ), jednak te wymienione powyżej poradziły sobie najlepiej.



Rysunek 6: Drzewo o rozbieżnej strukturze z oczekiwanymi wynikami

Niektóre powstałe drzewa różniły się bardzo od oczekiwanych wyników. Na wykresie 6 widzimy, że nawet Szympans zwyczajny i Szympans karłowaty są rozdzieleni węzłem z Gorylem. Pozycja człowieka na tym drzewie to kolejna zagadka, zdecydowanie za daleko mu do naszych najbliższych krewnych - szympansów. W tym miejscu warto podkreślić, że drzewo filogenetyczne to diagram przedstawiający możliwe ewolucyjne relacje między organizmami i są to hipotezy, a nie ostateczne fakty. Wzór rozgałęzień w drzewie filogenetycznym odzwierciedla sposób, w jaki gatunki lub inne grupy wyewoluowały z szeregu wspólnych przodków. W przypadku drzew dwa gatunki są bardziej spokrewnione, jeśli mają nowszego wspólnego przodka, a mniej spokrewnione, jeśli mają dawniejszego wspól-

nego przodka, a przynajmniej to jeden ze sposób na interpretacje podobieństw sekwencji pomiędzy różnymi gatunkami organizmów [5].



Rysunek 7: Drzewo o zbliżonym podziale do drzewa oczekiwanego

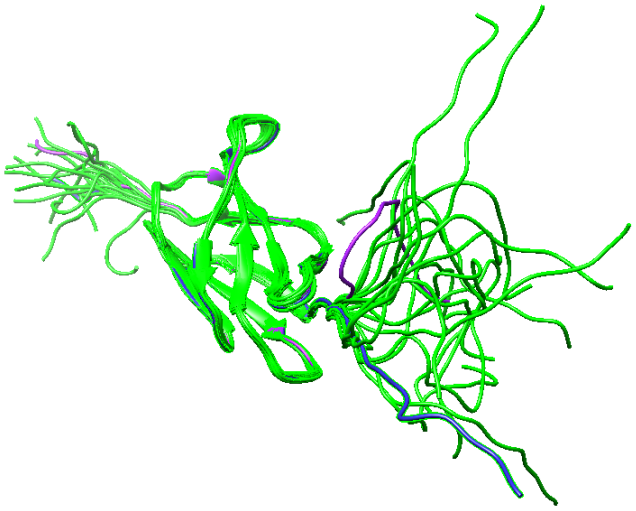
Z kolei na drugim drzewie 7 widzimy prawie dokładne odwzorowanie oczekiwanych wyników z dokładnością do długości poszczególnych gałęzi. Takich drzew na szczęście powstało znacznie więcej podczas naszej analizy, co może wskazywać na to, że nasze finalne drzewo konsensusowe, będzie w jakimś stopniu zbliżone do drzew budowanych na podstawie aktualnej wiedzy. Wskazuje też to na to, że niektóre sekwencje białek prawdopodobnie zostały lepiej poklastrowane, lub po prostu lepiej nadawały się do zadania budowania drzewa filogenetycznego z wielu różnych względów. Powyższy diagram powstał na przykład z analizy sekwencji izoform wariantów białka (Nuclear factor of activated T-cells 5 - rys. 8) kodowanego przez gen **NF-AT5** dla wszystkich naczelnych (https://www.uniprot.org/uniprot/O94916 - przykład dla człowieka). Białko kodowane przez gen NFAT5 odgrywa rolę w układzie odpornościowym i patogenezie chorób autoimmunologicznych [6].

Rysunek 8: Struktura białka człowieka kodowanego przez NF-AT5 uzyskana przy użyciu programu Chimera

Nagłówki plików FASTA sekwencji białek dla drzewa 7 (zbieżnego):

- NP_006590.1 nuclear factor of activated T-cells 5 isoform c [*Homo sapiens*]
- XP_016785574.1 nuclear factor of activated T-cells 5 isoform X3 [*Pan troglodytes*]
- XP_003814589.2 nuclear factor of activated T-cells 5 isoform X3 [*Pan paniscus*]
- XP_030858415.1 NFAT5 isoform X3 [*Gorilla gorilla gorilla*]
- PNJ61558.1 NFAT5 isoform 5 [*Pongo abelii*]
- XP_032013674.1 nuclear factor of activated T-cells 5 isoform X1 [*Hylobates moloch*]
- AFH30855.1 nuclear factor of activated T-cells 5 isoform c [*Macaca mulatta*]
- XP_020138201.1 NFAT5 X8 [*Microcebus murinus*]

Co ciekawe drzewo 6 także zostało zbudowane na podstawie jednego wariantu białka, a konkretnie fosfatazy tyrozynowo-białkowej typu receptora delta - rys. 9, która jest enzymem kodowanym u ludzi przez gen **PTPRD** (https://www.uniprot.org/uniprot/P23468 - przykład dla człowieka). Białko kodowane przez ten gen należy do rodziny białek fosfatazy tyrozynowej (PTP). Wiadomo, że PTP są enzymami, które regulują różne procesy komórkowe, w tym wzrost komórek, różnicowanie, cykl mitotyczny i transformację onkogenną [7]. Łatwo można ocenić jakość klastrowania po tym, że większość klastrów zawiera właśnie jeden wariant białka dla każdego naczelnego, co może wskazywać na duże podobieństwo w ramach jednej takiej grupy.



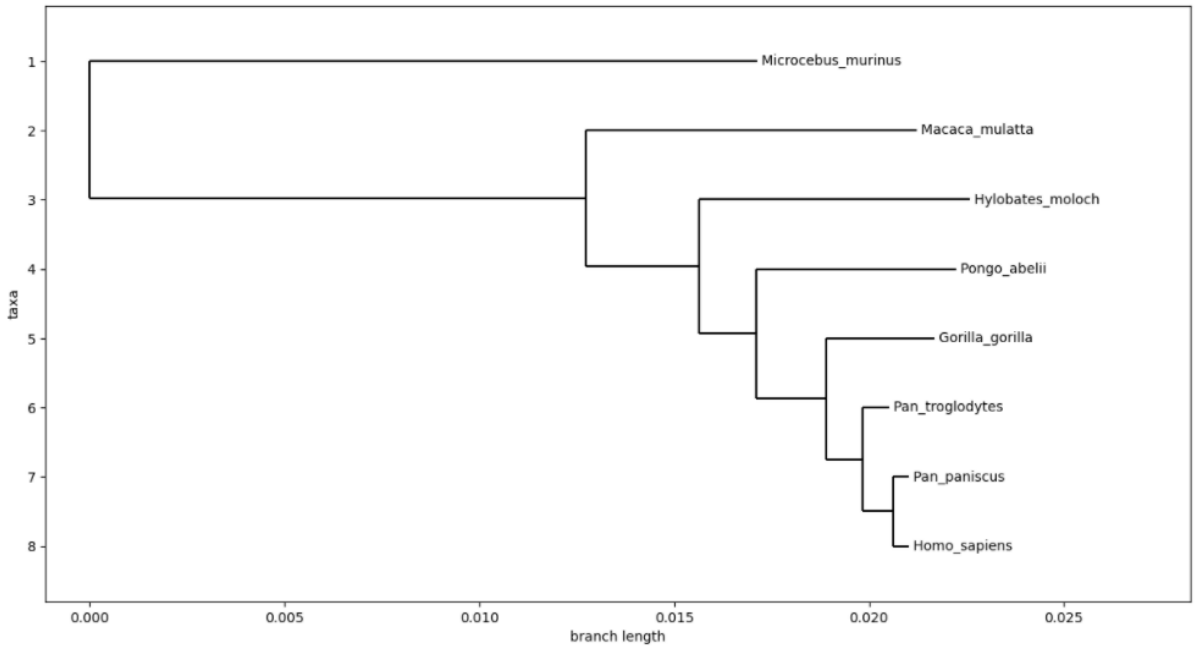
Rysunek 9: Struktura białka człowieka kodowanego przez PTPRD uzyskana przy użyciu programu Chimera

Nagłówki plików FASTA sekwencji białek dla drzewa 6 (rozbieżnego):

- XP_016870480.1 R-PTP-delta isoform X36 [*Homo sapiens*]
- XP_016817838.2 R-PTP-delta isoform X35 [*Pan troglodytes*]
- XP_034785632.1 R-PTP-delta isoform X38 [*Pan paniscus*]
- XP_030870265.1 R-PTP-delta isoform X30 [*Gorilla gorilla gorilla*]
- PNJ81640.1 PTPRD isoform 8 [*Pongo abelii*]
- XP_031995050.1 R-PTP-delta isoform X33 [*Hylobates moloch*]
- XP_028690288.1 R-PTP-delta isoform X32 [*Macaca mulatta*]
- XP_020144775.1 R-PTP-delta isoform X41 [*Microcebus murinus*]

2.4 Drzewo konsensusowe

Wygenerowane wcześniej drzewa posłużyły nam do stworzenia drzewa konsensusowego, które podsumuje nasze dotychczasowe analizy na jednym diagramie. Istnieją różne metody obliczania drzewa konsensusowego dla danego zbioru drzew filogenetycznych. Najbardziej znanymi typami są strict consensus tree, majority consensus tree i extended majority consensus tree. Strict consensus tree zawiera tylko krawędzie, które są wspólne dla wszystkich drzew wejściowych. Drzewa majority consensus zawierają krawędzie, które są obecne w ponad 50% drzew wejściowych, chociaż można również wziąć pod uwagę wyższe wartości procentowe. Zgodnie z regułą extended majority consensus drzewo konsensusu obejmuje wszystkie krawędzie większości, do których stopniowo dodawane są zgodne krawędzie resztkowe, zaczynając od najczęściej występujących. Drzewa majority consensus są jednymi z najczęściej używanymi drzewami konsensusowymi w biologii ewolucyjnej [8]. W naszej analizie wybraliśmy właśnie **majority consensus tree** do podsumowania wyników.



Rysunek 10: Drzewo konsensusowe zbudowane na podstawie powstałych klastrów

Na diagramie 10 widać, że może poza zmianą położenia szympansów i człowieka, wiele poprawek nie trzeba, by to drzewo filogenetyczne wyglądało podobnie do oczekiwanego. Zgodnie z oczekiwaniami daleką odległość pomiędzy Lemurkiem myszatym, a resztą naczelnych można zinterpretować jako obecność bardzo dawno żyjącego wspólnego przodka dla lemurowatych i reszty zwierząt. Dodatkowo widzimy bardzo zbliżone pozycje szympansów i człowieka, co wskazuje na stosunkowo niedawno żyjącego przodka i rzeczywiście tak jest. Pozycja orangutana także wskazuje na bliskie pokrewieństwo z gorylem i szympansami pomimo, aktualnie innych miejsc występowania tych gatunków. Widoczne jest też wydzielenie się małpiatek - Lemurek myszaty.

3 Dyskusja wyników i podsumowanie

3.1 Analiza wyników

Dzięki powyższej analizie uzyskaliśmy dość zbliżone wyniki do oczekiwanych, co może wskazywać, że metodologia naszego badania była w jakimś stopniu sensowna. Duży wpływ na jakość wyników miała liczba rozważanych danych - zwiększenie liczby analizowanych sekwencji białkowych zdecydowanie pozytywnie wpływało na otrzymywane wyniki, ważne było też unikanie dysproporcji w danych ze względu na poszczególne gatunki. Innym znaczącym aspektem był wybór poszczególnych metod analizy danych - wpływ na końcowy wynik miały na przykład użyte macierze punktowania i algorytmy budowy drzew oraz drzew konsensusowych, a także metoda klastrowania i jej parametry. Dzięki wyżej wymienionym działaniom nasze finalne drzewo nie odbiegało znacząco od przyjętego za odniesienie drzewa zbudowanego przy użyciu wielu różnych bardziej zaawansowanych metod użytych na podstawie aktualnej ludzkiej wiedzy. Prawdopodobnie gdybyśmy rozważyli większą ilość wysokiej jakości danych, otrzymalibyśmy jeszcze bliższe wyniki.

3.2 Podsumowanie

Aby zbudować dokładne, najbardziej prawdopodobne drzewa, biolodzy często używają wielu różnych cech i metod (zmniejszając prawdopodobieństwo, że jakikolwiek niedoskonały fragment danych może doprowadzić do niewłaściwych wniosków). Mimo to drzewa filogenetyczne są hipotezami, a nie ostatecznymi odpowiedziami, i mogą być tak dobre, jak dane, których użyliśmy. Drzewa są z czasem aktualizowane, gdy nowe dane stają się dostępne i można je dodać do analizy. Jest to szczególnie prawdziwe dzisiaj, ponieważ nowe metody zwiększają naszą zdolność do porównywania genów czy właśnie sekwencji białek między gatunkami [5].

(a) Lemurek myszaty. Źródło: http://najdziwniejsze-zwierzeta-na-planecie.com

(b) Lemurek myszaty - grafika z filmu animowanego DreamWorks Animation - ”Madagascar”

4 Bibliografia

Literatura

[1] https://drive5.com/usearch/manual/uclust_algo.html

[2] https://www.ebi.ac.uk/Tools/msa/muscle/

[3] http://www.cs.tau.ac.il/ rshamir/algmb/98/scribe/html/lec03/node10.html

[4] https://en.wikipedia.org/wiki/UPGMA#: :text=UPGMA

[5] https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees

[6] https://www.frontiersin.org/articles/10.3389/fimmu.2019.00270/full

[7] https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTPRD

[8] Tahiri, N., Willems, M. & Makarenkov, V. A new fast method for inferring multiple consensus trees using k-medoids. BMC Evol Biol