Filogenetyczny pipeline do analizy Bifidobacterium

Ada Hryniewicka

26 stycznia 2023

1 Wstęp teoretyczny

Bifidobakterie znajdują się w ludzkim przewodzie żołądkowo-jelitowym i promują ustanowienie zdrowego konsorcjum drobnoustrojów w jelitach niemowląt. Podstawowe pytania dotyczące występowania, różnorodności i funkcji różnych gatunków Bifidobacterium w ludzkim przewodzie pokarmowym wymagają umiejętności dokładnego i spójnego przypisania filogenezy. W literaturze obecne są różnorakie podejścia do klasyfikacji tych organizmów. Przedstawione są analizy na podstawie 16S rRNA [1] [2], która niestety może mieć problem z reprodukowalnością wyników ze względu na nowo rozbieżne gatunki, które przeszły intensywną presję ewolucyjną i mogą posiadać bardzo podobne sekwencje genów 16S rRNA, które mogą ignorować szeroką lukę filogenetyczną między tymi taksonami. Podobna sytuacja jest przy wykorzystaniu metod opierających się na hybrydyzacji DNA-DNA. Rozwiązaniem tego problemu może być podejście używające analizy rdzeniowego genomu (core-genome) i genomika porównawcza. Jest to jednak wymagający obliczeniowo proces. Natomiast wyniki autorów w publikacji [1] świadczą o skuteczności analizy białek związanych z glikolizą. Takie podejście zapewniło poprawne wyniki znacznie obniżając koszt obliczeniowy. Celem tego projektu jest sprawdzenie jak użycie całych proteomów dla różnych gatunków Bifidobakterii wpływa na poprawne określenie ich filogenezy.

2 Metodyka

2.1 Dane

Dane do analizy proteomów Bifidobakterii pobrane były z bazy Uniprot.org. Dostępnych jest tam 47 referencyjnych genomów. Pierwszym krokiem było skorzystanie ze wszystkich dostępnych danych jednak dawało to bardzo niską wartość klastrów 1-1. Otrzymałam 34 klastry 1-1 na 23 tysiące wszystkich. Ze względu na niewystarczającą ilość do dalszej analizy postanowiłam zredukować liczbę wykorzystanych organizmów do takich jak występują również w publikacji [1], aby zapewnić porównanie rezultatów. Wybrałam 23 organizmy wspólne i po klastrowaniu miałam 258 klastrów 1-1 na 11 tysięcy. Ilość białek zawartych

w poszczególnych proteomach była w przedziale 1500-2500. Następnym etapem było wykonanie uliniowienia i zbudowanie nieukorzeninych drzew metodą Neighbour Joining (NJ). Otrzymany plik w formacie .newick był wykorzystany do zbudowania superdrzew oraz nieukorzenionych drzew konsensusu dla p= 0.5. Program wykonujący superdrzewa przyjmował drzewa nieukorzenione natomiast zwracał ukorzenione. Biorąc to pod uwagę drzewo do porównania było odkorzenione. Dodatkowymi krokami był bootstrap przy tworzeniu drzew NJ oraz stworzenie superdrzewa dla paralogów, a następnie analiza otrzymanych wyników. Ilość otrzymanych drzew dla paralogów wyniosła 500, wykorzystano wszystkie klastry z ilością zawartych białek powyżej 3. Drzewa porównałam z drzewem znajdującym się w publikacji z przepisanymi tylko wspólnymi organizmami. Miarą porównawczą była znormalizowana odległość Robinsona-Fouldsa (RF).

2.2 Narzędzia

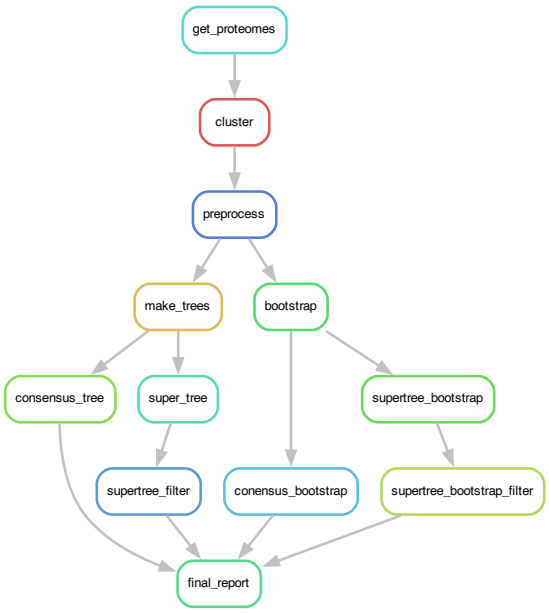
Do pobrania danych wykorzystywałam skrypty napisane w języku Python oraz API z bazy Uniport.org. Dodatkowe operacje na obróbce danych takich jak segregowanie klastrów i dostosowanie nazw wykorzystywałam Python. Jeśli chodzi o klastrowanie użyłam programu mmseqs. Wszystkie obliczenia związane ściśle z filogenetyką takie jak uliniowienie, budowa drzew NJ oraz bootsrap wykonywane były w skryptach w języku R z wykorzystaniem pakietów mse, ape, phytools. Drzewo konsensusu wykonywane było dla p= 0.5. Superdzrzewa budowane były za pomocą programu fasturec. Do zintegtowania całego pipeline'u użyłam snakemake'a, gdzie w pliku Snakefile zawarte są polecenia dotyczące uruchamiania skryptów.

2.3 Przepływ pracy pipeline'u

Pierwszym krokiem jest pobranie danych (get proteomes), które zapisują wszystkie białka do jednego pliku .fasta, który następnie jest wykorzystywany przez mmseqs (cluster). Po klastrowaniu dane były odpowiednio przygotowane przez wybranie klastrów 1-1 w jednym folderze oraz wszystkim klastrami w oddzielnych plikach w drugim folderze dla późniejszego obliczania drzewa paralogów. Następnie wykonywane są podstawowe drzewa NJ. Skrypty make trees i bootsrap (dla 100 powtórzeń) odpowiednio dbają o wykonanie uliniowienia, zbudowanie drzew i zapisanie ich do plików. Skrypt zawierający bootstap dodatkowo filtruje i odrzuca drzewa słabo wspierane ze średnią wartością $< 70\%$. Ilość drzew spełniających ten restrykcyjny warunek to 27. Schemat działania pipeline'u przedstawiono na rys. 1.

3 Wyniki

Pierwszym porównaniem było sprawdzenie drzewa z bazy timetree.org przedstawione na rys.2. Odległość RF jest duża, mówiąca o silnej niezgodności drzew. Jednak trzeba wziąć pod uwagę wybór danych używanych w drzewach timetree.



Rysunek 1: Graf pracy wygenerowany przez snakemake.

Jak wspominane zostało na wstępnie badanie drzew na podstawie 16S rRNA niosą pewne problemu w przypadku przypisania filogenezy Bifidobakterii, a których wyniki mogły być przekazane przez timetree. Z tego powodu do porównania wyników pracy pipeline'u korzystano tylko z drzewa z publikacji referencyjnej.[1]

Dla analizy klastrów 1-1 wykonałam drzewa konsensusu rys.3 oraz superdrzewo rys.4. Istotnym etapem w klasyfikacji organizmów było nadanie koloru dla organizmów z tej samej grupy wyznaczonej przez analizę rdzeniowego genomu i traktowaną jako 'ground truth' dla tego eksperymentu. Podział był następujący:
-Bifidobacterium longum group (czerwony)
-Bifidobacterium adolescentis group (niebieski)
-Bifidobacterium pseudolongum group (zielony)
-Bifidobacterium pollurom group (różowy)
-Bifidobacterium boum (fioletowy)

Istotnie najmniejsza wartość RF jest otrzymana dla drzewa konsensusu dla ortologów, równa 0.4, ponieważ drzewo referencyjne również prezentuje drzewo konsensusu.

Kolejnym etapem było przygotowanie analogicznych drzew z wykorzystaniem bootstrapu. Dla drzewa konsensusowego odległość RF zwiększyła się z 0.4 do 0.5. Natomiast jeśli chodzi o superdrzewo widoczna jest poprawa z 0.57 do 0.47.Konsensus przedstawiony jest na rys.5, a superdrzewo na rys.6.

Kolejnym etapem było przygotowanie drzewa dla paralogów, zaprezentowanym na rys.7. W tym wypadku wartość RF=0.42.

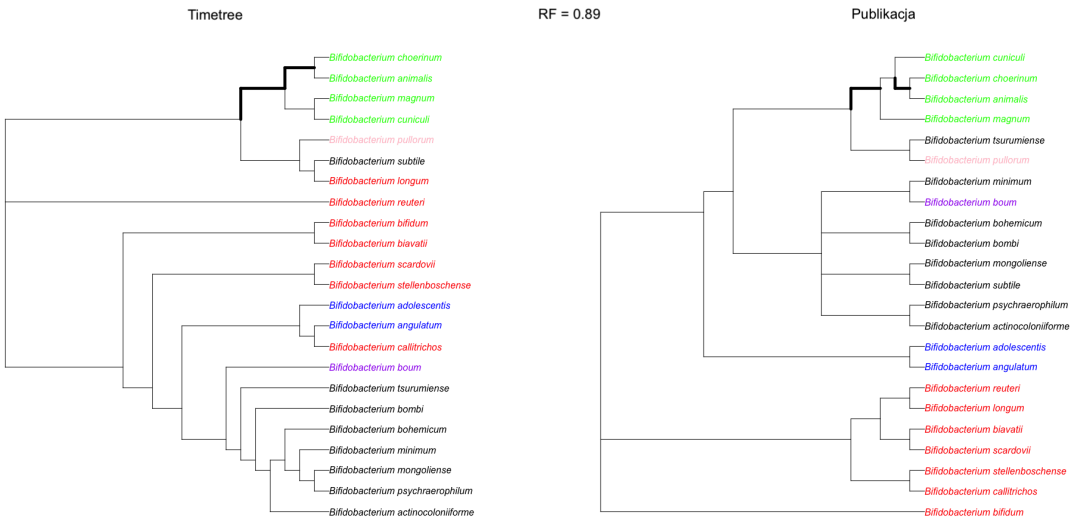
4 Podsumowanie

Wykonaną procedurę tworzenia drzew filogenetycznych można uznać za poprawną, ponieważ większość organizmów została odpowiednio sklasyfikowana, a wyniki mają odległość RF < 0.5. Występujące różnice mogą wiązać się różnicami w wyborze metod wykonywania poszczególnych kroków jedną z nich był sposób budowania drzew, ja użyłam NJ, a autorzy ML. Co implikuje, że budowanie drzew NJ może wpływać na pogorszenie wyników. Autorzy publikacji [1] uwzględnili wiedzę na temat tradycyjnego biochemicznego szlaku glikolizy. Wykorzystanie całych proteomów dało w najlepszym przypadku podobieństwo mierzone w znormalizowanym RF na poziomie 0.4. Dalsze udoskonalanie metod analiz całych proteomów może przynieść lepszą dokładność tak samo jak w przypadku zaprezentowanym w artykule [1], gdzie odpowiedni wybór zestawu białek prowadził do poprawnej klasyfikacji filogenetycznej nowych szczepów lub gatunków w krótszym czasie i przy mniejszej ilości danych niż genom rdzeniowy. Takie podejście zapewnia wysoką rozdzielczość, niską przepustowość, przystępną cenę i dokładność.

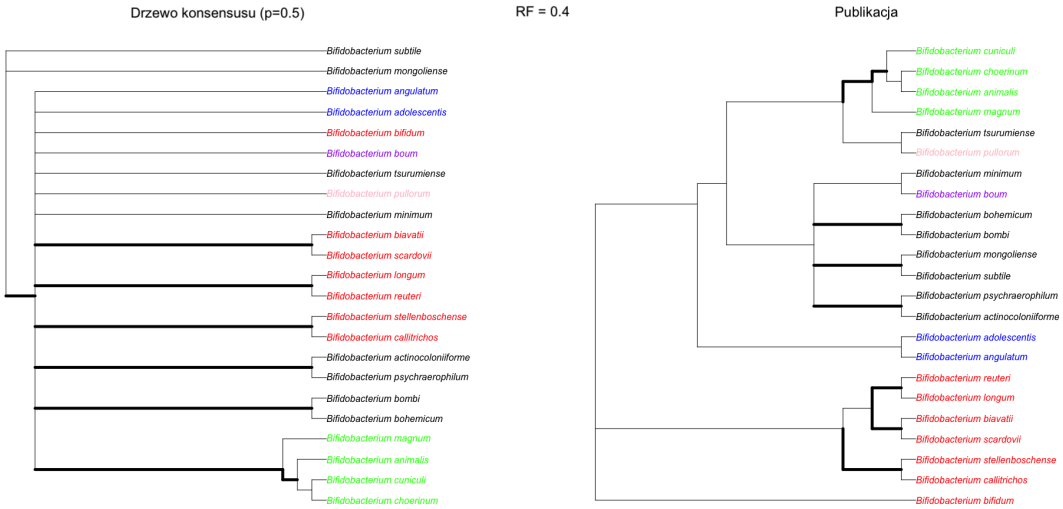
Literatura

[1] K. Brandt and R. Barrangou. Phylogenetic analysis of the bifidobacterium genus using glycolysis enzyme sequences. *Frontiers in Microbiology*, 7, 2016.

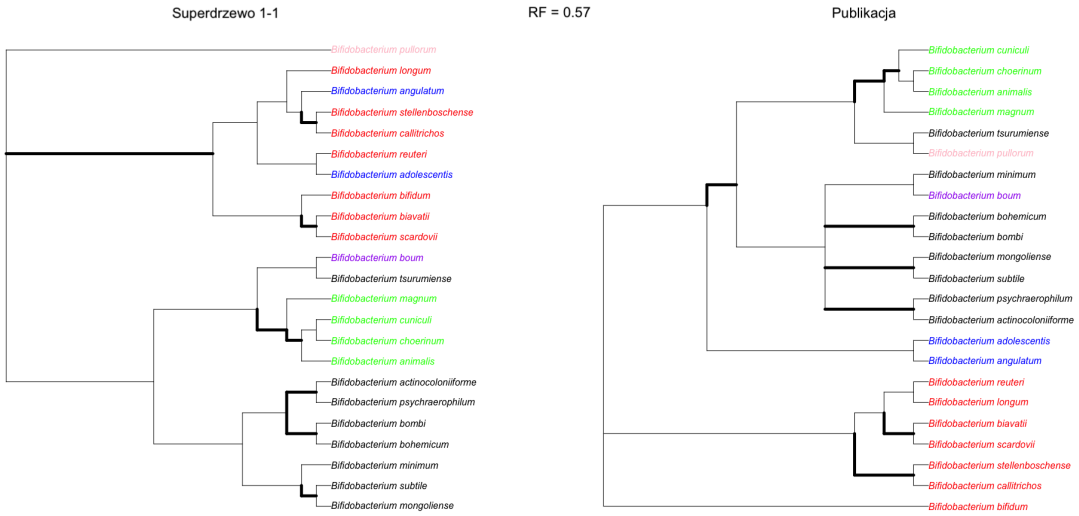
[2] G. A. Lugli, C. Milani, S. Duranti, L. Mancabelli, M. Mangifesta, F. Turroni, A. Viappiani, D. van Sinderen, and M. Ventura. Tracking the taxonomy of the genus bifidobacterium based on a phylogenomic approach. *Applied and Environmental Microbiology*, 84(4):e02249–17, 2018.



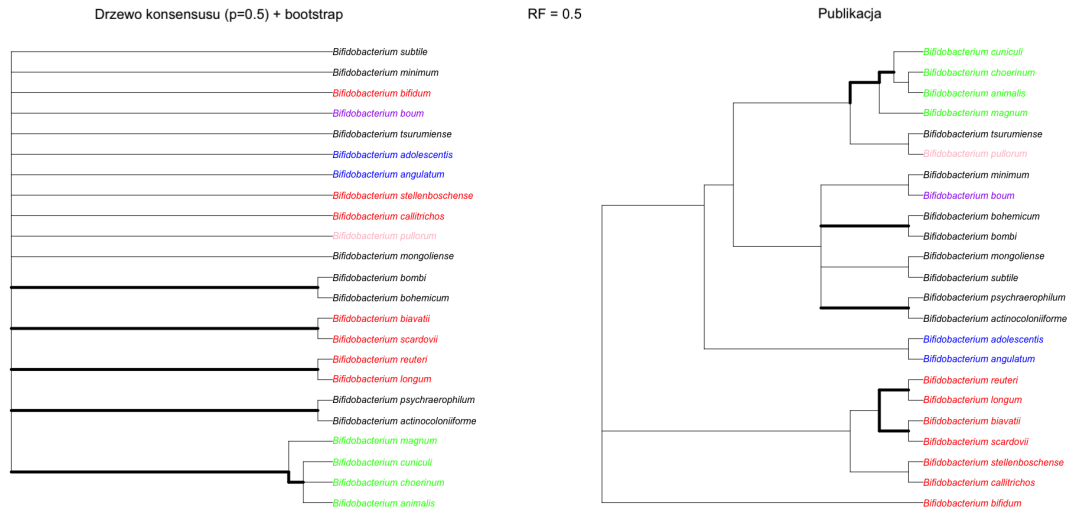
Rysunek 2: Drzewo z timetree.org i publikacji [1]



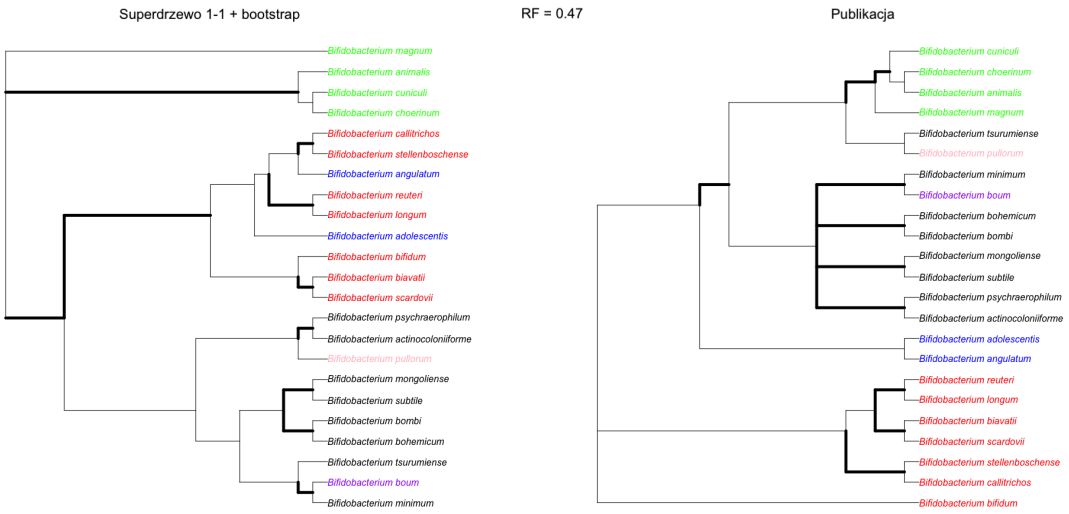
Rysunek 3: Drzewo konsensusu klastrów 1-1.



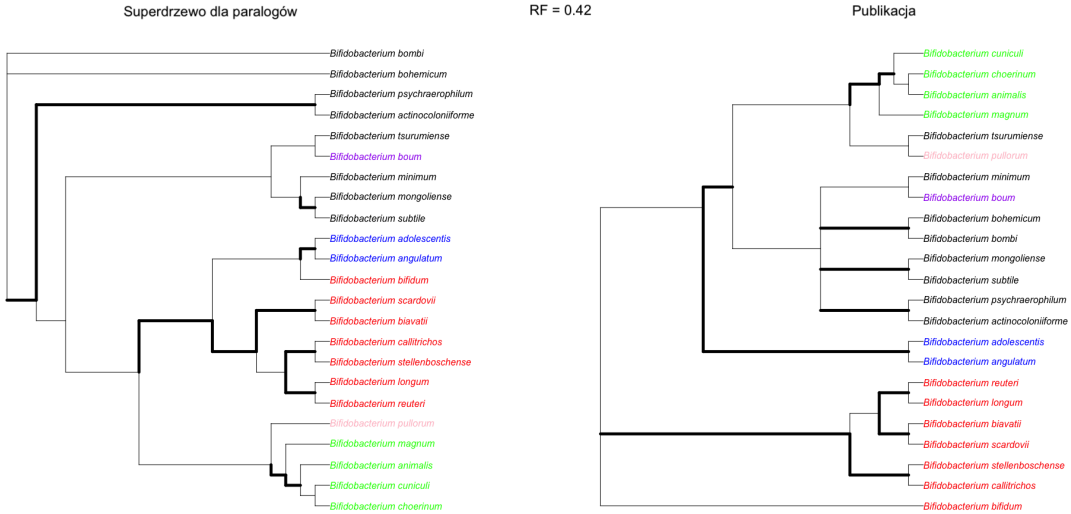
Rysunek 4: Superdrzewo dla klastrów 1-1.



Rysunek 5: Drzewo konsensusu z bootstrapem dla klastrów 1-1.



Rysunek 6: Superdrzewo z bootstrapem dla klastrów 1-1.



Rysunek 7: Superdrzewo dla paralogów.