Genomika porównawcza – projekt zaliczeniowy

1. Cel

Celem projektu jest implementacja pipeline'u filogenetycznego prowadzącego do obliczenia zbioru drzew genów oraz drzewa genomów na podstawie proteomów, a następnie wykorzystanie go do analizy filogenetycznej wybranej grupy organizmów.

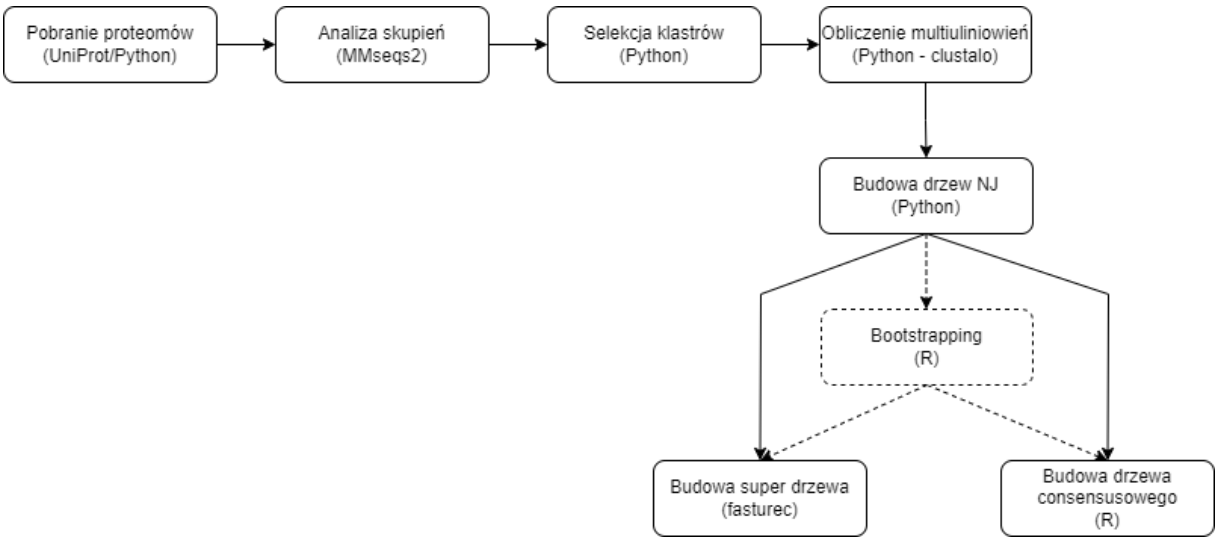
2. Wybór organizmów

Bakterie z klasy *Clostridia* stanowią bardzo heterogeniczny zbiór, który nie tworzy grupy spójnej filogenetycznie. Szczególnym problemem jest określenie, które gatunki klasy *Clostridia* należą do rodzaju *Clostridium*. Początkowo do tego rodzaju klasyfikowano beztlenowe, Gram-dodatnie laseczki tworzące przetrwalniki. Następnie zaproponowano, aby za gatunki należące do rodzaju *Clostridium* uznać jedynie te gatunki, które tworzą wyraźne skupisko w drzewie 16S rRNA. Jednakże to skupisko zostało zdefiniowane jedynie w kategoriach filogenetycznych i nie była znana żadna biochemiczna, molekularna lub fenotypowa cecha łącząca wszystkie gatunki należące do tego skupiska. Wobec tego analiza proteomów tych bakterii może doprowadzić do nowych, ciekawych wniosków na temat bakterii należących do rodzaju *Clostridium.* Jest to szczególnie pożądane, ze względu na fakt, że wiele gatunków bakterii należących do tego rodzaju jest ważna w medycynie, ponieważ stanowi groźne patogeny.

Zbiór gatunków bakterii do analizy został wybrany na podstawie artykułu „*Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I)*" autorstwa R. S. Gupta oraz B. Gao. W tabeli 1. znajdującej się na kolejnej stronie zebrano informacje na temat wybranych gatunków.

3. Pipeline filogenetyczny

Na rys. 1 przedstawiono ogólny schemat skonstruowanego pipelinu filogenetycznego. Liniami przerywanymi oznaczone są opcjonalne kroki. Pipeline jest obsługiwany za pomocą programu snakemake, który stanowi system zarządzania przepływem pracy.



Rysunek 1. Schemat pipelinu filogenetyczneg.

3.1. Pobieranie proteomów

Pierwszym etapem w pipeline jest pobranie proteomów z bazy danych UniProt. Pobieranie jest zrealizowane za pomocą API UniProt w języku Python. Proteomy są pobierane na podstawie pliku wejściowego zawierającego w kolejnych linijkach Proteome ID każdego proteomu oraz nazwy gatunku, który ma zostać pobrany.

Tabela 1. Zbiór organizmów wybranych do analizy filogenetycznej.

Typ	Klasa	Rząd	Rodzina	Rodzaj	**Gatunek**	Rozmiar proteomu	Ciekawostka
Tenericutes	*Mollicutes*	*Mycoplasmatales*	*Mycoplasmataceae*	*Ureaplasma*	***Ureaplasma parvum***	611	patogen: choroby układu moczowego i dróg rodnych
				Mycoplasma	***Mycoplasma mycoides***	978	-
Actinomycetota	*Actinomycetes*	*Streptomycetales*	*Streptomycetaceae*	*Streptomyces*	***Streptomyces coelicolor***		-
		Corynebacteriales	*Mycobacteriaceae*	*Mycobacterium*	***Mycobacterium tuberculosis***	4001	patogen: gruźlica
Bacillota	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	***Lactobacillus johnsonii***	1809	składnik probiotyków Nestle
		Bacillales	*Bacillaceae*	*Bacillus*	***Bacillus subtilis***	4264	odpowiada za psucie się pieczywa
			Staphylococcaceae	*Staphylococcus*	***Staphylococcus aureus***	2889	patogen: zapalenie płuc i inne
	Clostridia	*Thermoanaerobacterales*	*Thermoanaerobacterales Family III. Incertae Sedis*	*Caldicellulosiruptor*	***Caldicellulosiruptor saccharolyticus***	2629	-
			Thermoanaerobacteraceae	*Caldanaerobacter*	***Caldanaerobacter subterraneus***	2545	-
				Thermoanaerobacter	***Thermoanaerobacter pseudethanolicus***	2193	-
				Carboxydothermus	***Carboxydothermus hydrogenoformans***	2615	-
				Moorella	***Moorella thermoacetica***	2482	-
		Eubacteriales	*Lachnospiraceae*	*Lachnoclostridium*	***Lachnoclostridium phytofermentans***	3892	przekształcanie celulozy w etanol
			Peptococcaceae	*Desulforamulus*	***Desulforamulus reducens***	3221	-
			Desulfitobacteriaceae	*Desulfitobacterium*	***Desulfitobacterium hafniense***	5014	-
			Symbiobacteriaceae	*Symbiobacterium*	***Symbiobacterium thermophilum***	3313	-
			Oscillospiraceae	*Acetivibrio*	***Acetivibrio thermocellus***	4383	przekształcanie celulozy w etanol
			Peptostreptococcaceae	*Clostridioides*	***Clostridioides difficile***	3762	patogen: rzekomobłoniaste zapalenie jelit
			Clostridiaceae	*Alkaliphilus*	***Alkaliphilus oremlandii***	2828	-
					Alkaliphilus metalliredigens	4467	-
				Clostridium	***Clostridium kluyveri***	3828	-
					Clostridium beijerinckii	5398	produkcja butanolu, acetonu
					Clostridium tetani	2415	produkcja neurotoksyny: tetanospazminę
					Clostridium acetobutylicum	3847	produkcja acetonu, etanolu i n-butanolu ze skrobii
					Clostridium perfringens SM101 / ATCC 13124 / 13	2558 / 2873 / 2721	patogen: zgorzela gazowa
					Clostridium botulinum A / F	3590 / 3653	produkcja toksyny botulinowej (jadu kiełbasianego)
			Syntrophomonadaceae	*Syntrophomonas*	***Syntrophomonas wolfei***	2473	-

3.2. Analiza skupień

Drugim etapem jest analiza skupień wszystkich pobranych sekwencji białkowych (połączonych w jeden plik FASTA) za pomocą programu MMseqs2. Przetestowano kilka wartości parametru -c (parametr mówiący o minimalnym podobieństwie sekwencji trafiających do jednego klastra) i wybrano wartość 0.5 jako najbardziej optymalną dla badanego zbioru sekwencji.

3.3. Selekcja klastrów

W kolejnym kroku przeprowadzono selekcję otrzymanych klastrów. Selekcja klastrów umożliwia wyodrębnienie klastrów o określonym rozmiarze (lub zakresie rozmiarów) stanowiących rodziny ortologów (sekwencje o jednoznacznych nazwach genomów), a także klastrów stanowiących rodziny paralogów. W przypadku rodzin paralogów możliwe jest przekazanie parametru, który określi maksymalną akceptowalną częstość występowania najczęściej występującego organizmu w danym klastrze. Przypadek rodzin ortologicznych można rozszerzyć o sztuczną konstrukcję dodatkowych ortologicznych klastrów poprzez wybranie klastrów zawierających paralogi i wylosowanie z nich po 1 sekwencji dla każdego powtarzającego się organizmu – tak aby stworzyć klaster zawierający jedynie po 1 sekwencji z każdego organizmu (nowa rodzina ortologów). Opcja ta może pomóc rozwiązywać problem małej ilości klastrów ortologicznych zawierających po dokładnie 1 sekwencji z każdego z analizowanych organizmów. Szczegóły opisane są w README projektu. Po selekcji odpowiednich klastrów (rodzin genów) nazwy genów zostają zastąpione nazwami genomów.

3.4. Obliczenie multiuliniowień

W następnym etapie dla wyselekcjonowanych klastrów obliczane są multiuliniowienia sekwencji białkowych za pomocą programu ClustalO w języku Python.

3.5. Budowa drzew Neighbor Joining (NJ)

Na podstawie otrzymanych multiuliniowień sekwencji, konstruowane są drzewa NJ za pomocą pakietu Biopython.

3.6. Bootstrapping

Kolejnym, opcjonalnym krokiem analizy jest wykonanie bootstrapping dla wszystkich otrzymanych drzew NJ. W tym celu wykorzystane zostały biblioteki ‘ape’ oraz ‘phangorn’ w języku R. Po wykonaniu 100-krotnego bootstrappingu dla każdego z drzew NJ, obliczany jest średni wynik (średnia arytmetyczna wartości otrzymanych dla każdego wierzchołka). Słabo wspierane drzewa NJ z końcowym wynikiem poniżej 50% są odrzucane z dalszej analizy. Próg odrzucenia stanowi parametr piepline’u i można go regulować w pliku konfiguracyjnym.

3.7. Budowa super drzewa

Otrzymane (i opcjonalnie odfiltrowane za pomocą bootstrappingu) drzewa NJ są wykorzystywane do budowy super drzewa. W tym celu wykorzystano program fasturec. Do analiz oraz porównań z drzewem referencyjnym otrzymane super drzewa zostały odkorzenione.

3.8. Budowa drzewa konsensusowego

Otrzymane (i opcjonalnie odfiltrowane za pomocą bootstrappingu) drzewa NJ są wykorzystywane do budowy drzewa konsensusowego. W tym celu wykorzystano bibliotekę ‘ape’ w języku R. Wykorzystano wersję dla drzew nieukorzenionych.

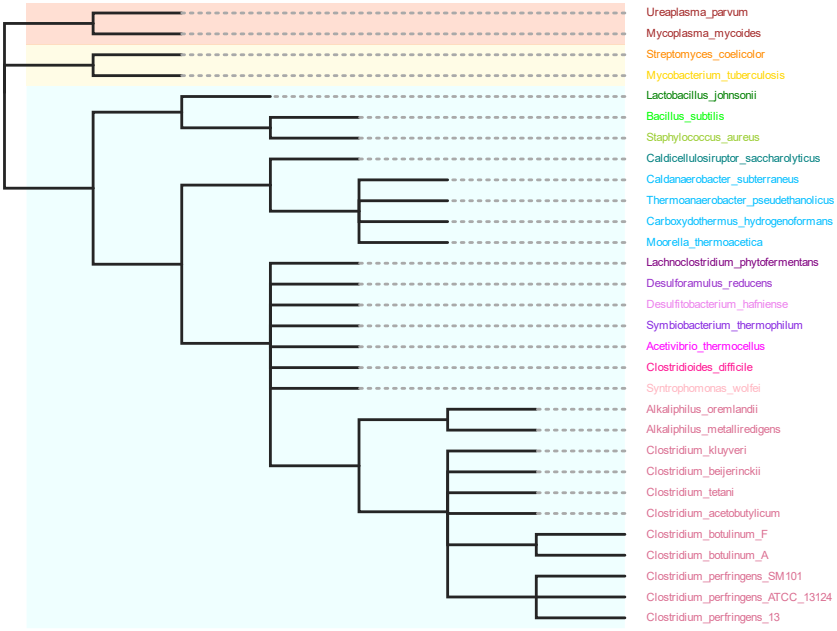
3.9. Przetwarzanie wstępne wyników

Otrzymane wyniki przetworzono z wykorzystaniem pakietu toytree w języku Python. Liście każdego drzewa zostały pokolorowane według rodziny, do której należy dany gatunek. Otrzymane drzewa genomów porównano z drzewem referencyjnym (drzewo pochodzące z artykułu) oraz z taksonomią gatunków NCBI. Obliczono odległość Robinsona-Fouldsa w wersji nieukorzenionej dla otrzymanych drzew oraz drzewa referencyjnego.

4. Analiza wyników

W wyniku analiz otrzymano ok. 70 różnych super drzew i drzew konsensuowych dla różnych konfiguracji parametrów. W dalszej części raportu przedstawiono jedynie wybrane drzewa.

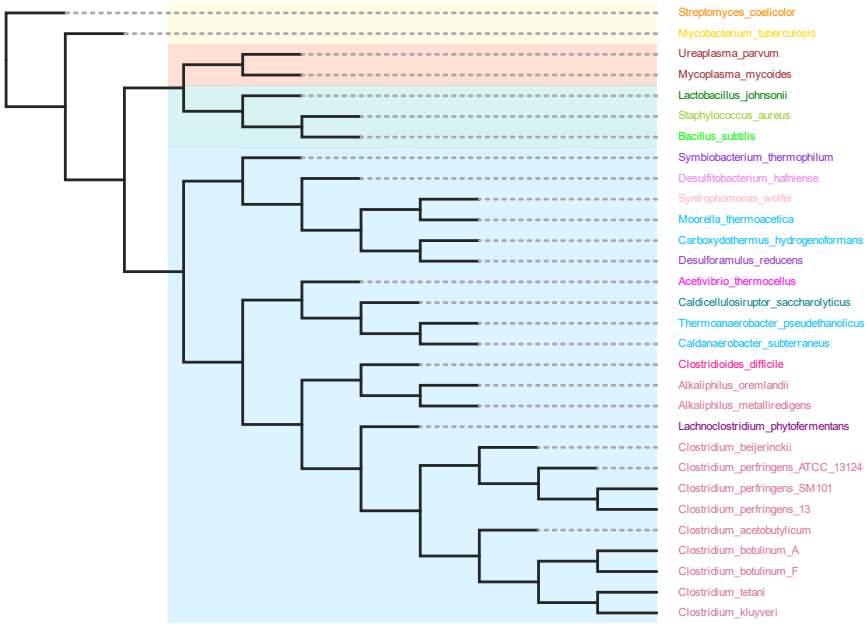
4.1. Taksonomia NCBI



Rysunek 2. Taksonomia NCBI

Na rys. 2 przedstawiono drzewo taksonomiczne NCBI. Kolor liści wskazuje na przynależność danych gatunków do tej samej rodziny. Kolorowe prostokąty wskazują na przynależność gatunków do tego samego typu

4.2. Drzewo referencyjne



Rysunek 3. Drzewo referencyjne

Na rys. 3 przedstawiono referencyjne drzewo, stanowiące wynik analiz przeprowadzonych przez R. S. Gupta oraz B. Gao. W drzewie referencyjnym ponownie ten sam kolor liści wskazuje na przynależność do tej samej rodziny, natomiast kolorowe prostokąty wskazują na przynależność gatunków do tej samej klasy (jeżeli połączymy organizmy z prostokątów zielonego i niebieskiego, otrzymamy przynależności do typów).

4.3. Analiza skupień

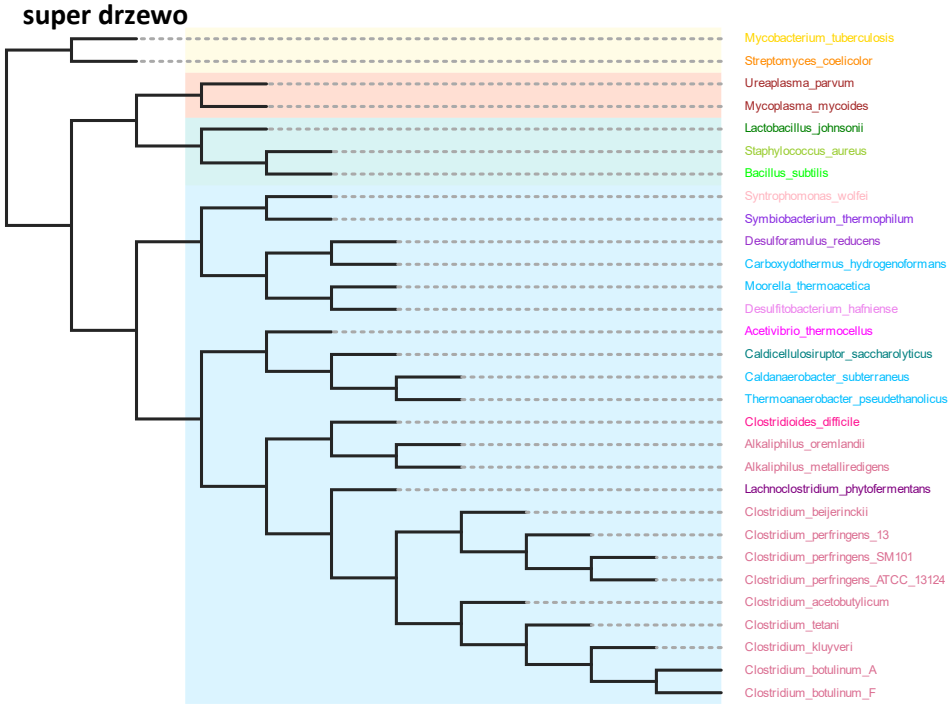
W wyniku przeprowadzonej analizy skupień dla sekwencji białkowych pochodzących od wszystkich analizowanych organizmów otrzymano 25764 klastrów, z czego 8896 klastrów zawierało więcej niż 1 sekwencję, a 3253 więcej niż 5 sekwencji. Najliczniejszy klaster zawierał 808 sekwencji.

4.4. Przypadek klastrów ortologicznych

4.4.1. Klastry zawierające dokładnie 30 sekwencji białkowych

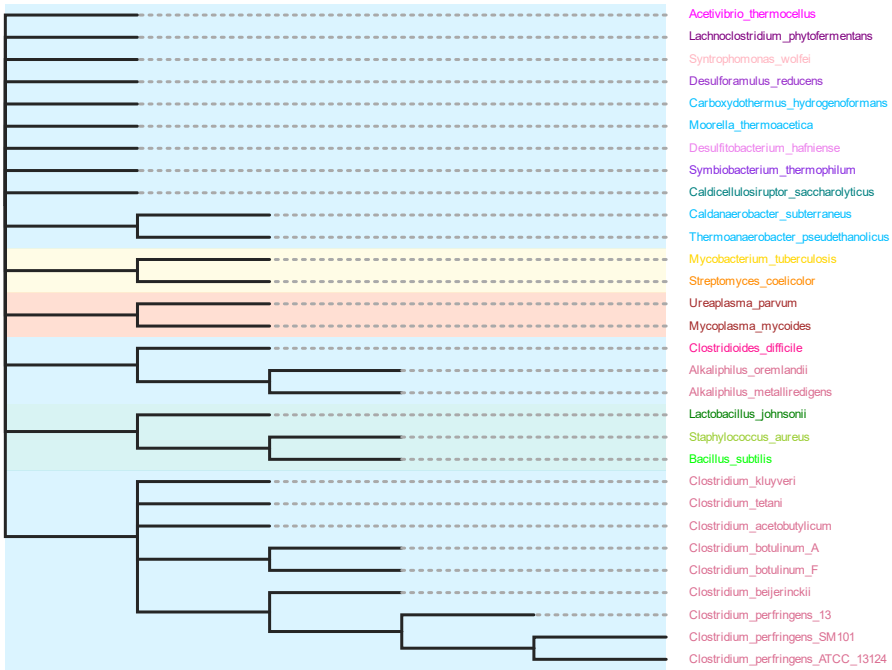
Znaleziono 52 klastry, które zawierały dokładnie po 1 sekwencji z każdego analizowanego organizmu (w sumie 30 sekwencji). Na ich podstawie obliczono multiuliniowienia sekwencji, zbudowano drzewa genów NJ, których użyto do konstrukcji drzew genomów dwiema metodami: super drzewo (rys. 4) oraz drzewo konsensusowe (rys. 5).

- **super drzewo**



Rysunek 4. Super drzewo dla rodzin ortologów 1 do 1

- **drzewo konsensusowe**



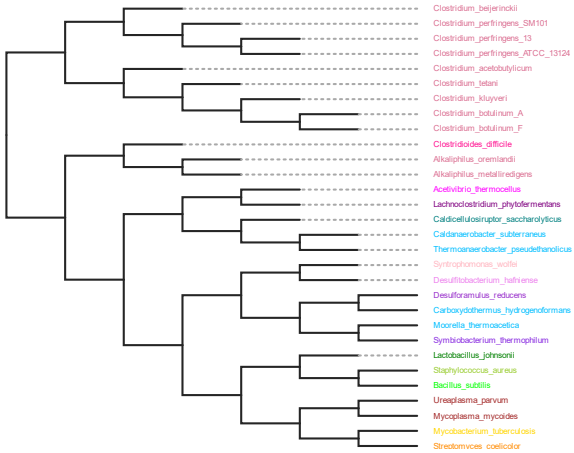
Rysunek 5. Drzewo konsensusowe dla rodzin ortologów 1 do 1

Dla obu drzew obliczono odległość Robinsona-Fouldsa (wersja dla drzew nieukorzenionych) względem drzewa referencyjnego. Dla super drzewa odległość ta wyniosła 10 (znormalizowany RF: 0.185), natomiast dla drzewa konsensusowego – 17 (znormalizowany RF: 0.436). Należy zauważyć, że otrzymane drzewo konsensusowe nie jest drzewem binarnym – występuje w nim wiele multifurkacji. Drzewo konsensusowe bardzo słabo oddaje referencyjne relacje między poszczególnymi wierzchołkami wewnętrznymi. O wiele lepsze pod tym względem jest otrzymane super drzewo, w którym prawidłowo został odtworzony podział organizmów na klasy (a także na typy), obserwowany zarówno w drzewie referencyjnym, jak i w taksonomii NCBI. W otrzymanym superdrzewie obserwujemy również prawidłowo odtworzoną referencyjną, a zarazem taksonomiczną relację między organizmami należącymi do klasy *Bacilli*. Pewne niezgodności względem drzewa referencyjnego pojawiają się w węzłach odpowiadającym organizmom z gatunku *Clostridium perfringens* oraz *Clostridium tetani* i *Clostridium kluyveri*. Poprawnie zostały odtworzone relacje gatunków z rodzaju *Alkaliphilus* oraz gatunku *Lachnoclostridium phytofermentans* względem siebie oraz względem organizmów z rodzaju *Clostridium*. Węzeł zawierający te organizmy pozostał w prawidłowej relacji z węzłem zawierającym organizmy *Thermoanaerobacter pseudethanolicus***,** *Caldanaerobacter subterraneus***,** *Caldicellulosiruptor saccharolyticus* oraz *Acetivibrio thermocellus*, które również pozostają w prawidłowej relacji między sobą. Największą ilość niezgodności możemy zaobserwować w węźle, pod którym znajdują się pozostałe organizmy z klasy *Clostridia,* jednak niezgodności nie są duże. Podsumowując – otrzymane super drzewo dobrze odwzorowało relacje obserwowane w drzewie referencyjnym.

Ze względu na fakt, że liczba znalezionych klastrów zawierających dokładnie po 1 sekwencji białkowej z każdego analizowanego organizmu nie jest bardzo duża, sztucznie stworzono większą liczbę ortologicznych klastrów. W tym celu znaleziono klastry paralogiczne zawierające co najmniej 1 sekwencję z każdego z analizowanych organizmów i dla powtarzających się organizmów wylosowano po 1 sekwencji dla każdego organizmu. W ten sposób otrzymano 66 dodatkowych klastrów reprezentujących rodziny ortologów 1 do 1, które wykorzystano do ponownego zbudowania super drzewa i drzewa konsensusowego. Okazało się jednak, że ta strategia nie przyniosła pozytywnych rezultatów – otrzymane drzewo konsensusowe pozostało niezmienione, natomiast w przypadku super drzewa wynik uległ znacznemu pogorszeniu (RF=22).

4.4.2. Klastry zawierające mniej niż 30 sekwencji białkowych

W celu sprawdzenia czy uzyskane super drzewo można udoskonalić przez wykorzystanie większej liczby ortologicznych klastrów, znaleziono wszystkie ortologiczne klastry o wielkościach w przedziałach 5-30 (liczba klastrów: 1228), 10-30 (liczba klastrów: 523) oraz 15-30 (liczba klastrów: 360) sekwencji. Na ich podstawie zbudowano super drzewa. Odległości RF względem drzewa referencyjnego wyniosły odpowiednio: 20, 18, 18. Po prawej stronie zamieszczono drzewo zbudowane na podstawie klastrów o liczności 15-30 (rys. 6). Możemy zauważyć, że odwzorowanie relacji między organizmami uległo znaczącemu pogorszeniu względem drzewa referencyjnego. Może to wynikać z faktu, że uwzględniając dużo dadatkowych klastrów nie zawierających sekwencji od wszystkich analizowanych organizmów, dodaliśmy dużo zbędnych informacji, co źle wpłynęło na końcowy wynik. Należy w tym miejscu również dodać, że referencyjne drzewo było konstowane jedynie na



Rysunek 6. Super drzewo dla rodzin ortologów o rozmiarach 15-30 sekwencji.

podstawie 37 wysoce konserwowanych wśród bakterii białek, wobec tego możliwe, że dodawanie większej ilości białek, która nie jest wspólna dla wszystkich organizmów, będzie oddalało nas od referencyjnego rozwiązania.

4.4.3. Bootstraping

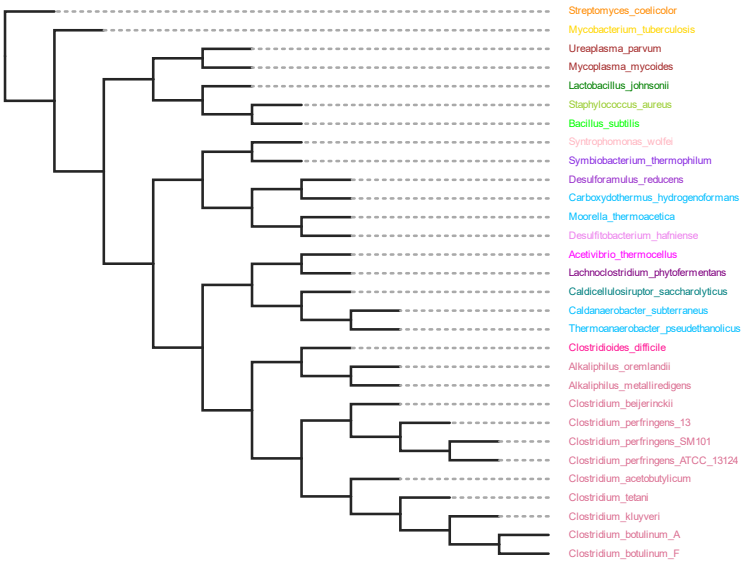
W celu wyeliminowania słabo wspieranych drzew NJ dla każdego z powyżej omówionych przypadków wykonano bootstraping. Liczba drzew NJ otrzymanych po bootstrapingu:

- rodziny ortologów 1-1: 40
- rodziny ortologów 1-1 z losowaniem: 96
- rodziny ortologów o zmniejszonym rozmiarze 5-30: 1032
- rodziny ortologów o zmniejszonym rozmiarze 10-30: 413
- rodziny ortologów o zmniejszonym rozmiarze 15-30: 283

Na podstawie otrzymanych drzew ponownie zbudowano super drzewa oraz drzewa konsensusowe (dla przypadków rodzin ortologów 1-1). Poniżej zamieszczono obliczone odległości RF (w nawiasach znormalizowane RF) względem drzewa referencyjnego:

- super drzewo rodziny ortologów 1-1: 16 (0.296)
- drzewo konsensusowe rodziny ortologów 1-1: 17 (0.436)
- super drzewo rodziny ortologów 1-1 z losowaniem: 16 (0.296)
- drzewo konsensusowe rodziny ortologów 1-1 z losowaniem: 17 (0.436)
- super drzewo rodziny ortologów o zmniejszonym rozmiarze 5-30: 22 (0.407)
- super drzewo rodziny ortologów o zmniejszonym rozmiarze 10-30: 18 (0.333)
- super drzewo rodziny ortologów o zmniejszonym rozmiarze 15-30: 18 (0.333)

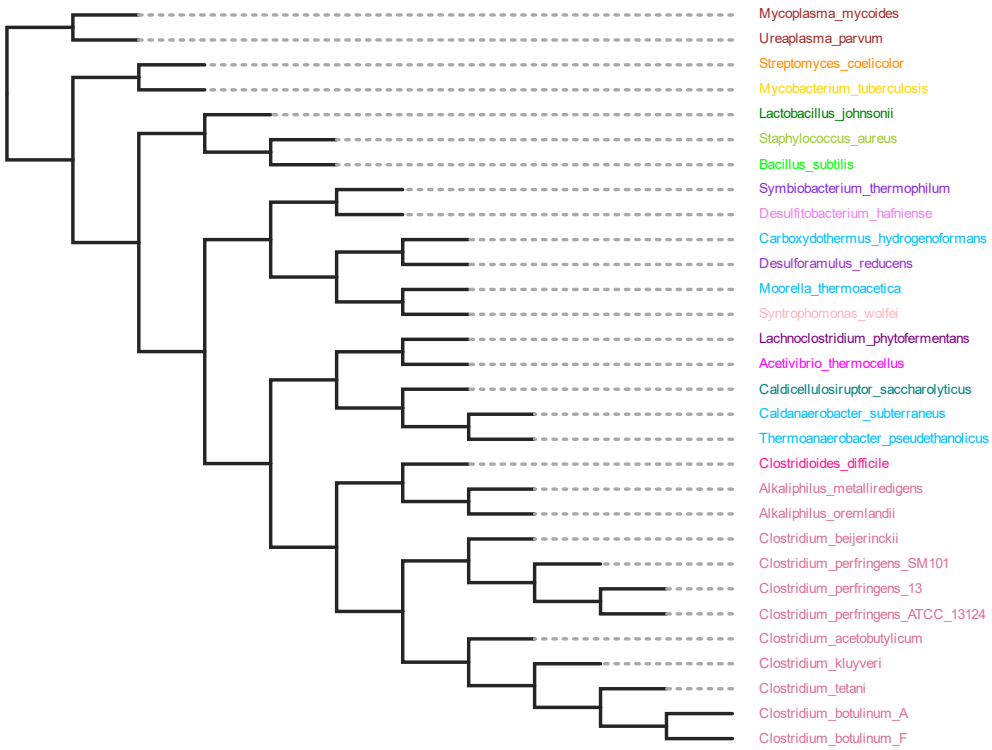
Jedynym przypadkiem, gdy proces bootstrapingu poprawił wynik jest super drzewo dla rodziny ortologów 1-1 z losowaniem, gdzie bez eliminacji słabo wspieranych drzew za pomocą bootstrappingu odległość RF wynosiła 20. W pozostałych przypadkach otrzymane wyniki są porównywalne lub gorsze po eliminacji części drzew NJ. W przypadku drzew konsensusowych drzewa przed oraz po bootstrappingu były takie same. W przypadku super drzew zaobserwowano pewne różnice. Na rys. 7 przedstawiono super drzewo rodziny ortologów skonstruowane na podstawie drzew NJ po bootstrappingu.



Rysunek 7. Super drzewo dla rodzin ortologów 1 do 1 skonstrowane na podstawie drzew NJ odfiltrowanych techniką bootstrappingu.

4.5. Przypadek rodzin paralogów

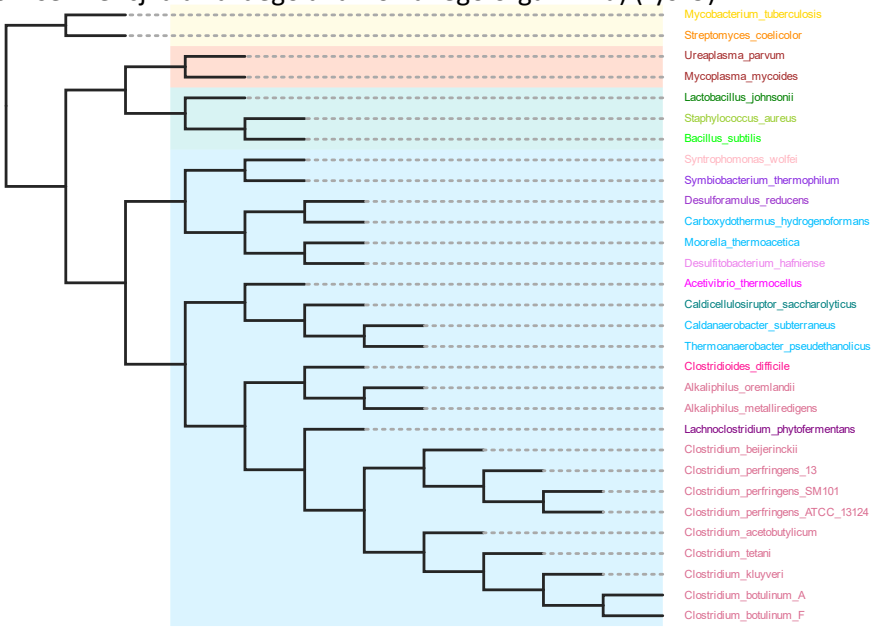
Dodatkowo przeprowadzono analizy dla rodzin paralogów. Wybrano kilka różnych zbiorów rodzin paralogów za pomocą manipulacji parametrami min_size (minimalna wielkość klastra), max_size (maksymalna wielkość klastra), max_repeat (maksymalna akceptowalna częstość występowania najczęściej występującego organizmu w danym klastrze) i zbudowano dla nich super drzewa. Najlepsze wyniki otrzymano dla parametrów min_size=25, max_size=150 oraz max_repeat=0.2. Dla tak zadanych parametrów otrzymano 789 klastrów. Odległość RF otrzymanego super drzewa (rys. 8) względem drzewa referencyjnego wyniosła 14 (znormalizowany RF: 0.26). Mimo gorszego wyniku względem super drzewa dla rodzin ortologów 1 do 1, można zauważyć, że część relacji w otrzymanym super drzewie uległa poprawie (porównując do drzewa referencyjnego). Ma to miejsce na przykład dla pary gatunków *Syntrophomonas wolfei* oraz *Morella thermoacetica*, a także dla *hafniense* i *Carboxydothermus hydrogenoformans*. Błędnie natomiast, w porównaniu do drzewa referencyjnego, została przedstawiona relacja typów bakterii – w otrzymanym super drzewie typ *Tenericutes* został zamieniony z typem *Actinomycetota*, co stanowi błąd także względem dostępnej wiedzy na temat ewolucji typów bakterii.



Rysunek 8. Super drzewo dla rodzin paralogów o rozmiarach od 25 do 30 sekwencji zawierających do 20% sekwencji pochodzącego od jednego organizmu.

4.6. Wnioski

Na podstawie powyższych analiz drzewem najlepiej odwzorowującym relacje zawarte w drzewie referencyjnym jest super drzewo dla przypadku rodzin ortologów 1 do 1 (zawierających dokładnie po 1 sekwencji dla każdego analizowanego organizmu) (rys. 9).



Rysunek 9. Super drzewo dla rodzin ortologów 1 do 1

Należy jednak wziąć pod uwagę, że drzewo referencyjne zostało zbudowane na podstawie 36 rodzin wysoko konserwowanych białek i również może zawierać niezgodne z rzeczywistością relacje między organizmami. Na przykład, na podstawie taksonomii NCBI oraz wyników otrzymanych dla wszystkich pozostałych drzew, wydaje się, że relacja między organizmami rodziny *Clostridium* a gatunkiem *Lachnoclostridium phytofermentas* jest niepoprawna w drzewie referencyjnym oraz super drzewie dla rodzin ortologów 1 do 1. Na podstawie pozostałych otrzymanych drzew można by stwierdzić, że organizmy należące do rodziny *Alkaliphilus* oraz gatunek *Clostridioides difficiae* znajdują się ewolucyjnie bliżej rodziny *Clostridium* niż gatunkiem *Lachnoclostridium phytofermentas.*

Kolejną ciekawą kwestią, którą możemy zaobserwować na większości otrzymanych drzew, jest rozdzielenie się typów bakterii. Możemy przypuszczać, że początkowo bakterie rozdzieliły się na dwa typy *Actinomycetota* oraz *Tenericutes+Bacillota*, następnie osobno ewoluowały i dopiero po pewnym czasie typ *Tenericutes* oddzielił się od klasy *Bacilli* poprzez utratę części genomu (tezę o utracie genomu można poprzeć faktem, że rozmiar proteomu organizmów należących do typu *Bacillota* wynosi ok. 2-4 tysiące białek, natomiast proteomy organizmów należących do typu *Tenericutes* składają się z mniej niż tysiąca białek).

Dodatkowo we wszystkich otrzymanych drzewach oraz w drzewie referencyjnym i taksonomii NCBI organizmy należące do klasy *Bacilli* znajdują się w identycznych relacjach względem siebie oraz względem klasy *Clostridia*. Może to świadczyć o dobrze zakonserwowanych zmianach w genomach organizmów należących do obu klas, które pozwalają na ich wyraźne rozdzielenie.

Analogicznie we wszystkich otrzymanych drzewach, a także w drzewie referencyjnym, wszystkie organizmy należące do rodzaju *Clostridium* są zawsze zgrupowane pod wspólnym wierzchołkiem, pod którym nie znajdują się organizmy należące do innych rodzajów. To także może świadczyć o dobrze zakonserwowanych zmianach w genomach (a tym samym proteomach) organizmów należących do tego rodzaju względem organizmów należących do innych rodzajów. Ta obserwacja może stanowić podstawę do poszukiwania wzorca molekularnego charakterystycznego jedynie dla organizmów z rodzaju *Clostridium*, co umożliwi znalezienie wspólnych cech biochemicznych tych organizmów i będzie stanowiło dodatkową pomoc w problemie klasyfikacji organizmów do tego rodzaju.