Rekonstrukcja drzewa genomów Gammaproteobacteria

Justyna Kowalska

20 stycznia 2025

1 Wstęp

Rekonstrukcja relacji filogenetycznych na podstawie danych genomowych jest jednym z kluczowych zadań współczesnej filogenomiki. W przypadku bakterii wyzwanie to jest szczególnie istotne, ponieważ różne fragmenty genomu mogą mieć odmienną historię ewolucyjną, a transfer poziomy genów może prowadzić do konfliktów pomiędzy drzewami wyznaczanymi z pojedynczych markerów. W konsekwencji, filogenezy oparte wyłącznie na 16S rRNA lub na pojedynczych białkach bywają słabo rozdzielone i niekiedy dają rozbieżne wnioski taksonomiczne, co było jednym z głównych motywów pracy Sharma i in. (2022) [1], w której autorzy podjęli próbę rewizji i uściślenia relacji wewnątrz złożonego typu Proteobacteria przy użyciu podejść filogenomicznych.

Przedmiotem niniejszej pracy jest klasa Gammaproteobacteria, która stanowi jedną z najliczniejszych i najbardziej zróżnicowanych fenotypowo grup w obrębie Proteobacteria. Zbiór danych wejściowych obejmuje 27 kompletnych proteomów, stanowiących reprezentatywny podzbiór taksonów analizowanych w pracy Sharma i in. (2022) [1]. Znalazły się w nim m.in. przedstawiciele *Enterobacterales* (*Escherichia coli K-12*, *Salmonella typhimurium*, *Yersinia pestis*), a także taksony z *Vibrionales*, *Alteromonadales*, *Legionellales*, *Pseudomonadales*, *Xanthomonadales* oraz linii związanych z bakteriami siarkowymi. Tak dobrany zestaw organizmów pozwala na przeprowadzenie wielopoziomowej analizy filogenetycznej obejmującej zarówno weryfikację pokrewieństwa blisko związanych taksonów jak i rekonstrukcję głębokich węzłów ewolucyjnych w obrębie całej badanej klasy.

Punktem odniesienia dla interpretacji wyników jest analiza Sharma i in. (2022), w której autorzy rekonstruowali filogenezę Proteobacteria kilkoma podejściami, obejmującymi drzewa oparte na 16S rRNA, drzewa pojedynczych białek konserwatywnych oraz drzewo z konkatenacji 85 genów housekeeping, po dopasowaniu sekwencji metodą MAFFT w środowisku NGPhylogeny i filtracji bloków BMGE. Równolegle zastosowano metody genomowe niewymagające dopasowania, w tym CVTree oraz drzewo na podstawie macierzy AAI. Szczególnie istotnym elementem ich metodologii było wyznaczenie drzewa konsensusowego (przedstawionego w artykule na Ryc. 5). Nie powstało ono bezpośrednio z sekwencji, lecz zostało obliczone przy użyciu pakietu PHYLIP jako konsensus topologii uzyskanych wszystkimi wyżej wymienionymi metodami, uzupełnionych dodatkowo o indywidualne drzewa dla sześciu wybranych genów markerowych: dnaK, gyrA, rplP, rplR, pheT oraz yidC. Opis metod w pracy Sharma i in. (2022) [1] koncentruje się na doborze danych i ogólnej strategii rekonstrukcji, natomiast część szczegółów implementacyjnych dotyczących m.in metody optymalizacji topologii drzew czy wartości zastosowanych parametrów, nie została doprecyzowana. Takie podejście miało na celu ograniczenie wpływu artefaktów pojedynczych metod i wyłonienie kladów stabilnych między rekonstrukcjami. Ich analiza pokazała, że dla Proteobacteria, w tym dla

Gammaproteobacteria, wyniki mogą zależeć od doboru markerów i metody wnioskowania, a część relacji wykazuje konflikt sygnału między rekonstrukcjami.

W obrębie Gammaproteobacteria oczekuje się, że część rzędów odtworzy się jako stabilne, monofiletyczne klady zgodne z aktualną taksonomią, w szczególności Enterobacterales. Jednocześnie literatura wskazuje na obszary topologicznie wrażliwe, gdzie różne geny i różne metody mogą prowadzić do alternatywnych ustawień kladów. W projekcie przyjęto zatem hipotezy robocze, że Enterobacterales zostanie odtworzony jako spójny klad w większości wariantów analizy, natomiast pozycja*Pseudomonas aeruginosa* względem części linii morskich i halofilnych może wykazywać większą zmienność zależną od doboru rodzin genów oraz metody konstrukcji drzewa genomów.

Niniejszy projekt koncentruje się na odtworzeniu podrzewa relacji filogenetycznych w obrębie Gammaproteobacteria z pracy Sharma i in. (2022) [1], z wykorzystaniem w pełni zautomatyzowanego pipeline'u, który rozpoczyna się od kompletnych proteomów, następnie wyznacza rodziny genów, rekonstruuje drzewa genowe, a finalnie wnioskuje drzewo genomów metodami konsensusowymi i superdrzewowymi, a następnie porównano wyniki z topologią referencyjną z pracy Sharma i in. (2022) [1] oraz z topologią z Timetree[2].

2 Metody

2.1 Dane wejściowe

Zestaw danych obejmował 27 genomów Gammaproteobacteria (RefSeq, accession GCF) wybranych na podstawie Sharma i in. (2022). Dla każdego genomu pobrano z NCBI[3] kompletny proteom jako sekwencje białkowe w formacie FASTA, z użyciem NCBI Datasets CLI. Ze zbioru wykluczono taksony z fragmentarycznymi proteomami oraz takie, których liczba sekwencji białkowych była wyraźnie niższa niż w pozostałych analizowanych genomach, co mogło wskazywać na niekompletność assembly lub niespójność adnotacji. Wszystkim białkom nadano globalnie unikalne identyfikatory genome_id__protein_id, a dodatkowo utworzono pliki mapujące białka do genomów i scalony FASTA wszystkich sekwencji, użyty jako wejście do klastrowania.

2.2 Klastrowanie i wybór rodzin

Rodziny genów wyznaczono przez klastrowanie sekwencji białkowych w trybie all-vs-all z użyciem MMseqs2[4, 5, 6], (workflow `easy-cluster`) na scalonym pliku `merged_proteins.faa`. Klastrowanie uruchamiano z równolegleniem (`-threads`), a kontrolę nakładania długości sekwencji zapewniał parametr pokrycia `-c`. Przed wyborem konfiguracji użytej w analizie końcowej przetestowano kilka ustawień parametrów klastrowania, w tym wartości minimalnego pokrycia (`-c`) oraz minimalnej identyczności sekwencji (`-min-seq-id`), oceniając ich wpływ na liczbę klastrów oraz liczebność rodzin spełniających kryterium 1:1. W analizie końcowej zastosowano `-c=0.8`, a `-min-seq-id` pozostawiono domyślne. W wyniku klastrowania otrzymano łącznie 30 278 klastrów.

Na podstawie wyników klastrowania wygenerowano pliki FASTA odpowiadające poszczególnym rodzinom genów. Do dalszej analizy przygotowano dwa warianty danych wejściowych do części gatunkowej:

- zbiór rodzin ortologicznych, który zawiera klastry zawierające dokładnie jedną sekwencję z każdego genomu (1:1), z pełnym pokryciem zestawu taksonów, 196 rodzin, wykorzystywane w analizach konsensusowych oraz supertree

- zbiór rodzin paralogicznych, czli klastry, gdzie z zachowaniem wszystkich kopii z danego gatunku, 535 rodzin, wykorzystywane w analizach superdrzewowych

2.3 Multiuliniowienie

Dla każdej rodziny genów obliczono wielokrotne wyrównanie sekwencji (MSA) z użyciem MAFFT[7] w trybie automatycznego doboru strategii. Wyrównania wykonywano równolegle dla wielu rodzin, a liczba wątków MAFFT była parametryzowana. Wyniki zapisywano jako osobne pliki MSA.

2.4 Drzewa genowe

Dla każdego MSA zrekonstruowano drzewo filogenetyczne metodą Maximum Likelihood z użyciem IQ-TREE2[8]. Metoda ML polega na znalezieniu topologii oraz długości gałęzi, które maksymalizują prawdopodobieństwo zaobserwowanego wyrównania przy założonym modelu ewolucji sekwencji. W porównaniu z metodami odległościowymi zwykle zapewnia lepszą estymację topologii, kosztem większej złożoności obliczeniowej, co w tym projekcie było akceptowalne dzięki równolegleniu obliczeń. Model substytucji nie był z góry narzucony. Dla każdej rodziny dobierano go automatycznie z użyciem ModelFinder (MFP), co oznacza wybór najlepiej dopasowanego modelu aminokwasowego (wraz z wariantem heterogeniczności szybkości ewolucji między pozycjami) osobno dla każdego MSA. Wynikowe drzewa (`.treefile`) scalono do jednego pliku `all_trees.nwk` (jedno drzewo na linię) na potrzeby dalszych analiz drzewa genomów.

2.5 Drzewa genomów

Drzewa genomów wyznaczono 2 metodami, na podstawie zbiorów wcześniej uzyskanych drzew genowych.

W podejściu konsensusowym wykorzystano drzewa z rodzin 1:1, co zapewniało identyczny zestaw taksonów we wszystkich drzewach. Drzewa traktowano jako nieukorzenione, a następnie obliczono drzewa konsensusowe w dwóch wariantach: (i) majority-rule z progiem $p = 0.5$ za pomocą `ape::consensus` [9], gdzie zachowywano jedynie klady występujące w ponad 50% drzew, oraz (ii) wariant greedy (extended majority-rule) z użyciem IQ-TREE2 [8] w trybie budowy konsensusu (`-con`) z parametrem `-minsup 0.0`, co oznacza brak minimalnego progu częstości kladu w drzewie konsensusu. W obu wariantach częstości kladów wśród drzew wejściowych (w %) zaprezentowano na wykresach jako etykiety węzłów, do czego wykorzystano pakiet `ape` [9].

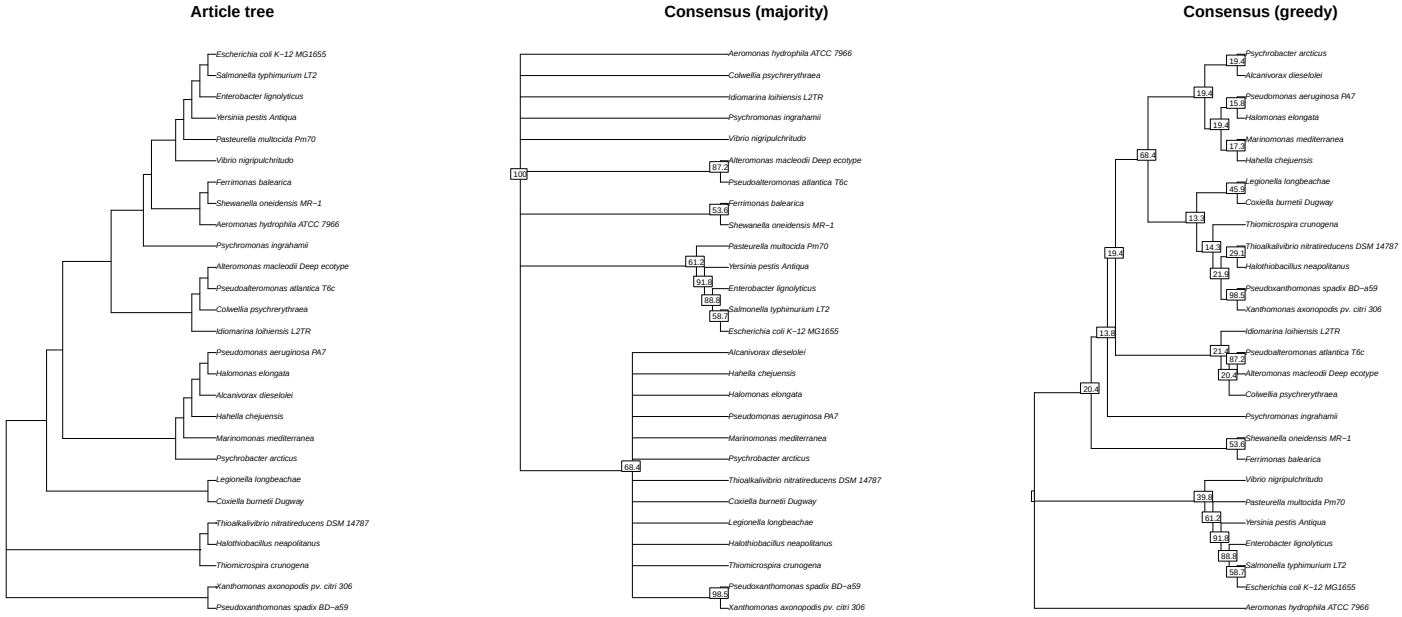
W podejściu supertree drzewo genomów rekonstruowano programem Fasturec (v2.07) [10, 11, 12] w dwóch wariantach: (i) na podstawie drzew z rodzin 1:1 oraz (ii) na podstawie drzew z rodzin paralogicznych. Uruchomienie Fasturec zautomatyzowano wrapperem, który standaryzował format wejścia, a w wariancie z paralogami upraszczał etykiety liści przez usunięcie sufiksów identyfikujących kopie (format `genom__ID`), tak aby wszystkie kopie z jednego genomu były traktowane jako ten sam takson. Fasturec uruchamiano na drzewach nieukorzenionych, a drzewo genomów estymowano przez minimalizację kosztu rekonsyliacji w wariancie duplication–loss (DL). Optymalizację prowadzono heurystycznie metodą hill-climbing i stosując lokalne modyfikacje topologii (NNI, SPR oraz TSW). Jako wynik wybierano drzewo o minimalnym koszcie raportowanym przez program, zapisując je w formacie Newick.

2.6 Środowisko obliczeniowe i automatyzacja

Pipeline zaimplementowano jako zestaw skryptów Python, Bash i R odpalany za pomocą bashowego skryptu main.sh. Zrównoleglenie zastosowano na poziomie klastrowania (MMseqs2), budowy MSA oraz drzew genowych. Obliczenia wykonano na komputerze z procesorem AMD Ryzen 7 4700U (8 wątków) oraz 16 GB pamięci RAM. Całkowity czas wykonania pipeline'u wyniósł 8h27min.

3 Wyniki

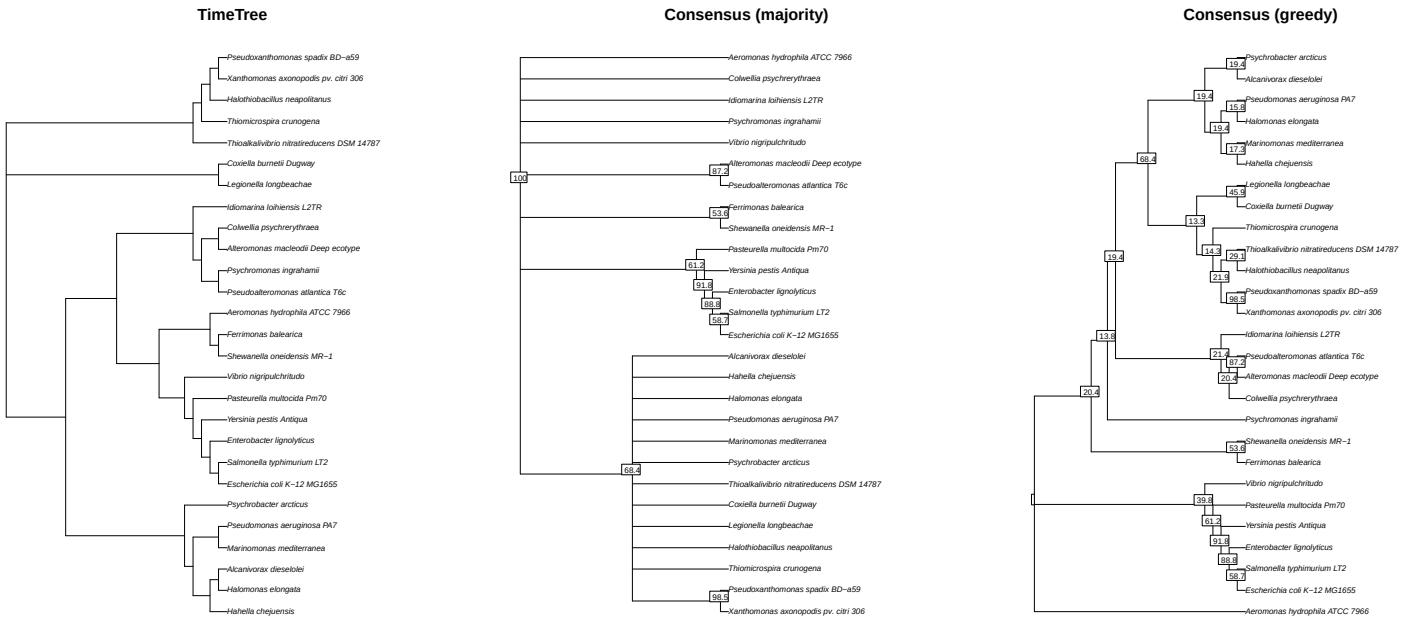
W wyniku analiz otrzymano cztery drzewa genomów uzyskane w pipeline'ie: (i) drzewo konsensusowe majority-rule z rodzin 1:1, (ii) drzewo konsensusowe greedy z rodzin 1:1, (iii) supertree na drzewach z rodzin 1:1, (iv) supertree na drzewach z rodzin paralogicznych. Jako punkty odniesienia wykorzystano drzewo z artykułu Sharma i in. (2022) [1] oraz drzewo pozyskane z serwisu TimeTree.org [2]. Wszystkie porównania miar odległości i podobieństwa wykonywano na drzewach nieukorzenionych.



Rysunek 1: Porównanie drzewa referencyjnego z literatury oraz drzew consensus majority rule i greedy

Wyniki przedstawiono zarówno wizualnie (Rys. 1, 2, 3 and 4), jak i ilościowo w tab. 1. Podobieństwo topologii oceniono dwiema miarami opartymi o porównanie podziałów: (i) Odległość Robinsona–Fouldsa (RF) porównuje dwie topologie poprzez ich zbiory splitów indukowanych przez krawędzie wewnętrzne. RF to liczba takich podziałów, które nie są wspólne dla obu drzew. Znormalizowana wersja (nRF) dzieli tę liczbę przez maksymalną możliwą liczbę niezgodnych splitów

dla porównywanych drzew, dzięki czemu wynik mieści się w przedziale $[0, 1]$, gdzie 0 oznacza identyczne topologie, a wartości bliższe 1 oznaczają większą niezgodność. (ii) Znormalizowane Mutual Clustering Information (nMCI) [13] jest miarą podobieństwa opartą na teorii informacji, która wyraża, ile informacji o grupowaniu taksonów jest wspólne dla obu drzew. Po normalizacji wartości są skalowane do zakresu $[0, 1]$, gdzie 0 oznacza brak wspólnej informacji klastrującej, natomiast wartości bliższe 1 oznaczają większą zgodność relacji grupowania między drzewami (1 odpowiada pełnej zgodności topologii).



Rysunek 2: Porównanie drzewa z Timetree.org oraz drzew consensus majority rule i greedy

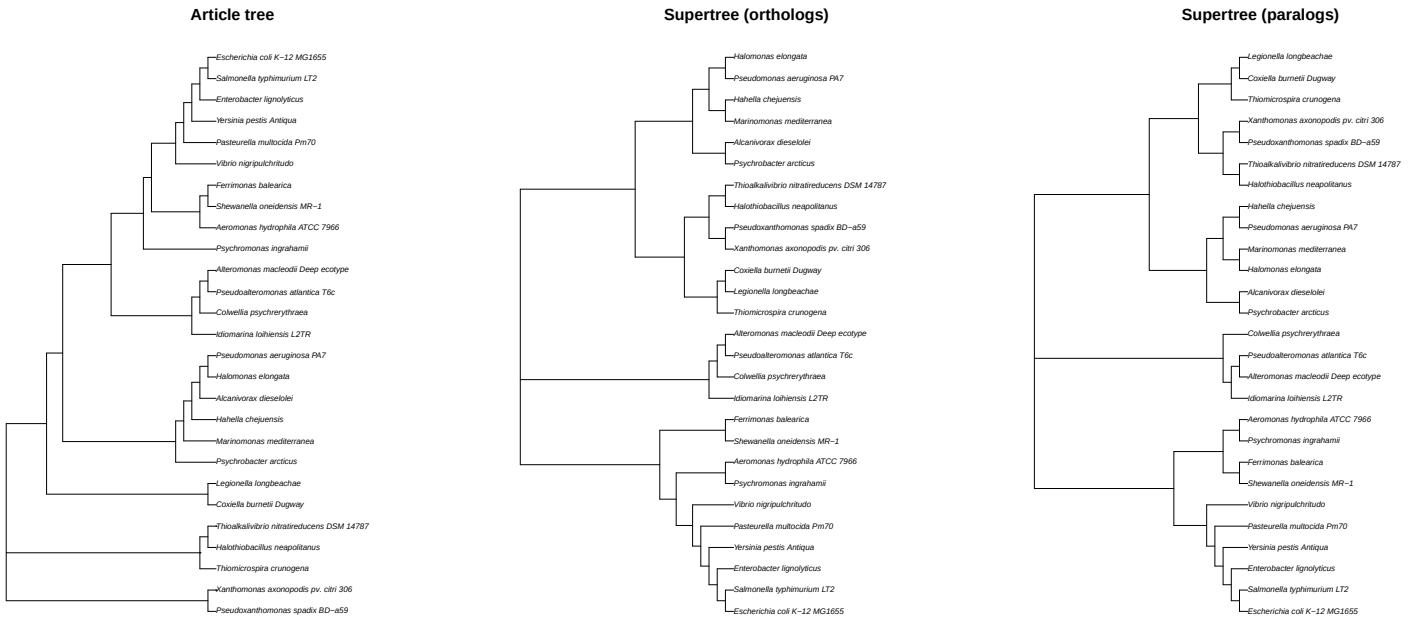
3.1 Porównanie wizualne drzew genomów

Na rys. 1 zestawiono drzewo referencyjne z pracy Sharma i in. (2022) z dwoma wariantami drzewa konsensusowego, majority-rule oraz greedy. We wszystkich trzech topologiach widoczne jest zachowanie kilku stabilnych kladów, w szczególności zwartego kladu Enterobacterales z parą *Escherichia coli* i *Salmonella typhimurium* oraz dołączającymi *Enterobacter lignolyticus* i *Yersinia pestis*, a także pary *Xanthomonas axonopodis* i *Pseudoxanthomonas spadix*. Wspólne jest również bliskie grupowanie *Ferrimonas balearica* i *Shewanella oneidensis*. Drzewo majority-rule ma ograniczoną liczbę rozgałęzień binarnych, ponieważ próg 50% eliminuje klady rzadziej obserwowane wśród drzew genowych, co prowadzi do licznych nierozdzielonych węzłów na głębszych poziomach drzewa. Wariant greedy daje topologię w pełni rozstrzygniętą, ale większość głębokich węzłów ma bardzo niskie częstości kladów (często rzędu kilkunastu lub kilkudziesięciu procent),co sugeruje ograniczoną zgodność sygnału filogenetycznego pomiędzy rodzinami genów i utrudnia interpretację relacji na

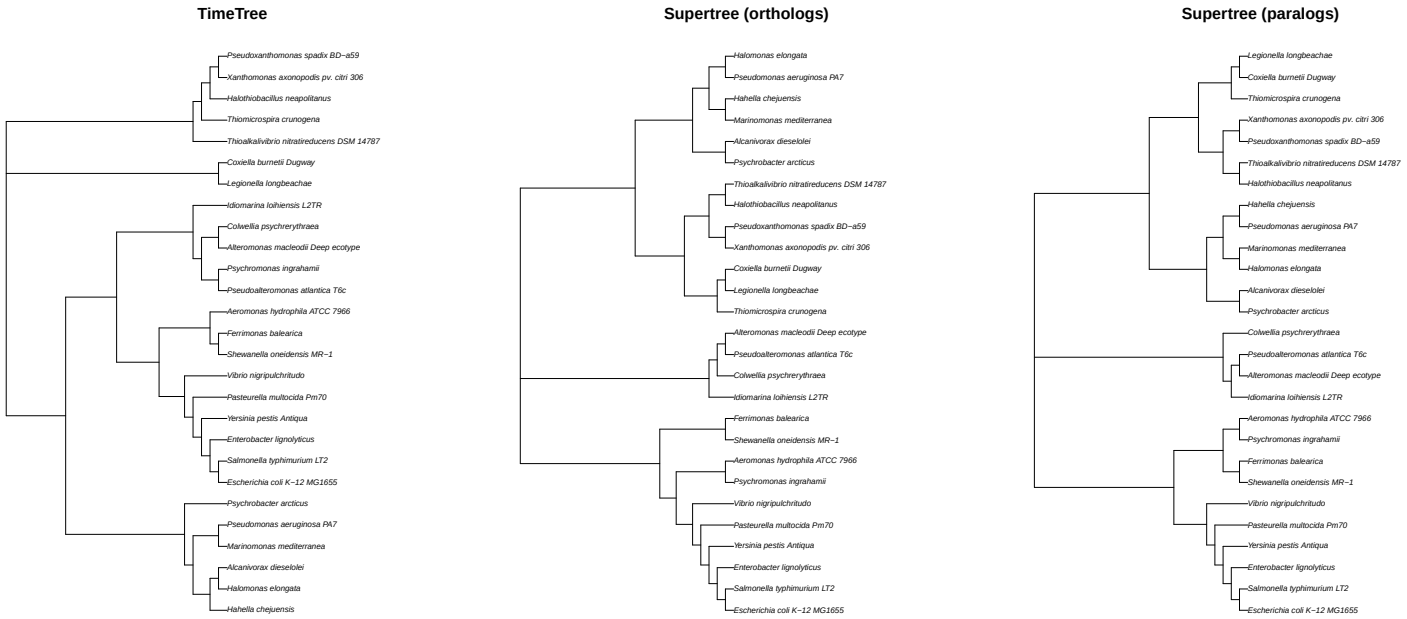
głębszych poziomach drzewa. Z kolei klady płytkie o wysokich częstościach, takie jak *Xanthomonas–Pseudoxanthomonas* oraz wewnętrzna struktura Enterobacterales, pozostają stabilne niezależnie od wariantu konsensusu.

Na rys. 2 porównano drzewo TimeTree z tymi samymi dwoma wariantami konsensusu. Również w tym zestawieniu zachowane są stabilne relacje płytkie, w tym Enterobacterales oraz para *Xanthomonas–Pseudoxanthomonas*. TimeTree grupuje ponadto taksony związane z bakteriami siarkowymi (*Halothiobacillus*, *Thiomicrospira*, *Thioalkalivibrio*) w sąsiedztwie kladu *Xanthomonas–Pseudoxanthomonas*, co jest zgodne z wariantem greedy, natomiast w majority-rule relacje w tej części drzewa pozostają w dużej mierze nierozdzielone. Podobnie, wariant greedy wyznacza uporządkowaną topologię dla grupy taksonów morskich i halofilnych (m.in. *Halomonas*, *Marinomonas*, *Hahella*, *Alcanivorax*, *Pseudomonas*), jednak niskie częstości kladów dla głębokich węzłów wskazują na konflikt sygnału między rodzinami genów i ograniczają interpretację tych relacji.

Na rys. 3 porównano drzewo z artykułu z dwoma superdrzewami, odpowiednio dla wariantu 1:1 oraz wariantu paralogicznego. W obu superdrzewach zachowane są relacje płytkie, w tym zwarty klad Enterobacterales z parą *Escherichia coli–Salmonella typhimurium* oraz dołączającymi *Enterobacter* i *Yersinia*. W obu rekonstrukcjach obecne są również małe klady zgodne z topologią z artykułu, m.in. *Xanthomonas axonopodis–Pseudoxanthomonas spadix*, *Coxiella burnetii–Legionella longbeachae* oraz *Alteromonas macleodii–Pseudoalteromonas atlantica*. Różnice między superdrzewami dotyczą m.in. położenia *Ferrimonas balearica* i *Shewanella oneidensis*, które w wariancie paralogicznym tworzą wspólną gałąź, a w wariancie 1:1 rozdzielają się na wczesnych odgałęzieniach. W wariancie paralogicznym widoczna jest też zmiana położenia *Psychrobacter arcticus* względem bloku taksonów morskich i halofilnych.



Rysunek 3: Porównanie drzewa z artykułu z superdrzewami (wariant ortologiczny oraz paralogiczny)



Rysunek 4: Porównanie drzewa z TimeTree.org z superdrzewami (wariant ortologiczny oraz paralogiczny)

Na rys. 4 przedstawiono analogiczne zestawienie, lecz z drzewem TimeTree jako punktem odniesienia. Także w tym porównaniu oba superdrzewa są bardziej zbieżne z TimeTree w obrębie małych kladów (np. Enterobacterales oraz para *Xanthomonas–Pseudoxanthomonas*), natomiast rozbieżności dotyczą przede wszystkim relacji głębokich. W szczególności wariant paralogiczny silniej odbiega od TimeTree w położeniu *Psychrobacter arcticus*, natomiast wariant 1:1 lepiej zachowuje jego przynależność do grupy taksonów morskich/halofilnych.

3.2 Porównanie ilościowe

W tab. 1 zestawiono porównanie topologii drzew genomów względem dwóch referencji, drzewa z artykułu oraz TimeTree, z użyciem miar nRF i nMCI. Niższe wartości odległośći nRF oznaczają większe podobieństwo topologii, natomiast wyższe wartości nMCI wskazują na większą zgodność podziałów kladystycznych. Dodatkowo podano porównanie samych referencji, tj. TimeTree względem drzewa z artykułu.

Największe podobieństwo do drzewa z artykułu uzyskano dla konsensusu greedy z rodzin 1:1 (nRF=0.208, nMCI=0.829). Konsensus majority-rule był wyraźnie mniej zgodny z referencją (nRF=0.500, nMCI=0.442), co jest spójne z jego mniej rozdzieloną topologią wynikającą z progu 50%. Drzewa supertree osiągnęły wartości pośrednie względem drzewa z artykułu, przy czym supertree zbudowane na podstawie rodzin ortologicznych było bliżej referencji (nRF=0.292, nMCI=0.792) niż supertree oparte na rodzinach paralogicznych (nRF=0.375, nMCI=0.766).

W porównaniu do TimeTree oba superdrzewa miały identyczne nRF (0.542) oraz zbliżone nMCI

(0.717 i 0.732). Konsensus majority-rule był w tym zestawieniu bardziej odległy według nRF (0.562) i jednocześnie wykazywał niską zgodność według nMCI (0.458), natomiast konsensus greedy był wyraźnie bliższy TimeTree (nRF=0.458, nMCI=0.752).

Tabela 1: Porównanie topologii drzew genomów względem referencji (miary nRF i nMCI)

compared / reference	consensus (majority)		consensus (greedy)		supertree orthologs		supertree paralogs		TimeTree	
	nRF	nMCI	nRF	nMCI	nRF	nMCI	nRF	nMCI	nRF	nMCI
article	0.500	0.442	0.208	0.829	0.292	0.792	0.375	0.766	0.375	0.787
TimeTree	0.562	0.458	0.458	0.752	0.542	0.717	0.542	0.732	–	–

4 Wnioski

W niniejszym projekcie zaimplementowano w pełni zautomatyzowany pipeline filogenomiki, który od pobrania proteomów z NCBI poprzez klastrowanie sekwencji, budowę MSA i rekonstrukcję drzew genowych metodą ML, prowadzi do wnioskowania drzewa genomów podejściem consensus oraz supertree. Uzyskano 4 drzewa genomów, consensus majority-rule, consensus greedy z rodzin 1:1, supertree z rodzin 1:1 oraz supertree z rodzin paralogicznych, a następnie porównano je z topologią referencyjną z literatury oraz z drzewem z TimeTree.org na drzewach nieukorzenionych.

Wyniki wskazują, że relacje na bardziej płytkich poziomach w obrębie badanego zestawu Gammaproteobacteria są odtwarzane stabilniej niż relacje głębsze. We wszystkich wariantach rekonstrukcji zachowane były powtarzalne klady o krótkich dystansach ewolucyjnych, w szczególności zwarta grupa Enterobacterales (z parą *Escherichia coli–Salmonella typhimurium* i dołączającymi *Enterobacter lignolyticus* oraz *Yersinia pestis*) oraz para *Xanthomonas axonopodis–Pseudoxanthomonas spadix*. Różnice pomiędzy drzewami koncentrowały się głównie na wzajemnym ułożeniu większych linii w obrębie klasy, co jest zgodne z obserwacjami Sharma i in. (2022) [1], że dla Proteobacteria sygnał filogenetyczny może być niespójny między markerami i metodami, szczególnie dla węzłów głębokich.

Porównanie ilościowe (tab. 1) wskazuje, że największą zgodność z drzewem z artykułu uzyskano dla konsensusu greedy. Wyższa zgodność konsensusu greedy z drzewem referencyjnym można wyjaśnić zarówno własnościami samej procedury konsensusowej, jak i charakterem punktu odniesienia. Wariant greedy, który poza kladami o częstości powyżej $\geq 50\%$ może zachowywać również dodatkowe, wzajemnie kompatybilne podziały obserwowane w mniejszej części drzew genowych, co prowadzi do większej liczby rozgałęzień i może sprzyjać współdzieleniu splitów z topologią referencyjną. Jednocześnie drzewo z pracy Sharma i in. (2022) [1] nie jest pojedynczą rekonstrukcją z jednego markera, lecz konsensusem topologii uzyskanych wieloma metodami i na różnych typach danych, co sprzyja zgodności z podejściami agregującymi informację z wielu drzew. Z tego względu lepsze dopasowanie konsensusu greedy do referencji nie musi oznaczać jednoznacznie większej wiarygodności wszystkich relacji w drzewie, szczególnie na poziomie głębokich węzłów o niskich częstościach kladów, lecz raczej większą zgodność zbioru splitów z topologią konsensusową przyjętą jako punkt odniesienia. Superdrzewa otrzymały pośrednie wyniki, natomiast konsensus majority-rule był istotnie mniej zgodny z referencją, co jest spójne z faktem, że próg 50% eliminuje część

kladów i pozostawia więcej nierozdzielonych rozgałęzień. Należy przy tym zauważyć, że metody supertree optymalizują funkcję celu związaną z kosztem rekonsyliacji, a nie bezpośrednio zgodność z zewnętrzną topologią referencyjną, dlatego niższe dopasowanie do drzewa z artykułu nie jest w tym przypadku zaskakujące i może odzwierciedlać konflikt sygnału między rodzinami genów oraz różnice między kryteriami optymalizacji. Wariant superdrzewa oparty na rodzinach paralogicznych nie poprawił zgodności z drzewem z artykułu, co wskazuje, że uwzględnienie wielu kopii genów nie zwiększyło spójności sygnału filogenetycznego w skali całego zestawu. Jednym z możliwych wyjaśnień jest to, że obecność paralogów oraz konieczne uproszczenie etykiet liści do poziomu genomu zwiększają heterogeniczność historii genowych, a tym samym utrudniają znalezienie topologii minimalizującej koszt rekonsyliacji w sposób zgodny z topologią referencyjną.

Istotnym elementem interpretacji jest niezgodność między samymi referencjami, drzewem z artykułu i drzewem z TimeTree.org. Taki wynik wskazuje, że część węzłów głębokich nie jest stabilna także między niezależnymi źródłami odniesienia, dlatego zgodność drzew uzyskanych w pipeline'ie należy oceniać z ostrożnością i nie ograniczać interpretacji do jednego punktu odniesienia oraz rozdzielać wnioski dotyczące kladów płytkich od wniosków o relacjach głębokich. W tym kontekście konsensus greedy, mimo najlepszych wartości nRF i nMCI, wymaga ostrożnej interpretacji dla głębokich rozgałęzień, ponieważ wiele odpowiadających im częstości kladów jest niska, co wskazuje na ograniczoną zgodność sygnału między rodzinami genów.

Uzyskane topologie są zgodne z obserwacjami Sharma i in. (2022) [1], że w obrębie Gammaproteobacteria pozycja *Pseudomonas aeruginosa* bywa niestabilna, a *Pseudomonadales* nie zawsze tworzą klad monofiletyczny. Autorzy raportują, że *P. aeruginosa* wykazuje tendencję do grupowania z liniami Oceanospirillales w wielu rekonstrukcjach. W analizowanym podzbiorze taksonów *P. aeruginosa* również lokuje się w sąsiedztwie taksonów morskich i halofilnych, takich jak *Halomonas elongata*, *Alcanivorax dieselolei*, *Marinomonas mediterranea* oraz *Hahella chejuensis*, co wskazuje, że sygnał dla tej relacji utrzymuje się niezależnie od tego, czy superdrzewo wyznaczano na rodzinach 1:1, czy na rodzinach paralogicznych.

Jednocześnie porównanie z topologią z TimeTree.org [2] pokazuje, że różnice między źródłami odniesienia koncentrują się na relacjach głębokich, natomiast zgodność jest wyraźniejsza w obrębie małych kladów. W konsekwencji interpretację położenia większych linii w obrębie Gammaproteobacteria należy traktować jako mniej jednoznaczną niż wnioski dotyczące relacji w obrębie blisko spokrewnionych grup.

Ograniczenia analizy wynikają głównie z zakresu danych oraz z przyjętego zestawu metod. Zastosowano podzbiór 27 taksonów, co ogranicza rozdzielczość części relacji głębokich w obrębie Gammaproteobacteria. Dodatkowo, drzewa genowe rekonstruowano bez bootstrappingu i nie przeprowadzono filtracji rodzin na podstawie stabilności topologii. Ponadto rekonstrukcję drzewa genomów oparto na jednym typie danych i jednym źródle sygnału filogenetycznego, tj. drzewach rodzin genów wyznaczonych na podstawie klastrowania sekwencji białkowych z pełnych proteomów. Podczas gdy Sharma i in. zestawiali kilka niezależnych podejść m.in 16S rRNA, drzewo na podstawie AAI, drzewa markerów białkowych oraz konkatenację genów konserwatywnych i interpretowali zgodność kladów między metodami Sharma i in. (2022) [1].

W dalszej pracy nad pipeline'em najbardziej uzasadnione byłoby: (i) wprowadzenie bootstrappingu w drzewach genowych oraz filtracji rodzin o niskiej stabilności topologii, (ii) rozszerzenie rekonstrukcji drzewa genomów o dodatkowe, niezależne podejścia, analogicznie do strategii zastosowanej przez Sharma i in. (2022) [1], w tym wariant konkatenacyjny dla konserwatywnych markerów.

Bibliografia

[1] Vaibhav Sharma i in. "Phylogenomics of the Phylum Proteobacteria: Resolving the Complex Relationships". W: *Current Microbiology* 79 (2022). DOI: 10.1007/s00284-022-02910-9.

[2] Sudhir Kumar i in. "TimeTree 5: An Expanded Resource for Species Divergence Times". W: *Molecular Biology and Evolution* 39.8 (2022), msac174. DOI: 10.1093/molbev/msac174. URL: https://timetree.org/.

[3] Nuala A. O'Leary i in. "Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets". W: *Scientific Data* 11.1 (2024), s. 732. DOI: 10.1038/s41597-024-03571-y.

[4] Martin Steinegger i Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". W: *Nature Biotechnology* 35.11 (2017), s. 1026–1028. DOI: 10.1038/nbt.3988.

[5] Martin Steinegger i Johannes Söding. "Clustering huge protein sequence sets in linear time". W: *Nature Communications* 9 (2018), s. 2542. DOI: 10.1038/s41467-018-04964-5.

[6] Felix Kallenborn i in. "GPU-accelerated homology search with MMseqs2". W: *Nature Methods* 22 (2025), s. 2024–2027. DOI: 10.1038/s41592-025-02819-8.

[7] Kazutaka Katoh i Daron M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". W: *Molecular Biology and Evolution* (2013). DOI: 10.1093/molbev/mst010.

[8] Bui Quang Minh i in. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era". W: *Molecular Biology and Evolution* (2020). DOI: 10.1093/molbev/msaa015.

[9] Emmanuel Paradis, Julien Claude i Korbinian Strimmer. "APE: Analyses of Phylogenetics and Evolution in R language". W: *Bioinformatics* 20.2 (2004), s. 289–290. DOI: 10.1093/bioinformatics/btg412.

[10] Paweł Górecki, J. Gordon Burleigh i Oliver Eulenstein. "GTP Supertrees from Unrooted Gene Trees: Linear Time Algorithms for NNI Based Local Searches". W: *Bioinformatics Research and Applications (ISBRA 2012)*. T. 7292. Lecture Notes in Computer Science. Springer, 2012, s. 102–114. DOI: 10.1007/978-3-642-30191-9_11.

[11] Paweł Górecki i Jerzy Tiuryn. "Inferring phylogeny from whole genomes". W: *Bioinformatics* 23.2 (2007), e116–e122. DOI: 10.1093/bioinformatics/btl296.

[12] Paweł Górecki i Jerzy Tiuryn. "URec: a system for unrooted reconciliation". W: *Bioinformatics* 23.4 (2007), s. 511–512. DOI: 10.1093/bioinformatics/btl634.

[13] Martin R. Smith. "Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees". W: *Bioinformatics* 36.7 (2020), s. 2047–2055. DOI: 10.1093/bioinformatics/btz875.