Pipeline filogenetyczny

Julia Świątkowska

1. Wstęp

Raport przedstawia proces filogenetycznej analizy porównawczej opartej o dane genomowe, Analiza została przeprowadzona dla 26 gatunków należących do rzędu *Enterobacterales*. W ostatnich latach taksonomia bakterii uległa istotnej zmianie - od klasyfikacji opartej na cechach fenotypowych i pojedynczych markerach molekularnych, takich jak gen 16S rRNA, w kierunku filogenomiki wykorzystującej dane całogenomowe. Rząd *Enterobacterales*, obejmujący liczne organizmy o znaczeniu klinicznym i środowiskowym, stanowi dobry model tej transformacji, ponieważ tradycyjne analizy 16S rRNA często nie pozwalały na jednoznaczne rozdzielenie głównych linii ewolucyjnych w obrębie tej grupy.

Punktem wyjścia dla niniejszego projektu była praca pt. “Genome-based phylogeny and taxonomy of the ‘Enterobacteriales’” [Adeolu et al., 2016], w której zaproponowano gruntowną rewizję taksonomii rzędu Enterobacterales w oparciu o analizę filogenomiczną genomów.

W celu rekonstrukcji relacji ewolucyjnych autorzy zastosowali dwa komplementarne podejścia. Po pierwsze, skonstruowali drzewa filogenetyczne na podstawie: (i) konkatenacji 1548 białek rdzeniowych, (ii) 53 konserwatywnych białek rybosomalnych oraz (iii) czterech genów wykorzystywanych w analizach MLSA (gyrB, rpoB, atpD, infB). Wszystkie analizy wykazały istnienie siedmiu głównych, silnie wspieranych kladów. Po drugie jako niezależne potwierdzenie monofiletyzmu tych grup, zidentyfikowano specyficzne konserwowane insercje i delecje (CSI) w sekwencjach białkowych - wspólne dla całego rzędu oraz charakterystyczne dla poszczególnych kladów.

Na podstawie połączenia danych filogenomicznych i cech molekularnych zaproponowano podział rzędu *Enterobacterales* na siedem rodzin: *Enterobacteriaceae, Erwiniaceae, Pectobacteriaceae, Yersiniaceae, Hafniaceae, Morganellaceae* oraz *Budviciaceae*.

Krótki opis bakterii wybranych z publikacji:

Rodzina *Enterobacteriaceae*

- *Salmonella enterica:* Patogen jelitowy ludzi/zwierząt, zatrucia pokarmowe.
- *Escherichia coli:* Komensal jelit; niektóre szczepy to patogeny (zatrucia, zakażenia dróg moczowych).
- *Enterobacter cloacae:* Oportunistyczny patogen szpitalny (zakażenia płuc, ran, sepsa).
- *Klebsiella pneumoniae:* Oportunistyczny patogen szpitalny (zapalenie płuc), znany z oporności na antybiotyki.

Rodzina *Erwiniacea*

- *Erwinia amylovora:* Sprawca zarazy ogniowej jabłoni i grusz (niszczycielska choroba roślin).
- *Pantoea agglomerans:* Bakteria wszechobecna (rośliny, środowisko); może być patogenem roślin i oportunistą u ludzi.
- *Tatumella ptyseos:* Rzadki oportunista, izolowany głównie z dróg oddechowych.
- *Erwinia billingiae:* Zazwyczaj niepatogeniczny komensal/endofit roślin.

Rodzina *Pectobacteriaceae*

- *Pectobacterium carotovorum:* Fitopatogen powodujący mokrą zgniliznę warzyw (np. ziemniaka, marchwi).
- *Pectobacterium atrosepticum:* Wyspecjalizowany patogen ziemniaka (czarna nóżka, zgnilizna bulw).
- *Dickeya chrysanthemi / dadantii:* Agresywne fitopatogeny powodujące zgniliznę miękką wielu roślin.

Rodzina *Yersiniaceae*

- *Ewingella americana:* Rzadki oportunista człowieka (zakażenia krwi).
- *Yersinia bercovieri:* Niepatogenicznna dla człowieka, izolowana ze środowiska.
- *Serratia marcescens:* Oportunistyczny patogen szpitalny (zakażenia dróg moczowych, ran). Często wytwarza czerwony pigment.
- *Yersinia pestis:* Czynnik etiologiczny dżumy. Przenoszony przez pchły.

Rodzina *Hafniaceae*

- *Hafnia alvei:* Bakteria jelitowa, rzadki oportunista.
- *Edwardsiella tarda:* Patogen ryb i gadów; u ludzi powoduje głównie problemy żołądkowo-jelitowe.
- *Edwardsiella ictaluri:* Śmiertelny patogen ryb (posocznica u sumów).
- *Obesumbacterium proteus:* Zanieczyszczenie piwa.

Rodzina *Morganellaceae*

- Proteus mirabilis: Częsty sprawca skomplikowanych zakażeń układu moczowego (zwłaszcza z kamicą).
- *Morganella morganii:* Oportunistyczny patogen (zakażenia układu moczowego, ran).
- *Providencia stuartii:* Patogen szpitalny, problematyczny w zakażeniach z cewnikowaniem.
- *Xenorhabdus nematophila:* Symbiont nicieni owadobójczych;

pomaga zabijać owady.

Rodzina *Budviciaceae*

- *Leminorella grimontii:* Rzadki oportunista, izolowany z próbek klinicznych.
- *Pragia fontium:* Bakteria środowiskowa, izolowana z wody.

2.1. Metody

2.1.1. Pobieranie proteomów bakteryjnych z bazy danych NCBI

Proteomy bakteryjne pobrano z bazy NCBI przy użyciu interfejsu Entrez API z pakietu BioPython [Cock et al., 2009]. Dla każdego z 26 gatunków z listy wejściowej, skrypt najpierw wyszukiwał najnowsze, referencyjne genomy w RefSeq, a w przypadku ich braku korzystał z ogólnej bazy Assembly (w tym GenBank), zapewniając maksymalną dostępność i jakość danych.
Pobrane pliki proteomów (*_protein.faa.gz) były dekompresowane, a nagłówki FASTA standaryzowane poprzez dodanie prefiksu z nazwą gatunku (np. >Escherichia_coli_), co umożliwiało jednoznaczne przypisanie sekwencji do gatunku w dalszych etapach analizy. Skrypt obsługiwał błędy dla poszczególnych gatunków oraz wprowadzał krótką przerwę między zapytaniami do NCBI, aby uniknąć przeciążenia serwerów.
Wszystkie proteomy zapisano w pojedynczym zbiorczym pliku FASTA. Łącznie pobrano 103 504 sekwencje białkowe z kompletnych genomów.

2.1.2. Klastrowanie sekwencji

Analizę wszystkich pobranych białek przeprowadzono przy użyciu narzędzia MMseqs2 [Steinegger & Söding, 2017]. Na początku przygotowano bazę danych, tworząc ją z pliku FASTA zawierającego wszystkie proteomy za pomocą polecenia `mmseqs createdb`. Następnie wykonano porównanie all-vs-all sekwencji przy zastosowaniu progu istotności statystycznej e-value ≤ 1×10^{-5}, wykorzystując do tego cztery wątki CPU. Kolejnym etapem była klasteryzacja sekwencji z minimalnym podobieństwem 80% i minimalnym pokryciem alignmentu również 80%, przy użyciu algorytmu greedy set cover oraz trybu pokrycia, w którym procent pokrycia obliczany jest jako stosunek długości alignmentu do długości dłuższej sekwencji. Na końcu wygenerowano plik TSV z przypisaniem sekwencji do poszczególnych klastrów. Pliki tymczasowe były automatycznie usuwane zarówno przed rozpoczęciem analizy (stare bazy danych), jak i po jej zakończeniu (cały katalog roboczy), co pozwoliło zoptymalizować wykorzystanie przestrzeni dyskowej. W wyniku całego procesu powstało 60 805 klastrów sekwencji o wysokim podobieństwie (≥80% identyczności), które reprezentują grupy potencjalnych ortologów, a końcowym wynikiem był plik TSV przypisujący sekwencje do reprezentatywnych klastrów.

2.1.3 Rodziny genów (przypadek 1-1)

Z przeprowadzonych analiz klasteryzacji wyselekcjonowano rodziny ortologiczne typu 1:1. W pierwszym kroku przetworzono strukturę klastrów, tworząc pełną listę grup homologicznych

i identyfikując genom pochodzenia każdego genu na podstawie jego identyfikatora (np. `Salmonella_enterica` z `Salmonella_enterica_NP_123456`).
Kryteria selekcji rodzin ortologicznych obejmowały:

- unikalność genów w genomie - każdy genom mógł być reprezentowany w klastrze przez dokładnie jeden gen,
- kompletność reprezentacji - klastry musiały zawierać geny odpowiadające wszystkim analizowanym gatunkom,
- jednoznaczność przypisania - odrzucono klastry zawierające duplikaty lub brakujące genomy, aby zapewnić spójność rodzin 1:1.

W wyniku analizy zidentyfikowano 229 rodzin ortologicznych obecnych we wszystkich 26 gatunkach bakteryjnych. Skrypt wygenerował cztery pliki wynikowe zawierające przypisanie genów do rodzin ortologicznych, kompozycje genomową każdej rodziny, sekwencje reprezentatywne rodzin, wszystkie geny z wybranych rodzin przygotowane do wielosekwencyjnego wyrównania sekwencji.

2.1.4. Ekstrakcja sekwencji i tworzenie plików FASTA dla rodzin ortologicznych

Dla każdej z rodzin ortologicznych wygenerowano oddzielny plik FASTA zawierający sekwencje aminokwasowe genów należących do danej rodziny.
Sekwencje odpowiadające wszystkim zebranym identyfikatorom genowym wyekstrahowano z pliku na podstawie identyfikatorów występujących w nagłówkach FASTA. Odnalezione sekwencje zapisano w strukturze mapującej identyfikator genu na sekwencję aminokwasową, przy jednoczesnym raportowaniu brakujących wpisów.
Dla każdej rodziny utworzono osobny plik FASTA w dedykowanym katalogu nazwany sekwencyjnie (family_0001.fasta, family_0002.fasta, …). Pliki zawierały linię komentarza z numerem rodziny i liczbą reprezentowanych gatunków oraz sekwencje w formacie FASTA z nagłówkami ograniczonymi do nazw gatunków wyekstrahowanych z identyfikatorów genów. Rodziny, dla których nie znaleziono żadnych sekwencji, były automatycznie pomijane.
W wyniku działania skryptu utworzono 229 plików FASTA, które stanowiły dane wejściowe do dalszych analiz porównawczych i wielosekwencyjnego wyrównania sekwencji.

2.1.5. Wielosekwencyjne wyrównanie rodzin ortologicznych

Wielosekwencyjne wyrównania (MSA) dla wszystkich rodzin ortologicznych wykonano przy użyciu programu MAFFT [Katoh, 2002] w trybie automatycznym, który pozwala programowi samodzielnie dobrać optymalną strategię wyrównania w zależności od liczby i podobieństwa sekwencji w każdej rodzinie. Na początku przygotowano dane wejściowe, wczytując wszystkie pliki FASTA zawierające sekwencje aminokwasowe poszczególnych rodzin ortologicznych. Następnie utworzono dedykowany katalog tymczasowy, w którym przechowywano pliki robocze generowane przez MAFFT [Katoh, 2002]. Kolejnym etapem było wykonanie wyrównań – każdy

plik FASTA przetworzono niezależnie w trybie równoległym z wykorzystaniem czterech wątków, a wynikowe pliki zapisano w wyznaczonym katalogu pod nazwami `aligned_family_XXXX.fasta`. Po zakończeniu wszystkich wyrównań katalog tymczasowy został automatycznie usunięty. W efekcie całego procesu powstało 229 poprawnie wyrównanych plików FASTA.

2.1.6. Konstrukcja drzew filogenetycznych rodzin genowych

Dla każdej z wyrównanych rodzin ortologicznych zrekonstruowano drzewa filogenetyczne metodą maksymalnej wiarygodności (ML) przy użyciu programu FastTree. Na początku przygotowano dane wejściowe, wczytując wszystkie pliki wyrównań znajdujące się w katalogu zawierającym wielosekwencyjne dopasowania aminokwasowe rodzin ortologicznych. Następnie każdy plik wyrównania przetworzono niezależnie w trybie równoległym z wykorzystaniem czterech wątków. FastTree zastosowano z modelem ewolucyjnym LG [Le & Gascuel, 2008], uznawanym za domyślny dla dużych i zróżnicowanych zestawów sekwencji białkowych, oraz z uwzględnieniem zmienności tempa ewolucji w różnych pozycjach sekwencji poprzez przyjęcie rozkładu gamma. Wynikowe drzewa zapisano w formacie Newick do tymczasowych plików, które następnie scalono w jeden zbiorczy plik, a pliki tymczasowe zostały usunięte w celu oszczędności przestrzeni dyskowej. W efekcie całego procesu powstało 229 drzew filogenetycznych połączonych w jeden plik.

2.1.7. Konstrukcja drzewa konsenusowego zachłannego

Aby uzyskać reprezentatywne drzewo konsensusowe z zestawu wcześniej zrekonstruowanych drzew filogenetycznych, zastosowano metodę zachłannego konsensusu (greedy consensus) przy użyciu programu IQ-TREE [Nguyen et al., 2015]. Na początku przygotowano dane wejściowe, wczytując plik, który zawierał wszystkie indywidualne drzewa filogenetyczne wygenerowane w poprzednim etapie. Następnie program analizował te drzewa i generował jedno drzewo zachłannego konsensusu, uwzględniając wszystkie klady i zachowując spójność struktury filogenetycznej; proces ten został przeprowadzony z użyciem opcji `-con`, odpowiadającej budowie drzewa majority-rule extended (greedy consensus). Powstałe drzewo konsensusowe zapisano w formacie Newick, dzięki czemu otrzymano jedno, reprezentatywne drzewo filogenetyczne obejmujące cały zbiór analizowanych rodzin genowych.

2.1.8. Konstrukcja drzewa majory-rule oraz analiza drzew konsenusowych

W kolejnym etapie przeprowadzono analizę i wizualizację drzew konsensusowych przy użyciu języka R oraz pakietów ape [Paradis and Schliep, 2019], phangorn [Schliep, 2011] i TreeDist [Smith, 2020]. Wczytano wszystkie drzewa wejściowe i ujednolicono zestawy taksonów, pozostawiając wspólne taksony i przycinając drzewa w celu zachowania zgodności. Następnie zbudowano drzewo majority-rule z progiem 50% i przypisano wartości wsparcia do węzłów, a jego strukturę zapisano w formacie Newick i zwizualizowano.

Dodatkowo wczytano wcześniej utworzone drzewo greedy consensus, które również wizualizowano i porównano z drzewem referencyjnym z publikacji [Adeolu et al., 2016] oraz pobranym z TimeTree. Porównania wykonano przy użyciu metryki Robinson-Foulds (RF), zarówno w formie surowej, jak i znormalizowanej, po wcześniejszym przycięciu drzew do wspólnych taksonów. W wyniku procesu powstały wizualizacje dla obu drzew konsenusowych,i plik Newick dla drzewa majority-rule oraz zestaw metryk RF umożliwiających ocenę spójności i różnic strukturalnych między drzewami.

2.1.9. Konstrukcja supertree

Na podstawie zrekonstruowanych drzew rodzin genowych utworzono superdrzewo gatunków przy użyciu programu ASTRAL. Wszystkie drzewa rodzin genowych zapisano w pliku wejściowym, umieszczając jedno drzewo w każdej linii. Następnie zbudowano superdrzewo w trybie obsługi długości gałęzi (`branch_length_mode = 2`), co pozwala oszacować wartości wsparcia dla poszczególnych węzłów w oparciu o zestaw drzew rodzinnych. Skrypt automatycznie wykrywał lokalizację pliku JAR ASTRAL lub korzystał z instalacji dostępnej w PATH, przydzielając 8 GB pamięci dla procesu Java. Wygenerowane superdrzewo zapisano w formacie Newick, a cały proces był monitorowany pod kątem błędów, przy czym tymczasowe pliki używane przez ASTRAL były automatycznie usuwane po zakończeniu obliczeń.

2.1.10. Analiza supertee

W kolejnym etapie oceniono topologiczną zgodność superdrzewa z drzewem referencyjnym z publikacji [Adeolu et al., 2016] oraz pobranym z TimeTree przy użyciu języka R oraz pakietów ape [Paradis and Schliep, 2019], phangorn [Schliep, 2011] i TreeDist [Smith, 2020]. Na początku superdrzewo oraz drzewa referencyjne przekształcono do wspólnego zestawu taksonów, tak aby wszystkie drzewa obejmowały te same gatunki. Następnie obliczono odległość Robinsona-Fouldsa (RF), mierząc topologiczną zgodność superdrzewa względem drzew referencyjnych oraz timetree, przy czym uwzględniono zarówno odległość surową, jak i znormalizowaną. Superdrzewo i drzewo referencyjne zapisano w formie wykresów w katalogu z wizualizacjami, a węzły superdrzewa oznaczono wartościami wsparcia przeskalowanymi względem liczby drzew rodzinnych, co umożliwia szybką ocenę stabilności poszczególnych kladów. W efekcie całego procesu powstał raport topologicznej zgodności superdrzewa z drzewami referencyjnymi oraz zestaw wizualizacji pozwalający ocenić spójność i stabilność kladów w odniesieniu do istniejącej filogenezy.

2.2. Opis techniczny

Wszystkie obliczenia zostały wykonane na komputerze MacBook Air (MacBookAir10,1) z procesorem Apple M1 × 8 (4 rdzenie wydajnościowe i 4 rdzenie efektywne) i 16 GiB pamięci RAM.

Czas działania poszczególnych etapów pipeline'u został przedstawiony w Tabeli 1. Cały pipeline obejmuje skrypty napisane zarówno w języku Python oraz R. Pipeline można wywołać za pomocą pojedynczego skryptu sh (run_all.sh).

Poszczególne etapy	Czas działania
Pobieranie proteomów bakteryjnych z bazy danych NCBI	895 s
Klastrowanie sekwencji	1809 s
Rodziny genów (przypadek 1-1)	22 s
Ekstrakcja sekwencji i tworzenie plików FASTA dla rodzin ortologicznych	<1 s
Wielosekwencyjne wyrównanie rodzin ortologicznych	35 s
Konstrukcja drzew filogenetycznych rodzin genowych	17 s
Konstrukcja drzewa konsenusowego zachłannego	1 s
Konstrukcja drzewa majory-rule oraz analiza drzew konsenusowych	1 s
Konstrukcja supertree	1 s
Analiza supetree	2 s

Tabela 1. Czas działania poszczególnych etapów pipeline'u.

3. Wyniki

Na podstawie 229 indywidualnych drzew genowych wygenerowano drzewo konsensusowe większościowe, zachłanne drzewo konsensusowe oraz superdrzewo. Następnie porównano topologie uzyskanych drzew z drzewem referencyjnym z publikacji [Adeolu et al., 2016] oraz drzewem pobranym z bazy TimeTree. Porównania przeprowadzono dla wspólnego zbioru taksonów. W przypadku drzewa czasowego analiza obejmowała 24 gatunki, ponieważ *Ewingella americana* oraz *Erwinia billingiae* nie były dostępne w bazie TimeTree. Dla pozostałych analiz uwzględniono pełny zestaw 26 gatunków. Wyniki porównań topologicznych porównano ilościowo wyrażone za pomocą odległości Robinsona-Fouldsa (RF), zestawione w Tabeli 2. oraz jakościowo porównując wizualizacje drzew. Surowa odległość RF odpowiada liczbie niespójnych podziałów (bipartycji) pomiędzy porównywanymi drzewami i stanowi bezpośrednią miarę różnic topologicznych. Znormalizowana odległość RF została obliczona jako stosunek wartości RF do maksymalnej możliwej odległości dla danej liczby taksonów.

Drzewo konsensusowe większościowe charakteryzowało się wyższymi wartościami znormalizowanej odległości RF w porównaniu z drzewem referencyjnym i drzewem czasowym. Należy jednak zaznaczyć, że podczas konstrukcji tego drzewa pojawiał się komunikat: „Some

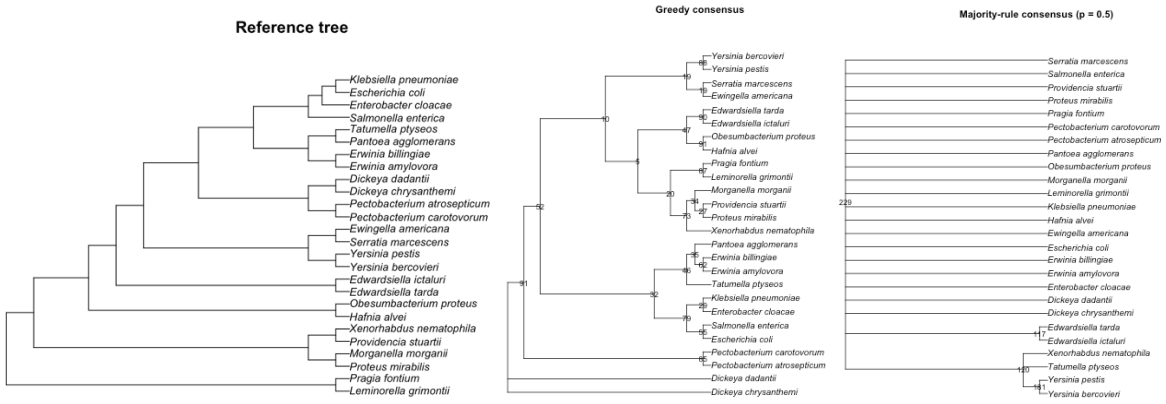
trees are not binary. Result may not what you expect!", co oznacza, że RF należy traktować ostrożnie i nie interpretować go dosłownie, ponieważ obecność politomii może sztucznie zawyżać odległości. Najniższe wartości RF uzyskano w porównaniach superdrzewa i zachłannego drzewa konsensusowego z drzewem referencyjnym, co wskazuje na dużą zgodność topologiczną. Wyższe wartości RF obserwowano natomiast w porównaniach z drzewem czasowym, co odzwierciedla różnice wynikające z ograniczeń danych czasowych i braku niektórych taksonów w bazie TimeTree.

Porównywane drzewa	**Liczba wspólnych taksonów**	**Surowe RF**	**Znormalizowane RF**	**Maksymalne możliwe RF**
Konsensus większościowy vs drzewo referencyjne	26	22	0.846	46
Konsensus większościowy vs drzewo czasowe	24	20	0.833	42
Konsensus zachłanny vs drzewo referencyjne	26	12	0.261	46
Konsensus zachłanny vs drzewo czasowe	24	24	0.571	42
Supertree vs drzewo referencyjne	26	12	0.261	46
Supertree vs drzewo czasowe	24	24	0.571	42

Tabela 2. Porównanie topologii drzew filogenetycznych za pomocą odległości Robinsona-Fouldsa (RF).

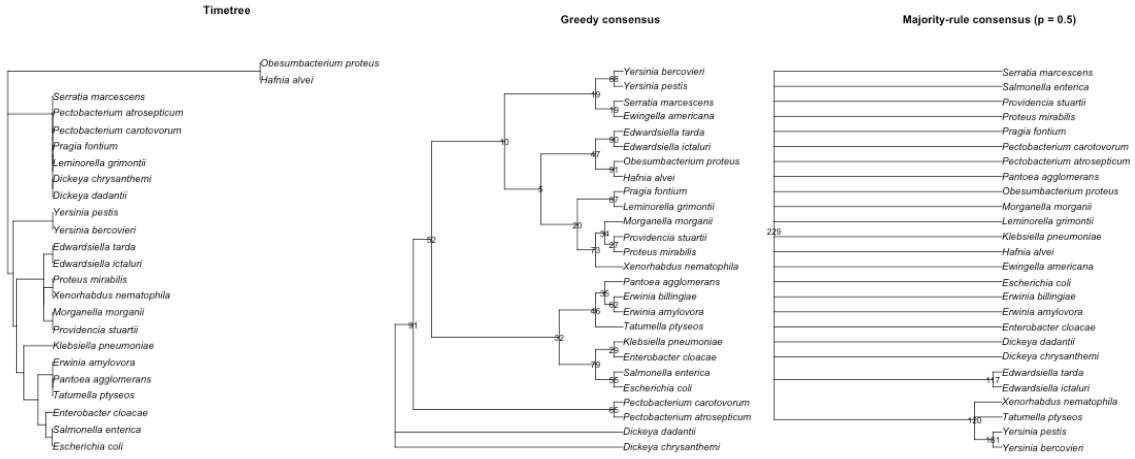
Analiza wizualna topologii ujawnia zarówno zgodność, jak i istotne rozbieżności w odtwarzaniu relacji filogenetycznych pomiędzy drzewem referencyjnym a konsensusami. Drzewo konsenusowe większościowe charakteryzuje się bardzo niską rozdzielczością i obecnością rozległych politomii (węzłów, z których wychodzi więcej niż dwóch potomków). Struktura drzewa jest w znacznej mierze „płaska" (grzebieniasta), co oznacza, że dla większości głębokich węzłów nie udało się uzyskać poparcia w ponad 50% pojedynczych drzew genowych. Drzewo poprawnie zgrupowało jedynie pary gatunków takie jak *Yersinia pestis* i *Yersinia bercovieri* oraz *Edwardsiella tarda* i *Edwardsiella ictaluri.*

Główna różnica między drzewem konsensusowym zachłannym a referencyjnym dotyczy rodziny Pectobacteriaceae, która jest rozdzielona w drzewie zachłannym. Para *Dickeya dadantii* i *Dickeya chrysanthemi* stanowi najbardziej bazalną linię ewolucyjną całego drzewa (grupę zewnętrzną dla wszystkich pozostałych gatunków). Pomijając ten wyjątek drzewo zachłanne wykazuje wysoki stopień zgodności w zakresie składu i monofiletyzmu niemal wszystkich pozostałych rodzin rzędu *Enterobacterales.*



Rysunek 1. Porównanie drzew konsensusowych (zachłannego i większościowego) z drzewem z publikacji [Adeolu et al., 2016].

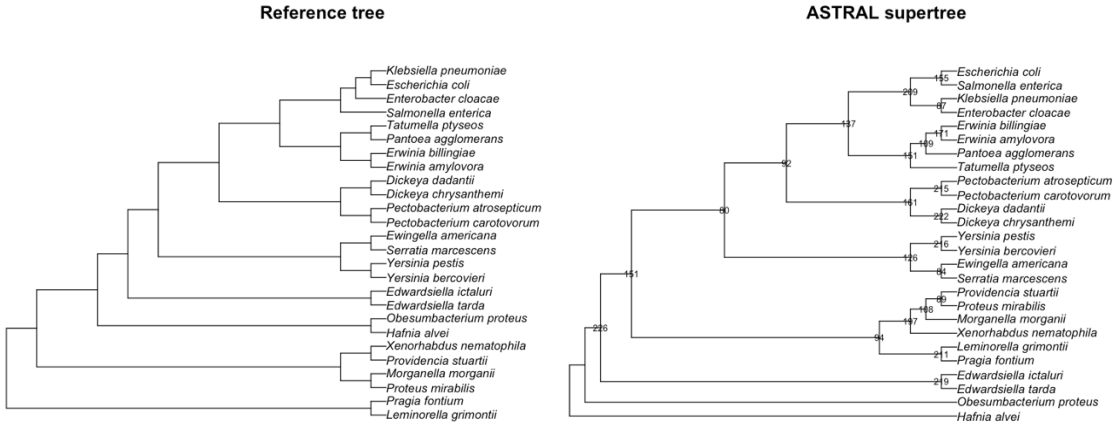
W bazie TimeTree brakowało dwóch taksonów: *Ewingella americana* oraz *Erwinia billingiae.* Różnice pomiędzy TimeTree, a drzewami konsensusowymi są znacznie większe. Podobnie jak w poprzednim porównaniu, drzewo większościowe wykazuje bardzo niską zgodność z drzewem TimeTree. Główną robieżnością jest nierozdzielony blok mieszany (*Serratia - Pectobacteriaceae - Budviciaceae*) w drzewie TimeTree. Gatunki *Serratia marcescens*, *Pectobacterium* (*P. carotovorum, P. atrosepticum*), *Dickeya* (*D. dadantii, D. chrysanthemi*) oraz przedstawiciele rodziny *Budviciaceae* (*Pragia fontium, Leminorella grimontii*) tworzą jedną, wspólną grupę wychodzącą z głównej osi drzewa. W drzewie TimeTree *Klebsiella pneumoniae* zajmuje pozycję bazalną względem grupy obejmującej m.in. *Erwinia* i *Pantoea*, co sugeruje polifiletyzm rodziny *Enterobacteriaceae* w tym ujęciu. W drzewie zachłannym *Klebsiella* jest poprawnie zagnieżdżona wewnątrz kladu *Enterobacteriaceae.*



Rysunek 2. Porównanie drzew konsensusowych (zachłannego i większościowego) z drzewem z TimeTree.

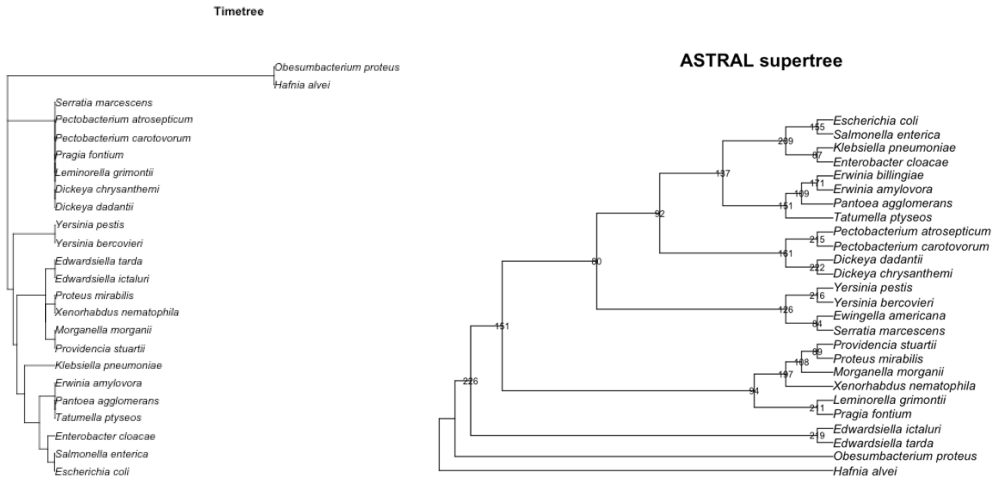
Metoda superdrzewowa, w przeciwieństwie do konsensusowej (zachłannej) poradziła sobie gorzej z odwzorowaniem rodziny *Hafniaceae - Hafnia alvei* i *Obesumbacterium* występują jako

kolejne, osobne odgałęzienia, co jest niezgodne z układem referencyjnym. Jednak zdecydowanie lepiej rekonstruuje rodzinę *Pectobacteriaceae* - rodzaje *Dickeya* (*D. dadantii, D. chrysanthemi*) oraz *Pectobacterium* (*P. carotovorum, P. atrosepticum*) tworzą jeden wspólny klad.



Rysunek 3. Porównanie drzewa supertree z drzewem z publikacji [Adeolu et al., 2016].

Zasadnicza różnica dotyczy rozdzielczości głębokich węzłów: superdrzewo poprawnie segreguje rodziny *Pectobacteriaceae*, *Yersiniaceae* i *Budviciaceae*, które w bazie TimeTree tworzą jedną, nierozwiązaną politomię. Superdrzewo wiarygodniej umieszcza *Klebsiella pneumoniae* wewnątrz rodziny *Enterobacteriaceae*, korygując jej nietypową, odseparowaną pozycję widoczną w bazie TimeTree. Z kolei TimeTree połowicznie lepiej odzwierciedla relacje wewnątrz rodziny *Hafniaceae*, łącząc *Hafnia alvei* i *Obesumbacterium proteus* w parę siostrzaną (czego superdrzewo nie odtworzyło), jednak w obu przypadkach rodzina ta nie zachowuje monofiletyzmu, gdyż rodzaj *Edwardsiella* został umieszczony na odrębnej gałęzi.



Rysunek 4. Porównanie drzewa supertree z drzewem z TimeTree.

4. Wnioski

Przeprowadzona rekonstrukcja relacji ewolucyjnych dla 26 gatunków z rzędu *Enterobacterales*, oparta na analizie 229 rodzin ortologicznych, wykazała zróżnicowaną skuteczność testowanych metod. Najbardziej wiarygodne rezultaty, zbliżone do współczesnej wiedzy taksonomicznej, uzyskano przy użyciu metody superdrzewa (ASTRAL) oraz konsensusu zachłannego (greedy). Podejścia te osiągnęły najwyższą zgodność topologiczną z drzewem referencyjnym (odległość RF wynosząca 0.261), poprawnie odtwarzając kluczowe rodziny.

Obserwowane rozbieżności między tymi wynikami a rewizją taksonomiczną z publikacji [Adeolu et al., 2016] wynikają z fundamentalnej różnicy w zakresie i głębi analizy. Podczas gdy niniejszy projekt oparł się na pojedynczej analizie ortologów, praca referencyjna zastosowała wielopoziomowe, komplementarne podejście, obejmujące: (i) analizę ilościową 1548 białek rdzeniowych, (ii) niezależne weryfikacje na podstawie 53 białek rybosomalnych oraz czterech genów MLSA, oraz (iii) identyfikację jakościowych synapomorfii molekularnych w postaci specyficznych konserwowanych insercyj i delecji (CSI). Ta kompleksowa strategia zapewniła niezwykle solidne podstawy dla rewizji taksonomicznej. Selekcja 229 rodzin ortologicznych typu 1:1, choć zapewniała porównywalność genomów, mogła nie być wystarczająco restrykcyjna pod względem odporności na horyzontalny transfer genów (HGT). Geny spełniające kryterium "obecności u wszystkich gatunków" niekoniecznie muszą być genami "rdzeniowymi" o niskiej skłonności do transferu poziomego.

Zdecydowanie najsłabszym elementem analizy okazało się drzewo konsensusowe większościowe (majority-rule), które przyjęło strukturę grzebieniastą o minimalnej rozdzielczości. Niepowodzenie to wynika bezpośrednio z mechanizmu działania tej metody, która wymaga, aby dany węzeł (podział) był obecny w ponad 50% wszystkich drzew genowych, aby trafić do drzewa

wynikowego. W analizowanym zbiorze danych sygnał filogenetyczny dla głębokich węzłów był prawdopodobnie zbyt słaby lub niejednorodny pomiędzy poszczególnymi genami, przez co żaden układ nie przekroczył wymaganego progu poparcia.

Obserwowane rozbieżności między wynikami projektu a literaturą referencyjną wynikają przede wszystkim z różnicy w wolumenie danych wejściowych. Praca Adeolu i wsp. (2016) opierała się m.in. na wykorzystaniu 1548 białkach rdzeniowych, podczas gdy niniejsza analiza wykorzystała 229 rodzin genowych. Mniejsza liczba markerów przełożyła się na trudności w rekonstrukcji niektórych relacji. Należy jednak podkreślić, że część różnic (np. względem TimeTree) działa na korzyść przeprowadzonej analizy, ujawniając braki w bazach publicznych, takie jak brakujące taksony (*Ewingella americana*, *Erwinia billingiae*) czy błędne grupowanie linii bazalnych.

W celu poprawy rozdzielczości drzew, a zwłaszcza naprawy wyników konsensusu większościowego, kluczowe byłoby zastosowanie pełnej analizy bootstrap. Wprowadzenie standardowego bootstrapu nieparametrycznego pozwoliłoby na ocenę stabilności statystycznej węzłów w każdym z 229 drzew genowych przed ich syntezą.

Zastosowanie bootstrapu umożliwiłoby „zwijanie” (ang. collapsing) gałęzi o niskim wsparciu (np. <50% lub <70% BS) do politomii w drzewach wejściowych. Dzięki temu proces konstrukcji konsensusu nie byłby zaburzany przez przypadkowe, słabo poparte relacje (szum filogenetyczny). Gdyby z drzew genowych usunięto niepewne gałęzie, pozostawiając tylko te stabilne, możliwe, że metoda konsensusu większościowego łatwiej odnalazłaby wspólny, silny sygnał dla głównych rodzin, co prawdopodobnie wyeliminowałoby problem „płaskiego” drzewa.

5. Literatura

[Adeolu et al., 2016] Adeolu M, Alnajar S, Naushad S, S Gupta R. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. Int J Syst Evol Microbiol. 2016 Dec;66(12):5575-5599. doi: 10.1099/ijsem.0.001485. Epub 2016 Sep 11. PMID: 27620848.

[Cock et al., 2009] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422-3. doi: 10.1093/bioinformatics/btp163. Epub 2009 Mar 20. PMID: 19304878; PMCID: PMC2682512.

[Katoh, 2002] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059-66. doi: 10.1093/nar/gkf436. PMID: 12136088; PMCID: PMC135756.

[Le & Gascuel, 2008] Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008 Jul;25(7):1307-20. doi: 10.1093/molbev/msn067. Epub 2008 Mar 26. PMID: 18367465.

[Nguyen et al., 2015] Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268-74. doi: 10.1093/molbev/msu300. Epub 2014 Nov 3. PMID: 25371430; PMCID: PMC4271533.

[Paradis and Schliep, 2019] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019 Feb 1;35(3):526-528. doi: 10.1093/bioinformatics/bty633. PMID: 30016406.

[Schliep, 2011] Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011 Feb 15;27(4):592-3. doi: 10.1093/bioinformatics/btq706. Epub 2010 Dec 17. PMID: 21169378; PMCID: PMC3035803.

[Smith, 2020] Smith MR. Information theoretic generalized Robinson-Foulds metrics for comparing phylogenetic trees. Bioinformatics. 2020 Dec 22;36(20):5007-5013. doi: 10.1093/bioinformatics/btaa614. Erratum in: Bioinformatics. 2021 Aug 4;37(14):2077-2078. doi: 10.1093/bioinformatics/btab200. PMID: 32619004.

[Steinegger & Söding, 2017] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov;35(11):1026-1028. doi: 10.1038/nbt.3988. Epub 2017 Oct 16. PMID: 29035372.