Analiza Filogenetyczna Mikrobiomu w Chorobach Autoimmunologicznych

Jakub Aleksander Giezgała

21 stycznia 2026

Streszczenie

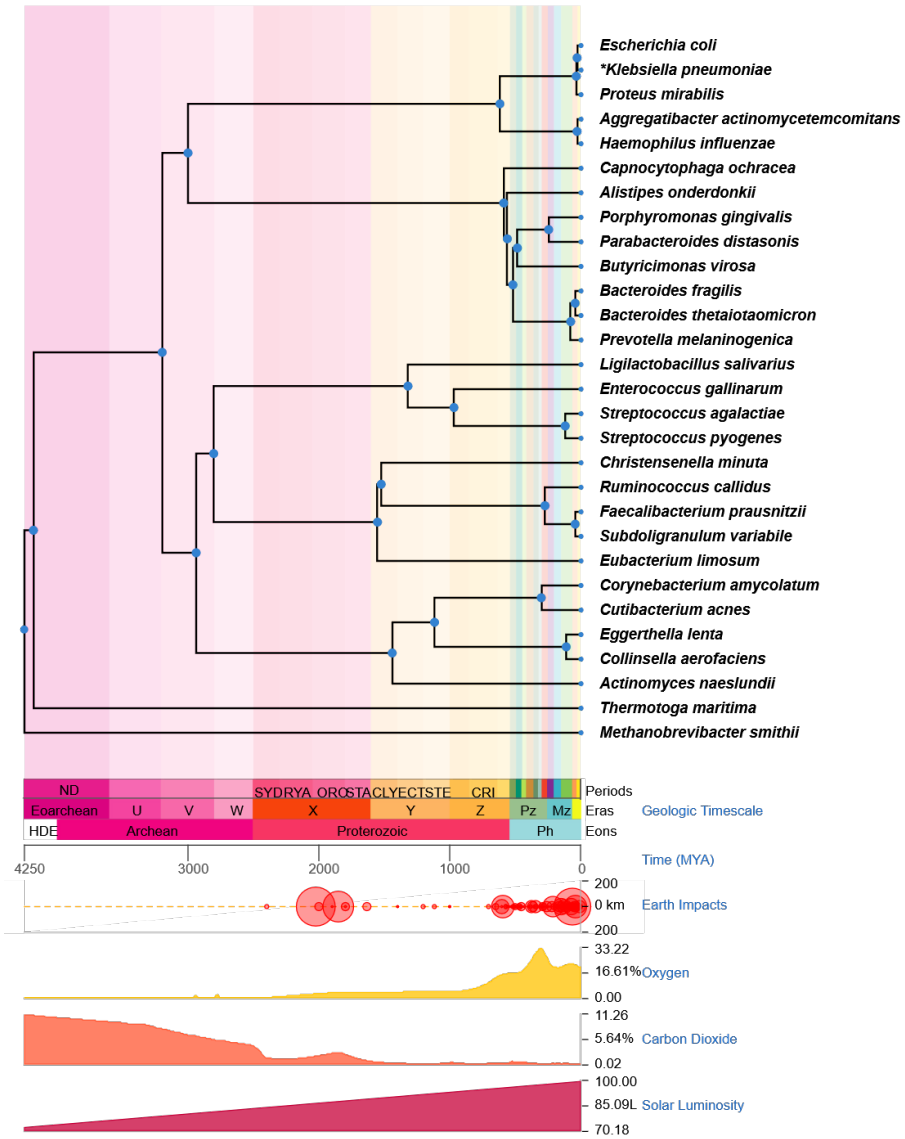
Raport przedstawia szczegółowy opis analizy filogenetycznej bakterii komensalnych wykazujących molekularne powiązania z ludzkimi chorobami autoimmunologicznymi. Na podstawie szerokiej weryfikacji literaturowej, w tym kluczowej pracy [4], wyselekcjonowano 30 taksonów (bakterie i archeony) o rozmiarze genomu poniżej 5000 genów. Ich proteomy wykorzystano jako wejście do pipeline'u opartego na narzędziu Snakemake [7] aby stworzyć narzedzie zdolne do weryfikacji hipotezy mimikry molekularnej (Ro60, GNS/FLNA) oraz ortologii.

1 Wstęp

Etiologia chorób autoimmunologicznych, takich jak toczeń rumieniowaty układowy (SLE) i reumatoidalne zapalenie stawów (RA), jest coraz częściej wiązana z mikrobiomem jelitowym, jamy ustnej i skóry. Wiodącą hipotezą wyjaśniającą ten związek jest mechanizm mimikry molekularnej, w którym antygeny drobnoustrojowe wykazują podobieństwo strukturalne lub sekwencyjne do autoantygenów gospodarza, przełamując tolerancję immunologiczną [10].

Fundamentem dla niniejszego badania jest publikacja [4], która zidentyfikowała ortologi ludzkiego autoantygenu **Ro60** u komensali, takich jak *Bacteroides thetaiotaomicron* czy *Cutibacterium acnes*. W kontekście RA, kluczowe znaczenie mają badania nad gatunkami *Sagatella copri* (dawniej *Prevotella*) [9] oraz nowo zidentyfikowanym *Subdoligranulum variabile* [1], które mogą indukować odpowiedź przeciwko białkom cytrulinowanym lub wykazywać mimikrę do filaminy A (FLNA).

Celem niniejszego raportu jest dostarczenie zweryfikowanego zestawu danych genomowych oraz drzewa referencyjnego, które posłużą do weryfikacji hipotez konwergencji pojedynczych białek ludzkiego mikrobiomu. Autorski pipeline, wykorzystujący takie biblioteki, jak na przykład Biopython [2] ma na celu porównanie wyselekcjonowanych genomów z referencyjnym drzewem życia [16] jako próba negatywna dla poszczególnych prac naukowych.



Rysunek 1: Referencyjne drzewo pozyskane ze strony https://timetree.org/.

ID	Gatunek	Szczep	Accession	Typ/Związek
G01	*Methanobrevibacter smithii*	OF4	GCF_000016525.1	Outgroup (Archaea)
G02	*Bacteroides thetaiotaomicron*	VPI 5482	GCF_000011065.1	SLE
G03	*Bacteroides fragilis*	NCTC 9343	GCF_000025985.1	SLE / Komensal
G04	*Prevotella copri*	FDAARGOS 1573	GCF_020735445.1	RA Marker
G05	*Prevotella melaninogenica*	ATCC 25845	GCF_000144405.1	RA (Perio)
G06	*Parabacteroides distasonis*	ATCC 8503	GCF_000012845.1	SLE
G07	*Butyricimonas virosa*	DSM 23226	GCF_025148635.1	SLE
G08	*Porphyromonas gingivalis*	ATCC 33277	GCF_000010505.1	RA
G09	*Capnocytophaga ochracea*	DSM 7271	GCF_000023285.1	SLE (Ro60)
G10	*Alistipes onderdonkii*	3BBH6	GCF_006542645.1	SLE Potential
G11	*Cutibacterium acnes*	NBRC 107605	GCF_006739385.1	SLE
G12	*Corynebacterium amycolatum*	FDAARGOS 1189	GCF_016889425.1	SLE
G13	*Eggerthella lenta*	APC055-529-1D	GCF_021378605.1	RA
G14	*Collinsella aerofaciens*	JCM 10188	GCF_010509075.1	RA y
G15	*Actinomyces naeslundii*	FDAARGOS 1037	GCF_016127855.1	RA

ID	Gatunek	Szczep	Accession	Typ/Związek
G16	*Enterococcus gallinarum*	EG81	GCF_021496385.1	SLE
G17	*Ruminococcus callidus*	ATCC 27760	GCF_049562855.1	SLE
G18	*Subdoligranulum variabile*	DSM 15176	GCF_025152575.1	RA
G19	*Eubacterium limosum*	ATCC 8486	GCF_000807675.2	Komensal
G20	*Faecalibacterium prausnitzii*	M21/2	GCF_000154385.1	Komensal
G21	*Ligilactobacillus salivarius*	B4311	GCF_035231985.1	RA
G22	*Streptococcus pyogenes*	NCTC12064	GCF_900475035.1	RA
G23	*Streptococcus agalactiae*	NGBS128	GCF_001552035.1	SLE
G24	*Christensenella minuta*	DSM 22607	GCF_003628755.1	Komensal
G25	*Escherichia coli*	K-12 substr.MG1655	GCF_000005845.2	RA
G26	*Proteus mirabilis*	HI4320	GCF_000069965.1	RA
G27	*Klebsiella pneumoniae*	HS11286	GCF_000240185.1	RA
G28	*Aggregatibacter actinomycetemcomitans*	PA7	GCF_000017205.1	RA
G29	*Haemophilus influenzae*	FDAARGOS 1560	GCF_020736045.1	RA
G30	*Thermotoga maritima*	JGI	GCF_000230655.2	Outgroup (Bacteria)

Tabela 1: Lista analizowanych genomów została wybrana na podstawie ich związku z chorobą autoimmunologiczną, potwierdzoną we wcześniej wspomnianej literaturze. Ograniczono się do organizmów, których proteom liczy do 5000 białek, co pozwala na efektywne obliczenia (MSA przy użyciu MAFFT [5], drzewa genów w IQ-TREE [8]) oraz obecnej, wysokiej jakości genomów w bazie NCBI.

2 Metodyka

Całość analizy, poza wygenerowaniem wykresów przeprowadzono w jednym potoku bioinformatycznym zaprojektowanym w środowisku Snakemake, co zapewnia powtarzalność i skalowalność obliczeń. Dodatkowo przez tworzenie punktów zapisu, bez zmiany kodu można wznawiać obliczenia.

2.0.1 Pobranie Sekwencji

W pierwszym kroku pobrano kompletne proteomy dla wszystkich 30 wyselekcjonowanych gatunków. Wykorzystano do tego narzędzie wiersza poleceń **NCBI Datasets CLI**, filtrując dane pod kątem kompletności złożeń (assembly level: complete/chromosome) oraz dostępności adnotacji białkowych.

2.0.2 Klastrowanie i wybór ortologów

Klastrowanie przeprowadzono przy użyciu oprogramowania **MMseqs2** w trybie `easy-linclust` z parametrami: `--min-seq-id 0.25`, `-c 0.6`, `--cov-mode 0`. Literatura uznaje taki wybór za ryzykowny, ale w przypadku różnych rodzin bakterii jest dopuszczalny [10]. Uzyskano łącznie 21 474 klastrów (w tym 12 958 singletonów).

W procedurze wyodrębniania klastrów z jednoznacznym genem, z kazdego gatunku (SCO), kluczowy był wybór sekwencji reprezentatywnej. **W pierwszym przypadku, jako reprezentatywne uznawane było białko o największym podobieństwie do pozostałych białek w klastrze.** Podobieństwo to mierzono jako sumę wyników *bit-score* z porównań **BLASTP** z każdym z pozostałych elementów klastra. *Bit-score* definiuje się jako zlogarytmowany rozmiar bazy danych, w której bieżące dopasowanie mogłoby zostać znalezione przypadkowo. Zastosowanie tej metryki pozwoliło na precyzyjną identyfikację reprezentanta i redukcję klastra do pojedynczego ortologa, zamiast odrzucania całych klastrów. Ostatecznie do analizy włączono 187 klastrów SCO oraz 199 klastrów z uwzględnieniem paralogów.

2.1 Dopasowanie wielu sekwencji (MSA)

Dla każdego klastra przeprowadzono dopasowanie sekwencji (Multiple Sequence Alignment) przy użyciu programu **MAFFT** [5] z parametrem `--auto`, która automatycznie dobiera algorytm (L-INS-i/FFT-NS-2) w zależności od rozmiaru danych. Następnie dopasowania zostały oczyszczone z regionów o niskiej jakości filogenetycznej przy użyciu programu **TrimAl** [15]. Zastosowano heurystykę `-automated1`, zoptymalizowaną pod kątem inferencji drzew metodą Największej Wiarygodności (ML), którą wykorzystujemy.

2.2 Inferencja drzew genów

Rekonstrukcję drzew filogenetycznych dla poszczególnych rodzin genów przeprowadzono dwiema metodami, aby porównać ich wpływ na finalną topologię:

1. **Neighbor Joining (NJ) / Aproksymacja ML:** Wykorzystano program **FastTree** [**?**] z flagą `-fastest`. Metoda ta stosuje test Shimodaira-Hasegawa [17] do oceny lokalnego wsparcia gałęzi. Drzewa poddano filtracji, odrzucając te ze średnim wsparciem S-H < 0.85, co zredukowało zbiór do 1784 drzew.

2. **Maximum Likelihood (ML):** Wykorzystano program **IQ-TREE** [8]. W przeciwieństwie do aproksymacji `ultrafast`, przeprowadzono standardowy bootstrap, aby uzyskać bardziej konserwatywne oszacowania wsparcia. Ze względu na wysoki koszt obliczeniowy, z analizy wykluczono klastry zawierające mniej niż 8 taksonów (nie licząc grupy zewnętrznej). Przyjęto próg akceptacji bootstrapu na poziomie 0.50, aby urealnić porównanie z często zawyżonymi wartościami S-H z FastTree.

2.3 Inferencja drzewa gatunkowego

2.3.1 Drzewo konsensusowe

Obliczono drzewa konsensusu większościowego (Majority Rule Consensus) przy użyciu IQ-TREE z parametrem `--minsup 0.50`. Oznacza to, że w finalnym drzewie znalazły się tylko te klady, które występowały w ponad 50% drzew pojedynczych genów.

2.3.2 Superdrzewo (Fasturec)

Główną analizę oparto na metodzie superdrzew. Wykorzystano program **Fasturec** [3], który implementuje algorytm lokalnego przeszukiwania (hill-climbing) do znalezienia optymalnego superdrzewa na podstawie zestawu nieukorzenionych drzew genowych. Użyto flagi `-Y` (heurystyka NNI). Superdrzewa obliczono w dwóch wariantach:

- **SCO:** Wyłącznie na podstawie klastrów nie zawierających ortologów z jednego gatunku.
- **PAR:** Z uwzględnieniem paralogów (wszystkie zidentyfikowane klady).

Wszystkie wynikowe drzewa zostały ukorzenione na grupie zewnętrznej *Methanobrevibacter smithii*.

2.4 Wydajność obliczeniowa

Analizę wydajności przeprowadzono na podstawie logów (Tabela 2). Etapy wstępne (pobieranie, klastrowanie, ekstrakcja) charakteryzowały się zbliżonym czasem wykonania dla obu ścieżek. Największą różnicę odnotowano w etapie inferencji drzew genów. Metoda NJ (FastTree) pozwoliła na obliczenie tysięcy drzew w czasie niespełna 2,5 godziny, podczas gdy metoda ML (IQ-TREE) wymagała łącznie ponad 6 godzin, co czyni ją rzędu trzykrotnie bardziej kosztowną obliczeniowo.

Tabela 2: Porównanie czasu i zasobów dla poszczególnych etapów pipeline'u (dane na podstawie plików benchmarkowych). Czas dla drzew genów stanowi sumę czasów dla wszystkich przetworzonych rodzin.

Proces	**Ścieżka NJ**		**Ścieżka ML**	
	Czas [s]	RAM [MB]	Czas [s]	RAM [MB]
Pobieranie genomów (`download`)	80.94	86.41	65.09	86.97
Klastrowanie (`clustering`)	113.28	981.21	115.71	984.41
Ekstrakcja SCO (`extract_sco`)	88.03	674.39	45.44	662.54
Ekstrakcja Paralogów (`extract_par`)	18.80	550.05	7.31	550.82
Inferencja Drzew Genów (Suma)	**8 783.23**	**80.62**	**23729.55**	**213.92**
(w przybliżeniu)	*2h 26m*		*6h 30m*	

2.5 Wykorzystane zasoby sprzętowe i orkiestracja

Obliczenia wykonano na komputerze IdeaPad Gaming 3-15IHU6 (Type 82K1) wyposażonym w procesor 11th Gen Intel(R) Core(TM) i7-11370H @ 3.30GHz.

Zarządzanie procesem obliczeniowym odbywało się poprzez system **Snakemake**, który optymalizował wykorzystanie dostępnych rdzeni poprzez zrównoleglenie zadań. Konfiguracja zasobów (*threads*) dla poszczególnych reguł została zdefiniowana w pliku `Snakefile` następująco:

- **4 wątki:** Przypisano do zadań wymagających dużej mocy obliczeniowej lub operacji na dużych zbiorach danych, takich jak klastrowanie (`clustering`), ekstrakcja ortologów (`orthologs`) oraz inferencja superdrzewa (`supertree`).
- **1 wątek:** Przypisano do zadań, które były masowo zrównoleglane na poziomie procesów (uruchamianie wielu instancji jednocześnie przez Snakemake), takich jak dopasowanie (`mafft`), przycinanie (`trimal`) oraz inferencja pojedynczych drzew genów (`genetree`).

Taka strategia pozwoliła na efektywne wykorzystanie architektury procesora przy jednoczesnym uniknięciu przeciążenia pamięci RAM.

3 Wyniki

Analiza porównawcza wygenerowanych drzew z referencyjnym drzewem czasu (TimeTree 1) pozwoliła na ocenę skuteczności przyjętych strategii (Tabela 3). W tabeli uwzględniono metryki topologiczne (Dystans Robinsona-Fouldsa - RF) oraz informacyjne (Shared Phylogenetic Information - SPI).

Tabela 3: Porównanie wygenerowanych drzew z referencją (TimeTree). SCO - Single Copy Orthologs, PAR - z paralogami.

Lp.	Plik / Metoda	Taxa	RF Dist	Norm RF	SPI Bits	SPI Norm	Jaccard	Path Dist
1	ML_sco_filtered_consensus.treefile	30	17	0.3148	288.22	0.5260	0.4500	80.94
2	**ML_sco_filtered_supertree.newick**	30	16	0.2963	350.04	0.6388	0.5676	**29.93**
3	ML_sco_raw_consensus.treefile	30	17	0.3148	237.63	0.4336	0.3810	80.94
4	ML_sco_raw_supertree.newick	30	24	0.4444	288.63	0.5267	0.4146	33.76
5	NJ_par_filtered_supertree.newick	30	18	0.3333	**354.65**	**0.6472**	0.5263	34.13
6	NJ_par_raw_supertree.newick	30	28	0.5185	223.68	0.4082	0.3488	44.87
7	NJ_sco_filtered_consensus.treefile	30	**15**	**0.2778**	237.63	0.4336	0.3810	82.07
8	NJ_sco_filtered_supertree.newick	30	26	0.4815	229.20	0.4183	0.3810	40.80
11	*timetree.nwk (Referencja)*	30	0	0.0000	547.99	1.0000	1.0000	0.00

Analiza wyników:

Jak pokazuje Tabela 3, najlepszy wynik pod względem topologicznym (najniższy dystans Robinsona-Fouldsa, RF = 15) uzyskało drzewo konsensusowe oparte na metodzie NJ i filtrowanych ortologach (wiersz 7). Uwzględniając inne metryki można wywnioskowac, że jest to wynik mylący.

Biorąc pod uwagę stabilność struktury (Path Distance) oraz ilość zachowanej informacji filogenetycznej (SPI), **superdrzewo oparte na metodzie ML (ML_sco_filtered_supertree)** wydaje się być najbardziej wiarygodną rekonstrukcją (wiersz 2). Osiągnęło ono najniższy dystans ścieżek (29.93) i bardzo wysoki wskaźnik SPI Norm (0.64), co wskazuje na wierniejsze odwzorowanie głębokich relacji ewolucyjnych, a nie tylko powierzchownej topologii.

Warto również odnotować wynik superdrzewa `NJ_par_filtered` (wiersz 5), które uzyskało najwyższy wskaźnik znormalizowanej informacji SPI (0.6472). Sugeruje to, że włączenie paralogów – przy odpowiedniej filtracji szumu – może wzbogacić sygnał filogenetyczny tam, gdzie brakuje ortologów jednokopijnych.

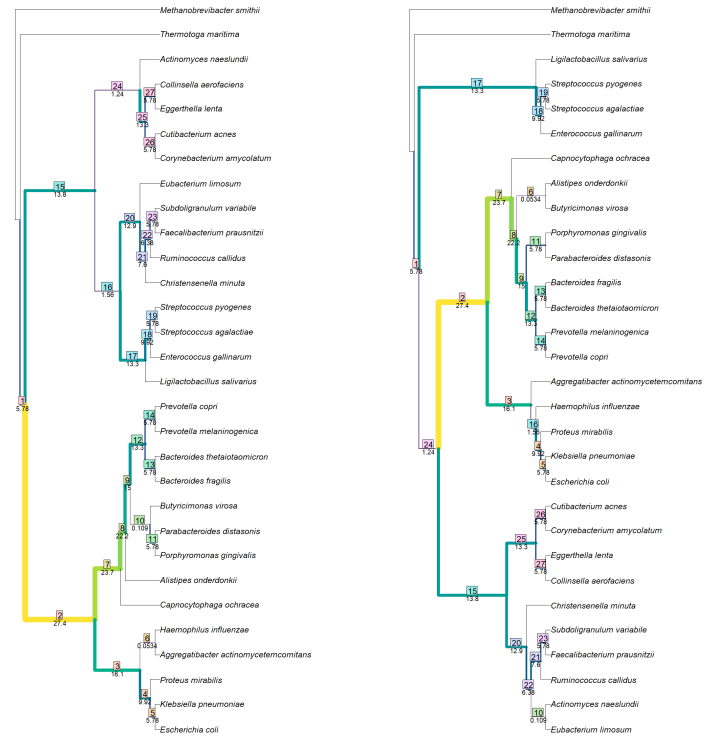
3.1 Dalsze etapy

Analiza wykazała, że w pojedynczych przypadkach (np. *Prevotella/Sagatella*) obserwuje się anomalie topologiczne względem referencji. Mogą one wynikać z mimikry molekularnej (homoplazji), gdzie presja ewolucyjna ze strony układu odpornościowego gospodarza wymusza konwergencję sekwencji białkowych bakterii z białkami ludzkimi (np. w SLE). Obecny pipeline działa efektywnie, ale wymaga rozbudowy o:

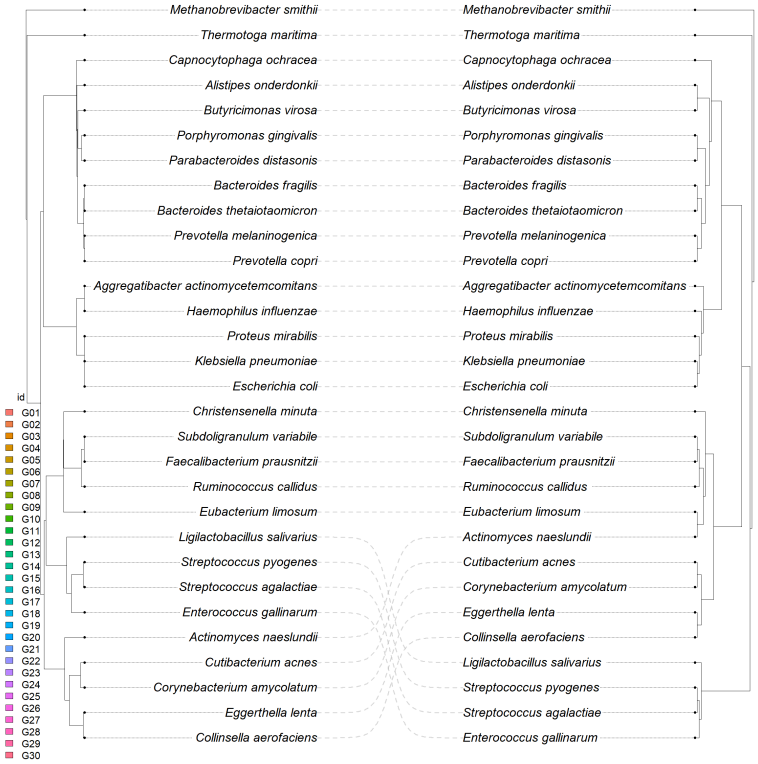
1. Algorytm *backtracingu*, który pozwoliłby zidentyfikować konkretne rodziny genów odpowiedzialne za te przesunięcia topologiczne.

2. Wdrożenie metod Największej Oszczędności (Maximum Parsimony, MP) lub porównawcze użycie **RAxML** w celu weryfikacji stabilności klastrów metodami alternatywnymi do IQ-TREE.

3. Mimo dłuższego czasu obliczeń, metoda ML powinna być stosowana jako standard dla całego genomu, gdyż jak pokazuje wynik `ML_sco_filtered`, prowadzi ona do najwierniejszego odwzorowania odległości ewolucyjnych.



Shared phylogenetic information (SPI)



(b) Tanglegram

Rysunek 2: Porównanie wizualne. Po lewej każdej grafiki znajduje się referencyjne drzewo filogenetyczne. Po prawej natomiast najlepsze drzewo wyłonionego w analizie (`ML_sco_filtered_supertree`). Linie na tanglegramie, poniżej łączą odpowiadające sobie taksony; ich splątanie wskazuje na różnice topologiczne.

Literatura

[1] Chriswell, M. E., et al. (2022). Clonal IgA and IgG autoantibodies from individuals at risk for rheumatoid arthritis identify an arthritogenic strain of Subdoligranulum. *Science Translational Medicine*, 14(668). https://pubmed.ncbi.nlm.nih.gov/36288282/

[2] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

[3] Górecki, P., Burleigh, J. G., & Eulenstein, O. (2012). GTP Supertrees from Unrooted Gene Trees: Linear Time Algorithms for NNI Based Local Searches. In *Bioinformatics Research and Applications* (Vol. 7292, pp. 102–114). Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-30191-9_11

[4] Greiling, T. M., et al. (2018). Commensal orthologs of the human autoantigen Ro60 as triggers of autoimmunity in lupus. *Science Translational Medicine*, 10(434). https://pubmed.ncbi.nlm.nih.gov/29593104/

[5] Katoh, K. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066. https://pubmed.ncbi.nlm.nih.gov/12136088/

[6] Konig, M. F., et al. (2016). Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. *Science Translational Medicine*, 8(369). https://pubmed.ncbi.nlm.nih.gov/27974664/

[7] Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, 10, 33. https://pubmed.ncbi.nlm.nih.gov/34035898/

[8] Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274. https://pubmed.ncbi.nlm.nih.gov/25371430/

[9] Pianta, A., et al. (2017). Evidence of the Immune Relevance of Prevotella copri, a Gut Microbe, in Patients With Rheumatoid Arthritis. *Arthritis & Rheumatology*, 69(5). https://pubmed.ncbi.nlm.nih.gov/27863183/

[10] Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2), 85–94. https://pubmed.ncbi.nlm.nih.gov/10195279/

[11] Scher, J. U., et al. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *eLife*, 2, e01202. https://pubmed.ncbi.nlm.nih.gov/24192039/

[12] Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4), 592–593. https://pubmed.ncbi.nlm.nih.gov/21169378/

[13] Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11), 1026–1028. https://pubmed.ncbi.nlm.nih.gov/29035372/

[14] Zhang, X., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nature Medicine*, 21(8), 895–905. https://pubmed.ncbi.nlm.nih.gov/26214836/

[15] Capella-Gutiérrez, S., José M Silla-Martínez, Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 8;25(15):1972–1973. https://pmc.ncbi.nlm.nih.gov/articles/PMC2712344/

[16] Kumar S., et al. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, Volume 39, Issue 8. https://academic.oup.com/mbe/article/39/8/msac174/6657692

[17] Morgan N. Price, Paramvir S. Dehal, Adam P Arkin (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. https://pmc.ncbi.nlm.nih.gov/articles/PMC2693737/