Pipeline Filogenetyczny

Contents

1. Wstęp

Genomy koronawirusów wybrałam z publikacji [1], dostając finalnie 28 genomów. Według pracy, koronawirusy dzielą się na Alfa, Beta oraz Gamma koronawirusy, przy czym beta koronawirusy modzielone są głębiej na grupy A, B, C i D. Wyłania się również grupa Delta koronawirusów.

Drzewo filogenetyczne w powyższej pracy powstało na bazie analizy RNA-zależnej polimerazy RNA (Pol) oraz całkowitych genomów wirusów. Do jego wyliczenia użyto metody NJ i bootstrapowania z 1000 drzew. Takie drzewo zostało następnie ukorzenione z użyciem Pol dla Breda wirusa.

W moim podejściu, wykorzystałam wcześniejszy fragment artykułu, wymieniający geny S, N, HE, PL (PL1), CPL (PL2), Hel, Pol oraz 3CL. Proteomy pozyskałam z sekwencji wymienionych w pracy, używając baz nucleotide oraz protein, ze względu na zawarcie większości genów w ORF1, bez informacji w bazie nucleotide. Resztę podejścia opisuję w kolejnej sekcji.

2. Metody

Moje podejście opierało się na:

1. Pobraniu proteomów dla zadanych id dostępu (Python).
2. Normalizacji nazw genów, ponieważ były one bardzo różne we wszystkich sekwencjach (co również zostało wspomniane jako problem w [1]) (Python).
3. Pogrupowaniu sekwencji według genów (Python).
4. Poklastrowaniu rodzin genów przy użyciu programu MMSeqs2 z parametrami: `easy-cluster --min-seq-id 0.98 --cov-mode 0 --alignment-mode 4 -v 1 --add-self-matches 1 --cluster-mode 1 -c 0.8` [2]. W wyniku, dostałam łącznie 125 klastrów. (MMSeqs2).
5. Podzieliłam klastry na osobne foldery (Python).
6. Multiuliniowieniu klastrów genów przy użyciu programu Muscle z domyślnymi opcjami dla `-align` (Muscle).
7. Stworzyłam naiwne drzewa o głębokości 1, ponieważ IQ-Tree nie pozwala na tworzenie drzew dla < 3 sekwencji (Python).
8. Obliczyłam drzewa rodzin genów metodą ML używając programu IQ-Tree z parametrami `-T AUTO -s`.
9. Kroki podobne do 6-8 wykonałam również dla reprezentantów klastrów. Tu niestety wykorzystałam wyłącznie pierwszego reprezentanta, jednak wartoby sprawdzić, czy wybór alternatywnych reprezentantów nie dałby lepszych wyników.
10. Podstawiłam drzewa rodzin genów za reprezentantów klastrów, rozszerzając je jednocześnie o brakujące nazwy genomów (np. dla HE, Hel) oraz dodając tymczasowy takson “DUMMY” w

celu wymuszenia na IQ-Tree, żeby traktował te drzewa jako nieukorzenione, gdyż rzucał on błędy przy liczeniu drzewa konsensusowego przy mieszanych drzewach un-/rooted.

11. Policzyłam drzewo konsensusowe z użyciem IQ-Tree z parametrami `-T AUTO --sup-min 0.3`. Parametr `--sup-min` dobrałam eksperymentalnie porównując wyniki dla `0.5, 0.4, ..., 0.0`. Usunęłam następnie taxon "DUMMY" i porównałam z drzewem z [1], używając unormowanej miary RF.
12. Policzyłam superdrzewo przy użyciu Clann z `criterion=dfit` (porównywałam `sfit, dfit, qfit` i dfit dało najlepszy balans wynikowo-wydajnościowy), `hs nreps=5 nsteps=20 sample=1000 swap=spr` i porównałam z [1].

Do organizacji zadań użyłam makefile'u oraz skryptu shellowego `run.sh`.

W wielu miejscach wykorzystywałam program gnu parallel, co pozwoliło na znaczne zrównoleglenie całego pipeline'u, szczególnie dodając opcję `-T AUTO` do IQ-Tree.

Cały pipeline zajmuje ok. 30 min., ze zużyciem procesora sięgającym 100% na 16-rdzeniowym procesorze z 32GB RAMu.

3. Wyniki

3.1. Rozkład wielkości klastrów

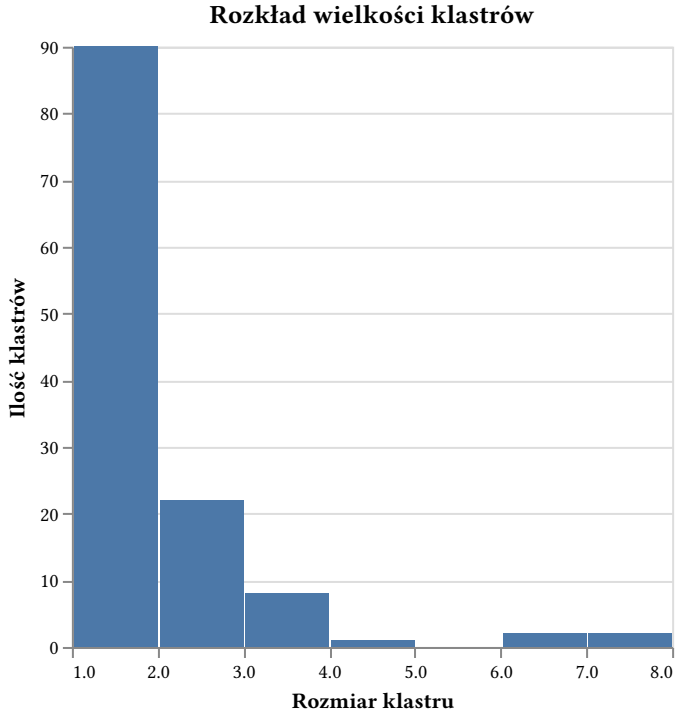

Figure 1: Rozkład wielkości klastrów.

3.2. Drzewo konsensusowe

Niestety nie ma zachowania relacji pomiędzy typami koronawirusów w każdym przypadku, jednak widać zachowanie spójności oraz bliskości wewnątrz typów, pomijając beta koronawirusy z grupy A.

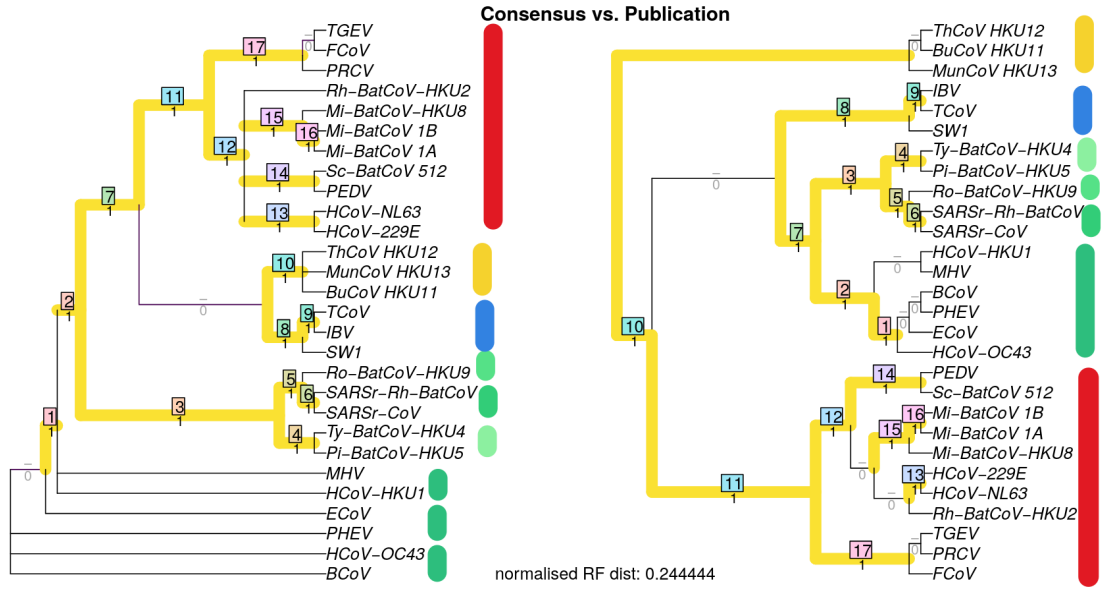

Figure 2: Porównanie wynikowego drzewa konsensusowego (po lewej) z drzewem z [1] (po prawej) z użyciem biblioteki TreeDist. Kolorem czerwonym oznaczyłam alfa koronawirusy, zielonym - beta koronowarisy (najciemniejszym grupę A, B, D, i najjaśniejszym - C), niebieskim - gamma koronawirusy oraz żółtym - deltakoronawirusy.

3.3. Superdrzewo

Obserwujemy tu podobne zachowanie, jak w drzewie konsensusowym, czyli zachowanie spójności oraz bliskości wewnątrz grup, jednak z zachowaniem go dla wszyskich grup beta koronawirusów.

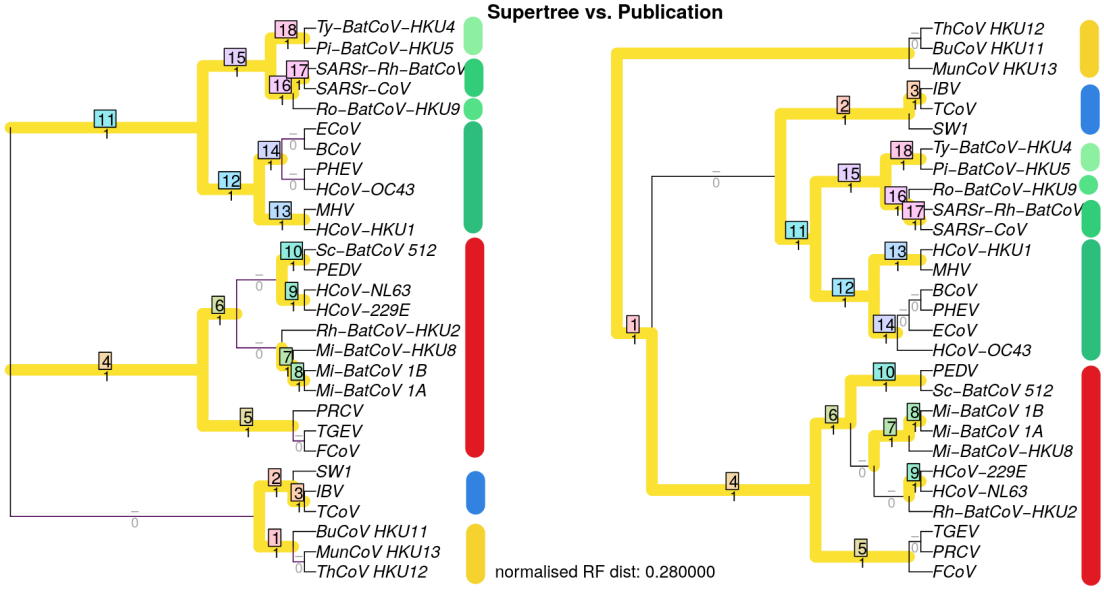

Figure 3: Porównanie wynikowego superdrzewa (po lewej) z drzewem z [1] (po prawej) z użyciem biblioteki TreeDist. Oznaczenia takie same, jak dla drzewa konsensusowego.

4. Wnioski

Uzyskane drzewa filogenetyczne są niestety dalekie od idealnych, jednakże nie są pozbawione meritum: mamy zachowane większość rodzin koronawirusów.

W celu polepszenia jego wyników, wartoby rozważyć:

1. Użycie lepiej zbadanej metody dla tworzenia drzew rodzin genów, ponieważ metoda użyta przeze mnie wydaje się nie do końca optymalna, ani typowa. Szczególny upadek po kątem optymalności widać przy budowaniu drzewa ML dla reprezentantów klastrów dla genu S oraz ilości przerw w multiuliniowieniach jego sekwencji.
2. W przypadku pozostania przy metodzie budowania drzew, warto rozpatrzyć używanie alternatywnych reprezentantów klastrów. Wiązałoby się to jednak z dłuższym czasem wykonania.
3. W moim rozwiązaniu wymuszałam użycie drzew nieukorzenionych, jednak wydaje mi się, że drzewa ukorzenione mogłby prowadzić potencjalnie do lepszych wyników, szczególnie dla drzewa konsensusowego.

Bibliography

[1] L. S. Y. K. Woo PCY Huang Y, "Coronavirus Genomics and Bioinformatics Analysis," *Viruses. 2010 Aug;2(8):1804-1820.*

[2] T. Ç. Oğuzoğlu and B. T. Koç, "Global Phylogenetic Analysis of the CDV Hemagglutinin Gene Reveals Positive Selection on Key Receptor-Binding Sites," *Viruses*, vol. 17, no. 9, 2025, doi: 10.3390/v17091149.