Pipeline filogenetyczny

Anna Szymik

5 lutego 2025

1 Wstęp

Niniejsza praca przedstawia pipeline do analizy filogenetycznej organizmów wyższych, wraz z wynikami analizy dla wybranych gatunków muszek z rodzaju *Drosophila*. Stosowane narzędzia są zatem dopasowane do charakterystyki genomów tych organizmów.

Rodzaj *Drosophila* jest jednym z najbardziej zróżnicowanych taksonów wśród owadów, obejmując ponad 1,500 gatunków, różniących się pod względem morfologii i zachowania. Zajmują one różnorodne nisze ekologiczne, a ze względu na krótki czas życia i duże wielkości populacji charakteryzują się wysokim tempem specjacji. Są zatem idealnym taksonem do badania mechanizmów ewolucji.

Genomy *Drosophila* są stosunkowo niewielkie w porównaniu z innymi organizmami wyższymi – przykładowo genom modelowego organizmu *Drosophila melanogaster* to około 175 milionów par zasad – co ułatwia ich badanie. Liczba genów kodujących białka dla tego taksonu wynosi około 13,000.

Niniejsza praca bazuje na wynikach przedstawionych w publikacji [Li et al., 2022], której autorzy opracowali bazę danych dla 46 gatunków *Drosophila* i jednego blisko spokrewnionego *Zaprionus*. W wyniku ich prac piętnaście genomów zostało na nowo zsekwencjonowanych, a dwadzieścia kolejnych – udoskonalonych dzięki dodatkowemu sekwencjonowaniu. Ich analizy filogenomiczne rozwiązały kilka wcześniej niejednoznacznych relacji, szczególnie w grupie gatunków *melanogaster*.

Drzewa zostały przez nich zrekonstruowane z dwóch zbiorów danych – sekwencji nukleotydowych całych genomów i sekwencji aminokwasowych genów ortologicznych, w którym znalazło się 8550 ortologów. Drzewa gatunków zostały z nich wywnioskowane dwiema metodami (konkatenacyjną i koalescencyjną), przy użyciu programów ExaML [Kozlov et al., 2015] i ASTRAL [Zhang et al., 2018]. Z wszystkich drzew obliczony został konsensus, również przy pomocy pakietu ASTRAL [Zhang et al., 2018].

Prezentowana w tej pracy analiza skupia się na 15 genomach nowo zsekwencjonowanych w wyniku badania [Li et al., 2022]: *Drosophila bipectinata, Drosophila ironensis, Drosophila pseudoananassae, Drosophila setifemur, Drosophila birchii, Drosophila jambulina, Drosophila pseudotakahashii, Drosophila sulfurigaster, Drosophila bunnanda, Drosophila kikkawai, Drosophila rubida, Zaprionus bogoriensis, Drosophila immigrans, Drosophila pandora* oraz *Drosophila serrata*. Są one dostępne są w bazie danych NCBI pod numerem projektu PRJNA736147.

2 Metody

2.1 Etapy analizy

2.1.1 Pobranie sekwencji

Pierwszym etapem pipeline'u było pobranie sekwencji wszystkich białek wybranych organizmów z bazy danych NCBI. Jako że były one zdeponowane pod wspólnym numerem *BioProject ID* (PRJNA736147), wystarczyło posłużyć się tym numerem dostępu w celu pobrania wszystkich sekwencji z bazy danych Assembly. Zostało to zrobione przy użyciu modułu Entrez z pakietu BioPython [Cock et al., 2009], a pobrane proteomy zapisane do osobnych plików FASTA.

Mimo że w porównaniu z innymi organizmami wyższymi, genomy *Drosophila* są niewielkie, to analizy proteomów składających się z kilkunastu tysięcy białek są dosyć kosztowne obliczeniowo, tak więc pipeline optymalizowany był pod kątem szybkości działania używanych algorytmów.

2.1.2 Klastrowanie

Pobrane sekwencje zostały następnie poklastrowane przy użyciu programu MMSeqs2 (Many-against-Many sequence searching) [Steinegger & Söding, 2017] z flagą `easy-linclust`, metodą zoptymalizowaną pod kątem szybkości i pamięci, przez co dedykowana jest dużym zestawom danych, z jakiego przykładem mamy do czynienia. Próg identyczności został ustanowiony na 35%, jako że w analizach filogenetycznych i ewolucyjnych często przyjmuje się, że sekwencje o $\geq$ 30–40% identyczności należą do tej samej rodziny białkowej [Rost, 1999].

W wyniku tej analizy powstało 37 022 surowych klastrów, z czego 16 273 jednoelementowych. Następnym etapem było zatem wybranie klastrów zawierających ortologi wszystkich analizowanych organizmów. Zostało to zrobione na dwa sposoby – po ich usunięciu paralogów, pozostawiając po jednym reprezentancie w klastrze dla każdego organizmu, oraz z zachowaniem wszystkich paralogów. W pierwszym przypadku, jako reprezentatywne uznawane było białko o największym podobieństwie do pozostałych białek w klastrze. Owo podobieństwo mierzone było jako suma wyników *bit-score* z porównań BLASTP z każdym z pozostałych elementów klastra – *bit-score* jest to zlogarytmowany wymagany rozmiar bazy danych, w której bieżące dopasowanie można znaleźć przypadkowo, nie zależy zatem od rozmiaru obecnej bazy danych.

Klastrów, które zawierały po co najmniej jednej sekwencji z każdego organizmu, było 3 829, nie brano więc do dalszych analiz mniejszych klastrów.

2.1.3 Multiuliniowienia

Dla każdej rodziny genów policzone zostało multiuliniowienie z wykorzystaniem programu MAFFT [Katoh, 2002], szybkiego algorytmu, który liczy multiuliniowienia z wykorzystaniem szybkiej transformaty Fouriera.

2.1.4 Drzewa rodzin

Następnie, z uliniowienia dla każdej rodziny obliczone zostały drzewa genów. Zostało to zrobione metodą Neighbor-Joining przy pomocy pakietu `phangorn` [Schliep, 2011]. Metoda ta jest obarczona pewnymi ograniczeniami, przede wszystkim założeniem addytywności macierzy odległości ewolucyjnych, co może prowadzić do błędnych wniosków w przypadku sekwencji o nierównomiernych szybkościach ewolucyjnych, w tym także do ujemnych długości krawędzi. Niemniej jednak, została ona użyta ze względu na efektywność obliczeniową – pozwala na szybkie konstruowanie drzew nawet dla dużych zbiorów danych, co jest kluczowe w kontekście analizy tak wielu rodzin genów. Dodatkowo, metoda NJ dostarcza rozsądnej aproksymacji topologii drzewa, co czyni ją odpowiednim narzędziem na wstępnym etapie analizy filogenetycznej.

Wykorzystanym do obliczenia macierzy odlgłości modelem substytucji był JTT (Jones-Taylor-Thornton), który jest często używany do porównań sekwencji białkowych.

Na tym etapie przeprowadzona była również analiza bootstrap ze stu powtórzeniami dla każdego drzewa genów. Drzewa, dla których przynajmniej jeden klad miał wsparcie niższe niż 60, były w tym przypadku usuwane z dalszej analizy. W ten sposób z 3829 klastrów wziętych pod uwagę w tej części analizy było jedynie 135 drzew. Porównanie drzew otrzymanych z i bez analizy wsparcia bootstrap znajduje się w Rozdziale 3.

2.1.5 Drzewo genomów

Z rodzin genów zostało obliczone drzewo genomów metodą konsensusową i superdrzewową. Konsensus został policzony z klastrów ortologicznych na dwa sposoby – jako konsensus większościowy i zachłanny – przy pomocy programu IQ-TREE [Nguyen et al., 2015]. Superdrzewa zostały obliczone dla drzew inferowanych zarówno z klastrów ortologicznych, jak i tych z zachowaniem paralogów. Wykorzystanym narzędziem był Fasturec [Górecki et al., 2012], program do inferencji superdrzewa z zestawu nieukorzenionych drzew genowych – czyli takich, jakie zwracane są przez algorytm NJ – implementujący lokalny algorytm wyszukiwania z wykorzystaniem heurystyk hill-climbing.

Drzewa zostały następnie ukorzenione przez klad zewnętrzny do pozostałych, czyli *Zaprionus bogoriensis*.

2.2 Opis techniczny

Wszystkie obliczenia zostały wykonane na komputerze HP ENVY Notebook 13-ab0XX z procesorem Intel Core i7-7500U CPU @ 2.70GHz $\times$ 4 i 8 GiB pamięci RAM. Czas działania poszczególnych etapów pipeline'u podsumowany jest w Tabeli 1. Całość pipeline'u została opakowana w jeden

plik przy wykorzystaniu narzędzia Snakemake [Mölder et al., 2021]. Wszystkie skrypty dostępne są w repozytorium https://github.com/aszymik/phylogenetic-pipeline.

Etap	**Czas działania**
Pobranie proteomów z bazy danych	1.89 min
Klastrowanie sekwencji	1.49 min
Poprawa klastrowania	58.72 min
Multiuliniowienia rodzin genów	18.20 min
Budowa drzew rodzin metodą NJ (bez analizy bootstrap)	1.52 min
Konstrukcja drzewa genomów metodą konsensusową	1.18 s
Konstrukcja drzewa genomów metodą superdrzewową	5.66 s

Tabela 1: Tabela etapów i czasu działania

3 Wyniki

Wynikowe drzewa zostały porównane z drzewem opublikowanym przez [Li et al., 2022], z użyciem zawartości informacji filogenecznej każdego podziału, zależnej od prawdopodobieństwa znalezienia tego podziału w losowo wybranym drzewie (proporcji drzew binarnych, które zawierają ten podział). Wspólna dla porównywanych drzew informacja filogenetyczna została następnie znormalizowana przez informację filogenetyczną drzewa referencyjengo, która wynosi 85.1452 bity. Wyniki tych porównań przedstawione są w Tabeli 2.

Metoda	**Wspólna informacja filogenetyczna**
Konsensus większościowy	0.643335
Konsensus zachłanny	0.9468724
Superdrzewo	0.9468724
Superdrzewo z zachowaniem paralogów	0.9468724
Konsensus większościowy z bootstrap	0.7520542
Konsensus zachłanny z bootstrap	1.0
Superdrzewo z bootstrap	1.0

Tabela 2: Porównanie ilościowe otrzymanych drzew z drzewem referencyjnym. Wspólna informacja filogenetyczna jest znormalizowana przez informację filogenetyczną drzewa referencyjnego.

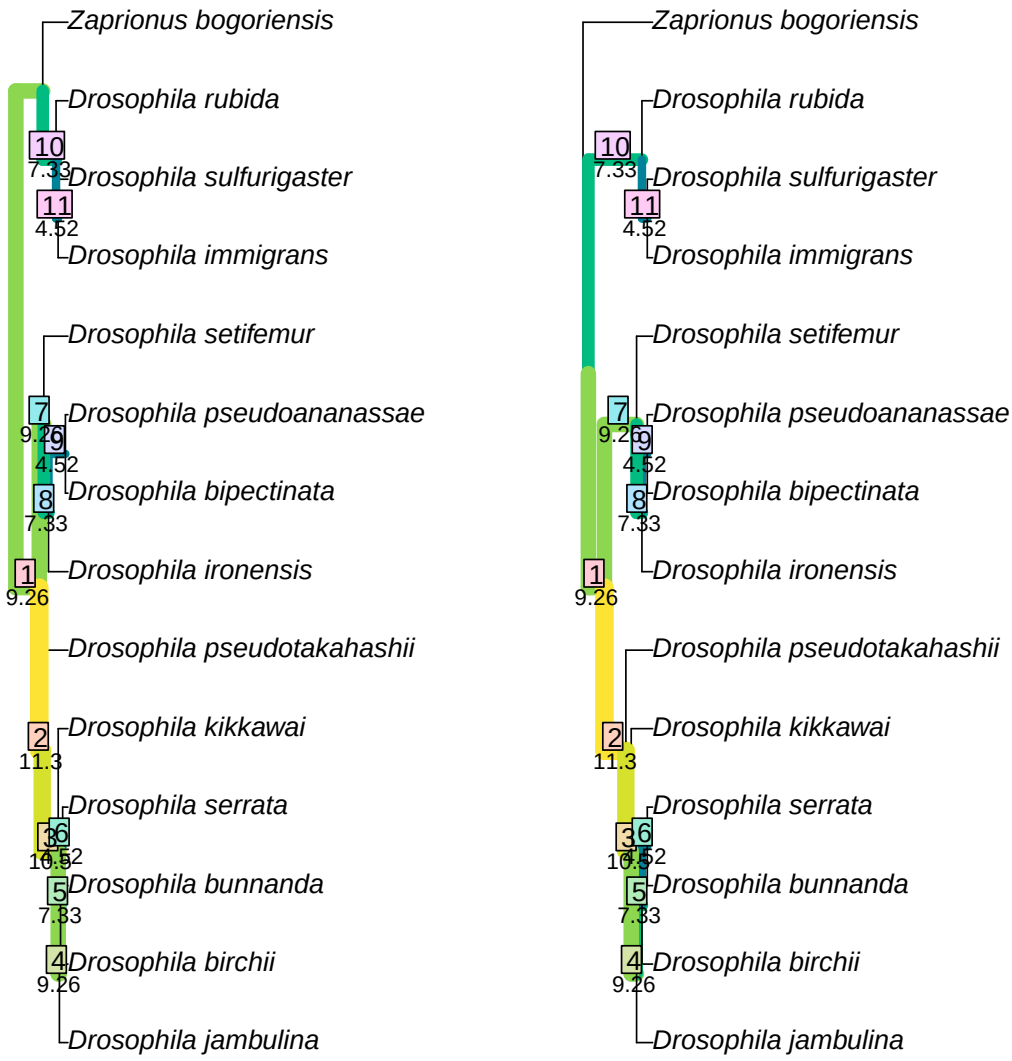
Jak widać, pozostawienie paralogów w klastrach nie poprawiło w żadnym stopniu wyników analiz (choć zgodność informacji filogenetycznej na poziomie 94.7% jest i tak bardzo wysoka), za to przeprowadzenie analiz jedynie na drzewach z silnym wsparciem bootstrap – tak. Konsensus większościowy ma co prawda najmniej zgodne z literaturą wyniki ze wszystkich – co nie dziwi ze względu na rygorystyczność tej metody i w związku z tym częste pozostawianie multifurkacji – natomiast zarówno konsensus zachłanny i superdrzewo obliczone na tym podzbiorze drzew zrekonstruowały drzewo identyczne z przedstawionym w publikacji [Li et al., 2022]. Jest ono przedstawione na Rysunku 1.

4 Wnioski

Niniejsza analiza opiera się na pipelinie automatyzującym pobieranie, przetwarzanie i analizę sekwencji białkowych wybranych gatunków z rodzaju *Drosophila*. W jej ramach przeprowadzone zostało klastrowanie sekwencji, multiuliniowienia, inferencja drzew genów oraz konstrukcja drzewa genomów przy użyciu metod konsensusowych i superdrzewowej.

Pipeline jest w pełni zautomatyzowany i efektywny czasowo, choć w przyszłości można byłoby poprawić skrypt do usuwania paralogów z klastrów. Obecny wielokrotnie otwiera pliki w celu znajdowania sekwencji białek do porównania przy użyciu programu BLASTP, co jest nieefektywne czasowo – poprawianie klastrów stanowi najdłużej trwającą część całego pipeline'u.

Można byłoby rozważyć również inferencję drzew genów inną metodą niż NJ. Choć jest ona bardzo wydajna, jej ograniczenia mogą skutkować błędnymi długościami gałęzi i niską rozdzielczością niektórych kladów. Można rozważyć zastosowanie metod ML lub bayesowskich, które mogą



Rysunek 1: Informacja filogenetyczna w drzewie uzyskanym metodą superdrzewową z drzew silnie wspierancyh przez bootstrap i drzewie referencyjnym.

lepiej modelować ewolucję białek (zwłaszcza w przypadku posiadaniu większych zasobów obliczeniowych), choć mając do czynienia z dużą ilością danych, takich jak genomy organizmów wyższych, nie wydaje się to niezbędne.

Choć w przypadku analizowanych danych nie było to problemem, w przypadku obliczania drzew metodą NJ, można byłoby w przyszłości dodać do skryptu asercję dotyczącą zwracanych długości krawędzi, tak aby drzewa z ujemnymi krawędziami nie były w ogóle brane pod uwagę w dalszych analizach.

Mimo pewnych ograniczeń, wyniki pokazują wysoką zgodność uzyskanych drzew z literaturą – szczególnie w przypadku metod konsensusu zachłannego oraz superdrzewowej dla drzew z wysokim wsparciem bootstrap, w przypadku których zrekonstruowane drzewa są identyczne z literaturowym drzewem referencyjnym – co potwierdza poprawność zastosowanego pipeline'u.

Literatura

[Cock et al., 2009] 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

[Górecki et al., 2012] Górecki, P., Burleigh, J. G., & Eulenstein, O. (2012). GTP Supertrees from Unrooted Gene Trees: Linear Time Algorithms for NNI Based Local Searches. In L. Bleris, I.

Măndoiu, R. Schwartz, & J. Wang (Eds.), *Bioinformatics Research and Applications* (Vol. 7292, pp. 102–114). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30191-9_11

[Katoh, 2002] 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://doi.org/10.1093/nar/gkf436

[Kozlov et al., 2015] Kozlov, A. M. , Aberer, A. J. , & Stamatakis, A. (2015). ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics, 31*(15), 2577–2579. 10.1093/bioinformatics/btv184

[Li et al., 2022] Li, F., Rane, R. V., Luria, V., Xiong, Z., Chen, J., Li, Z., Catullo, R. A., Griffin, P. C., Schiffer, M., Pearce, S., Lee, S. F., McElroy, K., Stocker, A., Shirriffs, J., Cockerell, F., Coppin, C., Sgrò, C. M., Karger, A., Cain, J. W., ... Zhang, G. (2022). Phylogenomic analyses of the genus Drosophila reveals genomic signals of climate adaptation. *Molecular Ecology Resources, 22*(4), 1559–1581. https://doi.org/10.1111/1755-0998.13561

[Mölder et al., 2021] 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. *F1000Res 10*, 33.

[Nguyen et al., 2015] 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://doi.org/10.1093/molbev/msu300

[Rost, 1999] Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection, 12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

[Schliep, 2011] Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics, 27*(4), 592–593. https://doi.org/10.1093/bioinformatics/btq706

[Steinegger & Söding, 2017] 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://doi.org/10.1038/nbt.3988

[Zhang et al., 2018] Zhang, C. , Rabiee, M. , Sayyari, E. , & Mirarab, S . (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics, 19*(6), 15–30. 10.1186/s12859-018-2129-y