Rekonstrukcja Wiarygodnych Relacji Między Genami a Genomami

autoreferat

Agnieszka Mykowiecka

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

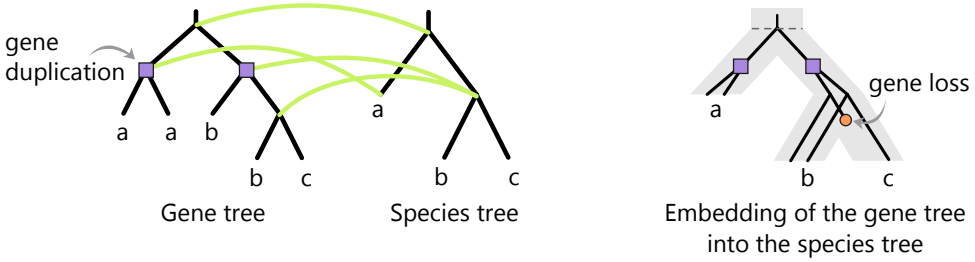
Filogenetyka, według najbardziej ogólnej definicji, jest dziedziną zajmującą się badaniem historii ewolucji i związków między pojedynczymi osobnikami, grupami organizmów lub genami i genomami. Hipotezy dotyczące przebiegu ich ewolucji można przedstawić w postaci drzew filogenetycznych, które ilustrują relacje przodek–potomek między badanymi jednostkami. Potencjalna obecność zdarzeń ewolucyjnych, takich jak duplikacje lub straty, w ewolucji rodzin genów może sprawiać, że rekonstruowane topologie drzewa genów i gatunków będą niezgodne. Do zlokalizowania zdarzeń powodujących różnice między drzewami, można zastosować metodę uzgadniania drzew.

Drzewa rekonstruowane na podstawie podobieństwa sekwencji mogą być nieprawidłowe z wielu powodów. Do najczęstszych należą błędy w sekwencjonowaniu DNA, duże podobieństwo badanych sekwencji oraz ograniczenia stosowanych metod. Błędne rekonstrukcje drzew mogą być także spowodowane obecnością zdarzeń retykulacyjnych, takich jak horyzontalny transfer genów (HGT) czy hybrydyzacje, wprowadzających zakłócenia do badanych sekwencji. Zaburzenia w topologiach drzew mogą przekładać się na problemy w dalszych krokach przeprowadzanej analizy danych. Poprawność zdarzeń ewolucyjnych, zrekonstruowanych za pomocą opartej na topologiach metody uzgadniania, jest w ogromnym stopniu zależna od poprawności uzgadnianych drzew. Duża wrażliwość na błędy i ograniczenia metody uzgadniania sprawiają, że wiarygodność zdarzeń ewolucyjnych oraz opracowanie wiarygodnych metod rekonstruowania zdarzeń retykulacyjnych, wciąż stanowią otwarty problem w dziedzinie filogenetyki.

Klasyczna metoda uzgadniania drzew

Uzgadnianie drzew to metoda oparta na mapowaniu najniższego wspólnego przodka (lca), która łączy dwie topologie drzew i wyjaśnia niezgodności między nimi. Nieformalnie rzecz ujmując, metoda uzgadniania umożliwia poznanie scenariusza ewolucyjnego dla danych drzew genów i gatunków poprzez wbudowanie drzewa genów w drzewo gatunków (Rysunek 1). Wbudowanie wyznacza duplikacje i straty genów potrzebne do dopasowania drzewa genów do drzewa gatunków, wskazując w drzewie genów węzły, w których te zdarzenia prawdopodobnie miały miejsce. Koncepcja uzgadniania drzew została zaproponowana przez Goodmana w pracy [Goodman *et al.* (1979)], a później sformalizowana przez Page'a w [Page (1994)]. Algorytm uzgadniania poszukuje wbudowania o minimalnym koszcie, w związku z czym do jego działania należy określić model definiujący dozwolone zdarzenia ewolucyjne i ich koszty. Najczęściej stosowanym modelem jest model duplikacji i strat (DL), który dopuszcza dwa rodzaje zdarzeń: duplikacje i straty genów. Duplikacje genów są mechanizmem, uznaje się za jedno z

głównych źródeł nowego materiału genetycznego w procesie ewolucji molekularnej [Taylor and Raes (2004)], natomiast w wyniku strat genów część informacji zawarta w DNA jest tracona. Wpływ obu tych zjawisk na DNA sprawia, że model oparty na duplikacjach i stratach wydaje się być dobrym przybliżeniem ewolucji. Koszt otrzymanego scenariusza jest sumą kosztów zdarzeń ewolucyjnych wymaganych do uzgodnienia dwóch drzew.



Rysunek 1: *Przykład obrazujący mapowanie lca oraz scenariusz ewolucyjny. Po lewej:* Mapowanie lca między drzewem genów G a drzewem gatunków S (pominięto mapowanie liści). *Po prawej:* Wbudowanie reprezentujące scenariusz ewolucyjny, który odpowiada mapowaniu lca. Tutaj do uzgodnienia drzew G i S potrzebne są dwie duplikacje i jedna strata genu.

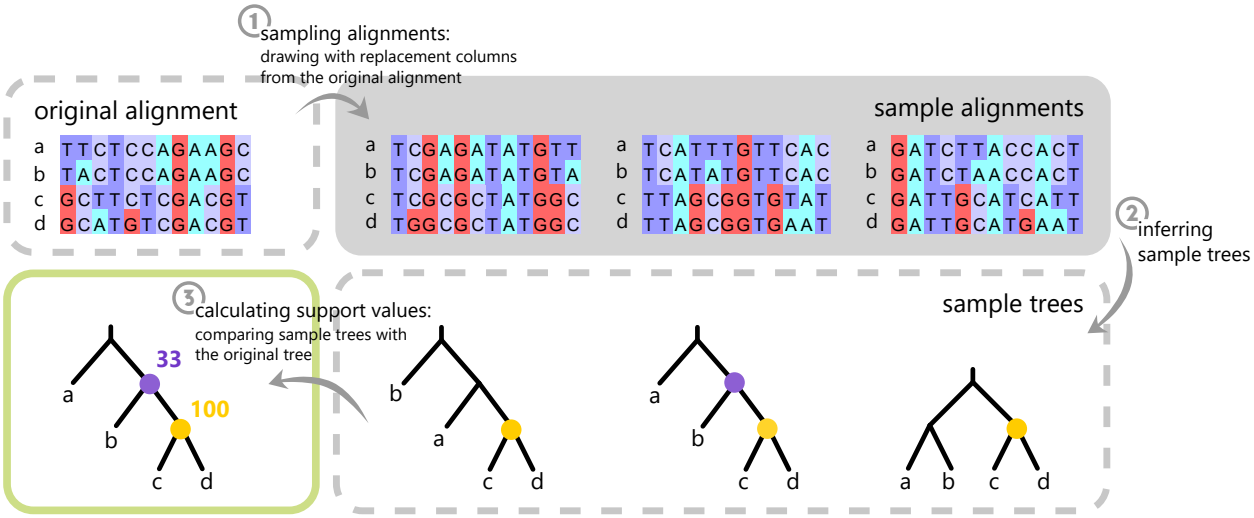
Ocena poprawności topologii drzew

Drzewa filogenetyczne zrekonstruowane na podstawie podobieństwa sekwencji genetycznych przedstawiają jedynie hipotetyczne zależności ewolucyjne. Ze względu na ograniczenia metod rekonstrukcji oraz potencjalne błędy w danych, wiarygodność drzew może być w niektórych przypadkach wątpliwa. Analizowane dane mogą również zawierać bardzo podobne do siebie sekwencje lub być zaburzone przez zdarzenia retykulacyjne, co skutkuje rekonstrukcją drzew o nieprawidłowych topologiach.

Rozwiązanie kwestii wiarygodności drzew filogenetycznych zostało zaproponowane przez Felsensteina [Felsenstein (1985)], który w swojej metodzie wykorzystał nieparametryczny bootstrap. W początkowym etapie tej metody tworzony jest zestaw drzew próbkowych, zrekonstruowanych na podstawie losowo zmienionych sekwencji genów. Następnie oryginalne drzewo porównuje się z drzewami próbkowymi i dla każdej gałęzi lub klastra oblicza się procent drzew próbkowych, w których znaleziono identyczną gałąź lub klaster, czyli zbiór wszystkich liści w danym poddrzewie. Otrzymane wyniki można interpretować jako wskazanie wpływu arbitralnych zmian nieprzypominających wzorca ewolucyjnego, takich jak błędy sekwencjonowania, na topologię rekonstruowanego drzewa filogenetycznego. Graficzne przedstawienie schematu metody bootstrapu oraz wyliczone wartości wsparcia dla przykładowego drzewa przedstawione są na Rysunku 2.

Wiarygodność Zdarzeń Duplikacji i Specjacji

Zdarzenia duplikacji i start w historii ewolucyjnej rodziny genów można określić za pomocą metody uzgadniania drzew. Jednakże, podczas gdy klasyczny model uzgadniania stosuje się tylko do drzew ukorzenionych, większość standardowych metod rekonstrukcji zwraca drzewa nieukorzenione. Jednym z rozwiązań tego problemu jest ukorzenianie badanych drzew. Istnieje kilka metod ukorzeniania, jednak problem identyfikacji wiarygodnego miejsca ukorzeniania jest nietrywialny [Górecki and Eulenstein (2012)]. Ukorzenianie metodą *outgroup*,



Rysunek 2: Schemat przedstawiający trzy etapy obliczania wartości wsparcia dla drzew filogenetycznych przy użyciu metody bootstrap.

w przypadku gdy drzewo jest heterogeniczne, może prowadzić do błędnego ukorzenienia. Innymi stosowanymi rozwiązaniami są metody oparte o wykorzystanie założenia o zegarze molekularnym lub punktu środkowego (*midpoint*). W obu przypadkach wyniki ukorzeniania mogą być jednak nieprawidłowe, jeśli w drzewie występuje zmienność tempa ewolucji [Holland *et al.* (2003); Huelsenbeck *et al.* (2002)]. W pracach [Chaudhary *et al.* (2012); Durand *et al.* (2006); Górecki and Eulenstein (2012)] zaproponowano rozwiązanie polegające na wykorzystaniu operacji edycji drzew w celu poprawienia topologii drzew przed uzgadnianiem. Z kolei metoda proponowana w [Beretta and Dondi (2014); Swenson *et al.* (2012); Dondi *et al.* (2014)] polegała na wstępnym przetworzeniu zbioru drzew genów poprzez usunięcie z niego drzew zawierających węzły powodujące niespójności.

Podejście do problemu z zupełnie innej strony polega na rozszerzeniu metody uzgadniania drzew w taki sposób, aby możliwe było uzgodnienie nieukorzenionego drzewa genów z ukorzenionym drzewem gatunków. Takie rozwiązania zostały przedstawione w pracach [Górecki and Tiuryn (2007); Yu *et al.* (2011)], gdzie metoda uzgadniania poszukuje takiego ukorzenienia drzewa, które indukuje minimalny koszt duplikacji i strat. Opisana metoda nie rozwiązuje jednak problemu innych błędów jakie mogą występować w topologiach drzew, co sprawia, że wiarygodność wyników uzgadniania pozostaje istotnym zagadnieniem w badaniach filogenetycznych.

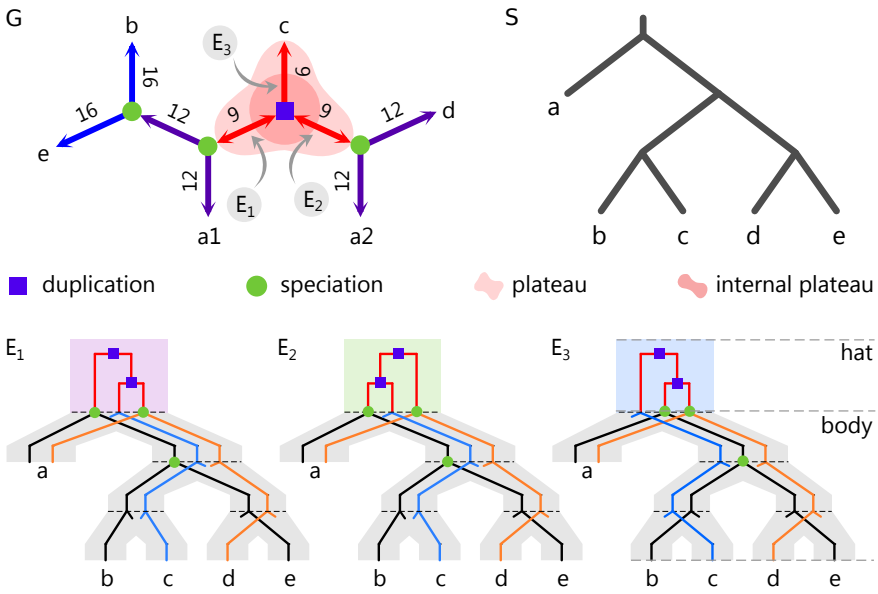
W Rozdziale 3 rozprawy doktorskiej przedstawiamy wyniki opublikowane w pracy [Mykowiecka and Górecki (2018)], w której poruszamy kwestię wiarygodności zdarzeń ewolucyjnych oraz opisujemy naszą metodę, opartą na połączeniu uzgadniania nieukorzenionych drzew i nieparametrycznego bootstrapu, która pozwala rozwiązać ten problem.

Uzgadnianie nieukorzenionych drzew

Klasyczny model uzgadniania drzew można rozszerzyć o uzgadnianie nieukorzenionego drzewa genów z ukorzenionym drzewem gatunków, poprzez poszukiwanie ukorzenienia drzewa genów z minimalnym kosztem duplikacji i strat [Górecki and Tiuryn (2007); Yu *et al.* (2011)].

Zbiór wszystkich krawędzi o minimalnym koszcie duplikacji i strat nazywamy *plateau*.

Z twierdzenia zaczerpniętego z [Górecki and Tiuryn (2007)] oraz topologicznych własności drzew wynika, że graf indukowany przez plateau jest nieukorzenionym drzewem binarnym. Przykład nieukorzenionego uzgadniania jest przedstawiony na Rysunku 3.



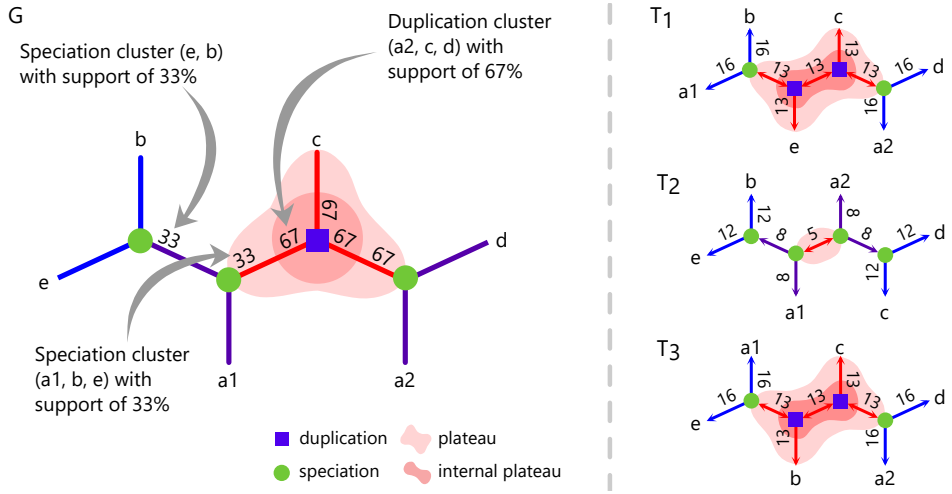
Rysunek 3: *Na górze*: Przykład nieukorzenionego drzewa genów G, uzgodnionego z drzewem gatunków S. Przy każdej krawędzi umieszczony koszt duplikacji i strat uzgodnienia drzewa G, ukorzenionego na danej krawędzi, z drzewem S (optymalny koszt wynosi 9). W G występują trzy klastry specjacyjne bez liści (zaznaczone zielonymi kółkami) oraz dwa klastry duplikacyjne (jeden zaznaczony ciemnofioletowym kwadratem i klaster korzenia). Każda krawędź należąca do plateau jest pokolorowana na czerwono i ma etykietę E1, E2 lub E3, odpowiadającą jednemu z widocznych poniżej wbudowań. *Na dole*: Trzy wbudowania przedstawiające wszystkie optymalne scenariusze dla G i S. Odpowiadające sobie krawędzie w poszczególnych wbudowaniach są oznaczone odpowiednimi kolorami.

Wartości wsparcia dla zdarzeń ewolucyjnych

W rozprawie zaprezentowaliśmy nowe podejście do oceny wiarygodności duplikacji genów i zdarzeń specjacyjnych w ukorzenionych i nieukorzenionych drzewach genów. Wśród opisanych wyników teoretycznych przedstawiamy w szczególności liniowy algorytm obliczający wartości wsparcia dla zdarzeń ewolucyjnych. Zaproponowana przez nas koncepcja wsparcia dla duplikacji i specjacji, pozwala na weryfikację wiarygodności scenariuszy ewolucyjnych uzyskiwanych przy użyciu metody uzgadniania drzew. Przykład przedstawiający obliczone wartości wsparcia dla duplikacji i specjacji w drzewie genów z Rysunku 3, znajduje się na Rysunku 4.

Jednym z możliwych zastosowań wartości wsparcia zdarzeń ewolucyjnych jest wykorzystanie ich w problemie rekonstrukcji superdrzewa. Zaobserwowaliśmy, że jakość otrzymywanego superdrzewa można poprawić poprzez usunięcie z badanego zbioru drzew genów o słabym wsparciu dla zdarzeń ewolucyjnych. Superdrzewo zrekonstruowane na podstawie tak oczyszczonego zbioru danych było również bardziej spójne biologicznie.

Drugim zaprezentowanym zastosowaniem było użycie naszej metody do oceny poprawności ukorzenienia drzew. Przeprowadziliśmy porównanie kilku metod ukorzeniania drzew i uzyskane wyniki pokazały, że większość metod zwraca nieprawidłowo ukorzenione drzewa.



Rysunek 4: *Przykład wartości wsparcia dla drzewa z Rysunku 3. Po lewej*: drzewo genów G z podanymi wartościami wsparcia duplikacji i specjacji dla klastrów nie będących liśćmi, obecnych w optymalnym ukorzenieniu. *Po prawej*: drzewa T_1, T_2 i T_3 - drzewa próbkowe otrzymane metodą bootstrap dla G. Przy krawędziach drzew T_i znajdują się koszty duplikacji i strat uzyskane dla odpowiadających im ukorzenień.

Rekonstrukcja Wiarygodnych i Spójnych Czasowo Horyzontalnych Transferów Genów

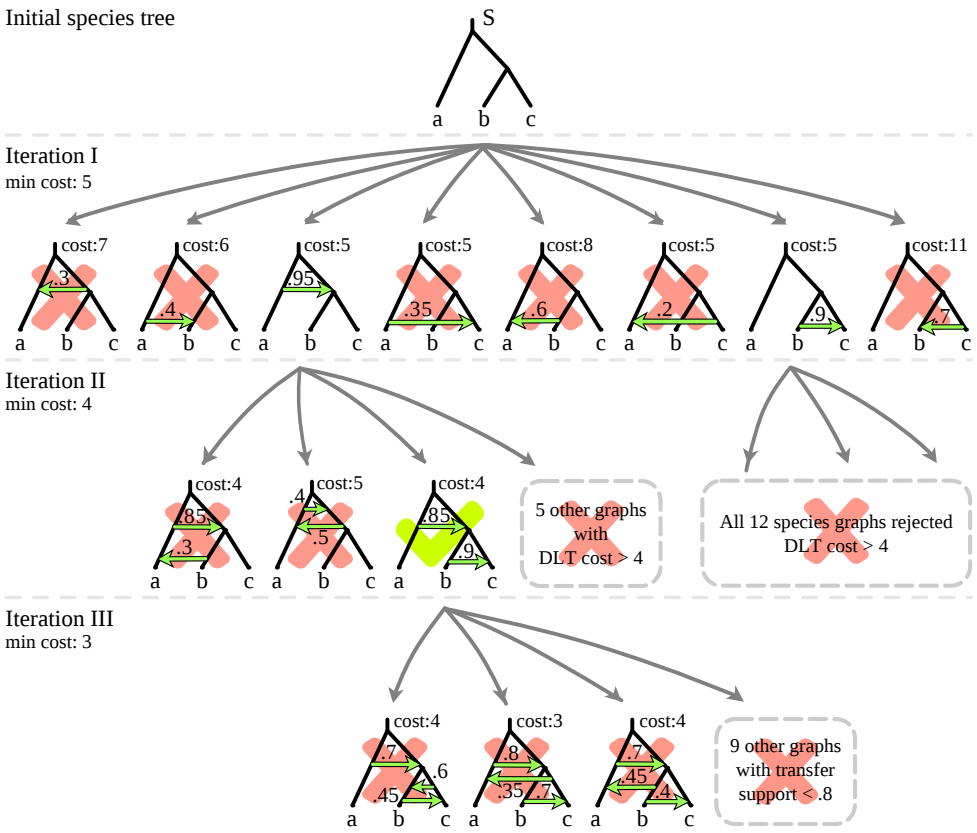
Ewolucję pojedynczej rodziny genów lub grupy gatunków można zazwyczaj przedstawić w postaci drzewa filogenetycznego, jednak w przypadku analiz metagenomicznych, model zdarzeń ewolucyjnych powinien uwzględniać także zdarzenia retykulacyjne, takie jak horyzontalny transfer genów, który jest istotnym czynnikiem wprowadzającym zmienność genetyczną. Trudność problemu, w przypadku zdarzeń HGT, polega między innymi na tym, że uwzględniające je modele są dużo bardziej złożone i wymagające obliczeniowo [Nguyen *et al.* (2013); Scornavacca *et al.* (2014); Sjöstrand *et al.* (2014); Szöllősi *et al.* (2013); Tofigh *et al.* (2011)].

Proponowanych było kilka strategii podejścia do problemu. Jednym z możliwych sposobów zmniejszenia złożoności czasowej do wielomianu jest zezwolenie na tworzenia cykli przez transfery. W takim przypadku zadanie może być rozwiązane przy pomocy programowania dynamicznego [Bansal *et al.* (2012); Mykowiecka *et al.* (2017); Tofigh *et al.* (2011)], jednak otrzymane rozwiązanie może nie być biologicznie poprawne.

Inna możliwość polega na przypisaniu niektórym lub wszystkim węzłom drzewa gatunków czasu rozejścia. Wprowadzony zostaje w ten sposób dodatkowy wymóg uporządkowania czasowego w scenariuszu, co ułatwia obliczenia [Ranwez *et al.* (2015)] i pozwala na rozwiązanie problemu w czasie wielomianowym [Bansal *et al.* (2012); Doyon *et al.* (2010)]. Dostępność datowanych drzew gatunków jest jednak bardzo ograniczona, zwłaszcza w przypadku gatunków bakterii, u których częste są horyzontalne transfery genów.

Kolejnym sposobem jest wstawienie potencjalnych krawędzi transferowych bezpośrednio do drzewa gatunków. Otrzymany w ten sposób graf skierowany, zwany *grafem gatunków*, reprezentuje ewolucję gatunków i zawiera zestaw poziomych krawędzi, które mogą być wykorzystane przez linie ewolucyjne genów jako horyzontalne transfery. W przypadku, gdy proponowane transfery nie tworzą cykli, złożoność czasowa takiego modelu jest wielomianowa [Górecki (2004); Scornavacca *et al.* (2017)].

W Rozdziale 4 rozprawy zajęliśmy się kwestią znajdowania wysoko wspieranych zdarzeń transferowych. Do weryfikacji wiarygodności znajdowanych transferów, zaproponowaliśmy nową miarę opartą na nieparametrycznym bootstrapie, zwaną wsparciem transferów. Następnie, wykorzystaliśmy tę miarę do zaprojektowania wydajnego algorytmu iteracyjnego, szukającego acyklicznych i dobrze wspieranych zdarzeń HGT. Złożoność czasowa i pamięciowa stworzonego algorytmu, w kroku obliczania minimalnego kosztu uzgodnienia binarnego drzewa genów G z drzewem gatunków S, wynosi $O(|G||S|)$. Mimo że nasz algorytm jest heurystyką, której złożoność zależy od zastosowanych warunków zatrzymania, drzew wejściowych, liczby znajdowanych optymalnych scenariuszy HGT, oraz innych parametrów (np. liczby iteracji), to nasze testy wykazały jego wysoką wydajność na danych empirycznych. Podsumowując, nasza metoda jest nowym podejściem do problemu poszukiwania zdarzeń HGT w ewolucji genów i gatunków, która postuluje najbardziej prawdopodobne miejsca na podstawie wiarygodności znajdowanych transferów. Schemat przedstawiający przebieg algorytmu jest przedstawiony na Rysunku 5. Opisane w Rozdziale 4 badania i ich wyniki zostały opublikowane w [Mykowiecka *et al.* (2018)].



Rysunek 5: Przykładowy przebiegu algorytmu szukającego wiarygodnych i spójnych czasowo zdarzeń HGT. *Od góry:* wejściowe drzewo gatunków $(a,(b,c))$ i trzy iteracje pętli głównej z potencjalnymi grafami gatunków. Dla każdego grafu gatunków wartości wsparcia transferu są podane na odpowiednich krawędziach transferowych, natomiast koszt DLT, tj. suma kosztów duplikacji, strat i HGT, jest pokazany po prawej stronie korzenia. W tym przykładzie przyjmujemy, że transfery w grafie gatunków są dobrze wspierane, jeśli wsparcie każdego z nich jest większe niż .8. Odrzucone grafy są zaznaczone czerwonymi krzyżykami. Zgodnie z tymi kryteriami nasz algorytm zwraca zaznaczony na zielono graf o koszcie 4 z drugiej iteracji.

W celu sprawdzenia skuteczności naszej metody, przeprowadziliśmy eksperymenty na dwóch empirycznych zbiorach danych, z których jeden zawierał blisko, a drugi daleko spokrewnione grupy gatunków. Oba eksperymenty wykazały, że metoda ta może być wykorzystywana do wspierania znanych lub weryfikowania rozważanych hipotez dotyczących transferów, choć musi być stosowana ze świadomością istnienia problemu ukorzeniania. Dokładność algorytmu została zweryfikowana w eksperymentach z wykorzystaniem symulowanych danych. Wyniki pokazały, że algorytm osiąga wysoki odsetek poprawnie wskazanych transferów zarówno dla drzew z jednym, jak i z dwoma HGT.

Rekonstrukcja Relacji Gen–Gatunek

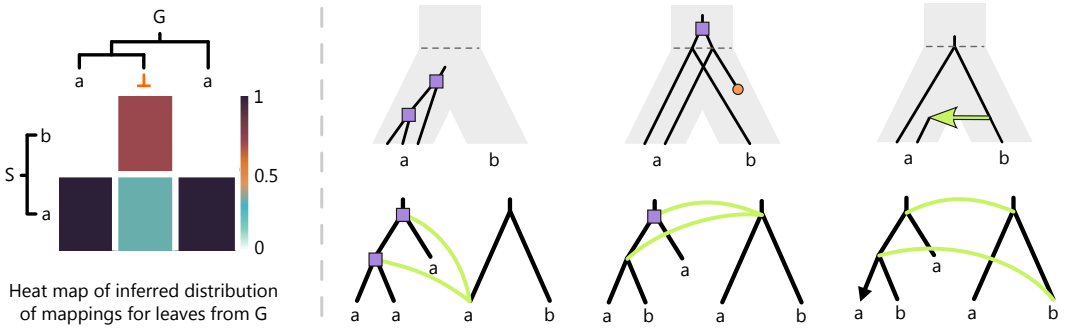
Złożoność fizjologii i biochemii drobnoustrojów sprawia, że często niemożliwe jest stworzenie jasnego obrazu pojedynczych gatunków należących do grupy drobnoustrojów. Najczęściej stosowana w badaniach metagenomicznych metoda sekwencjonowania, czyli metoda typu *shotgun*, w której genomy wszystkich mikroorganizmów są sekwencjonowane razem, sprawia że tracone są informacje o przynależności genów do konkretnych gatunków.

Podejście stosowane równolegle do metagenomiki typu shotgun, w którym sekwencjonowane są amplikony genów markerowych, zapewnia dokładniejsze zobrazowanie różnorodności biologicznej [Weisburg *et al.* (1991)]. Niestety jedynie nieliczne drzewa rodzin genów, takich jak rybosomalny 16S lub gen białkowy RecA, swą topologią odpowiadają drzewu gatunków [Thompson *et al.* (2004)].

Do rozwiązania problemu rekonstrukcji relacji gen–gatunek, zaproponowaliśmy podejście oparte na uzgadnianiu drzew z uwzględnieniem zdarzeń HGT. Problem ten może być w ogólności zdefiniowany w następujący sposób: *dla danego drzewa genów z częściowym etykietowaniem i drzewa gatunków, przypisz wszystkie brakujące etykiety w drzewie genów w taki sposób, aby koszt uzgodnienia drzew był minimalny.*

Według naszej wiedzy, dotychczasowe rozwiązania problemu rekonstrukcji relacji gen–gatunek obejmowały jedynie metody oparte na koszcie *głębokiej koalescencji* oraz koszcie duplikacji i strat, natomiast rozszerzenia obejmujące zdarzenia HGT nie były stosowane [Maddison (1997); Zhang and Cui (2010); Bafna *et al.* (2000); Betkier *et al.* (2015)]. W Rozdziale 5 przedstawiamy stworzone przez nas wydajne algorytmy, służące do obliczania optymalnych kosztów i do rekonstrukcji relacji gen–gatunek przy zastosowaniu ważonych funkcji kosztów uwzględniających duplikacje i straty genów oraz zdarzenia HGT [Mykowiecka *et al.* (2017)]. Nasze rozwiązanie obejmuje dwa modele ze zdarzeniami transferowymi: spójny czasowo (tcDTL) i ogólny (DTL). Algorytm dla modelu DTL działa w czasie $O(|G||S|)$, jeśli zarówno drzewo genów G, jak i drzewo gatunków S są binarne W przypadku gdy S zawiera multifurkacje, a drzewo genów jest binarne złożoność wynosi $O(|G||S|\Delta S)$, gdzie ΔS jest maksymalnym stopniem węzłów z S. Dla modelu spójnego czasowo opisujemy algorytm o złożoności czasowej $O(|G||S|^2)$ i proponujemy usprawnienie bazujące na strukturach danych z [Bansal *et al.* (2012), które działa w czasie $O(|G||S|log|S|)$. Przykład rekonstrukcji relacji gen-gatunek przedstawiono na Rysunku 6.

Działanie naszego algorytmu przetestowaliśmy na zbiorze danych z [Betkier *et al.* (2015)], reprezentującym typowy scenariusz analizy amplikonów, a do obliczeń zastosowaliśmy kilka zestawów kosztów zdarzeń ewolucyjnych. Dla niektórych zestawów przyporządkowania były dość jednoznaczne, podczas gdy dla innych, zwłaszcza dla zerowego kosztu straty genów, były one rozproszone po wszystkich liściach drzewa gatunków. Obserwacja ta wskazuje na istotny wpływ kosztów zdarzeń na metody uzgadniania drzew i stawia pytanie o dobór odpowiednich



Rysunek 6: *Rekonstrukcja relacji gen–gatunek dla przykładowego drzewa genów G i gatunków S. Po lewej:* Grafika przedstawiająca rozkłady mapowań liści, czyli informacje o częstości mapowania danego genu na poszczególne gatunki. *Po prawej:* Optymalne scenariusze ewolucyjne. Brakujące przyporządkowanie liści w drzewie genów oznaczone jest symbolem "⊥". W tym przykładzie, przy założeniu, że zdarzenie HGT ma koszt 2 razy większy niż duplikacje i straty, istnieją trzy optymalne scenariusze ewolucji. Zakładając, że każdy optymalny scenariusz jest równie prawdopodobny, prawdopodobieństwo, że ⊥ jest a wynosi $\frac{1}{3}$, natomiast dla b jest równe $\frac{2}{3}$.

parametrów. Niemniej jednak, ogólne wyniki wykazały zdolność naszego podejścia do wzmocnienia taksonomicznego przyporządkowania sekwencji w badaniach metagenomicznych. Na Rysunku 7 przedstawiona jest rekonstrukcja relacji gen–gatunek przeprowadzona dla kilku zestawów kosztów zdarzeń ewolucyjnych.

Metody Oparte na Sieciach Filogenetycznych

Struktura drzew zdaje się być najbardziej naturalnym sposobem przedstawiania historii ewolucji gatunków i chociaż w wielu przypadkach są one wystarczające, to nie zawsze relacje ewolucyjne można przedstawić za pomocą struktury drzewowej. W przypadku zdarzeń retykulacyjnych, takich jak rekombinacja, hybrydyzacja lub horyzontalny transfer genów, do pokazania nowych i bardziej złożonych relacji potrzebne są dodatkowe rozgałęzienia i nowe typy węzłów. Potrzebne elementy struktury można znaleźć w sieciach filogenetycznych, które są coraz częściej wykorzystywane w badaniach z dziedziny filogenetyki. Niewątpliwą zaletą sieci filogenetycznych jest możliwość pokazania wielu możliwych ścieżek ewolucji. Podczas badania bardzo blisko spokrewnionych sekwencji pochodzących z mikroorganizmów, pojedynczych komórek lub ras zwierząt, wartości wsparcia bootstrap dla rekonstruowanych drzew są zazwyczaj bardzo niskie. Dzieje się tak ponieważ zbyt duże podobieństwo sekwencji uniemożliwia jednoznaczne rozstrzygnięcie stopni pokrewieństwa sekwencji. Podczas gdy struktura drzewa wymusza wybranie jednej, być może nieprawidłowej topologii, sieci umożliwiają przedstawienie wszystkich potencjalnych ścieżek. Przeprowadzana analiza danych może być dzięki temu pełniejsza i bardziej wnikliwa.

W Rozdziale 6 zaproponowaliśmy zastosowanie podejścia opartego na sieciach do analizy danych nowotworowych. Analizowane zbiory zawierały sekwencje receptorów BCR z komórek limfocytów B, będących częścią humoralnego układu odpornościowego, uzyskane od pacjentów z chłoniakiem pęcherzykowym (FL). Sieci zrekonstruowane dla powyższych danych pozwalają na modelowanie ewolucji guza i obserwację selekcji subklonalnej indukowanej mutacjami BCR. Uzyskane wyniki potwierdziły istotną rolę mutacji BCR w rozwoju chło-

niaka grudkowego, jednak konieczne są dalsze badania nad czynnikami, które mogą wpływać na zachowanie mutacji. Lepsze zrozumienie i poznanie procesów związanych z progresją FL powinno pozytywnie wpłynąć na rozwój metod leczenia i pomóc w odkryciu przyczyn nowotworu. Prace nad projektem wciąż trwają, a wstępne wyniki zostały opublikowane w [van Bergen *et al.* (2019)]. Przykłady sieci zrekonstruowanej za pomocą naszego protokołu WILLOW są przedstawione na Rysunku 8.

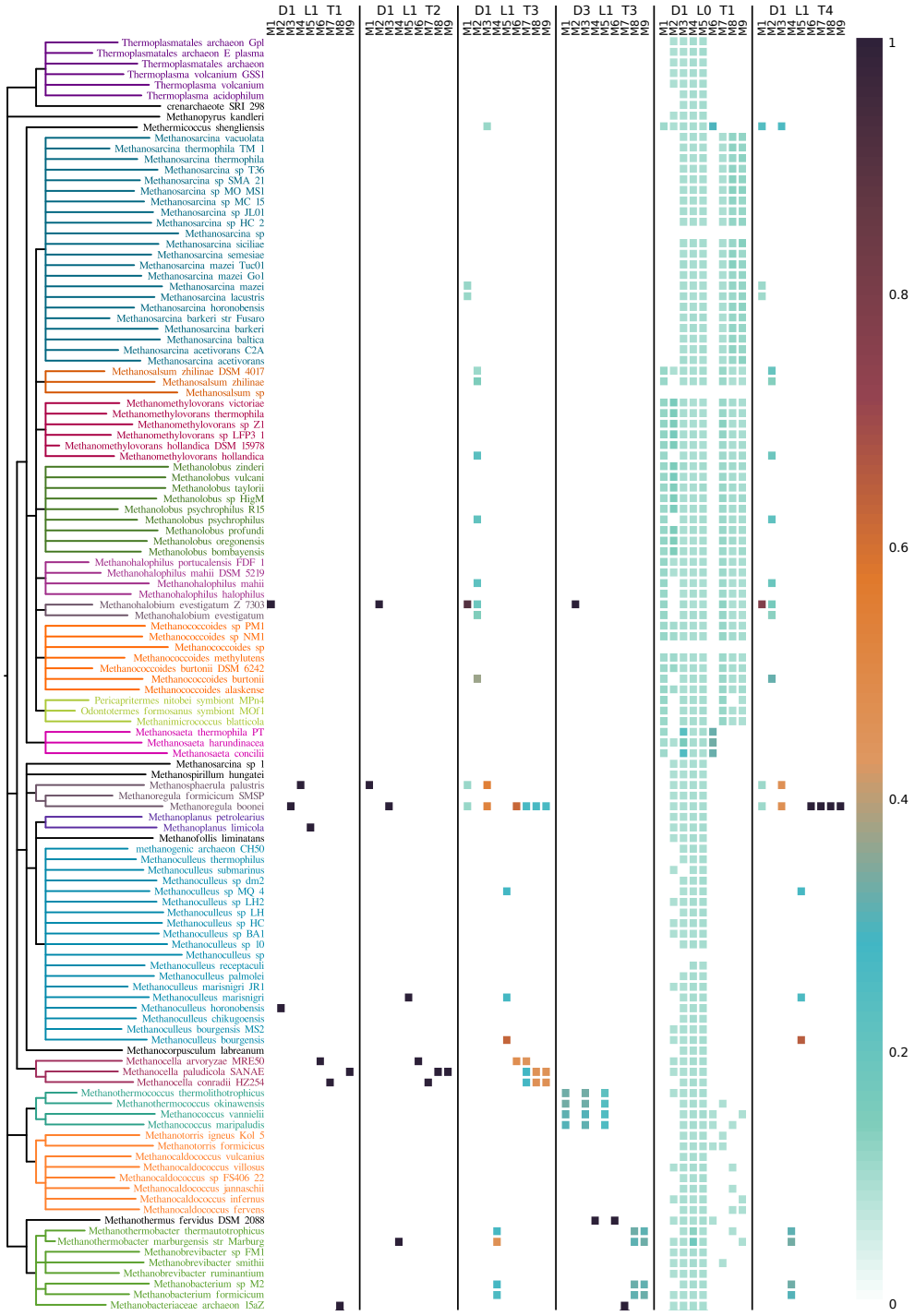
Podsumowanie

W niniejszej rozprawie skupiliśmy się na problemach z zakresu filogenetyki i teorii grafów, ze szczególnym uwzględnieniem rekonstrukcji wiarygodnych zdarzeń ewolucyjnych w drzewach genów i gatunków. Przedstawiliśmy zarówno wyniki teoretyczne, jak i eksperymentalne, w tym nowe twierdzenia, lematy i algorytmy wraz z dowodami poprawności, własności prezentowanych struktur oraz wyniki pokazujące na danych rzeczywistych i symulowanych potencjalne zastosowanie i wydajność naszych metod i algorytmów.

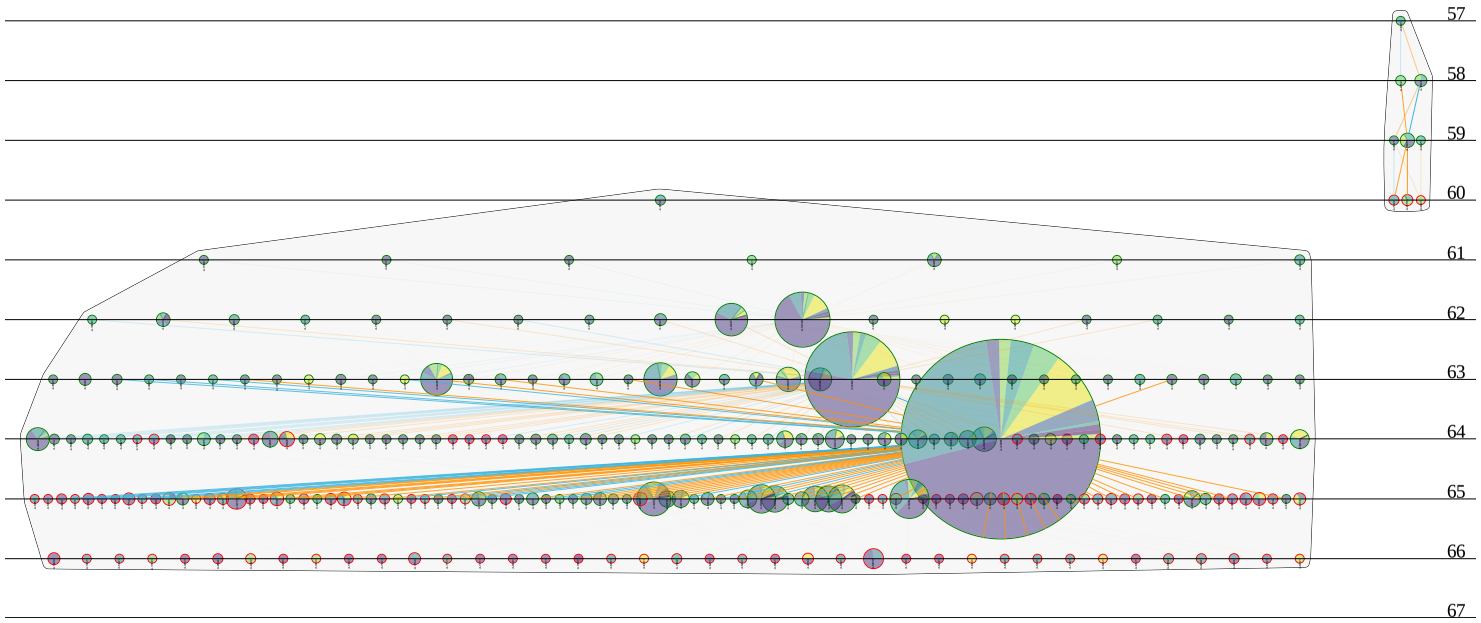
Proponowane rozwiązania są opracowanymi i gotowymi do wykorzystania metodami, które zostały zastosowane do różnego rodzaju zbiorów danych i problemów. Nadal istnieje jednak wiele potencjalnych sposobów na ich udoskonalenie, rozszerzenie i dalszy rozwój. Zaproponowane wartości wsparcia, zarówno dla duplikacji i specjacji zdefiniowane w Rozdziale 3, jak i dla zdarzeń HGT z Rozdziału 4, można rozszerzyć wprowadzając nieco inną definicję wsparcia, opartą na topologii poddrzew, a nie obecności klastrów. Przy takim podejściu uwzględniana byłaby nie tylko obecność liści, ale także ich rozkład topologiczny w drzewie. Pozwoliłoby to uchwycić bardziej szczegółowe zależności między drzewami genów, jednak należy się liczyć z tym, że uzyskiwane w ten sposób wartości wsparcia byłyby niższe niż te oparte na klastrach.

W przypadku algorytmu rozwiązującego problem rekonstrukcji relacji gen–gatunek potencjalne rozszerzenia obejmują metody analizy całych próbek metagenomicznych, które mogą zawierać sekwencje z wielu rodzin genów. Zależności między drzewami genów mogą pomóc we wprowadzeniu dodatkowych ograniczeń do mapowań nieznanych etykiet w drzewach.

Podejście oparte na sieciach filogenetycznych zastosowane do danych z projektu dotyczącego chłoniaka pęcherzykowego pozwoliło na zobrazowanie złożonych zależności między sekwencjami receptorów BCR pochodzącymi z limfocytów B. Byliśmy w stanie wymodelować ewolucję guza i zaobserwować selekcję subklonalną indukowaną przez mutacje BCR. Istnieje jednak jeszcze wiele możliwości rozszerzenia struktury sieci oraz wiele danych, takich jak sekwencje spoza regionu BCR, które można włączyć do informacji zawartych w sieci i które pozwolą poszerzyć naszą wiedzę na temat chłoniaka pęcherzykowego.



Rysunek 7: *Rekonstrukcja relacji gen–gatunek Po lewej:* Fragment drzewa gatunków z bazy SILVA z gatunkami *Euryarchaeota* obecnymi w zrekonstruowanych przyporządkowaniach genów do gatunków. *Po prawej:* Sześć grafik przedstawiających rozkład mapowań 9 nieznanych etykiet z drzewa genów (M1-M9). Każda z grafik odpowiada jednemu z zestawów kosztów zdarzeń ewolucyjnych. Koszty duplikacji (D), starty (L) i HGT (T) genów są przedstawione na górze. Minimalne koszty dla sześciu eksperymentów wynosiły odpowiednio 194, 220, 240, 325, 59 i 249.



Rysunek 8: Dwa podgrafy z przykładowej sieci WILLOW. Sekwencje subklonów to połączone łańcuchy ciężkie i lekkie regionu zmiennego receptora BCR. Sieć pokazuje relacje między subklonami, które różnią się tylko jedną pozycją w sekwencji. Kolor krawędzi uzależniony jest od tego, czy mutacja została wykryta w łańcuchu ciężkim (niebieski), czy lekkim (pomarańczowy). Krawędzie między większymi i bardziej istotnymi węzłami są grubsze niż pozostałe, natomiast krawędzie między małymi węzłami są usuwane. Kolorowe obramowania węzłów oznaczają, że węzeł jest liściem (czerwony) lub węzłem wewnętrznym (zielony). Każdy węzeł zawiera wykres kołowy przedstawiający rozkład ekspresji genów dla danego podklonu.

Wykaz publikacji głównych wyników przedstawionych w rozprawie:

Mykowiecka, A., & Górecki, P. (2016). Bootstrapping algorithms for gene duplication and speciation events. *In International Conference on Algorithms for Computational Biology* (pp. 106-118). Springer, Cham.

Mykowiecka, A., Szczęsny, P., & Górecki, P. (2017). Inferring gene-species assignments in the presence of horizontal gene transfer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5), 1571-1578.

Mykowiecka, A., & Górecki, P. (2018). Credibility of evolutionary events in gene trees. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3), 713-726.

Mykowiecka, A., Muszewska, A., & Górecki, P. (2018). Inferring time-consistent and well-supported horizontal gene transfers. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 79-83). IEEE.

van Bergen, C. A., Koning, M. T., Quinten, E., Mykowiecka, A., Sepulveda, J., Monajemi, R., et al. & Veelken, H. (2019). High-Throughput BCR Sequencing and Single-Cell Transcriptomics Reveal Distinct Transcriptional Profiles Associated with Subclonal Evolution of Follicular Lymphoma. *Blood*, 134, 298.

Wykaz wybranych publikacji w zakresie filogenetyki:

Szczesny, P., Mykowiecka, A., Pawłowski, K., & Grynberg, M. (2013). Distinct protein classes in human red cell proteome revealed by similarity of phylogenetic profiles. *PloS one*, 8(1), e54471.

Górecki, P., Paszek, J., & Mykowiecka, A. (2016). Mean values of gene duplication and loss cost functions. *In International Symposium on Bioinformatics Research and Applications* (pp. 189-199). Springer, Cham.

Górecki, P., Mykowiecka, A., Paszek, J., & Eulenstein, O. (2019). Mathematical properties of the gene duplication cost. *Discrete Applied Mathematics*, 258, 114-122.

Wawerka, M., Dąbkowski, D., Rutecka, N., Mykowiecka, A., & Górecki, P. (2021). Conflict Resolution Algorithms for Deep Coalescence Phylogenetic Networks. In 21st International Workshop on Algorithms in Bioinformatics (WABI 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Wawerka, M., Dąbkowski, D., Rutecka, N., Mykowiecka, A., & Górecki, P. (2022). Embedding gene trees into phylogenetic networks by conflict resolution algorithms. *Algorithms for Molecular Biology*, 17(1), 1-23.

Literatura

Bafna, V., Hannenhalli, S., Rice, K. and Vawter, L. (2000). Ligand-Receptor pairing via tree comparison. *J Comput Biol*, **7**, 59–70.

Bansal, M. S., Alm, E. J. and Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28** (12), i283–i291.

Beretta, S. and Dondi, R. (2014). Gene tree correction by leaf removal and modification: Tractability and approximability. *LNCS*, **8493**, 42–52.

Betkier, A., Szczęsny, P. and Górecki, P. (2015). Fast algorithms for inferring gene-species associations. *LNCS*, **9096**, 36–47.

Chaudhary, R., Burleigh, J. G. and Eulenstein, O. (2012). Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC Bioinformatics*, **13 Suppl 10**, S11.

Dondi, R., El-Mabrouk, N. and Swenson, K. M. (2014). Gene tree correction for reconciliation and species tree inference: Complexity and algorithms. *Journal of Discrete Algorithms*, **25**, 51–65.

Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllősi, G. J., Ranwez, V. and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *LNCS*, pp. 93–108.

Durand, D., Halldórsson, B. V. and Vernot, B. (2006). A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*, **13** (2), 320–335.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E. and Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, **28** (2), 132–163.

Górecki, P. (2004). Reconciliation problems for duplication, loss and horizontal gene transfer. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, RECOMB '04, New York, NY, USA: ACM, pp. 316–325.

Górecki, P. and Eulenstein, O. (2012). Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. *BMC Bioinformatics*, **13** (Suppl 10), S14.

— and Tiuryn, J. (2007). Inferring phylogeny from whole genomes. *Bioinformatics*, **23** (2), e116–e122.

Holland, B., Penny, D. and Hendy, M. (2003). Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. *Syst Biol*, **52**, 229–238.

Huelsenbeck, J. P., Bollback, J. P. and Levine, A. M. (2002). Inferring the Root of a Phylogenetic Tree. *Syst Biol*, **51** (1), 32–43.

Maddison, W. (1997). Gene trees in species trees.

Mykowiecka, A. and Górecki, P. (2018). Credibility of evolutionary events in gene trees (accepted). *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

—, Muszewska, A. and Górecki, P. (2018). Inferring time-consistent and well-supported horizontal gene transfers. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, pp. 79–83.

—, Szczęsny, P. and Górecki, P. (2017). Inferring gene-species assignments in the presence of horizontal gene transfer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **15** (5), 1571–1578.

Nguyen, T. H., Ranwez, V., Pointet, S., Chifolleau, A.-M. A., Doyon, J.-P. and Berry, V. (2013). Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms for Molecular Biology*, **8** (1), 12.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, **43** (1), 58–77.

Ranwez, V., Scornavacca, C., Doyon, J.-P. and Berry, V. (2015). Inferring gene duplications, transfers and losses can be done in a discrete framework. *Journal of Mathematical Biology*, pp. 1–34.

Scornavacca, C., Jacox, E. and Szöllősi, G. J. (2014). Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, **31** (6), 841–848.

—, Mayol, J. C. P. and Cardona, G. (2017). Fast algorithm for the reconciliation of gene trees and lgt networks. *Journal of theoretical biology*, **418**, 129–137.

Sjöstrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B. and Lagergren, J. (2014). A bayesian method for analyzing lateral gene transfer. *Systematic Biology*, **63** (3), 409–420.

Swenson, K., Doroftei, A. and El-Mabrouk, N. (2012). Gene tree correction for reconciliation and species tree inference. *Algorithm Mol Biol*, **7** (1), 31.

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. and Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, **62** (6), 901–912.

Taylor, J. S. and Raes, J. (2004). Duplication and divergence: The evolution. *Annual Review of Genetics*, **38**, 615–43.

Thompson, C. C., Thompson, F. L., Vandemeulebroecke, K., Hoste, B., Dawyndt, P. and Swings, J. (2004). Use of recA as an alternative phylogenetic marker in the family Vibrionaceae. *Int J Syst Evol Micr*, **54** (3), 919–924.

Tofigh, A., Hallett, M. and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8** (2), 517–535.

van Bergen, C. A., Koning, M. T., Quinten, E., Mykowiecka, A., Sepulveda, J., Monajemi, R., De Groen, R. A., Vermaat, J., Kloet, S. L., Kielbasa, S. M. *et al.* (2019). High-throughput bcr sequencing and single-cell transcriptomics reveal distinct transcriptional profiles associated with subclonal evolution of follicular lymphoma. *Blood*, **134**, 298.

Weisburg, W. G., Barns, S. M., Pelletier, D. A. and Lane, D. J. (1991). 16s ribosomal dna amplification for phylogenetic study. *J Bacteriol*, **173** (2), 697–703.

Yu, Y., Warnow, T. and Nakhleh, L. (2011). Algorithms for MDC-based multi-locus phylogeny inference. *Research in Computational Molecular Biology*, pp. 531–545.

Zhang, L. and Cui, Y. (2010). An efficient method for DNA-based species assignment via gene tree and species tree reconciliation. *LNCS*, **6293**, 300–311.