Studium zdarzeń duplikacji w genomie

autoreferat

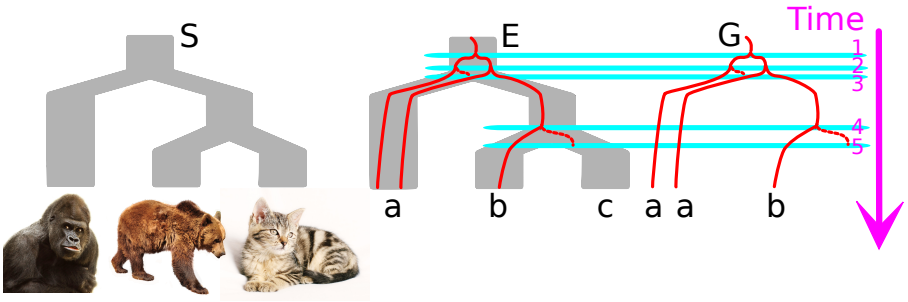
Jarosław Paszek

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

Jednym z fundamentalnych zagadnień w molekularnej biologii obliczeniowej jest wykrywanie zdarzeń duplikacji w genomie oraz określenie ich położenia w drzewie gatunków. Rekonstrukcja tych zdarzeń jest możliwa poprzez klastrowanie pojedynczych duplikacji genu, wyznaczonych przez uzgadnianie zbioru drzew genów ze zbiorem gatunków. Istniejące metody różnią się w dwóch zasadniczych kwestiach: (a) wyboru scenariuszy ewolucyjnych, które modelują dopuszczalne lokalizacje duplikacji w drzewie gatunków, oraz (b) określenia zasad klastrowania duplikacji genów z drzew genów w jedno zdarzenie wielokrotnej duplikacji, metod jak np. **episode clustering** lub **minimum episodes**. Analizując literaturę można wyróżnić kilka modeli opisujących jak duplikacje genów z drzew rodzin genów interpretować jako jedno zdarzenie, jednak wszystkie one dotyczą przypadku, gdy drzewa genów są ukorzenione. Warunek ten ogranicza możliwości zastosowań, gdyż to nieukorzenione drzewa genów są wynikiem popularnych metod filogenetycznych.

Klasyczny model ukorzenionego uzgadniania

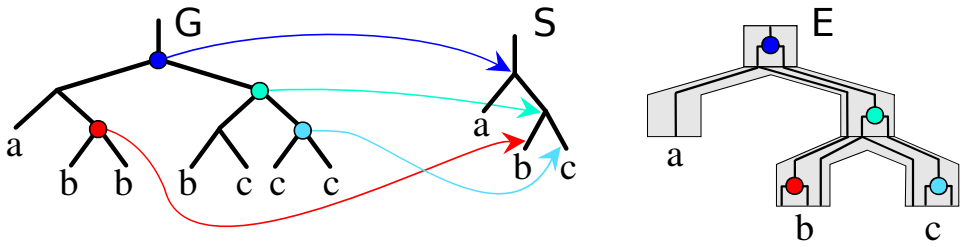
W niniejszej rozprawie proponujemy model uzgadniania, w którym ukorzenione drzewo genów jest uzgadniane z jego ukorzenionym drzewem gatunków. Goodman, w pracy [Goodman et al., 1979], wprowadza koncepcję uzgadniania, czyli wyjaśnienia różnic pomiędzy drzewem genów i drzewem gatunków za pomocą **mapowania** pomiędzy wierzchołkami tych drzew. Idea ta została sformalizowana w [Page, 1994, Page and Charleston, 1997], gdzie różnice pomiędzy drzewami wyjaśnione są przez zdarzenia ewolucyjne takie jak **duplikacja genu**, **strata genu** i **specjacja** (zob. Rysunek 1).



Rysunek 1. Przykład uzgadniania drzewa gatunków S i ukorzenionego drzewa rodziny genów G. Etykiety a, b, c przypisane są odpowiednio gatunkom małpy, niedźwiedzia i kota. Drzewo genów zawiera trzy geny: dwa od małpy i jeden od niedźwiedzia. W ukazanym przedziale czasu w scenariuszu ewolucyjnym miały miejsce następujące wydarzenia: 1 duplikacja genu, 2 specjacja, 3 strata genu, 4 specjacja i 5 strata genu. Scenariusz E ukazany jest jako wpisanie drzewa genów G w drzewo gatunków S zgodnie z zasadą, że "gatunki są pojemnikami na geny", którą uogólniamy do drzew.

Zdarzenia straty genów są częste zarówno u organizmów prokariotycznych jak i eukariotycznych [Koonin and Galperin, 2003, Sebat et al., 2004, Demuth et al., 2006], podczas gdy duplikacje uważa się za siłę napędową u organizmów eukariotycznych [Maere et al., 2005, Lynch and Conery, 2000, Lynch and Conery, 2003, Fischer et al., 2014].

W tej rozprawie, wykorzystujemy ideę interpretowania uzgadniania jako modelu biologicznie spójnych scenariuszy, które są wpisaniem drzewa genów w drzewo gatunków określającym lokalizacje zdarzeń ewolucyjnych w drzewie gatunków [Górecki and Tiuryn, 2006]. Scenariusz wyznaczony jest przez funkcję nazywaną **mapowaniem duplikacji**, która wierzchołkowi drzewa genów, interpretowanemu jako zdarzenie **pojedynczej duplikacji genu**, przypisuje wierzchołek drzewa gatunków [Guigó et al., 1996, Paszek and Górecki, 2016, Page and Cotton, 2002, Bansal and Eulenstein, 2008, Burleigh et al., 2008, Nøjgaard et al., 2017, Mettanant and Fakcharoenphol, 2008, Luo et al., 2011, Burleigh et al., 2010] (zob. Rysunek 2).



Rysunek 2. Przykład uzgadniania drzewa genów G z drzewem gatunków S za pomocą LCA-mapowania (ang. *least common ancestor*). LCA-mapowania są wskazane strzałkami. Wyznaczone są w oparciu o etykiety liści, a pełnią kluczową rolę w interpretacji zdarzeń makroewolucyjnych zlokalizowanych w drzewie genów i drzewie gatunków.

Kluczowymi własnościami modelu uzgadniania danego ukorzenionego drzewa genów i odpowiadającego mu drzewa gatunków są:

- uzgadnianie ma liniową złożoność czasową i pamięciową [Page, 1994, Ma et al., 2000],
- istnieje dokładnie jeden scenariusz oparty na LCA uzgadnianiu, który minimalizuje sumę zdarzeń duplikacji i strat genów [Bonizzoni et al., 2005, Górecki and Tiuryn, 2006],
- jednak tylko dla kosztu duplikacyjnego, taki scenariusz jest nieunikalny [Górecki and Tiuryn, 2006].
- jest nieskończenie wiele scenariuszy dla tych drzew [Górecki and Tiuryn, 2006].

LCA model składa się z jednego scenariusza, najbardziej parsymonicznego, który posiada minimalną sumę zdarzeń duplikacji i strat genów. Formalną definicję tego modelu zaprezentowaliśmy w [Paszek and Górecki, 2017a].

Zdarzenia duplikacji genomowych

W Rozdziale 1 rozprawy przedstawiamy biologiczne spojrzenie na ewolucję. Proces ten głównie przebiega spokojnie, bez zdarzeń, za to czasem przerywany jest wydarzeniami zwanymi *eksplozjami*, w których pojawia się wiele nowych gatunków. Jedna z

takich eksplozji odpowiedzialna jest za pojawienie się organizmów eukariotycznych, które posiadają dwa razy więcej genów niż organizmy prokariotyczne [Brown, 2002]. Ohno sugeruje, że przyczyną różnorodności organizmów są zdarzenia duplikacji na wielką skalę, po których następuje noefunkcjonalizacja lub straty genów [Ohno, 1970].

W niniejszej rozprawie analizujemy problemy dotyczące zdarzeń wielokrotnych duplikacji genów. Szczególnym przypadkiem takich zdarzeń są duplikacje całych genomów (ang. *whole-genome duplications*, WGD), które są częste u roślin. Poliploidalność czy hybrydyzacja jest szansą na uzyskanie nowych cech dlatego jest badana dla roślin uprawnych jak i chwastów. W Rozdziale 1 opisujemy wybrane wyniki badań. Podsumowując, wykrywanie duplikacji całych genomów jest pożądanym celem.

W Rozdziale 3 przedstawiamy metody detekcji duplikacji całych genomów, które możemy podzielić na trzy kategorie: bazujące na syntenii i porównywaniu kolinearności genomów [Kellis et al., 2004, Tang et al., 2008, Holloway et al., 2013], estymacji rozkładu wieku dla par paralogicznych genów [Vision et al., 2000, Lynch and Conery, 2000, Blanc and Wolfe, 2004], wnioskowaniu wspartym przez drzewa filogenetyczne [Bowers et al., 2003, Jiao et al., 2011, Rabier et al., 2014]. Metody przedstawione w ramach niniejszej rozprawy można określić jako filogenetyczne, co jest nowym podejściem.

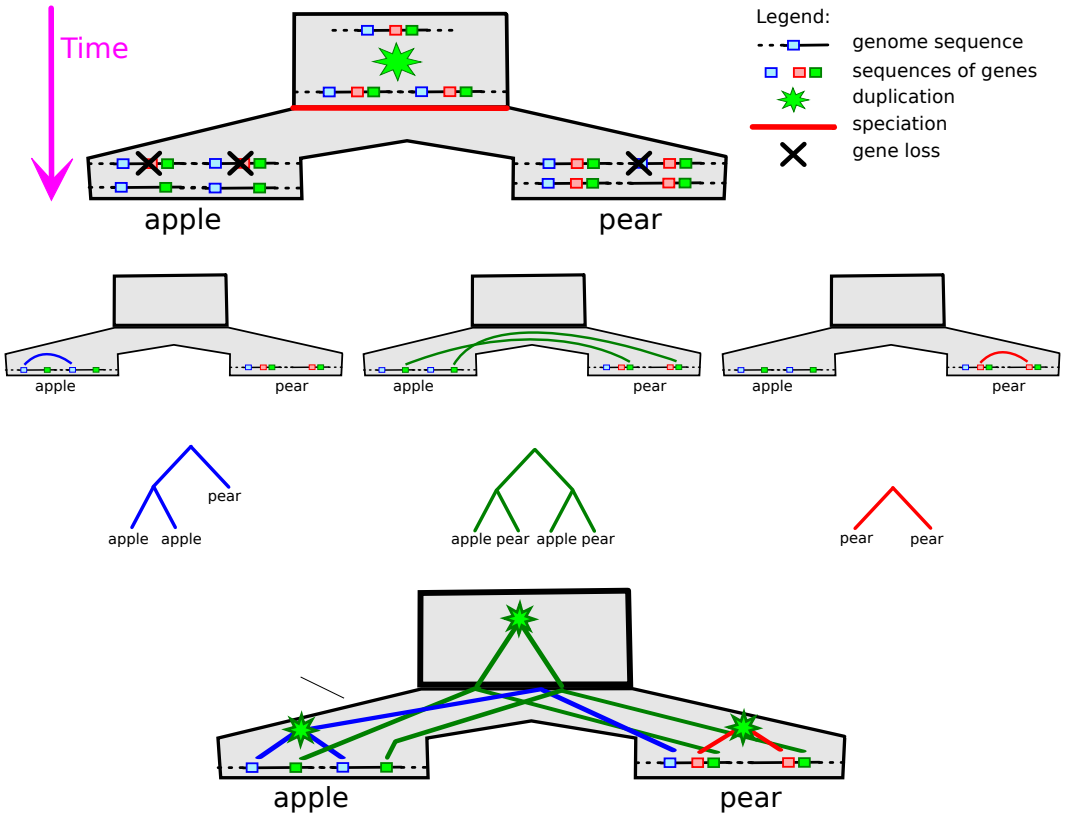
Duplikacje genomowe

Zagadnienie odkrywania lokalizacji pojedynczych i wielokrotnych duplikacji genów jest kluczowe do zrozumienia sposobu ewolucji rodzin genów i genomów. W szczególności interesujące są **duplikacje genomowe**, lub **wielokrotne duplikacje genów**, w których duplikacji ulegają części genomu i które mogą obejmować tysiące rodzin genów. W rzeczywistości po takim wydarzeniu często następuje wiele strat oraz rearanżacji genów i w konsekwencji rekonstrukcja takich wydarzeń może być trudna (zob. Rysunek 3). Uzgadnianie drzewa jednej rodziny genów z drzewem gatunków jest relatywnie proste z obliczeniowego punktu widzenia, jednak gdy skupiamy się na wielokrotnych duplikacjach genów, problem staje się bardziej złożony.

Wykorzystując uzgadnianie potrafimy określić lokalizację duplikacji genów w drzewie gatunków. Następnie, możemy wywnioskować zdarzenie duplikacji genomowej poprzez **grupowanie** pojedynczych zdarzeń duplikacji zlokalizowanych w tym samym wierzchołku drzewa gatunków. Teraz sformułujemy ogólną koncepcję problemu duplikacji genomowych, jako problemu klastrowania w następującej postaci:

Dane: zbiór drzew genów i odpowiadające im drzewo gatunków.
Znajdź: klastrowanie wszystkich pojedynczych duplikacji genów o minimalnym rozmiarze.

Powyższe sformułowanie problemu dla klasycznego opartego na LCA uzgadniania (zob. Rysunek 2), wymaga idealnych drzew z kompletnymi danymi bez błędów. W praktyce błędy w sekwencjonowaniu, ograniczenia obliczeniowe lub biologiczne procesy jak straty genów lub horyzontalny transfer genów sprawiają, że otrzymanie idealnego drzewa jest niemożliwym zadaniem. Dlatego, w ogólności, LCA uzgadnianie, które umiejscawia pojedyncze zdarzenie duplikacji na najniższym możliwym wierzchołku drzewa gatunków, nie jest odpowiednie do modelowania zdarzeń wielokrotnych duplikacji genów (zob. Rysunek 3).



Rysunek 3. Przykład ewolucji, gdy miało miejsce zdarzenie duplikacji wielu genów. **U góry:** Drzewo gatunków, w którym każdy wierzchołek reprezentuje gatunek, obrazuje relację pomiędzy trzema gatunkami: gruszką, jabłkiem oraz przodkiem gruszki i jabłka. Wewnątrz wierzchołka zaznaczony jest fragment sekwencji odpowiedniego gatunku. W wyniku wielokrotnej duplikacji (zaznaczonej przez zieloną gwiazdę w korzeniu drzewa), sekwencje trzech genów (oznaczonych kolorami czerwonym, zielonym, niebieskim) zostały zduplikowane u przodka gruszki i jabłka. Procesy odpowiedzialne za straty genów powodują, że obecne sekwencje genomów gruszki i jabłka zawierają różne kombinacje genów. **W środku:** Przykłady konstrukcji drzew genów. **Powyżej:** Drzewo gatunków, w którym wierzchołek reprezentuje gatunek. Wewnątrz liści zaznaczono fragmenty sekwencji obecnych gatunków. Liniami pokazano sekwencje zidentyfikowanych genów homologicznych, czyli posiadających wspólnego przodka. **Poniżej:** Ukorzenione drzewa genów uzyskane w oparciu o podobieństwo sekwencji. **Na dole:** Przykład scenariusza ewolucyjnego w LCA uzgadnianiu. Drzewa genów uzyskane z sekwencji sugerują trzy zdarzenia pojedynczych duplikacji genów zamiast jednego zdarzenia wielokrotnej duplikacji pokazanej w górnej części rysunku.

Modele dopuszczalnych scenariuszy ewolucyjnych

Guigo et al. w pionierskiej pracy [Guigó et al., 1996] opisuje podejście do wykrywania zdarzeń wielokrotnych duplikacji, w którym proponuje, aby rozluźnić LCA uzgadnianie poprzez dopuszczenie kilku dodatkowych lokalizacji dla pojedynczych zdarzeń duplikacji. W rezultacie proponowana metoda może prowadzić do zwiększenia kosztu uzgadniania.

W Rozdziale 3 opisujemy modele, które dla danych drzewa genów i drzewa gatunków, określają zbiór najlepszych scenariuszy ewolucyjnych, zwanych **dopuszczal-**

nymi scenariuszami, ocenionych według kryteriów zdefiniowanych przez model. Rozróżniamy następujące modele dopuszczalnych scenariuszy:

- FHS - opisany przez Fellows et al. w [Fellows et al., 1998], w którym każdy scenariusz dla G i S jest dopuszczalny.
- PG - zaproponowany przez Paszek and Górecki w [Paszek and Górecki, 2016], gdzie $\mathsf{PG}(G, S)$ jest zbiorem wszystkich scenariuszy dla G i S, które posiadają minimalną liczbę duplikacji genów.
- GMS - przedstawiony przez Guigó et al. [Guigó et al., 1996] ograniczony model. Niech g' jest rodzicem duplikacji g w G, a v z S jest LCA mapowaniem g'. W modelu tym nie są dopuszczalne scenariusze, w których duplikacja g z G mapuje się na v, lub przodka v w S. Zatem przykładowo scenariusze, gdzie g' jest mapowany powyżej v, a g jest mapowane na v, nie są dopuszczalne.
- LCA - najprostszy model, który opisuje tylko jeden scenariusz zdefiniowany przez LCA mapowanie.

Model interwałowy definiuje scenariusze poprzez opisanie dopuszczalnych lokalizacji duplikacji jako ścieżki w drzewie gatunków. Ponadto, wprowadza on ograniczenia takie jak monotoniczność, aby zapewnić spójność biologiczną. Duplikacja nie powinna być mapowana wyżej od mapowania swojego rodzica. Modele LCA, GMS i PG są modelami interwałowymi.

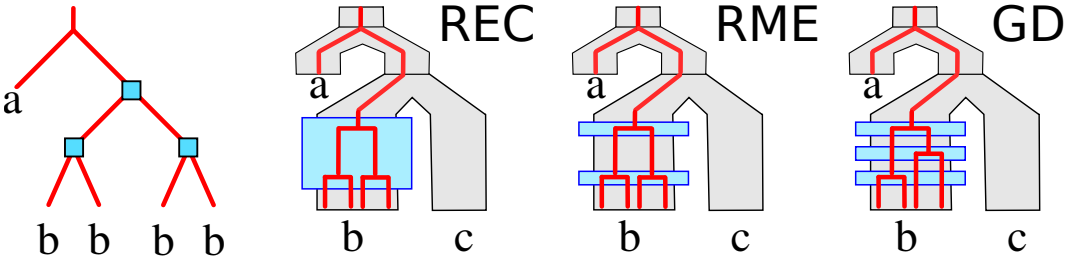
Reguły klastrowania pojedynczych duplikacji genów

Dwa fundamentalne aspekty problemów duplikacji genomów to: (1) model dopuszczalnych scenariuszy i (2) reguły klastrowania duplikacji genów z drzew genów w pojedyncze zdarzenie wielokrotnej duplikacji.

W celu zapewnienia dokładnego modelu dla wielu jednoczesnych duplikacji, potrzebujemy dodatkowych reguł, które określą, kiedy dwie pojedyncze duplikacje mogą być sklastrowane. Mając dany model dopuszczalnych scenariuszy, możemy wyróżnić trzy warianty **problemów wielokrotnych duplikacji genów** (zob. Rysunek 4), które różnią się regułami:

- **Episode Clustering** (REC) - duplikacje genów mogą być sklastrowane jeśli są zmapowane na ten sam wierzchołek drzewa gatunków [Guigó et al., 1996, Page and Cotton, 2002, Bansal and Eulenstein, 2008],
- **Minimum Episodes Clustering** (RME) - duplikacje z tego samego drzewa genów mogą być sklastrowane jeśli nie są porównywalne i są zmapowane na ten sam wierzchołek drzewa gatunków [Guigó et al., 1996, Bansal and Eulenstein, 2008],
- **Gene Duplication Clustering** (GD) - duplikacje występujące w tym samym drzewie genów nie mogą być sklastrowane [Fellows et al., 1998].

Powyższe reguły zdefiniowane są dla ukorzenionych drzew. W celu podkreślenia tego faktu, do skrótów klastrowań *episode clustering* i *minimum episodes* dodajemy prefiks R.



Rysunek 4. Przykład klastrowań duplikacji REC, RME i GD dla jednego drzewa genów. **Lewo:** drzewo genów z trzema duplikacjami zmapowanymi na *b*. **Prawo:** wpisania (scenariusze) drzewa genów w drzewo gatunków pokazujące rozwiązania problemów REC, RME, i GD, odpowiednio. W REC wszystkie duplikacje są sklastrowane razem, podczas gdy w RME górna duplikacja nie może być sklastrowana z jej dziećmi. Zatem rozwiązanie RME składa się z dwóch klastrów oznaczonych prostokątami. Klastrowanie GD daje 3 klastry, gdyż duplikacje z tego samego drzewa nie mogą być sklastrowane razem.

Problem REC można traktować jako uproszczoną wersję ogólnego problemu duplikacji genomowych, w którym celem jest znalezienie tylko minimalnej liczby wierzchołków w drzewie gatunków, gdzie wystąpiły zdarzenia wielokrotnej duplikacji genów. Innymi słowy, rozwiązania REC raczej stanowią pewne przybliżenie zdarzeń duplikacji genomowych. Przykładowo, jeśli dwa WGD znajdują się między dwoma kolejnymi specjacjami, to zostaną sklastrowane jako jedno zdarzenie wielokrotnej duplikacji. Z biologicznego punktu widzenia, najbardziej pożądane są rozwiązania RME (zob. Rysunek 4).

Przegląd powiązanych prac

Problem REC polega na znalezieniu scenariuszy ewolucyjnych o minimalnej liczbie lokalizacji zdarzeń duplikacji w drzewie gatunków. Innymi słowy, dwie duplikacje mogą być sklastrowane, jeśli mają tą samą lokalizację w drzewie gatunków. Problem ten został wprowadzony przez Guigó et al. [Guigó et al., 1996] wraz modelem GMS i heurystycznym rozwiązaniem. Page a Cotton [Page and Cotton, 2002] sformułowali problem lokalizacji zdarzeń duplikacji genów jako problem pokrycia zbioru i zaproponowali heurystykę. Bansal i Eulenstein [Bansal and Eulenstein, 2008] przedstawili wielomianowy algorytm dla problemu REC dla modelu GMS, który jest specjalnym przypadkiem problemu *Tree Interval Cover, (TIC)* [Burleigh et al., 2008]. Burleigh et al. [Burleigh et al., 2008] przedstawił wielomianowe rozwiązanie problemu TIC. Wreszcie, Luo et al. [Luo et al., 2011] zaproponował algorytm liniowy dla problemu TIC, który rozwiązuje REC dla każdego modelu interwałowego. Problem REC dla modelu FHS ma trywialne rozwiązanie z jednym klastrem.

Problem GD podobny jest do problemu REC z taką różnicą, że klaster nie może zawierać dwóch duplikacji z tego samego drzewa. Problem GD dla modelu FHS jest NP-trudny [Fellows et al., 1998].

Klastrowanie dla problemu REC można ulepszyć przez wykluczenie przypadków, w których duplikacja i jej przodek duplikacja, z tego samego drzewa genów, są sklastrowane razem. Takie sformułowanie problemu wielokrotnych duplikacji genów nazywamy problemem RME [Guigó et al., 1996, Bansal and Eulenstein, 2008] (zob. Rysunek 4).

Pierwszy wielomianowy algorytm dla problemu RME dla modelu GMS zaproponowano w [Bansal and Eulenstein, 2008], natomiast optymalny liniowy algorytm w [Mettanant and Fakcharoenphol, 2008, Luo et al., 2011]. Idea interwałów została wprowadzona w [Czabarka et al., 2012] dość ogólnie, bez założenia, że interwały indukują biologicznie spójne scenariusze ewolucyjne. Iteracyjny algorytm z [Czabarka et al., 2012] zaimplementowany w bezpośredni sposób ma złożoność $O(|S|^2|G|)$ ([Czabarka et al., 2012] sugeruje, że algorytm z [Bansal and Eulenstein, 2008] rozwiązuje instancje dla każdego modelu interwałowego, jednakże, jest on zaprojektowany dla modelu GMS i nie może zostać uogólniony). Podsumowując, nie istniał ogólny algorytm, który rozwiązuje RME. W Rozdziale 4 przedstawiamy rozwiązania RME dla wielu modeli, w szczególności, liniowy algorytm uniwersalny dla wszystkich modeli interwałowych.

Co więcej, nie istniały rozwiązania dla wersji problemów, kiedy wejściowe drzewa genów są nieukorzenione.

Podsumowując, często problemy biologii obliczeniowej są zdefiniowane w dwóch wersjach: dla ukorzenionych i nieukorzenionych drzew genów. Dla klasycznych problemów biologii obliczeniowej, tworzy się algorytmy dla ich wersji nieukorzenionej [Górecki and Eulenstein, 2011, Górecki and Eulenstein, 2012b, Górecki and Eulenstein, 2012a, Chang et al., 2013, Górecki and Eulenstein, 2014, Betkier et al., 2015]. Rozwiązanie problemu w wersji nieukorzenionej jest czasem bardziej pożądane, gdyż ukorzenienie drzewa genów może być trudne. W Rozdziale 5 i Rozdziale 6 przedstawiamy rozwiązania dla problemów **Unrooted Episode Clustering,** UEC, i **Unrooted Minimum Episodes**, UME (zob. [Paszek and Górecki, 2016, Paszek and Górecki, 2017b, Paszek and Górecki, 2018]). Kluczowym wynikiem, który umożliwił stworzenie efektywnego rozwiązania UME, był liniowy algorytm dla RME uniwersalny dla modeli interwałowych, opisany w Rozdziale 4 (zob. [Paszek and Górecki, 2017a]).

Rozwiązanie problemu RME

W Rozdziale 4 koncentrujemy się na problemie duplikacji genomowych, gdy wszystkie drzewa genów są ukorzenione. W szczególności, przedstawiamy liniowy algorytm dla problemu RME uniwersalny dla modeli interwałowych, w tym modeli GMS i PG. Następnie, opisujemy jego wykorzystanie do rozwiązania problemu RME dla najbardziej wymagającego obliczeniowo modelu FHS.

W pracy [Czabarka et al., 2012] jest opis rozwiązania problemu duplikacyjnego, jednak, proponowany model interwałów był zastosowany bez wymogu, aby interwały modelowały poprawne scenariusze ewolucyjne. Autorzy przedstawili ogólny iteracyjny algorytm do obliczania RME dla modelu interwałowego. W pracy [Czabarka et al., 2012] jest dowód poprawności i stwierdzenie, że algorytmu z [Bansal and Eulenstein, 2008] można użyć do efektywnego rozwiązania problemu. Jednakże, drugi algorytm jest opracowany do rozwiązania problemu RME dla modelu GMS i nie może być uogólniony z powodu prostego modelu interwałów (np. interwały dla porównywalnych duplikacji w GMS przecinają się w najwyżej jednym wierzchołku). Naiwna implementacja algorytmu [Czabarka et al., 2012] ma złożoność $O(|S||G|^2)$.

Podsumowując, dla problemu RME mamy następujące wyniki: złożoność dla FHS jest nieznana, dla GMS można go rozwiązać w czasie liniowym [Luo et al., 2011], podczas gdy dla PG w czasie $O(|S|^2|G|)$ przez naiwną implementację idei z [Czabarka et al., 2012].

W Rozdziale 4, proponujemy liniowy algorytm dla RME uniwersalny dla modeli interwałowych. Ponadto, opisujemy heurystykę efektywną na rzeczywistych danych biologicznych, która rozwiązuje RME dla FHS.

Rozdział 4 opisuje główne wyniki z pracy [Paszek and Górecki, 2017a].

Rozwiązanie problemu UEC

Uzgadnianie staje się bardziej złożone, gdy rozważamy nieukorzenione drzewa genów w miejsce ukorzenionych. Podobnie, problemy REC i RME są zdefiniowane dla drzew ukorzenionych, zatem, aby rozwiązać nieukorzeniony wariant problemu jak przykładowo UEC, musimy wybrać krawędź z nieukorzenionego drzewa genów, wyznaczyć ukorzenienie, i dopiero rozwiązać problem REC dla tego ukorzenienia. W Rozdziale 5 badamy problem UEC, w którym: (a) modelem dopuszczalnych scenariuszy jest PG, (b) reguły definiujące jak grupować duplikacje to *episode clustering* i (c) drzewa genów są nieukorzenione.

Prezentujemy nowe teorie z dziedziny nieukorzenionego uzgadniania, które można wykorzystać do rozwiązania UEC. Przedstawiamy pierwsze rozwiązanie otwartego problemu UEC zaproponowanego w [Burleigh et al., 2008]. Pokazujemy, że dla danego zbioru nieukorzenionych drzew genów i drzewa gatunków, możemy rozwiązać UEC poprzez redukcję do problemów REC, które mają złożoność liniową. Nasze rozwiązanie wymaga preprocessingu w czasie liniowym oraz wygenerowania maksymalnie $1 + 2^k$ zbiorów drzew ukorzenionych, czyli instancji REC, gdzie k jest liczbą wejściowych drzew genów posiadających specjalną topologię (formalnie, warunkiem jest wystąpienie dwóch gwiazd S2 [Górecki and Tiuryn, 2007]). Zazwyczaj k reprezentuje małą część całego wejścia i zastosowana redukcja znacznie poprawia złożoność. Innymi słowy, pokazujemy algorytm FPT (ang. fixed parameter tractable) dla problemu UEC.

Prezentujemy analizę eksperymentalną implementacji naszych algorytmów. W szeregu empirycznych obliczeniowych eksperymentów pokazujemy, że pomimo wykładniczej pesymistycznej złożoności, nasz algorytm rozwiązuje instancje problemu po weryfikacji co najwyżej dwóch ukorzenionych zbiorów drzew. W konsekwencji, nasze rozwiązanie można efektywnie wykorzystać do lokalizacji klastrów duplikacji dla zbiorów nieukorzenionych drzew genów.

Rozdział 5 bazuje na wynikach opublikowanych w [Paszek and Górecki, 2016].

Rozwiązanie problemu UME

W Rozdziale 6 badamy duplikacje genomowe dla nieukorzenionych drzew genów, klastrowania *minimum episodes*, oraz modelu dopuszczalnych scenariuszy PG, który zachowuje minimalną liczbę pojedynczych duplikacji genu. Rozwiązanie tego problemu dla ukorzenionych drzew genów (RME) opisane jest w Rozdziale 4.

Rozszerzyliśmy teorię nieukorzenionego uzgadniania, prezentując nowe własności plateau, czyli poddrzewa nieukorzenionego drzewa genów, które składa się z krawędzi, dla których ukorzenienie ma minimalny koszt duplikacyjny. Następnie, pokazujemy, że dzięki tym własnościom możemy dokonać dekompozycji nieukorzenionego drzewa genów, która pozwoli w znacznym stopniu ograniczyć przestrzeń poszukiwań.

Rozwiązanie problemu opisane jest w Rozdziale 6. Zgodnie z naszą wiedzą, złożoność UME jest nieznana. Pokazujemy, że każdą instancję UME można zastąpić

co najwyżej 5^k "prostszymi" instancjami, które można rozwiązać w czasie liniowym, gdzie k jest ograniczone z góry przez specjalne przypadki gwiazd S2 [Górecki and Tiuryn, 2007] w wejściowych drzewach. Następnie, proponujemy dwa liniowe algorytmy do obliczania granic rozwiązania. Ostatecznie, dla przypadku dużego k, przedstawiamy efektywny algorytm heurystyczny, który w praktyce pozwala na otrzymanie dokładnych rozwiązań dla empirycznych instancji składających się z tysięcy nieukorzenionych drzew genów.

Nasza eksperymentalna ewaluacja na rzeczywistych zbiorach danych potwierdziła kilka zdarzeń duplikacji genomowych z literatury i tym samym zademonstrowała, że algorytmu można z sukcesem wykorzystać. Przykładowo, znaleźliśmy 227 zdarzeń duplikacji dla zbioru TreeFam i wskazaliśmy wierzchołki (m.in. $D1$ i $D2$), które są potencjalnymi miejscami wielokrotnych duplikacji genów Dwa kolejne zdarzenia WGD w $D1$ oraz jedno zdarzenie WGD w $D2$ opisano w [Hufton et al., 2008, Inoue et al., 2015, Braasch and Postlethwait, 2012].

Rozdział 6 opiera się na wynikach z [Paszek and Górecki, 2018] i [Paszek and Górecki, 2017b].

Podsumowanie

W niniejszej rozprawie przedstawiliśmy nowe wyniki teoretyczne z zakresu nieukorzenionego uzgadniania, które zostały wykorzystane do opracowania rozwiązań dla kilku problemów algorytmicznych.

Ponadto, zaproponowaliśmy model dopuszczalnych scenariuszy ewolucyjnych, który zachowuje minimalną liczbę pojedynczych duplikacji genów. Przedstawiliśmy również biologiczną motywację dla modelu, wraz ze studium porównawczym z istniejącymi modelami.

Opracowaliśmy pierwszy liniowy algorytm dla RME uniwersalny modeli interwałowych, oraz rozwiązania dla otwartych problemów UME i UEC, które są zdefiniowane dla nieukorzenionych drzew genów i naszego modelu.

Wynikiem popularnych metod filogenetycznych są nieukorzenione drzewa rodzin genów, zatem nasze rozwiązanie poszerza możliwości stosowania metod klastrowania duplikacji. Co więcej, pokazaliśmy, że podejście nieukorzenione może ulepszyć znane wyniki wnioskowania duplikacji genomowych z ukorzenionych drzew.

Eksperymentalna ewaluacja na biologicznych zbiorach danych wykazała, że możemy wnieść nowe wyniki w wnioskowaniu duplikacji genomowych.

W przyszłości planujemy dalsze testowanie modeli dopuszczalnych scenariuszy. Naszym celem jest, przy współpracy z biologami, wykorzystanie implementacji stworzonych algorytmów do badań wielokrotnych duplikacji, jak duplikacje całych genomów (WGD). Obecnie potrafimy wykryć zarówno zdarzenia duplikacji całych genomów, jak i zdarzenia hybrydyzacji.

Wykaz publikacji głównych wyników przedstawionych w rozprawie:

Paszek, J. and Górecki, P. (2016). Genomic duplication problems for unrooted gene trees. *BMC Genomics*, 17(1):165–175. doi: 10.1186/s12864-015-2308-4.

Paszek, J. and Górecki, P. (2017a). Efficient algorithms for genomic duplication models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* doi: 10.1109/TCBB.2017.2706679.

Paszek, J. and Górecki, P. (2017b). New algorithms for the genomic duplication problem. *In: Meidanis J., Nakhleh L. (eds) Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science*, 10562:101–115. Springer, Cham.

Paszek, J. and Górecki, P. (2018). Inferring duplication episodes from unrooted gene trees. *BMC Genomics*, 19(5):288. doi: 10.1186/s12864-018-4623-z.

Wykaz wybranych publikacji w zakresie filogenetyki:

Górecki, P., Paszek, J., and Eulenstein, O. (2017). Unconstrained Diameters for Deep Coalescence. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(5):1002–1012. doi: 10.1109/TCBB.2016.2520937.

Górecki, P., Markin, A., Mykowiecka, A., Paszek, J., and Eulenstein, O. (2017). Phylogenetic tree reconciliation: Mean values for fixed gene trees. *LNCS*, 10330:234–245.

Górecki, P., Paszek, J., and Mykowiecka, A. (2016). Mean values of gene duplication and loss cost functions. *LNCS*, 9683:189–199.

Górecki, P., Paszek, J., and Eulenstein, O. (2014). Duplication Cost Diameters. *LNCS*, 8492:212–223.

Górecki, P., Paszek, J., and Eulenstein, O. (2014). Unconstrained gene tree diameters for deep coalescence. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14, pages 114–121. (best paper - nagroda przyznana przez ACM SIGBio w 2014)

Inne wybrane publikacje:

Gambin, T., Startek, M., Walczak, K., Paszek, J., Grzebelus, D., and Gambin, A. (2013). Tirfinder: A web tool for mining class ii transposons carrying terminal inverted repeats. *Evolutionary Bioinformatics*, 9:17–27.

Huczko, A., Lange, H., Paszek, J., Bystrzejewski, M., Cudziło, S., Gachet, S., Monthioux, M., Zhu, Y. Q., Kroto, H. W., and Walton, D. R. M. (2005). A simple route to new 1d nanostructures. *In: Intelligence in a Small Materials World. Selected Papers from IPMM-2003 The Fourth International Conference on Intelligent Processing and Manufacturing of Materials*, pages 721–729. DEStech Publications, Inc., Lancaster, Pennsylvania, USA.

Bibliografia

[Bansal and Eulenstein, 2008] Bansal, M. S. and Eulenstein, O. (2008). The multiple gene duplication problem revisited. *Bioinformatics*, 24(13):i132–8.

[Betkier et al., 2015] Betkier, A., Szczęsny, P., and Górecki, P. (2015). Fast algorithms for inferring gene-species associations. *Lecture Notes in Computer Science*, 9096:36–47.

[Blanc and Wolfe, 2004] Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell*, 16(7):1667–78.

[Bonizzoni et al., 2005] Bonizzoni, P., Della Vedova, G., and Dondi, R. (2005). Reconciling a gene tree to a species tree under the duplication cost model. *Theor Comput Sci*, 347(1-2):36–53.

[Bowers et al., 2003] Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–8.

[Braasch and Postlethwait, 2012] Braasch, I. and Postlethwait, J. H. (2012). *Polyploidy in Fish and the Teleost Genome Duplication*, pages 341–383. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Brown, 2002] Brown, T. (2002). *Genomes 2nd edition*. Oxford, United Kingdom: Wiley-Liss. available at https://www.ncbi.nlm.nih.gov/books/NBK21134/.

[Burleigh et al., 2010] Burleigh, J. G., Bansal, M. S., Eulenstein, O., and Vision, T. J. (2010). Inferring species trees from gene duplication episodes. *ACM BCB*, pages 198–203.

[Burleigh et al., 2008] Burleigh, J. G., Bansal, M. S., Wehe, A., and Eulenstein, O. (2008). Locating multiple gene duplications through reconciled trees. In Vingron, M. and Wong, L., editors, *RECOMB*, volume 4955 of *Lect Notes Comput Sc*, pages 273–284, Berlin, Germany. Springer.

[Chang et al., 2013] Chang, W., Górecki, P., and Eulenstein, O. (2013). Exact solutions for species tree inference from discordant gene trees. *J Bioinform Comput Bio*, 11(5).

[Czabarka et al., 2012] Czabarka, E., Székely, L., and Vision, T. (2012). Minimizing the number of episodes and gallai's theorem on intervals. *arXiv:1209.5699*.

[Demuth et al., 2006] Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N., and Hahn, M. W. (2006). The evolution of mammalian gene families. *PLoS One*, 1:e85.

[Fellows et al., 1998] Fellows, M., Hallet, M., and Stege, U. (1998). On the multiple gene duplication problem. In *9th International Symposium on Algorithms and Computation (ISAAC'98), Lecture Notes in Computer Science 1533*, pages 347–356, Taejon, Korea.

[Fischer et al., 2014] Fischer, I., Dainat, J., Ranwez, V., Glémin, S., Dufayard, J.-F., and Chantret, N. (2014). Impact of recurrent gene duplication on adaptation of plant genomes. *BMC Plant Biol*, 14:151.

[Goodman et al., 1979] 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. *Syst Zool*, 28(2):132–163.

[Górecki and Eulenstein, 2011] Górecki, P. and Eulenstein, O. (2011). A linear time algorithm for error-corrected reconciliation of unrooted gene trees. In *ISBRA*, pages 148–159.

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

[Górecki and Eulenstein, 2012b] Górecki, P. and Eulenstein, O. (2012b). GTP supertrees from unrooted gene trees: linear time algorithms for nni based local searches. *Lect Notes Comput Sc*, 7292:83–105.

[Górecki and Eulenstein, 2014] Górecki, P. and Eulenstein, O. (2014). Refining discordant gene trees. *BMC Bioinformatics*, 15(13):S3.

[Górecki and Tiuryn, 2006] Górecki, P. and Tiuryn, J. (2006). DLS-trees: A model of evolutionary scenarios. *Theor Comput Sci*, 359(1-3):378–399.

[Górecki and Tiuryn, 2007] Górecki, P. and Tiuryn, J. (2007). Inferring phylogeny from whole genomes. *Bioinformatics*, 23(2):e116–e122.

[Guigó et al., 1996] Guigó, R., Muchnik, I. B., and Smith, T. F. (1996). Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol*, 6(2):189–213.

[Holloway et al., 2013] Holloway, P., Swenson, K., Ardell, D., and El-Mabrouk, N. (2013). Ancestral genome organization: An alignment approach. *Journal of Computational Biology*, 20(4):280–295.

[Hufton et al., 2008] Hufton, A. L., Groth, D., Vingron, M., Lehrach, H., Poustka, A. J., and Panopoulou, G. (2008). Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res.*, 18(10):1582–1591.

[Inoue et al., 2015] Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., and Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 112(48):14918–14923.

[Jiao et al., 2011] Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., and dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100.

[Kellis et al., 2004] Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428:617–624.

[Koonin and Galperin, 2003] Koonin, E. and Galperin, M. (2003). *Sequence - evolution - function: computational approaches in comparative genomics*. Kluwer Academic.

[Luo et al., 2011] Luo, C.-W., Chen, M.-C., Chen, Y.-C., Yang, R. W. L., Liu, H.-F., and Chao, K.-M. (2011). Linear-time algorithms for the multiple gene duplication problems. *IEEE/ACM Trans Comput Biol Bioinform*, 8(1):260–265.

[Lynch and Conery, 2000] Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155.

[Lynch and Conery, 2003] Lynch, M. and Conery, J. S. (2003). The evolutionary demography of duplicate genes. *Journal of structural and functional genomics*, 3(1-4):35–44.

[Ma et al., 2000] Ma, B., Li, M., and Zhang, L. (2000). From gene trees to species trees. *SIAM J Comput*, 30(3):729–752.

[Maere et al., 2005] Maere, S., Bodt, S. D., Raes, J., Casneuf, T., Montagu, M. V., Kuiper, M., and de Peer, Y. V. (2005). Modeling gene and genome duplications in eukaryotes. *PNAS*, 102(15):5454–5459.

[Mettanant and Fakcharoenphol, 2008] Mettanant, V. and Fakcharoenphol, J. (2008). A linear-time algorithm for the multiple gene duplication problem. *NCSEC*, pages 198–203.

[Nøjgaard et al., 2017] Nøjgaard, N., Geiß, M., Merkle, D., Stadler, P. F., Wieseke, N., and Hellmuth, M. (2017). Forbidden time travel: Characterization of time-consistent tree reconciliation maps. In Schwartz, R. and Reinert, K., editors, *17th International Workshop on Algorithms in Bioinformatics, WABI 2017, August 21-23, 2017, Boston, MA, USA*, volume 88 of *LIPIcs*, pages 17:1–17:12. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

[Ohno, 1970] Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag.

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

[Page and Charleston, 1997] Page, R. D. M. and Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7:231–240.

[Page and Cotton, 2002] Page, R. D. M. and Cotton, J. A. (2002). Vertebrate phylogenomics: reconciled trees and gene duplications. *Pacific Symposium on Biocomputing*, pages 536–547.

[Paszek and Górecki, 2016] Paszek, J. and Górecki, P. (2016). Genomic duplication problems for unrooted gene trees. *BMC Genomics*, 17(1):165–175. doi: 10.1186/s12864-015-2308-4.

[Paszek and Górecki, 2017a] Paszek, J. and Górecki, P. (2017a). Efficient algorithms for genomic duplication models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi: 10.1109/TCBB.2017.2706679.

[Paszek and Górecki, 2017b] Paszek, J. and Górecki, P. (2017b). New algorithms for the genomic duplication problem. *In: Meidanis J., Nakhleh L. (eds) Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science*, 10562:101–115. Springer, Cham.

[Paszek and Górecki, 2018] Paszek, J. and Górecki, P. (2018). Inferring duplication episodes from unrooted gene trees. *BMC Genomics*, 19(5):288. doi: 10.1186/s12864-018-4623-z.

[Rabier et al., 2014] Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular biology and evolution*, 31(3):750–62.

[Sebat et al., 2004] Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528.

[Tang et al., 2008] Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, 320(5875):486–488.

[Vision et al., 2000] Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in Arabidopsis. *Science*, 290(5499):2114–2117.