POLITECHNIKA GDAŃSKA
Wydział Elektroniki, Telekomunikacji i Informatyki

Damian Bogdanowicz

Metody optymalizacji dyskretnej w analizie podobieństwa drzew filogenetycznych

Rozprawa doktorska

Promotor:
dr hab. inż. Krzysztof Giaro, prof. nadzw. PG
Wydział Elektroniki, Telekomunikacji
i Informatyki
Politechnika Gdańska

Gdańsk, 2012

Podziękowania

Chciałbym wyrazić gorące podziękowania dla mojego promotora dra hab. inż. Krzysztofa Giaro za inspirację do badań, cenne wskazówki, życzliwość i cierpliwość w trakcie realizowania niniejszej pracy.

Również bardzo gorąco chciałbym podziękować mojej żonie Agnieszce oraz synowi Adamowi. Dzięki ich cierpliwości, wyrozumiałości i wsparciu, na które zawsze mogłem liczyć, praca ta mogła powstać.

Spis treści

Wykaz skrótów i oznaczeń

$\|\|.\|\|_p$	—	norma L^p, $p \in \mathbb{R}_{\geq 1}$ zdefiniowana dla macierzy $M = [m_{ij}]$ o wymiarach $k \times l$ jako $\|\|M\|\|_p = \left(\sum_{i=1}^{k} \sum_{i=j}^{l} \|m_{ij}\|^p \right)^{1/p}$.
$\Delta_d(X)$	—	średnica zbioru X w metryce d, strona 12
MC	—	metryka skojarzeniowa dla drzew ukorzenionych, def. 3.3, strona 41.
ML	—	metoda konstrukcji drzew filogenetycznych wykorzystująca kryterium największej wiarygodności, strona 7.
MS	—	metryka skojarzeniowa dla drzew nieukorzenionych, def. 3.2, strona 40.
$N_d(x)$	—	sąsiedztwo punktu $x \in X$ w zbiorze X z metryką d, strona 12.
$N_d(x, \delta)$	—	zbiór punktów z X w odległości dokładnie δ od $x \in X$, strona 12.
ND	—	metryka węzłowa dla drzew nieukorzenionych, def. 2.4, strona 24.
NJ	—	metoda konstrukcji drzew filogenetycznych nazywana metodą przyłączania sąsiada, strona 6.
NNI	—	operacja edycyjna *Nearest Neighbour Interchange*, strona 26.
PD	—	metryka ścieżkowa, def. 2.3, strona 23.
QT	—	metryka kwartetowa, def. 2.6, strona 25.
RF	—	metryka Robinsona-Fouldsa dla drzew nieukorzenionych, def. 2.1, strona 22.
RFC	—	metryka Robinsona-Fouldsa dla drzew ukorzenionych, def. 2.2, strona 23.
R_L, R_n	—	rodziny ukorzenionych drzew filogenetycznych nad zbiorami liści odpowiednio L i $\{1, \ldots, n\}$.

R_L^B , R_n^B — rodziny ukorzenionych drzew filogenetycznych binarnych nad zbiorami liści odpowiednio L i $\{1, \ldots, n\}$.

SN — metryka węzłowa dla drzew ukorzenionych z normą L^2, def. 2.5, strona 25.

SPR — operacja edycyjna *Subtree Prune and Regraft*, strona 27.

TBR — operacja edycyjna *Tree Bisection and Reconnection*, strona 29.

TT — metryka tripletowa, def. 2.7, strona 26.

U_L, U_n — rodziny nieukorzenionych drzew filogenetycznych nad zbiorami liści odpowiednio L i $\{1, \ldots, n\}$.

U_L^B , U_n^B — rodziny nieukorzenionych drzew filogenetycznych binarnych nad zbiorami liści odpowiednio L i $\{1, \ldots, n\}$.

UM — model generacji losowych drzew filogenetycznych, w którym prawdopodobieństwo powstania każdego drzewa jest jednakowe, strona 103.

YM — model Yule'a generacji losowych drzew filogenetycznych, strona 103.

1 Wprowadzenie

Tematyka niniejszej pracy mieści się w przedmiocie badań względnie nowej dziedziny nauki jaką jest *bioinformatyka*. Istnieje wiele definicji tej dyscypliny. Poniżej przytoczony jest jeden z wariantów [60]:

„*Bioinformatyka jest interdyscyplinarną dziedziną nauki obejmującą wykorzystanie metod obliczeniowych do badania danych biologicznych*"

Ściślej mówiąc, rozważania zaprezentowane w pracy dotyczą filogenetyki, nauki wchodzącej w skład dyscypliny zwanej ewolucją molekularną. Ewolucja molekularna jest ściśle związana z bioinformatyką. Za jej narodziny jako nowej dziedziny nauki uznaje się czasami opublikowanie artykułu [117] Zauckerkandla i Paulinga w 1965 roku, gdzie po raz pierwszy wykorzystano sekwencje białek do konstrukcji drzewa filogenetycznego [60]. Warto zaznaczyć, że również w 1965 roku sformułowano prawo Moore'a [77], a komputery zaczęły odgrywać istotną rolę w badaniach naukowych.

1.1 Istota filogenetyki

Filogenetyka jest nauką o relacjach ewolucyjnych. Celem analizy filogenetycznej jest wysuwanie wniosków na temat tych relacji lub ich szacowanie [11]. Historia ewolucyjna, odtwarzana dzięki analizie filogenetycznej, na ogół przedstawiana jest w postaci diagramów przypominających drzewa, określanych jako *drzewa filogenetyczne*. Obiekty te obrazują ewolucyjne relacje podobieństwa pomiędzy gatunkami. Liście drzewa filogenetycznego odpowiadają istniejącym gatunkom, pozostałe wierzchołki reprezentują ich hipotetycznych przodków (rysunki 1.1, 1.2, 1.3). Dodatkowo, w przypadku drzew ukorzenionych jeden z wierzchołków niebędący liściem jest wyróżniony jako korzeń i reprezentuje wspólnego przodka wszystkich ga-

tunków z analizowanej grupy. Na ogół w procesie analizy filogenetycznej gatunki reprezentowane są przez sekwencje aminokwasów (białka) lub nukleotydów (DNA).

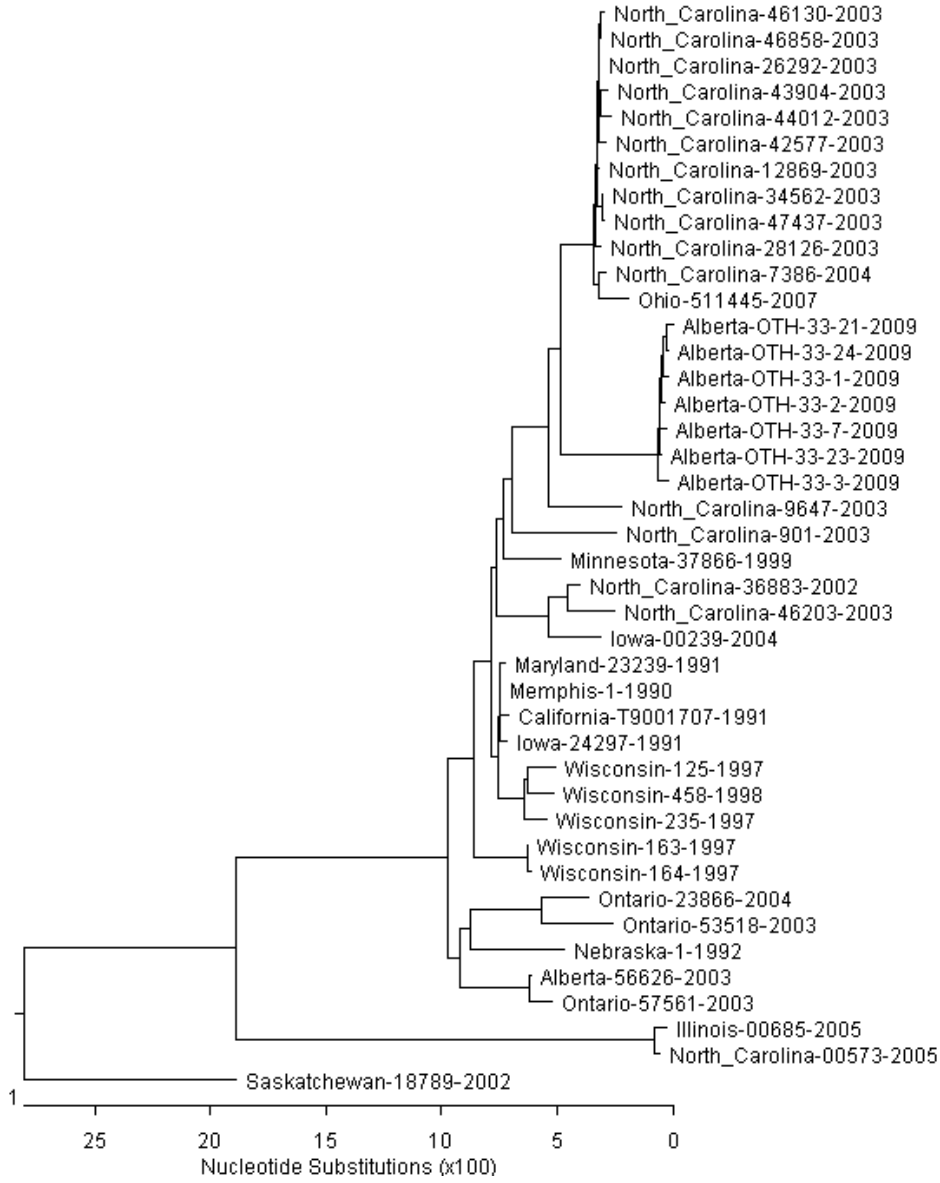
1.2 Metody tworzenia drzew filogenetycznych

Istnieje wiele metod konstrukcji drzew filogenetycznych, np. metody odległościowe, parsymonii, największej wiarygodności lub metody bayesowskie (por. [49]). Poniżej przedstawiona zostanie ich krótka charakterystyka. W większości metod niezbędna jest umiejętność wyznaczenia odległości ewolucyjnych pomiędzy sekwencjami lub ocena wiarygodności danego drzewa filogenetycznego. Aby móc ilościowo określić te wartości, wprowadza się różne modele substytucji określające koszt związany z podstawieniem danego elementu sekwencji przez inny. Wybór modelu podstawień wpływa na kształt tworzonego drzewa.

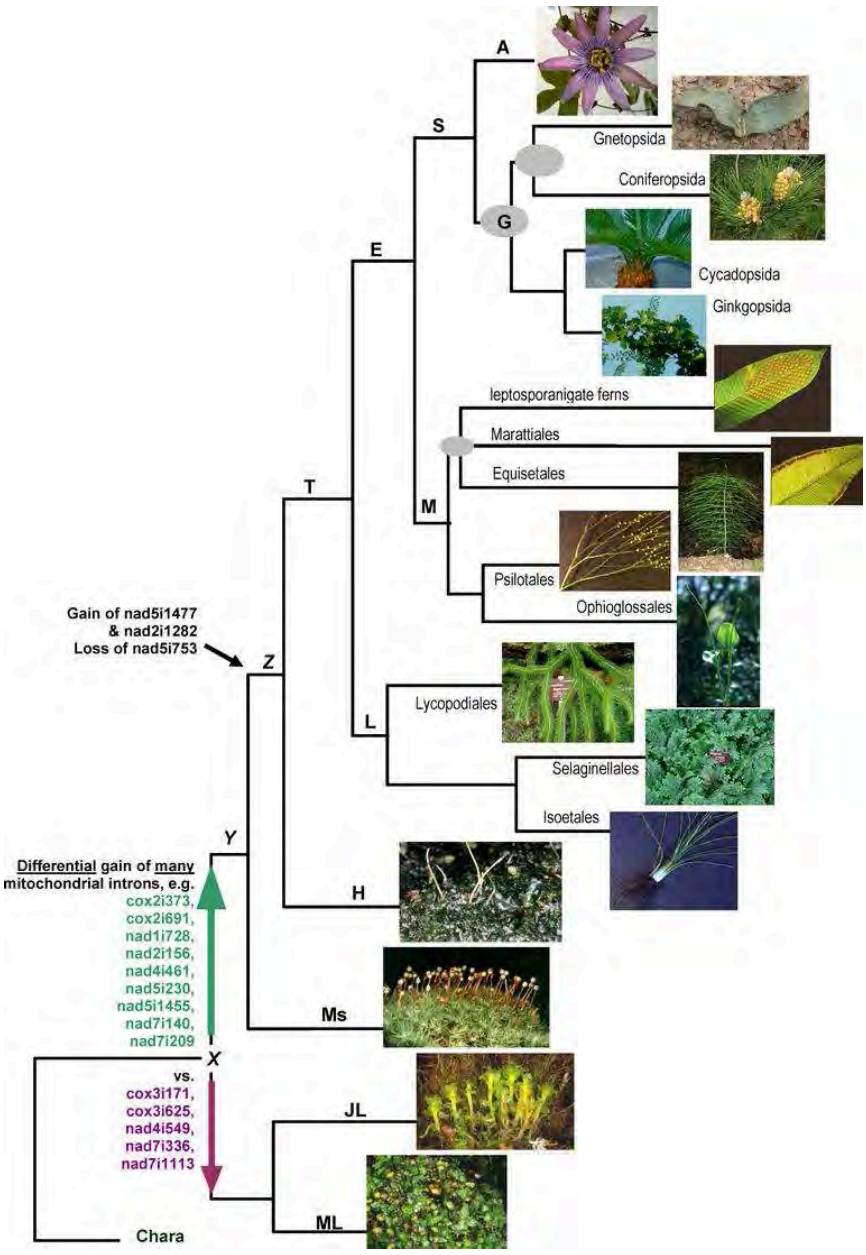
1.2.1 Metody odległościowe

W odległościowych metodach filogenetycznych w pierwszym kroku odpowiedniego algorytmu wyznaczana jest macierz odległości ewolucyjnych (według wybranego modelu substytucji) dla każdej pary sekwencji z analizowanego zbioru. Posiadając wyznaczoną w ten sposób macierz można przystąpić od budowy drzewa filogenetycznego na wiele sposobów. Wspólnym celem wszystkich algorytmów odległościowych jest konstrukcja drzewa posiadającego dodatnie wagi na krawędziach, które najlepiej odzwierciedla odległości zawarte w macierzy, czyli takiego by dla dowolnych dwóch gatunków ich odległość liczona wzdłuż ścieżki łączącej je w drzewie była w przybliżeniu równa odpowiedniej wartości w macierzy.

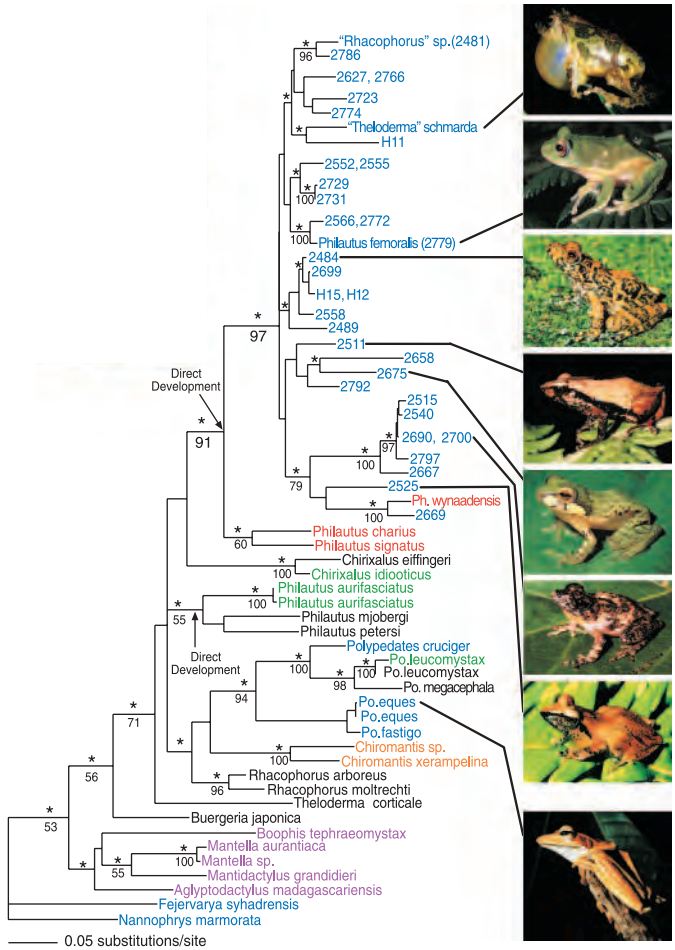
Najprostszym algorytmem stosowanym w tym celu jest metoda średnich połączań (UPGMA), której idea polega na hierarchicznej analizie skupisk sekwencji (klastrów) przy założeniu stałego tempa ewolucji różnych organizmów (tzw. *hipoteza zegara molekularnego*). W pierwszym kroku tego algorytmu łączone są skupiska zawierające najbliżej spokrewnione ewolu-



Rysunek 1.1: Drzewo filogenetyczne wirusa świńskiej grypy A(H1N1) zbudowane na podstawie 42 sekwencji białkowych hemaglutyniny wyizolowanych w latach 1990-2009 [100].



Rysunek 1.2: Przykład drzewa filogenetycznego roślin lądowych [68].



Rysunek 1.3: Drzewo filogenetyczne wybranych gatunków żab [75].

cyjnie sekwencje. W następnym kroku obliczana jest odległość nowo utworzonego klastra od pozostałych i ponownie dwa najbliższe klastry łączone są w nowe skupisko. Procedura ta powtarzana jest do momentu otrzymania jednego skupiska. W UPGMA odległość pomiędzy dwoma skupiskami definiuje się jako średnią odległość ewolucyjną między sekwencjami z jednego i drugiego skupiska. Ponieważ hipoteza zegara molekularnego jest rzadko spełniona dla rzeczywistych sekwencji, metoda UPGMA często daje błędne wyniki. Drzewa konstruowane przy użyciu UPGMA są

ukorzenione.

Kolejnym algorytmem z tej grupy jest metoda przyłączania sąsiada (NJ), w której drzewa są konstruowane przy założeniu addytywności macierzy wejściowej. Macierz odległości jest *addytywna*, jeśli możliwe jest wyznaczenie dla niej drzewa, w którym sumaryczna waga krawędzi łączących dwa dowolne gatunki jest równa odległości ewolucyjnej miedzy tymi gatunkami. Jeśli wejściowa macierz jest addytywna, to metoda NJ gwarantuje wyznaczenie dla niej poprawnego drzewa. W rzeczywistości jednak macierze odległości nie są dokładnie addytywne, stąd drzewa skonstruowane tą metodą mają przybliżony charakter. Metoda NJ konstruuje drzewa nieukorzenione.

1.2.2 Metoda parsymonii

Metoda parsymonii (MP), inaczej oszczędności, stanowi kryterium optymalizacyjne opierające się na zasadzie, że najlepsze rozwiązanie jest najprostsze. W odniesieniu do zbioru sekwencji kryterium parsymonii umożliwia wyznaczenie drzew, które opisują zmienność sekwencji za pomocą najmniejszej możliwej liczby podstawień. W modelu parsymonii z gatunkami są skojarzone zbiory cech. Każda cecha ma określoną liczbę stanów. Dany gatunek jest reprezentowany za pomocą wektora zawierającego wartości stanów przyjętych przez każdą z cech. Zamiana stanu cechy wzdłuż pewnej krawędzi drzewa odpowiada zmianie ewolucyjnej. Problem wyznaczenia topologii drzewa, które minimalizuje całkowitą liczbę zmian stanów wzdłuż swoich krawędzi, jest w ogólności NP-trudny, nawet gdy wszystkie cechy posiadają tylko dwa stany [41]. Istnieje jednak wiele algorytmów heurystycznych dla MP. Istotny jest fakt, że metoda MP prowadzi do konstrukcji nie jednego drzewa, lecz zbioru drzew o jednakowej wartości funkcji jakości.

1.2.3 Metoda największej wiarygodności

Na podstawie przyjętego modelu ewolucji sekwencji dla danego drzewa można obliczyć jego *wiarygodność*, czyli prawdopodobieństwo, że para-

metry tego drzewa opisują ewolucyjne związki między poszczególnymi sekwencjami. Istota metody wykorzystującej kryterium największej wiarygodności (ang. Maximum Likelihood, ML) sprowadza się do wyboru takiego drzewa, dla którego wiarygodność będzie największa. Jednak podobnie jak w przypadku metody parsymonii, znalezienie najbardziej wiarygodnego drzewa jest problemem NP-trudnym [35]. W praktyce w celu implementacji idei ML z powodzeniem stosuje się algorytmy heurystyczne.

1.2.4 Metody bayesowskie

Idea metod bayesowskich polega na przeszukiwaniu przestrzeni drzew filogenetycznych, podobnie jak w przypadku ML, lecz przy użyciu innego kryterium optymalizacji. Celem jest tu znalezienie drzewa T, które maksymalizuje prawdopodobieństwo warunkowe $\Pr(T|D)$, gdzie D odpowiada zdarzeniu polegającemu na pojawieniu się analizowanego zbioru sekwencji. W praktyce do wyznaczenia prawdopodobieństwa a posteriori wykorzystuje się metodę Monte Carlo dla łańcuchów Markowa (MCMC). Metoda MCMC umożliwia wygenerowanie zbioru drzew, w którym częstość występowania drzewa o określonej topologii jest proporcjonalna do wartości $\Pr(T|D)$. Podobnie jak w przypadku MP produktem metod bayesowskich są zbiory drzew.

1.3 Zastosowania

Głównym celem tworzenia drzew filogenetycznych jest poznanie i zrozumienie historii ewolucji badanej grupy organizmów. Drzewa filogenetyczne są jednak również wykorzystywane w biologii do innych celów, np. znajomość procesu ewolucji wirusa HIV może być wykorzystana do przewidywania jego reakcji na szczepionki lub nowe leki [91]. Wirus HIV charakteryzuje się dużą zmiennością, co oznacza, że wirusy nawet tego samego szczepu pochodzące od innych gospodarzy mogą posiadać istotnie różne genomy. W konsekwencji potencjalna szczepionka otrzymana na podstawie jednego materiału genetycznego może nie być skuteczna w przypadku

wirusów o innych genomach. Techniki filogenetyczne pozwalają jednak na znalezienie wspólnego przodka dla danej grupy wirusów, mogącego być lepszym kandydatem do projektowania szczepionki [53, 80]. Warto zauważyć, że stosunkowo niedawne badania (z 2010 roku) ewolucji wirusa HIV, przeprowadzone również za pomocą technik filogenetycznych, ujawniły istnienie związku miedzy genotypem wirusa a czasem trwania rozwoju infekcji w jej ostateczne stadium — AIDS [4]. Fakt ten stanowi istotny krok zbliżający badaczy do pełnego zrozumienia patogenezy wirusa HIV [4].

Złożoność zagadnienia rekonstrukcji nie pozwala jednak ciągle na wyłonienie lub określenie danej metody jako optymalnej, stąd też nadal rozwijane i testowane są nowe podejścia i implementacje (np. aplikacja FastTree2 [89]), opierające się często w swojej idei na wspominanych klasycznych algorytmach. Ponieważ istnieje wiele metod i często zdarza się, że zwracają one różne drzewa dla tych samych danych wejściowych, pojawia się potrzeba ilościowego określenia podobieństwa różnych drzew obrazujących historię ewolucji tej samej grupy gatunków. Naturalnym rozwiązaniem jest zdefiniowanie metryki w zbiorze wszystkich możliwych drzew filogenetycznych dla danego zbioru gatunków (liści). Stąd też wynika jedno z podstawowych zastosowań metryk filogenetycznych w biologii obliczeniowej — ilościowe określanie i porównywanie dokładności metod rekonstrukcji [70, 109]. Warto tu również wspomnieć pozycję [84], będącą jedną z pierwszych prac prezentujących zastosowanie metryk filogenetycznych, w której autorzy na podstawie porównywania drzew filogenetycznych otrzymanych z analizy sekwencji 5 białek dla 11 gatunków metodami dystansowymi potwierdzają poprawność tez teorii ewolucji.

Niektóre z metod rekonstrukcji (np. jedna z bardziej popularnych — metoda bayesowska) nie wyznaczają jednego drzewa, lecz zbiory drzew. W takim przypadku, w celu uzyskania biologicznie istotnych informacji, wykonuje się kolejne fazy przetwarzania. Istnieje wiele metod ekstrakcji wspólnej informacji reprezentowanej przez otrzymany zbiór drzew, polegających w głównej mierze na tworzeniu jednego drzewa konsensusu. W ostatniej dekadzie rozwinęły się również inne metody analizy wspomnianych zbiorów, u podstaw których leżą metryki, tj. metody wyko-

rzystujące klasteryzację zbioru drzew [107] oraz metody wizualizacji tej przestrzeni [62]. Szczegółowy przegląd zastosowań metryk w biologii jest przedstawiony w [85]. Umiejętność ilościowego określania podobieństwa drzew filogenetycznych okazuje się również nieodzowna przy przeszukiwaniu filogenetycznych baz danych (np. bazy TreeBASE) [111].

Zarówno drzewa filogenetyczne jak i metody ich porównywania okazują się bardzo przydatne także w innych dziedzinach nauki niezwiązanych z biologią. Techniki filogenetyczne znalazły zastosowanie w gałęzi informatyki zajmującej się badaniem i rozpoznawaniem wirusów komputerowych [67]. Metryki filogenetyczne mogą służyć również do porównywania hierarchicznych klasteryzacji, pojawiających się np. przy analizie danych z baz cząsteczek związków chemicznych [93]. W końcu metody i metryki filogenetyczne okazują się być wygodnymi narzędziami w badaniach związanych z lingwistyką i historią literatury, np. w [9] skonstruowano drzewo filogenetyczne obrazujące związki między 58 zachowanymi wersjami fragmentu „Opowieści kanterberyjskich" (ang. „The Canterbury Tales"), na podstawie którego potwierdzono przypuszczenia, że oryginalne dzieło mogło nigdy nie być kompletne i istnieć wyłącznie w wersji roboczej (zwierającej notatki, komentarze i przypisy autora). W [86, 87] wspomniane metody pozwoliły natomiast na konstrukcję i analizy drzew ewolucji języków.

Teza pracy

Istnieje ogólna, efektywna obliczeniowo metoda konstrukcji metryk w zbiorze drzew filogenetycznych, wykorzystująca ważone skojarzenia w grafach dwudzielnych, która umożliwia definiowanie odległości o intuicyjnych i pożądanych własnościach.

2 Definicje i pojęcia podstawowe

Podstawowe pojęcia i oznaczenia matematyczne przyjęte w pracy są zgodne z powszechnie stosowanym standardem (por. [92]):

- $\emptyset$ — zbiór pusty,
- $|A|$ — liczba elementów zbioru A,
- $A \times B = \{(a, b) : a \in A, b \in B\}$ — iloczyn kartezjański zbiorów,
- $f : A \to B$ — funkcja ze zbioru A w B.

Różnica symetryczna zbiorów A, B jest oznaczona w pracy przez $A \oplus B$, tj. $A \oplus B = (A \setminus B) \cup (B \setminus A)$. Dla zbioru A zbiór $2^A = \{B : B \subseteq A\}$ jest rodziną wszystkich podzbiorów A.

Zasadnicze znaczenie dla rozważań prowadzonych w kolejnych rozdziałach mają pojęcia metryki i przestrzeni metrycznej. Niech X będzie danym zbiorem. Funkcja $d : X \times X \to \mathbb{R}_{\geq 0}$ jest *metryką* w X wtedy i tylko wtedy, gdy spełnia poniższe warunki:

1. dla każdego $x, y \in X$ zachodzi $d(x, y) = 0 \Leftrightarrow x = y$,
2. dla każdego $x, y \in X$ jest spełnione $d(x, y) = d(y, x)$ — *symetria*,
3. dla każdego $x, y, z \in X$ prawdziwa jest zależność $d(x, y) + d(y, z) \geq d(x, z)$ — *nierówność trójkąta*.

Parę (X, d), gdzie d jest metryką w X, nazywamy *przestrzenią metryczną*. Wartość metryki d dla pary punktów $x, y \in X$ jest określana jako ich *odległość*.

Pod pojęciem *sąsiadów* w przestrzeni metrycznej będziemy rozumieć dwa elementy tej przestrzeni znajdujące się względem siebie w najmniejszej możliwej dodatniej odległości. *Sąsiedztwem* punktu $x \in X$ w przestrzeni metrycznej (X, d) jest zbiór $N_d(x)$ wszystkich elementów X, które są sąsiadami x. Dodatkowo niech $N_d(x, \delta) = \{y \in X : d(x, y) = \delta\}$ będzie zbiorem punktów z X w odległości dokładnie δ od x.

Średnicą $\Delta_d(X)$ zbioru X w metryce d nazywamy maksymalną możliwą odległość pomiędzy dwoma elementami zbioru X.

2.1 Podstawowe pojęcia teorii grafów

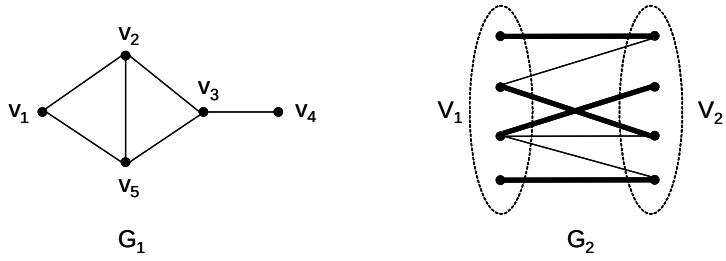
Podstawowe pojęcia i oznaczenia z zakresu teorii grafów używane w pracy są zgodne z powszechnie przyjętą terminologią proponowaną np. w [112].

Przez $G = (V, E)$ będziemy oznaczać *graf* prosty (tj. bez pętli i krawędzi wielokrotnych) o *zbiorze wierzchołków* V i *zbiorze krawędzi* E. *Krawędź* między dwoma wierzchołkami $u, v \in V$ reprezentowana jest jako dwuelementowy zbiór $\{u, v\}$. Jeśli w grafie istnieje krawędź $\{u, v\}$, to wierzchołki u, v są *sąsiadami*; oraz krawędź ta jest *incydentna* do u i v. Dla grafu G symbolem $V(G)$ będziemy oznaczać zbiór jego wierzchołków, zaś symbolem $E(G)$ jego zbiór krawędzi. Liczbę wierzchołków grafu G oznaczamy jako $n(G)$, zaś liczbę krawędzi jako $m(G)$. Dla wierzchołka $v \in V$ jego *stopień* $\deg_G(v)$ w G jest zdefiniowany jako liczba krawędzi w E incydentnych z v; minimalny stopień wierzchołka w grafie G oznaczamy przez $\delta(G) = \min_{v \in V(G)} \deg_G(v)$, zaś maksymalny przez $\Delta(G) = \max_{v \in V(G)} \deg_G(v)$ (por. rysunek 2.1).

Marszrutą o długości $k-1$ w grafie nazywamy sekwencję wierzchołków $(v_1, v_2, \ldots, v_k)$, taką że $\{v_i, v_{i+1}\} \in E$ dla $i = 1, \ldots, k-1$. *Marszruta zamknięta* to marszruta kończąca się w punkcie wyjścia, czyli taka, w której $v_1 = v_k$. *Cykl* to marszruta zamknięta, w której jedynym powtarzającym się wierzchołkiem jest jej początek (będący również jej końcem). Marszruta bez powtarzających się wierzchołków nazywana jest *ścieżką*. Dla dowolnej pary wierzchołków $u, v \in V$ odległość $\mathrm{dist}_G\{u, v\}$ pomiędzy u i v jest równa długości najkrótszej ścieżki w G łączącej u i v lub ∞, jeśli

ścieżka łącząca te wierzchołki nie istnieje.

Graf nazywamy *spójnym*, jeśli dla każdej pary wierzchołków istnieje łącząca je ścieżka. *Drzewem* nazywamy graf spójny bez cykli. Wierzchołki drzewa posiadające stopień równy 1 są określane jako *liście*. Graf G nazywamy *dwudzielnym*, jeśli jego zbiór wierzchołków V można rozdzielić na dwa rozłączne podzbiory (*partycje*) V_1, V_2, takie że $V_1 \cup V_2 = V$ oraz wszystkie krawędzie G posiadają jeden wierzchołek w zbiorze V_1, a drugi w V_2 (rysunek 2.1). Graf dwudzielny G o partycjach V_1 i V_2 będziemy też oznaczać jako trójkę $G = (V_1, V_2, E)$. Graf dwudzielny $G = (V_1, V_2, E)$, gdzie $n_1 = |V_1|$, $n_2 = |V_2|$, nazywamy *pełnym grafem dwudzielnym* i oznaczamy przez K_{n_1,n_2}, jeśli każdy wierzchołek z jednej partycji połączony jest krawędzią z każdym z wierzchołków drugiej partycji.



Rysunek 2.1: Dla grafu G_1 zachodzi: $n(G_1) = 5$, $m(G_1) = 6$, $\deg_{G_1}(v_3) = 3$, $\delta(G_1) = 1$, $\Delta(G_1) = 3$, $\mathrm{dist}_{G_1}\{v_1, v_3\} = 2$. Graf G_2 jest grafem dwudzielnym, w którym wyróżnione krawędzie tworzą doskonałe skojarzenie.

Definicja 2.1. *Skojarzeniem* w grafie $G = (V, E)$ nazywamy dowolny niezależny zbiór krawędzi $M \subseteq E$, tzn. taki, że dla dowolnych $e \neq f \in M$ krawędzie e i f nie mają wspólnego wierzchołka (por. rysunek 2.1).

Skojarzenie jest *doskonałe* jeśli pokrywa wszystkie wierzchołki grafu. Jeśli z krawędziami grafu G związemy funkcję wagową $w : E \to \mathbb{R}_{\geq 0}$, to *najlżejsze doskonałe skojarzenie* definiowane jest jako doskonałe skojarzenie o najmniejszej możliwej sumie wag krawędzi. Mimo że ilość doskonałych skojarzeń w grafie dwudzielnym $G = (V_1, V_2, E)$, $|V_1| = |V_2|$ może wynosić nawet $|V_1|!$, wyznaczenie najlżejszego doskonałego skojarzenia może być

dokonane efektywnie w czasie wielomianowym, np. za pomocą algorytmów o złożoności wynoszącej $O(|E|\sqrt{|V|}\log(|V|\max_{e\in E} w(e))$ [52, 83].

2.2 Podstawowe pojęcia z zakresu filogenetyki

2.2.1 Drzewa filogenetyczne

Definicja 2.2. *Nieukorzenione drzewo filogenetyczne* T nad zbiorem gatunków L jest drzewem bez wierzchołków stopnia 2, którego liście poetykietowane są wzajemnie jednoznacznie elementami zbioru L, a pozostałe wierzchołki zwane *wewnętrznymi* nie posiadają etykiet. Nieukorzenione drzewo filogenetyczne nazywamy *binarnym*, jeśli dodatkowo wszystkie jego wierzchołki wewnętrzne posiadają stopień równy 3.

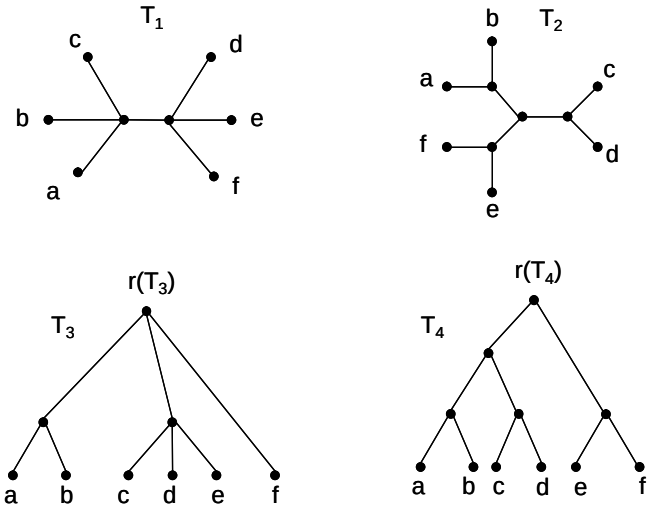
Dla uproszczenia w pracy utożsamia się zbiór L etykiet drzewa T z jego liśćmi, tj. $L \subseteq V(T)$. Liście reprezentują gatunki współczesne, natomiast wierzchołki wewnętrzne odpowiadają ich przodkom. Drzewo nieukorzenione stanowi zatem graficzną ilustrację relacji pokrewieństwa.

Definicja 2.3. *Ukorzenione drzewo filogenetyczne* T nad zbiorem gatunków L jest definiowane analogicznie jak drzewo nieukorzenione, z tą tylko różnicą, że w drzewie ukorzenionym istnieje dokładnie jeden wyróżniony wierzchołek wewnętrzny r zwany *korzeniem*, mogący posiadać stopień równy 2. Ukorzenione drzewo filogenetyczne nazywamy *binarnym*, jeśli jego korzeń posiada stopień 2 oraz wszystkie pozostałe wierzchołki wewnętrzne mają stopień równy 3.

Poprzez obecność korzenienia drzewo to oprócz wzajemnych relacji pokrewieństwa obrazuje porządek związany z przepływem czasu. Większość metod filogenetycznych umożliwia jednak wyznaczanie drzew nieukorzenionych. Transformację polegającą na przekształceniu drzewa nieukorzenionego w ukorzenione nazywamy *ukorzenianiem*. Operacja ta może być wykonana na dwa sposoby. Pierwszy sposób polega na wyróżnieniu jednego z wierzchołków wewnętrznych jako korzenia, w drugim zaś przypadku korzeń wprowadzany jest jako nowy wierzchołek stopnia dwa, rozdzielając

wybraną krawędź drzewa. Istnieje wiele metod pozwalających na ustalenie najlepszego miejsca dla wprowadzenia korzenia, np. metoda grupy zewnętrznej (ang. outgroup) lub metoda punktu środkowego (ang. midpoint method). Szerszy ich opis wraz z porównaniem i analizą można znaleźć w pracy [24].

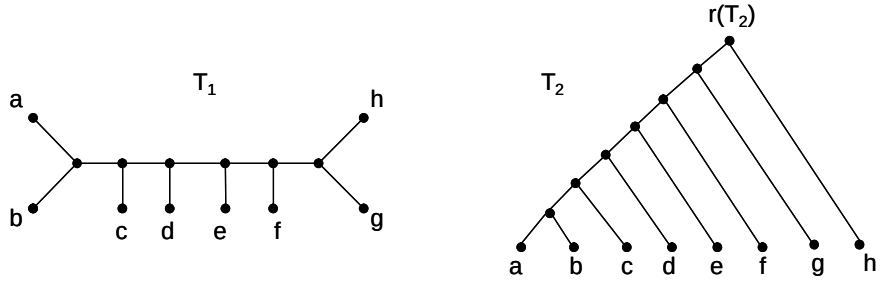
Obecność w drzewie wierzchołków o stopniu większym niż 3 (oraz większym niż 2 w przypadku korzenia drzewa), zwanych też *multifurkacjami*, świadczy na ogół o braku dostatecznej informacji niezbędnej do dokładnego określenia sposobu rozdzielania się linii ewolucyjnych (tj. *specjacji*). Zauważmy zatem, że najwięcej informacji filogenetycznej przedstawiają drzewa binarne, najmniej zaś, drzewa nazywane *gwiazdami*, które posiadają tylko jeden wierzchołek wewnętrzny.



Rysunek 2.2: Przykłady drzew filogenetycznych: T_1 — nieukorzenione niebinarne, T_2 — nieukorzenione binarne, T_3 — ukorzenione niebinarne, T_4 — ukorzenione binarne.

Pod pojęciem *topologii* drzewa filogenetycznego będziemy rozumieli wyłącznie graf zwiany z danym drzewem, bez etykiet na liściach. Krawędzie, których jeden z końców jest liściem nazywane są *wiszącymi*, zaś pozostałe tworzą zbiór *krawędzi wewnętrznych*.

Jednym ze standardowych przykładów binarnych drzew filogenetycznych są gąsienice. Nieukorzenione drzewo binarne nazywamy *gąsienicą*, jeśli wszystkie jego wierzchołki wewnętrzne leżą na jednej wspólnej ścieżce. *Gąsienica ukorzeniona* jest drzewem ukorzenionym binarnym, które powstaje z gąsienicy nieukorzenionej w wyniku wstawienia korzenia w postaci nowego wierzchołka stopnia 2 na jednej z czterech zewnętrznych krawędzi wiszących (tj. takich, których jednym z końców jest wierzchołek wewnętrzny sąsiadujący z dwoma liśćmi).



Rysunek 2.3: Przykłady gąsienic: nieukorzenionej T_1 oraz ukorzenionej T_2.

Zbiory U_L oraz U_L^B oznaczają odpowiednio zbiór wszystkich drzew filogenetycznych nieukorzenionych oraz zbiór wszystkich drzew filogenetycznych nieukorzenionych binarnych nad zbiorem liści L. Dla uproszczenia rozważań wygodnie jest ponumerować badane gatunki kolejnymi liczbami naturalnymi, tj. $L = \{1, \ldots, n\}$; w takiej sytuacji stosowany będzie skrócony zapis U_n oraz U_n^B. W przypadku drzew binarnych mamy $|U_n^B| = 1 \cdot 3 \cdot \ldots \cdot (2n-5) = (2n-5)!! = \frac{(2n-4)!}{(n-2)!2^{(n-2)}} \sim \frac{1}{2\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-2}$ [98]. Każde drzewo $T \in U_L^B$ posiada $|L| - 2$ wierzchołków wewnętrznych i $|L| - 3$ wewnętrznych krawędzi, dla drzew niebinarnych wielkości te są mniejsze, osiągając w skrajnym przypadku (tj. dla gwiazdy) odpowiednio 1 i 0.

Podobnie jak w przypadku drzew nieukorzenionych, R_L i R_L^B oznaczają odpowiednio zbiór wszystkich drzew ukorzenionych oraz zbiór wszystkich drzew ukorzenionych binarnych nad zbiorem liści L. W ukorzenionym drzewie filogenetycznym binarnym $T \in R_L^B$ znajduje się $|L| - 2$ krawędzi

wewnętrznych oraz $|L| - 1$ wierzchołków wewnętrznych. Dla drzew niebinarnych obie te liczby są mniejsze. Podobnie jak dla drzew nieukorzenionych, w przypadku gdy $L = \{1, \ldots, n\}$ stosuje się notację uproszczoną, tj. R_n oraz R_n^B, gdzie $|R_n^B| = (2n-3)!!$.

2.2.2 Rozbicia i klastry

W drzewie nieukorzenionym wprowadza się relację między krawędziami a rozbiciami zbioru liści. Nieuporządkowana para niepustych podzbiorów $A, B \subseteq L$ oznaczana jako $A|B$ (symbol ten traktujemy jako symetryczny, tj. $A|B = B|A$) jest *rozbiciem* zbioru L, jeśli $L = A \cup B$ i $A \cap B = \emptyset$. Rodzina wszystkich rozbić L jest oznaczona jako $Splits(L)$. Niech $\min(A|B) = \min\{|A|, |B|\}$. Jeśli $\min(A|B) = 1$, wówczas rozbicie $A|B$ nazywamy rozbiciem *trywialnym*; w przeciwnym przypadku rozbicie jest *nietrywialne* [26]. Zbiór robić trywialnych L oznaczamy jako $\beta_0(L) = \{x|L \setminus \{x\} : x \in L\}$.

Usunięcie krawędzi $e \in E(T)$ w drzewie $T \in U_L$ powoduje powstanie dwóch składowych spójności. Niech zbiory A i B oznaczają zbiory liści w obu tych składowych. Wówczas rozbicie $A|B$ jest rozbiciem *odpowiadającym* krawędzi e. Zbiór rozbić odpowiadających wszystkim krawędziom drzewa $T \in U_L$ jest oznaczony przez $\beta(T)$ [26], zatem zawiera on dokładnie $|L|$ rozbić trywialnych oraz $|\beta(T)| \leq 2|L| - 3$. Podzbiór $\beta(T)$ zawierający wyłącznie rozbicia nietrywialne oznacza się przez $\beta_*(T)$. Dla drzewa T_1 na rysunku 2.2 mamy $\beta(T_1) = \{a|bcdef, b|acdef, c|abdef, d|abcef, e|abcdf, f|abcde, abc|def\}$, $\beta_*(T_1) = \{abc|def\}$.

Definicja 2.4 ([98])**.** Dwa rozbicia $A_1|B_1$ i $A_2|B_2$ zbioru L są *kompatybilne*, jeśli jeden ze zbiorów: $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, $B_1 \cap B_2$ jest zbiorem pustym.

Związek między zbiorami rozbić a drzewami obrazuje następujące twierdzenie, dające podstawę do stosowania $\beta(T)$ jako niegrafowego opisu nieukorzenionego drzewa filogenetycznego.

Twierdzenie 2.1 ([29])**.** *Niech $A \subseteq Splits(L)$ będzie pewną rodziną rozbić zbioru L. Istnieje drzewo $T \in U_L$, takie że $A \cup \beta_0(L) = \beta(T)$ wtedy i tylko*

wtedy, gdy rozbicia z A są parami kompatybilne. Co więcej, może istnieć co najwyżej jedno takie drzewo.

Dowód tego twierdzenia można również znaleźć w [98] (tw. 3.1.4). Drzewo filogenetyczne nieukorzenione może być odtworzone na podstawie zbioru swoich rozbić w czasie liniowym [29, 57].

Ukorzenione drzewo T definiuje relację częściowego porządku (bycia przodkiem i potomkiem) na swoich wierzchołkach oznaczoną przez $\leq_T$. Dla $a, b \in V(T)$ zachodzi $a \leq_T b$, czyli a jest *potomkiem* b (równoważnie b jest *przodkiem* a), jeśli ścieżka w T łącząca a z korzeniem $r(T)$ przechodzi przez wierzchołek b. W szczególności $v \leq_T r(T)$ oraz $v \leq_T v$ dla każdego $v \in V(T)$. *Najniższym wspólnym przodkiem* (ang. *the Lowest Common Ancestor*) $LCA(A)$ zbioru wierzchołków $A \subseteq V(T)$ jest wierzchołek, który jest przodkiem wszystkich $v \in A$, taki że ścieżka łączącą go z korzeniem posiada maksymalną długość, inaczej mówiąc $LCA(A)$ jest kresem górnym A względem porządku $\leq_T$.

Drzewa ukorzenione podobnie jak nieukorzenione można opisać bez posługiwania się grafami. Z każdym wierzchołkiem v w drzewie ukorzenionym $T \in R_L$ kojarzymy zbiór $c(v) \subseteq L$ nazywany *klastrem* (lub *kladem*) zawierający liście (gatunki), które są potomkami v. W drzewie $T \in R_L$ znajduje się $|L| + 1$ klastrów *trywialnych*, $|L|$ z nich jest związanych z liśćmi $u \in L$ (wówczas $c(u) = \{u\}$), jeden zaś odpowiada korzeniowi $c(r(T)) = L(T)$. Pozostałe klastry określane są jako *nietrywialne*. Zbiór wszystkich klastrów w T oznaczany jest jako $\sigma(T)$, zaś zbiór wszystkich klastrów nietrywialnych w T przez $\sigma_*(T)$. Zatem dla $T \in R_L$ mamy $|\sigma(T)| \leq 2|L| - 1$, $|\sigma_*(T)| \leq |L| - 2$. Obie te nierówności stają się równościami dla drzew binarnych. Dla drzewa T_3 na rysunku 2.2 mamy $\sigma(T_3) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{a, b\}, \{c, d, e\}$ i $\sigma_*(T_3) = \{\{a, b\}, \{c, d, e\}\}$. Dodatkowo przez $\sigma_0(L) = \{\{x\} : x \in L\} \cup \{L\}$ oznaczymy zbiór klastrów trywialnych w $T \in R_L$.

Definicja 2.5. Dwa zbiory (klastry) $B, C \subseteq L$ są *kompatybilne*, jeśli zachodzi $B \cap C \in \{\emptyset, B, C\}$.

Każde drzewo ukorzenione T jest jednoznacznie wyznaczone przez zbiór $\sigma_*(T)$. Prawdziwe jest następujące twierdzenie.

Twierdzenie 2.2 ([98] tw. 3.5.2)**.** *Niech $A \subseteq 2^L$ będzie pewną rodziną niepustych podzbiorów L. Istnieje drzewo $T \in R_L$, takie że $A \cup \sigma_0(L) = \sigma(T)$ wtedy i tylko wtedy, gdy każde dwa zbiory (klastry) z A są parami kompatybilne. Co więcej, może istnieć co najwyżej jedno takie drzewo.*

Warunek podany w powyższym twierdzeniu jest określany jako *warunek kompatybilności* zbioru klastrów. Transformacja pomiędzy reprezentacjami drzewa w postaci jawnej oraz jako zbiór klastrów może być wykonana w czasie liniowym [57] (oraz [98] podrozdział 3.5).

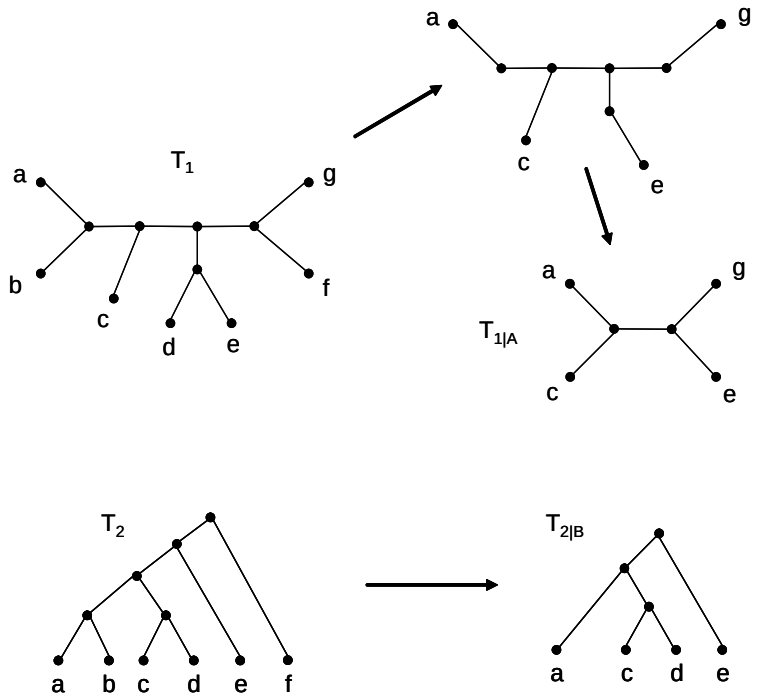
2.2.3 Poddrzewa nad podzbiorami liści

W celu ułatwienia opisu przekształceń dokonywanych na drzewach filogenetycznych T zdefiniujemy poniżej dwie wzajemnie odwrotne operacje dla wierzchołków v o stopniu dwa:

1. *wprowadzenie* wierzchołka v na krawędzi $e = \{u, w\}$ polega na usunięciu e z T i zastąpieniu jej dwoma krawędziami $\{u, v\}$, $\{v, w\}$,

2. *zdjęcie* lub *ściągnięcie* wierzchołka v stopnia dwa incydentnego do krawędzi $f = \{u, v\}$, $g = \{v, w\}$ polega na usunięciu v i zastąpieniu f i g jedną nową krawędzią $\{u, w\}$.

Rozważmy dowolne drzewo T o zbiorze liści L oraz zbiór $A \subseteq L$. Przez $T(A)$ oznaczymy najmniejszy spójny podgraf T, który zawiera wszystkie liście z A. W przypadku gdy T jest drzewem ukorzenionym, korzeniem w $T(A)$ jest jego wierzchołek najbliższy $r(T)$. Przez $T_{|A}$ oznaczymy *poddrzewo T indukowane przez A*, tzn. powstające z $T(A)$ w wyniku sekwencji operacji ściągnięcia kolejno wszystkich wierzchołków stopnia dwa (z wyjątkiem korzenia, jeśli operujemy na drzewach ukorzenionych) [26]. Obrazowo: zdejmujemy kolejno wierzchołki stopnia dwa z $T(A)$ za każdym razem „sklejając" wychodzące zeń krawędzie w jedną nową krawędź (por. rysunek 2.4). Drzewo $T_{|A}$ reprezentuje te same informacje odnośnie relacji pokrewieństwa co T, lecz tylko w obrębie zbioru liści A. Nie należy utożsamiać pojęcia poddrzewa indukowanego z *podgrafem indukowanym* znanym z teorii grafów.

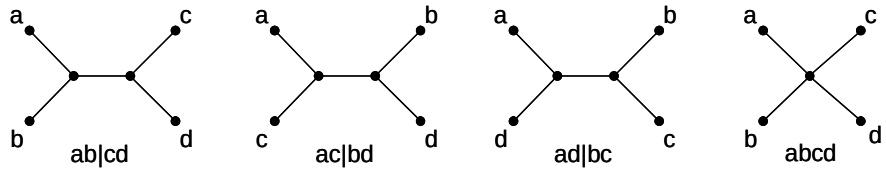


Rysunek 2.4: Konstrukcja poddrzew indukowanych: $A = \{a, c, e, g\}$, $B = \{a, c, d, e\}$. Dla drzewa nieukorzenionego T_1 przedstawiony został również stan pośredni w tej operacji.

Drzewo T nad zbiorem liści $A \subseteq L$ jest *poddrzewem zgodnym* dla $T_1, T_2 \in U_L$ (lub $T_1, T_2 \in R_L$), jeśli $T = T_{1|A} = T_{2|A}$. *Maksymalnym zgodnym poddrzewem*, w skrócie *MAST* (ang. *Maximum Agreement Subtree*), nazywamy poddrzewo zgodne posiadające maksymalną liczbę liści. MAST pozwala intuicyjnie zobrazować i wyróżnić te informacje dotyczące filogenezy gatunków, które są wspólne dla danego zbioru drzew. Przez $MAST(T_1, T_2)$ oznaczymy liczbę liści maksymalnego poddrzewa zgodnego dla T_1 i T_2. Problem polegający na wyznaczeniu MAST dla trzech lub więcej drzew jest NP-trudny [7]. Wyznaczenia MAST dla dwóch drzew nieukorzenionych (jak również i ukorzenionych) można dokonać w czasie $O(|L|^{1.5})$ [65], natomiast w przypadku, gdy drzewa te są binarne i ukorzenione znany jest szybszy algorytm $O(|L| \log |L|)$ [66, 37].

Dla danego drzewa T rozważmy drzewa posiadające mniej informacji filogenetycznej. Niech $e = \{u, v\}$ będzie krawędzią wewnętrzną w T. *Ściągnięcie* krawędzi e w T jest operacją, która przekształca T w T_e, polegającą na usunięciu krawędzi e i utożsamieniu wierzchołków u oraz v. Zauważmy, że wskutek tej transformacji ilość rozbić (lub klastrów, jeśli T jest ukorzenione) maleje o 1. Operacją odwrotną do ściągnięcia jest operacja *wprowadzenia* krawędzi, która odpowiada dołączaniu nowego rozbicia (lub klastra) kompatybilnego z pozostałymi. Drzewo T' jest *rozszerzeniem* drzewa T, jeśli T' może być otrzymane z T wskutek sekwencji operacji wprowadzenia krawędzi.

Istnieje dokładnie jedno drzewo nieukorzenione o trzech liściach oraz cztery drzewa posiadające 4 liście, które zostały przedstawione na rysunku 2.5. Trzy z nich są binarne, nazywamy je *kwartetami binarnymi*, natomiast drzewo niebinarne będziemy określać jako *kwartet nierozwiązany*. Kwartet

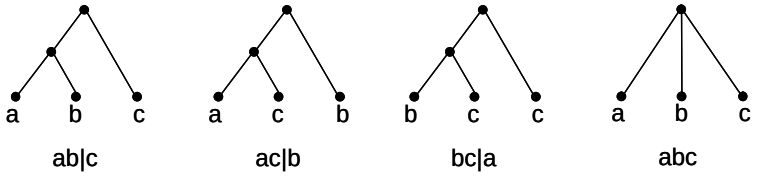


Rysunek 2.5: Drzewa nieukorzenione o 4 liściach.

q nad pewnym 4-elementowym podzbiorem $A \subseteq L$ *występuje* w drzewie $T \in U_L$, jeśli poddrzewo $T_{|A}$ jest tożsame z q.

Zbiór kwartetów binarnych (inaczej *rozwiązanych*) występujących w nieukorzenionym drzewie T oznaczymy przez $q_b(T)$, zaś zbiór kwartetów nierozwiązanych w T przez $q_u(T)$. Ponadto zbiór wszystkich kwartetów w T oznaczymy przez $qt(T) = q_b(T) \cup q_u(T)$. Nieukorzenione drzewo filogenetyczne $T \in U_L$ jest jednoznacznie określone przez zbiór jego kwartetów binarnych $q_b(T)$ [57].

W przypadku drzew ukorzenionych mamy jedno drzewo dwulistne oraz 4 możliwe drzewa trzylistne (rysunek 2.6). Konsekwentnie, trzylistne drzewa binarne nazywane są *tripletami binarnymi*, drzewo niebinarne zaś będziemy określać jako *triplet nierozwiązany*. Triplet t nad pewnym 3-elemento-



Rysunek 2.6: Drzewa ukorzenione o 3 liściach.

wym podzbiorem $A \subseteq L$ *występuje* w drzewie ukorzenionym $T \in R_L$, jeśli poddrzewo $T_{|A}$ jest tożsame z t. Zbiór wszystkich tripletów drzewa ukorzenionego T oznaczymy przez $tt(T)$, zbiór tripletów binarnych (inaczej *rozwiązanych*) przez $t_b(T)$, a tripletów nierozwiązanych w T przez $t_u(T)$.

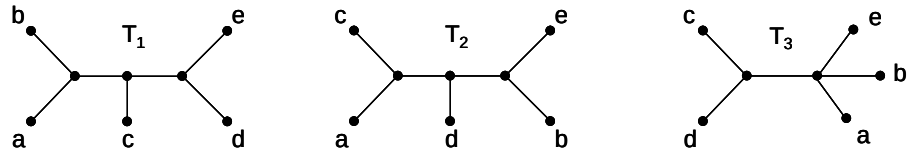
2.3 Klasyczne metryki filogenetyczne

Jedną z najczęściej używanych metod mierzenia podobieństwa drzew filogenetycznych jest odległość Robinsona-Fouldsa (RF) [95]. Istota tej metody polega na określeniu rozbić, które występują tylko w jednym z analizowanych drzew. Za wartość metryki RF dla drzew nad tym samym zbiorem liści przyjmuje się albo wprost moc zbioru $\beta(T_1) \oplus \beta(T_2)$ lub wartość tą przeskalowaną przez $\frac{1}{2}$. Zatem jeśli $T_1, T_2 \in U_L$, to wartość $|\beta(T_1) \oplus \beta(T_2)|$ może być interpretowana jako minimalna ilość operacji ściągnięcia i wprowadzania krawędzi, która jest potrzebna do przekształcenia drzewa T_1 w T_2. Dla drzew binarnych wartość ta jest zawsze liczbą parzystą, stąd wygodnie jest stosować w tym przypadku skalowanie przez $\frac{1}{2}$. Choć w pracy rozważane są zarówno drzewa binarne, jak i niebinarne, metrykę RF będziemy definiować konsekwentnie jako przeskalowaną.

Definicja 2.6. *Metryka Robinsona-Fouldsa* (RF) [95] dla drzew nieukorzenionych $T_1, T_2 \in U_L$ jest zdefiniowana następująco:

$$d_{RF}(T_1, T_2) = \frac{1}{2}|\beta(T_1) \oplus \beta(T_2)|. \tag{2.1}$$

Dla drzew przedstawionych na rysunku 2.7 otrzymujemy: $d_{RF}(T_1, T_2) = 2$, $d_{RF}(T_1, T_3) = 1.5$.

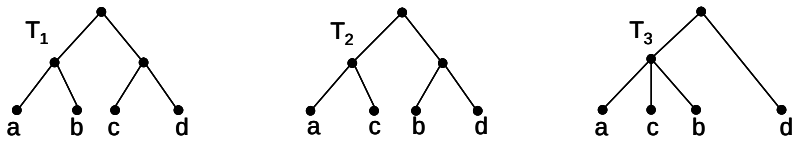


Rysunek 2.7: Przykłady drzew nieukorzenionych binarnych i niebinarnych.

Analogicznie definiuje się metrykę RF dla drzew ukorzenionych. Jedyna różnica w tym przypadku polega na użyciu w definicji zbioru klastrów w miejsce rozbić.

Definicja 2.7. *Metryka Robinsona-Fouldsa* (RFC) [95] dla drzew ukorzenionych $T_1, T_2 \in R_L$ jest zdefiniowana następująco:

$$d_{RFC}(T_1, T_2) = \frac{1}{2}|\sigma(T_1) \oplus \sigma(T_2)|. \tag{2.2}$$



Rysunek 2.8: Przykłady drzew ukorzenionych binarnych i niebinarnych.

Odległości dla drzew przedstawionych na rysunku 2.8 są następujące: $d_{RFC}(T_1, T_2) = 2$, $d_{RFC}(T_1, T_3) = 1.5$. Istotną zaletą metryki Robinsona-Fouldsa jest efektywność jej wyznaczania. Istnieje bowiem algorytm o złożoności $O(|L|)$ podany przez Day'a w [40].

Idea konstrukcji kolejnych metryk opiera się na wykorzystaniu różnic w odległości między parami liści w analizowanych drzewach. Niech $\Xi_T(i, j)$ oznacza liczbę krawędzi na ścieżce pomiędzy liśćmi i oraz j w drzewie $T \in U_L$, a $\Xi(T)$ będzie $\frac{|L|(|L|-1)}{2}$-elementowym wektorem tych odległości między wszystkimi parami liści w T o ustalonym porządku.

Definicja 2.8. *Metryka ścieżkowa* (PD — *Path Difference*) [106] dla drzew nieukorzenionych $T_1, T_2 \in U_L$ jest zdefiniowana następująco:

$$d_{PD}(T_1, T_2) = ||\Xi(T_1) - \Xi(T_2)||_2. \tag{2.3}$$

Zatem $d_{PD}(T_1, T_2)$ to pierwiastek kwadratowy z sumy kwadratów różnic odległości między poszczególnymi parami liści w porównywanych drzewach. Przykładowe odległości dla drzew na rysunku 2.7 są następujące: $d_{PD}(T_1, T_2) = \sqrt{14}$, $d_{PD}(T_1, T_3) = \sqrt{12}$. Poprawność tej definicji zapewnia twierdzenie Smolenskii'ego (1963 rok) [102], mówiące, że dwa drzewa nieukorzenione $T, T' \in U_L$ są izomorficzne ($T = T'$) wtedy i tylko wtedy, gdy dla każdej pary liści i, j odległości między i oraz j w T i T' są równe. Twierdzenie to zostało później rozszerzone przez Zaretskii'ego (1965 rok) [116], gdzie wprowadzono charakterystykę wektorów odległości między liśćmi drzewa nieukorzenionego w postaci tzw. *warunku czterech punktów*. Złożoność obliczeniowa wyznaczania wartości PD wynosi $O(|L|^2)$ [106].

Bardzo zbliżona metoda definiowania odległości, różniąca się jedynie użytą przy porównywaniu wektorów normą, została zaproponowana w pracy [47], natomiast algorytm wraz z podstawową analizą własności tej metryki pojawia się w [13].

Definicja 2.9. *Metryka węzłowa* (ND — *Nodal Distance*) [47], [13] dla drzew nieukorzenionych $T_1, T_2 \in U_L$ jest zdefiniowana następująco:

$$d_{ND}(T_1, T_2) = ||\Xi(T_1) - \Xi(T_2)||_1. \tag{2.4}$$

W miejsce normy L^2 pojawia się tu L^1. Odległość ND jest zatem równa sumie wartości bezwzględnych różnic w odległościach pomiędzy parami liści w analizowanych drzewach. Wartości metryki ND dla drzew na rysunku 2.7 są następujące: $d_{ND}(T_1, T_2) = 10$, $d_{ND}(T_1, T_3) = 8$.

Przeniesienie opisanej idei porównywania odległości między parami liści na drzewa ukorzenione jest bardziej skomplikowane. W pracy [33] z 2010 roku wykazano, że za pomocą wektora $\Xi(T)$ można jednoznacznie opisać tylko drzewa ukorzenione binarne. W przypadku drzew niebinarnych funkcje analogiczne do PD i ND nie są więc metrykami. Podejście zaproponowane w [33], polegające na rozbiciu długości ścieżki między dwoma liśćmi i oraz j w drzewie ukorzenionym na dwie części, z których jedną stanowi odległość od i do najbliższego wspólnego przodka $LCA(\{i, j\})$, a drugą odległość $LCA(\{i, j\})$ do j, pozwala na uniknięcie tego problemu.

Niech $\Xi_T^S(i,j)$ oznacza odległość między liściem i a $LCA(\{i,j\})$ w T, czyli zachodzi zależność $\Xi_T(i,j) = \Xi_T^S(i,j) + \Xi_T^S(j,i)$. Z dowolnym drzewem ukorzenionym $T \in R_n$ możemy zatem skojarzyć następującą macierz:

$$\Xi^S(T) = \begin{pmatrix} 0 & \Xi_T^S(1,2) & \cdots & \Xi_T^S(1,n) \\ \Xi_T^S(2,1) & 0 & \cdots & \Xi_T^S(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ \Xi_T^S(n,1) & \Xi_T^S(n,2) & \cdots & 0 \end{pmatrix}.$$

Definicja 2.10. Rodzina metryk *węzłowych* (SN — *Splitted Nodal Metrics*) [33] dla drzew ukorzenionych $T_1, T_2 \in R_L$ jest zdefiniowana następująco:

$$d_{SN}^p(T_1,T_2) = ||\Xi^S(T_1) - \Xi^S(T_2)||_p, \tag{2.5}$$

gdzie $||.||_p$ jest p-normą macierzy, $p \in R_{\geq 1}$, zdefiniowaną dla macierzy $M = [m_{ij}]$ o wymiarach $k \times l$ jako $||M||_p = \left(\sum_{i=1}^{k} \sum_{i=j}^{l} |m_{ij}|^p \right)^{1/p}$.

Rolę reprezentanta powyższej rodziny metryk w dalszych rozważaniach będzie pełnić funkcja d_{SN}^2, oznaczana dalej jako odległość SN. Wartości odległości SN dla drzew na rysunku 2.8 są następujące: $d_{SN}^2(T_1,T_2) = \sqrt{8}$, $d_{SN}^2(T_1,T_3) = \sqrt{7}$.

Kolejne dwie zbliżone w swojej konstrukcji metryki opierają się na zliczaniu różnych poddrzew 3- lub 4-listnych występujących w porównywanych drzewach.

Definicja 2.11. *Metryka kwartetowa* (QT) [45] dla drzew nieukorzenionych $T_1, T_2 \in U_L$ jest zdefiniowana następująco:

$$d_{QT}(T_1,T_2) = \frac{1}{2}|qt(T_1) \oplus qt(T_2)|. \tag{2.6}$$

Dla drzew przedstawionych na rysunku 2.7 otrzymujemy: $d_{QT}(T_1,T_2) = d_{QT}(T_1,T_3) = 4$. Dla drzew binarnych wartość metryki QT można wyznaczyć w czasie $O(|L| \log |L|)$ [25]. W przypadku drzew dowolnych najlepszy znany do tej pory algorytm, o złożoności niezależnej od stopni wierzchołków wynoszącej $O(|L|^{2.688})$, został podany stosunkowo niedawno (2011 rok) w pracy [81].

Definicja 2.12. *Metryka tripletowa* (TT) [38] dla drzew ukorzenionych $T_1, T_2 \in R_L$ jest zdefiniowana jako

$$d_{TT}(T_1, T_2) = \frac{1}{2}|tt(T_1) \oplus tt(T_2)|. \tag{2.7}$$

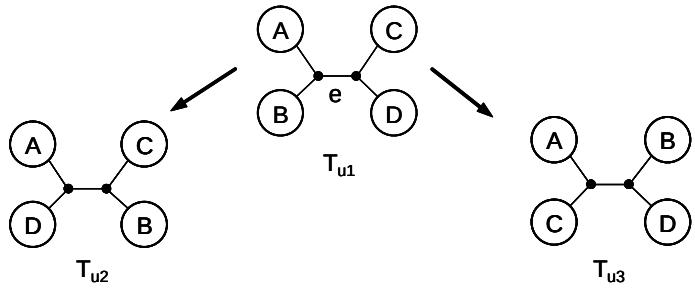
Wartości odległości TT dla drzew przedstawionych na rysunku 2.8 są następujące: $d_{TT}(T_1, T_2) = 4$, $d_{TT}(T_1, T_3) = 3$. Wartość metryki TT można wyznaczyć w czasie $O(|L|^2)$, zarówno w przypadku drzew binarnych, dla których możemy wykorzystać stosunkowo prosty algorytm zaprezentowany w [38], jak i dla drzew dowolnych, używając w tym przypadku nowszego (2011 rok) i dużo bardziej skomplikowanego algorytmu przedstawionego w [8].

2.4 Operacje edycyjne i indukowane przez nie metryki

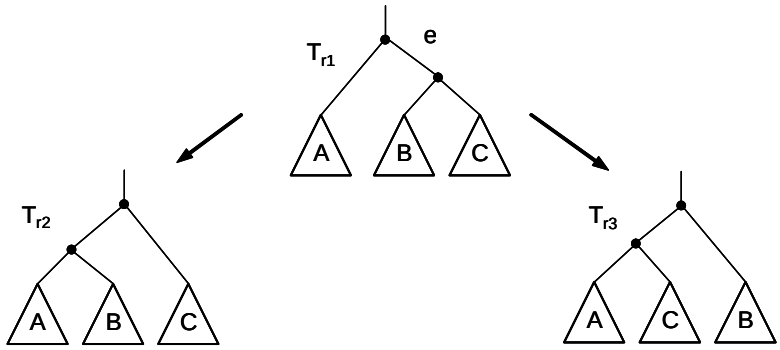
Operacje edycyjne są powszechnie stosowane w heurystykach przeszukujących zbiory drzew filogenetycznych używanych do przybliżonego rozwiązania wielu problemów optymalizacyjnych dotyczących filogenetyki. Za pomocą wspomnianych przekształceń można również określać odległości w zbiorach U_L^B i R_L^B.

Jedną z trzech podstawowych operacji edycyjnych definiowanych dla drzew filogenetycznych binarnych jest operacja *NNI* (ang. *Nearest Neighbour Interchange*), por. [6, 34]. Polega ona na zamianie miejscami dwóch poddrzew znajdujących się po przeciwnych stronach wspólnej krawędzi wewnętrznej (rys. 2.9, 2.10). Dla danej krawędzi wewnętrznej e możliwe są zawsze 2 operacje NNI tworzące różne drzewa.

Przez $d_{uNNI}(T_{u1}, T_{u1})$ (odpowiednio $d_{rNNI}(T_{r1}, T_{r2})$), gdzie $T_{u1}, T_{u2} \in U_L^B$ ($T_{r1}, T_{r2} \in R_L^B$) oznaczymy minimalną liczbę operacji uNNI (rNNI) niezbędną do transformacji drzewa T_{u1} (T_{r1}) w T_{u2} (T_{r2}). Ponieważ wykonując kolejno operacje NNI każde drzewo można przekształcić w dowolne inne [94], to funkcje d_{uNNI} i d_{rNNI} są dobrze określonymi metrykami w zbiorach odpowiednio U_L^B i R_L^B. Niestety wyznaczanie wartości metryk



Rysunek 2.9: Schemat operacji uNNI dla drzew nieukorzenionych. Koła reprezentują pojedyncze liście lub większe poddrzewa.



Rysunek 2.10: Schemat operacji rNNI dla drzew ukorzenionych. Trójkąty reprezentują pojedyncze liście lub większe poddrzewa ukorzenione. Wierzchołek umieszczony najwyżej w T_{r1}, T_{r2}, T_{r3} może być również korzeniem.

opartach na operacji NNI jest problemem NP-trudnym zarówno dla drzew nieukorzenionych jak i ukorzenionych [39].

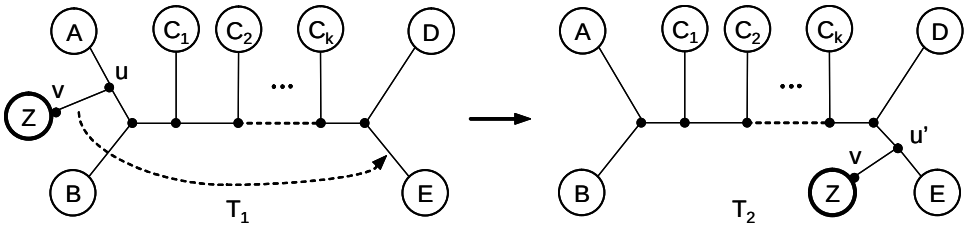
Kolejną istotną transformacją edycyjną jest operacja *SPR* (ang. *Subtree Prune and Regraft*). Niech $e = \{u, v\}$ będzie pewną krawędzią w $T \in U_L^B$, taką że u jest wierzchołkiem wewnętrznym. Idea operacji uSPR jest następująca:

1. Usuwamy krawędź e. Powoduje to rozpad drzewa T na dwie składowe T^u i T^v zawierające odpowiednio wierzchołki u i v.

2. Ściągamy wierzchołek u w T^u tworząc krawędź f.

3. Wybieramy dowolną krawędź w T^u (różną od f) i wprowadzamy na niej nowy wierzchołek u'.

4. Dołączamy poddrzewo T_v tworząc krawędź $\{v, u'\}$.



Rysunek 2.11: Drzewo T_2 powstaje z drzewa T_1 w wyniku pojedynczej operacji uSPR (SPR dla drzew nieukorzenionych). Koła reprezentują pojedyncze liście lub większe poddrzewa.

Podobnie jak NNI operacja SPR dla drzew nieukorzenionych (uSPR) ma swój odpowiednik dla drzew z korzeniem — rSPR. W przypadku operacji rSPR wygodnie jest rozszerzyć drzewo T o dodatkowy wierzchołek r' (pełniący wyłącznie rolę pomocniczą w definicji operacji rSPR) połączony krawędzią z $r(T)$. Niech $e = \{u, v\}$ będzie dowolną krawędzią, która nie jest incydentna z r', taką że wierzchołek u występuje na ścieżce łączącej v z r'. Operacja rSPR przebiega następująco (rys. 2.12):

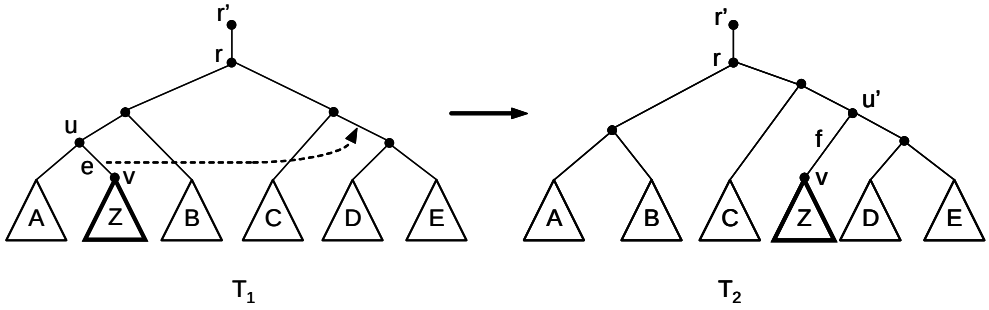
1. Usuwamy krawędź e. Powoduje to rozpad drzewa T na dwie składowe T^u i T^v zawierające odpowiednio wierzchołki u i v.

2. Wybieramy dowolną krawędź w T^u i wprowadzamy na niej nowy wierzchołek u'.

3. Dołączamy T^v łącząc wierzchołek v z u' nową krawędzią f.

4. Ściągamy u. Jeśli $u = r$, to korzeń drzewa ulegnie ściągnięciu, rolę korzenia w drzewie po transformacji będzie wówczas pełnić wierzchołek połączony krawędzią z r'.

5. Usuwamy wierzchołek pomocniczy r'



RYSUNEK 2.12: Drzewo T_2 powstaje z drzewa T_1 w wyniku pojedynczej operacji rSPR (SPR dla drzew ukorzenionych). Wierzchołek r' ma znaczenie pomocnicze.

Przez $d_{uSPR}(T_{u1}, T_{u2})$ oraz $d_{rSPR}(T_{r1}, T_{r2})$ oznaczymy minimalną liczbę odpowiednich operacji niezbędną do transformacji drzewa T_{u1} w T_{u2} oraz T_{r1} w T_{r2}. Podobnie jak w przypadku operacji NNI, d_{uSPR} oraz d_{rSPR} są metrykami w zbiorach odpowiednio U_L^B i R_L^B. Warto zauważyć, że operacja SPR jest szeroko wykorzystywana do modelowania i wykrywania horyzontalnego transferu genów (HGT) [12, 61, 15, 113]. Jedak obliczanie wartości metryki SPR jest problemem NP-trudnym, zarówno dla uSPR [59], jak i rSPR [23].

Wśród standardowych operacji edycyjnych oprócz NNI i SPR wymienia się również transformację *TBR* (ang. *Tree Bisection and Reconnection*). Niech $e = \{u, v\}$ będzie pewną krawędzią w $T \in U_L^B$, taką że u jest wierzchołkiem wewnętrznym. Operacja TBR dla drzewa nieukorzenionego T (uTBR) przebiega następująco (rys. 2.13):

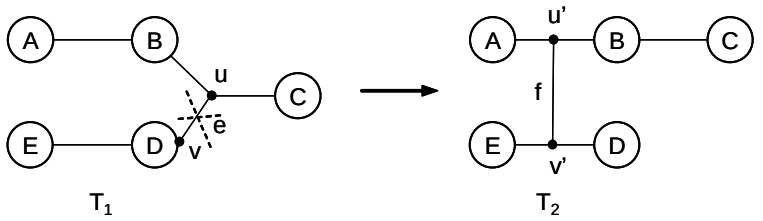
1. Usuwamy krawędź e. Powoduje to rozpad drzewa T na dwie składowe T^u i T^v zawierające odpowiednio wierzchołki u i v.

2. Ściągamy wierzchołek u w T^u oraz, o ile to możliwe, wierzchołek v (tj., gdy v ma stopień dwa).

3. Wybieramy dowolną krawędź w T^u i wprowadzamy na niej nowy wierzchołek u'.

4. Jeśli T^v jest pojedynczym liściem, wówczas $v' = v$, w przeciwnym razie wybieramy dowolną krawędź w T^v i wprowadzamy na niej nowy wierzchołek v'.

5. Łączymy składowe T^v i T^u tworząc krawędź $\{v', u'\}$.



Rysunek 2.13: Drzewo T_2 powstaje z drzewa T_1 w wyniku pojedynczej operacji uTBR (TBR dla drzew nieukorzenionych).

Operację TBR dla drzew ukorzenionych (rTBR) definiuje się podobnie jak rSPR z tą różnicą, że dopuszcza się możliwość ściągnięcia v i wprowadzenia nowego wierzchołka v' stopnia dwa jako korzenia w przemieszczanym poddrzewie Z (rysunk 2.12). Wówczas krawędź f łączy wierzchołek u' z v'. Wyznaczanie metryk d_{uTBR}, d_{rTBR}, których wartość jest zdefiniowana jako najmniejsza ilość operacji odpowiedniego typu dzieląca dane dwa drzewa, jest podobnie jak w poprzednich przypadkach problemem NP-trudnym [6].

Miedzy opisanymi operacjami edycyjnymi zachodzą następujące relacje:

Lemat 2.3 ([72, 6])**.**

1. *Operacja NNI jest szczególnym przypadkiem SPR, a SPR jest szczególnym przypadkiem operacji TBR.* [72]

2. *Dla dowolnych drzew* $T_1, T_2 \in U_L^B$ *zachodzi:*
   (a) $d_{uTBR}(T_1, T_2) \leq d_{uSPR}(T_1, T_2) \leq d_{uNNI}(T_1, T_2)$ [6],
   (b) $d_{uSPR}(T_1, T_2) \leq 2 \cdot d_{uTBR}(T_1, T_2)$ [6].

2.5 Podsumowanie

Metryki NNI, SPR i TBR, opierające się na operacjach edycyjnych, posiadają intuicyjną interpretację filogenetyczną, lecz ich wyznaczanie jest nieefektywne obliczeniowo. Metryki łatwe obliczeniowo, np. PD, NS, TT, QT, nie posiadają już tak jasnej interpretacji jak odległości wymienione poprzednio. Metryka RF ma dość intuicyjne uzasadnienie biologiczne, lecz jak zostanie to wykazane w kolejnych częściach pracy, ma również pewne istotne wady, wynikające głównie z prostoty jej konstrukcji.

W literaturze ciągle pojawiają się nowe definicje odległości, np. metryka algebraiczna dla drzew z korzeniem o liniowym względem liczby liści czasie wyznaczania [110, 2], metryki „podziałowa" i „ścieżkowa" zaproponowane w [20], czy metryka edycyjna dla drzew o dowolnym (niekoniecznie jednakowym) zbiorze liści [69]; niestety w przypadku trzech ostatnich efektywne algorytmy nie są znane. Niegasnące zainteresowanie badaczy metodami definiowania odległości w przestrzeni drzew filogenetycznych potwierdza aktualność problemu i dowodzi niemalejącego zapotrzebowania na nowe efektywne metody pomiaru.

Celem niniejszej pracy jest skonstruowanie metryk łatwych obliczeniowo, uogólniających odległości RF oraz RFC, lecz wolnych od ich podstawowych wad. Idea zaproponowanego rozwiązania polega na konstrukcji metryki między podzbiorami pewnej rodziny D, posługując się dowolną odległością h zdefiniowaną na jej elementach. Dla nieukorzenionych drzew filogenetycznych o etykietach liści z L zbiorem D będzie rodzina rozbić L, natomiast dla drzew z korzeniem będzie to rodzina podzbiorów L, czyli klastrów. Wykorzystując własności najlżejszego doskonałego skojarzenia w grafach dwudzielnych, dowolna metryka h zdefiniowana na zbiorach rozbić lub klastrów L może być rozszerzona na odpowiednie rodziny drzew filogenetycznych. Niezbędne pojęcia formalizujące powyższą ideę zostaną wprowadzone w następnym rozdziale.

3 Definicja metryk skojarzeniowych

3.1 Odległość podzbiorów przestrzeni metrycznej

Do określenia metryk skojarzeniowych na U_L i R_L przydatna będzie ogólna metoda definiowania odległości między podzbiorami, pozwalająca na przeniesienie metryki z punków na ich zbiory.

Lemat 3.1 ([17, 19])**.** *Dana jest przestrzeń metryczna (X, d), pełny graf dwudzielny $G(V_1, V_2, E)$, gdzie $|V_1| = |V_2| = n$ oraz funkcja $l : V_1 \cup V_2 \to X$ etykietująca wierzchołki G elementami tej przestrzeni. Krawędziom G przypisujemy wagi liczbowe w taki sposób, że $w(\{a, b\}) = d(l(a), l(b))$ dla $a \in V_1$, $b \in V_2$. Niech $a_1, \ldots, a_k \in V_1$, $b_1, \ldots, b_k \in V_2$. Jeśli $l(a_i) = l(b_i)$ dla $1 \leq i \leq k$, to istnieje najlżejsze doskonałe skojarzenie $M \subseteq E$ spełniające $\{a_i, b_i\} \in M$ dla $1 \leq i \leq k$.*

Dowód. Niech i będzie najmniejszą liczbą $1 \leq i \leq k$, taką że krawędź $\{a_i, b_i\}$ nie należy do najlżejszego doskonałego skojarzenia M. Istnieją zatem $x \in V_1 \setminus \{a_1, \ldots, a_i\}$, $y \in V_2 \setminus \{b_1, \ldots, b_i\}$, takie że krawędzie $\{a_i, y\}$, $\{x, b_i\}$ należą do M. Tworzymy nowe skojarzenie $M' = M \setminus \{\{a_i, y\}, \{x, b_i\}\} \cup \{\{x, y\}, \{a_i, b_i\}\}$. Na mocy nierówności trójkąta mamy $w(\{a_i, y\}) + w(\{x, b_i\}) \geq w(\{x, y\}) = w(\{x, y\}) + w(\{a_i, b_i\})$. Stąd M' podobnie jak M jest również najlżejszym doskonałym skojarzeniem. Jeśli istnieje taka potrzeba, opisana transformacja może być powtórzona dla kolejnych większych $i \leq k$. □

Definicja 3.1 ([16, 18, 17, 19])**.** Dane są: skończony zbiór D, wyróżniony element pomocniczy $O \notin D$ oraz metryka h w zbiorze $D \cup \{O\}$. Definiujemy metrykę $d_h : 2^D \times 2^D \to \mathbb{R}_{\geq 0}$, gdzie odległość $d_h(A, B)$ pomiędzy zbiorami $A, B \in 2^D$ jest równa wadze najlżejszego doskonałego skojarzenia w grafie dwudzielnym $G(V_1, V_2, E)$, $|V_1| = |V_2| = n$ określonym następująco:

- dla dowolnych s, t, takich że $s - t = |A| - |B|$ zbiory

$$V_1 = \{a_1, \ldots, a_{|A|}, a_{|A|+1}, \ldots, a_{|A|+t}\},$$
$$V_2 = \{b_1, \ldots, b_{|B|}, b_{|B|+1}, \ldots, b_{|B|+s}\}$$

  tworzą partycje grafu $G(V_1, V_2, E)$,

- etykiety wierzchołków wyznaczone są przez funkcję $l : V_1 \cup V_2 \to D \cup \{O\}$, tak że $A = \{l(a_i) : 1 \leq i \leq |A|\}$, $B = \{l(b_j) : 1 \leq j \leq |B|\}$ i $l(a_i) = l(b_j) = O$ dla $|A| + 1 \leq i \leq n$, $|B| + 1 \leq j \leq n$,

- wagi krawędzi określone są przez metrykę h dla $1 \leq i, j \leq n$ jako

$$w(\{a_i, b_j\}) = h(l(a_i), l(b_j)).$$

Lemat 3.2 ([19])**.** *Funkcja d_h jest metryką nad 2^D oraz wartość $d_h(A, B)$ nie zależy od wyboru s i t (dla $s - t = |A| - |B|$).*

Dowód. Ponieważ liczby t i s określają ilości wierzchołków związanych z elementem pomocniczym O odpowiednio w V_1 i V_2, to na mocy lematu 3.1 wartość d_h nie zależy od wyboru s i t .

Zauważmy, że $d_h(A, B) = 0$, w przypadku gdy wszystkie krawędzie najlżejszego doskonałego skojarzenia łączą wierzchołki o równych etykietach, tzn., gdy $A = B$. Symetria funkcji d_h jawnie wynika z definicji 3.1. Pozostaje zatem wykazanie nierówności trójkąta. Dane są zbiory $X_1, X_2, X_3 \in 2^D$. Niech $G_{ij} = (V_i, V_j, E_{ij})$, $1 \leq i < j \leq 3$ będą grafami użytymi do obliczania wartości d_h między zbiorami X_i, X_j. Na mocy własności udowodnionej powyżej możemy założyć, że $n = |V_1| = |V_2| = |V_3| =$

$\max\{|X_1|, |X_2|, |X_3|\}$ i $V_1 = \{a_1, \ldots, a_n\}$, $V_2 = \{b_1, \ldots, b_n\}$, oraz $V_3 = \{c_1, \ldots, c_n\}$. Co więcej, możemy również przyjąć, że $\{\{a_i, b_i\}\}_{i=1,\ldots,n} \subseteq E_{12}$ i $\{\{b_i, c_i\}\}_{i=1,\ldots,n} \subseteq E_{23}$ są najlżejszymi doskonałymi skojarzeniami w odpowiednich grafach. Z faktu, że h jest metryką w zbiorze $D \cup \{O\}$ wynika, iż $d_h(X_1, X_2) + d_h(X_2, X_3) = \sum_{i=1}^{n}(w(\{a_i, b_i\}) + w(\{b_i, c_i\})) \geq \sum_{i=1}^{n} w(\{a_i, c_i\}) \geq d_h(X_1, X_3)$. □

Na mocy powyższego lematu możemy zawsze przyjąć, że $\min\{s, t\} = 0$ i $\max\{s, t\} = ||A| - |B||$. Odległość $d_h(A, B)$ między zbiorami A i B może być interpretowana jako koszt najlepszego dopasowania (sparowania) elementów z obu zbiorów. Natomiast wartość $h(O, x)$ jest kosztem pozostawienia elementu x bez pary. W przypadku gdy moce obu zbiorów są równe, określenie odległości do elementu O jest zbędne.

Zauważmy też, że na mocy lematu 3.1 konstrukcję z definicji 3.1 wystarczy stosować dla zbiorów o niepustym przekroju, ponieważ dla $A, B \in 2^D$, $A \cap B \supseteq C$ mamy

$$d_h(A, B) = d_h(A \setminus C, B \setminus C). \tag{3.1}$$

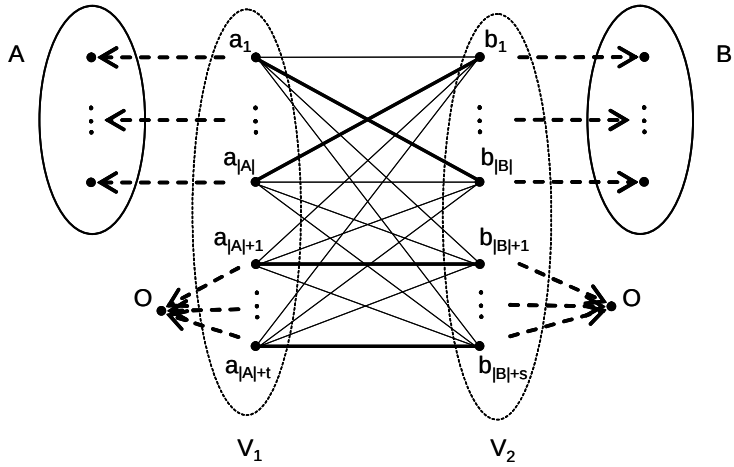
3.2 Metryki skojarzeniowe wykorzystujące rozbicia i klastry

Przedstawiona w poprzednim punkcie metoda pozwala wygodne konstruować metryki dla drzew filogenetycznych, zarówno w zbiorze drzew nieukorzenionych U_L, jak i ukorzenionych R_L.

Niech $Splits(L)_O = Splits(L) \cup \{O\}$ i $h_S : Splits(L)_O \times Splits(L)_O \to \mathbb{R}_{\geq 0}$ będzie dowolną metryką w zbiorze rozbić L uzupełnionym o element pomocniczy O. Wówczas funkcja $d^T_{h_S} : U_L \times U_L \to \mathbb{R}_{\geq 0}$ określona zależnością:

$$d^T_{h_S}(T_1, T_2) = d_{h_S}(\beta(T_1), \beta(T_2)) = d_{h_S}(\beta_*(T_1), \beta_*(T_2)), \tag{3.2}$$

na mocy lematu 3.1 oraz faktu, że zbiór $\beta(T)$ jednoznacznie opisuje $T \in U_L$, jest metryką w zbiorze nieukorzenionych drzew filogenetycznych U_L. Druga równość wynika z (3.1).



Rysunek 3.1: Ilustracja definicji metryki skojarzeniowej. Strzałki narysowane liniami przerywanymi odpowiadają etykietowaniu wierzchołków grafu wprowadzanemu przez funkcję l. Zbiory A i B mogą zawierać wspólne elementy, lecz dla przejrzystości rysunku umieszczono je osobno, po lewej i prawej stronie.

Zauważmy, iż za pomocą formuły (3.2) można łatwo opisać metrykę RF. W tym celu jako funkcję porównującą rozbicia wystarczy przyjąć metrykę $h_{RF} : Splits(L)_O \times Splits(L)_O \to \{0, \frac{1}{2}, 1\}$ zdefiniowaną następująco:

$$h_{RF}(s_1, s_2) = \begin{cases} 0, \text{ dla } s_1 = s_2, \\ 1, \text{ dla } s_1 \neq s_2 \text{ i } s_1, s_2 \in Splits(L), \\ \dfrac{1}{2}, \text{ w pozostałych przypadkach.} \end{cases} \tag{3.3}$$

Wniosek 3.3. *Niech $T_1, T_2 \in U_L$, wówczas:*

$$d_{RF}(T_1, T_2) = d^T_{h_{RF}}(T_1, T_2) = d_{h_{RF}}(\beta_*(T_1), \beta_*(T_2)). \tag{3.4}$$

Dowód. Graf dwudzielny $G(V_1, V_2, E)$ umożliwiający wyznaczanie wartości $d_{h_{RF}}(\beta_*(T_1), \beta_*(T_2))$ możemy zgodnie z definicją 3.1 skonstruować następująco:

- partycję V_1 tworzy $|\beta_*(T_1)|$ wierzchołków odpowiadających nietrywialnym rozbiciom w T_1 oraz dodatkowo $|\beta_*(T_2)| - |\beta_*(T_1) \cap \beta_*(T_2)|$ wierzchołków związanych z elementem O,
- w partycji V_2 znajduje się $|\beta_*(T_2)|$ wierzchołków odpowiadających nietrywialnym rozbiciom z T_2 i $|\beta_*(T_1)| - |\beta_*(T_1) \cap \beta_*(T_2)|$ wierzchołków odpowiadających elementowi O.

Niech $k = |\beta_*(T_1)| + |\beta_*(T_2)| - |\beta_*(T_1) \cap \beta_*(T_2)|$. Konstruujemy sparowanie $M = \{(s_i, t_i)\}$, $s_i \in \beta_*(T_1) \cup \{O\}$, $t_i \in \beta_*(T_2) \cup \{O\}$ dla $i = 1, \ldots, k$, w którym identyczne rozbicia są sparowane ze sobą, natomiast pozostałe występują w parze z elementem O. Zauważmy, że koszt sparowania M jest równy wadze najlżejszego doskonałego skojarzenia w G, gdyż dla każdej pary $(s_i, t_i) \in M$ związanej w wierzchołkami $u_i \in V_1$, $v_i \in V_2$ mamy $w(\{u_i, v_i\}) = \min_{x \in V_2} w(\{u_i, x\}) = \min_{x \in V_1} w(\{x, v_i\})$. Pozostaje wyznaczyć koszt M, który wynosi $\frac{1}{2}(|\beta_*(T_1)| - |\beta_*(T_1) \cap \beta_*(T_2)|) + \frac{1}{2}(|\beta_*(T_2)| - |\beta_*(T_1) \cap \beta_*(T_2)|) = \frac{1}{2}(|\beta_*(T_1)| + |\beta_*(T_2)| - 2|\beta_*(T_1) \cap \beta_*(T_2)|) = d_{RF}(T_1, T_2)$ i na mocy lematu 3.2 odpowiada wartości $d^T_{h_{RF}}(T_1, T_2)$. □

Przyjmijmy następnie za element pomocniczy zbiór pusty $O = \emptyset$. Niech $h_C : 2^L \times 2^L \to \mathbb{R}_{\geq 0}$ będzie dowolną metryką. Wówczas funkcja $d^T_{h_C} : R_L \times R_L \to \mathbb{R}_{\geq 0}$ określona zależnością:

$$d^T_{h_C}(T_1, T_2) = d_{h_C}(\sigma(T_1), \sigma(T_2)) = d_{h_C}(\sigma_*(T_1), \sigma_*(T_2)), \tag{3.5}$$

na mocy lematu 3.2, jest metryką w zbiorze ukorzenionych drzew filogenetycznych R_L. Analogicznie jak w przypadku RF metrykę RFC możemy opisać za pomocą zależności (3.5). W tym celu wystarczy zdefiniować funkcję $h_{RFC} : 2^L \times 2^L \to \{0, \frac{1}{2}, 1\}$ następująco:

$$h_{RFC}(c_1, c_2) = \begin{cases} 0, \text{ dla } c_1 = c_2, \\ 1, \text{ dla } c_1 \neq c_2 \text{ i } c_1, c_2 \in 2^L \setminus \{\emptyset\}, \\ \dfrac{1}{2}, \text{ w pozostałych przypadkach.} \end{cases} \tag{3.6}$$

Wniosek 3.4. *Niech $T_1, T_2 \in R_L$, wówczas:*

$$d_{RFC}(T_1, T_2) = d^T_{h_{RFC}}(T_1, T_2) = d_{h_{RFC}}(\sigma_*(T_1), \sigma_*(T_2)). \tag{3.7}$$

Dowód. Konstrukcja przebiega analogicznie jak w przypadku wniosku 3.3. □

Przedstawiona metoda definiowania odległości posiada wiele zalet, z których podstawową i bardzo istotną w praktycznych zastosowaniach jest możliwość konstrukcji metryk łatwych obliczeniowo (pod warunkiem, że można efektywnie wyznaczyć wartość funkcji h). Ponadto zauważmy, że odpowiednio definiując funkcje h_C oraz h_S możemy modyfikować własności powiązanych metryk na drzewach filogenetycznych, wpływając na takie cechy jak:

- zakres przyjmowanych wartości mający istotny wpływ na „rozdzielczość" metryki, np. maksymalna odległość w RF wynosi $|L| - 3$; zaś w dalszej części pracy zdefiniujemy metrykę, w której maksymalna wartość będzie wynosić $\Theta(|L^2|)$,

- sposób ilościowego określania podobieństwa klastrów lub rozbić, który np. w przypadku metryk RF i RFC jest bardzo uproszczony (binarny),

- sposób reakcji na modyfikacje topologii w zależności od ich umiejscowienia w drzewie, np. dla metryki RF wykonanie operacji uNNI zawsze oddala o 1, niezależnie od tego czy przemieszczane są tylko pojedyncze liście, czy duże poddrzewa.

Zauważmy również, że wprowadzenie koncepcji elementu dodatkowego pozwala na łatwe rozszerzanie definicji odległości na drzewa niebinarne.

W dalszej części pracy przebadany zostanie szczególny przypadek obu definicji ogólnych (3.2) i (3.5). Najbardziej naturalne wydaje się określenie funkcji h poprzez liczbę różnic w rozmieszczeniu elementów z L w porównywanych rozbiciach lub klastrach.

Zdefiniujemy teraz dwie metryki: po jednej dla drzew nieukorzenionych i ukorzenionych, których zalety i własności zostaną szczegółowo omówione

w dwóch następnych rozdziałach pracy. W myśl definicji określimy najpierw metrykę $h_{MS} : Splits(L)_O \times Splits(L)_O \to \mathbb{Z}_{\geq 0}$ mierzącą podobieństwo między rozbiciami jako

$$\begin{aligned} h_{MS}(A_1|B_1, A_2|B_2) &= \frac{1}{2}\min\{|A_1 \oplus A_2| + |B_1 \oplus B_2|, \\ &\qquad |A_1 \oplus B_2| + |B_1 \oplus A_2|\} \\ &= \min\{|A_1 \oplus A_2|, |A_1 \oplus B_2|\} \\ &= \min\{|A_1| + |A_2| - 2|A_1 \cap A_2|, \\ &\qquad |L| - (|A_1| + |A_2| - 2|A_1 \cap A_2|)\}, \\ h_{MS}(A|B, O) &= \min\{|A|, |B|\}. \end{aligned} \tag{3.8}$$

Wartość funkcji $h_{MS}(A|B, C|D)$ jest równa minimalnej liczbie operacji polegających na przeniesieniu pojedynczego liścia między zbiorami tworzącymi rozbicie, która wystarcza na przekształcenie rozbicia $A|B$ w $C|D$, np. $h_{MS}(abc|de, acd|be) = 2$, ponieważ wymagane są przynajmniej dwie takie operacje $abc|de \to ac|bde \to acd|be$.

Fakt 3.5. *Niech* $s_1, s_2 \in Splits(L)_O$ *oraz* $\min(O) = 0$*, wówczas:*

$$|\min(s_1) - \min(s_2)| \leq h_{MS}(s_1, s_2) \leq \min(s_1) + \min(s_2), \tag{3.9}$$

$$\max_{s_1, s_2 \in Splits(L)_O} h_{MS}(s_1, s_2) = \left\lfloor \frac{|L|}{2} \right\rfloor. \tag{3.10}$$

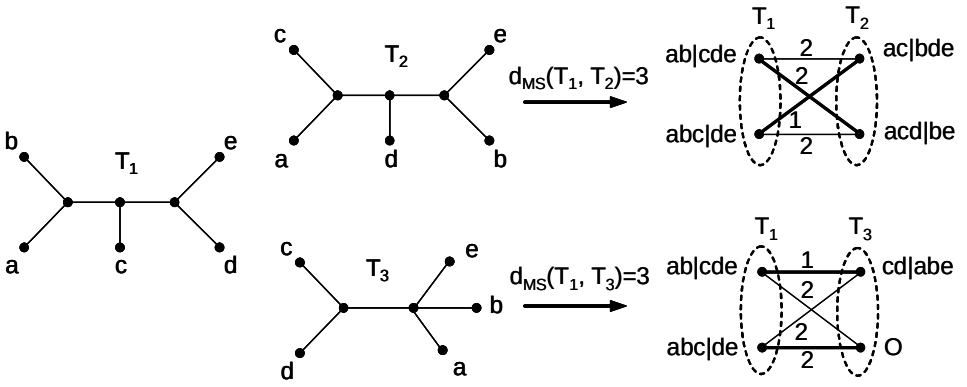
Lemat 3.6. *Funkcja* h_{MS} *jest metryką w zbiorze* $Splits(L)_O$*.*

Dowód. Zauważmy, że funkcję h_{MS} możemy otrzymać używając definicji 3.1 w następujący sposób. W celu uniknięcia niejednoznaczności element dodatkowy występujący w definicji funkcji h_{MS} na potrzeby tego dowodu oznaczymy jako $O_{h_{MS}}$. Każdemu elementowi $s \in Splits(L) \cup \{O_{h_{MS}}\}$ odpowiada partycja grafu G z definicji 3.1 posiadająca dwa wierzchołki v_1 i v_2 o następujących etykietach: $l(v_1) = A$, $l(v_2) = B$, jeśli $s = A|B$ jest rozbiciem L oraz $l(v_1) = \emptyset$, $l(v_2) = L$, jeśli $s = O_{h_{MS}}$. Metryka $h : 2^L \times 2^L \to \mathbb{R}_{\geq 0}$ jest zdefiniowana następująco: dla $A, B \subseteq L$,

$h(A,B) = \frac{1}{2}|A \oplus B|$. Definicja ta jest poprawna na mocy (3.8) i równości $h_{MS}(A|B, O_{h_{MS}}) = \frac{1}{2}\min\{|A \oplus \emptyset| + |B \oplus L|, |A \oplus L| + |B \oplus \emptyset|\}$. Zauważmy również, że w definicji funkcji h nie ma potrzeby uwzględniania elementu dodatkowego, gdyż w skonstruowanych grafach partycje są równoliczne i zawsze posiadają dokładnie dwa wierzchołki. Zatem na mocy lematu 3.2 funkcja h_{MS} jest metryką. □

Definicja 3.2 ([16, 18, 17, 19])**.** Niech $T_1, T_2 \in U_L$. Odległość *MS* (ang. *Matching Split distance*) pomiędzy drzewami T_1 i T_2 jest określona następująco:

$$d_{MS}(T_1, T_2) = d^T_{h_{MS}}(T_1, T_2) = d_{h_{MS}}(\beta_*(T_1), \beta_*(T_2)). \tag{3.11}$$



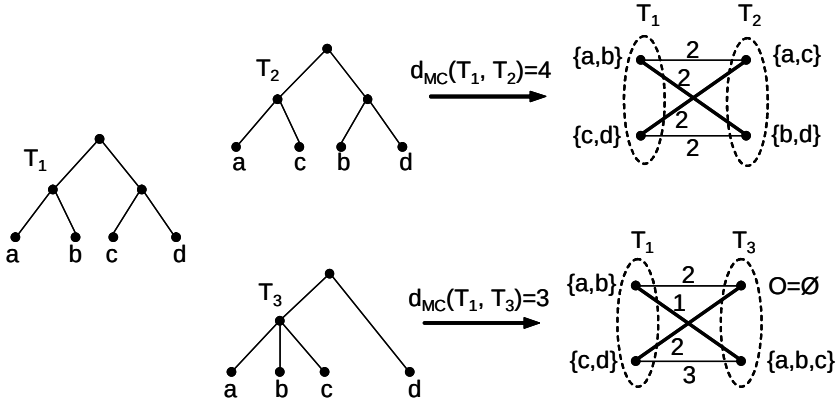
Rysunek 3.2: Przykład obliczania metryki MS.

W rozdziale 2 opisany został związek między drzewami ukorzenionymi a zbiorami klastrów związanych z ich wierzchołkami. Przy formułowaniu kolejnej metryki, tym razem dla drzew z R_L, naturalnym wydaje się wykorzystanie tej zależności. W celu określenia stopnia podobieństwa dwóch klastrów $A, B \subseteq L$ wykorzystamy ich różnicę symetryczną, czyli sumaryczną liczbę liści, które występują w jednym ze zbiorów, a nie występują w drugim. Ponieważ moc różnicy symetrycznej $A \oplus B$, dla $A, B \in 2^L$ jest

metryką w zbiorze 2^L, odległość skojarzeniową dla drzew ukorzenionych możemy zdefiniować następująco:

Definicja 3.3. Niech $T_1, T_2 \in R_L$, funkcja $h_{MC} : 2^L \times 2^L \to \mathbb{Z}_{\geq 0}$ będzie zdefiniowana jako $h_{MC}(A, B) = |A \oplus B|$ oraz $O = \emptyset$. Zgodnie z definicją 3.1 odległość *MC* (ang. *Matching Cluster distance*) $d_{MC} : R_L \times R_L \to \mathbb{Z}_{\geq 0}$ jest określona następującą zależnością:

$$d_{MC}(T_1, T_2) = d^T_{h_{MC}}(T_1, T_2) = d_{h_{MC}}(\sigma_*(T_1), \sigma_*(T_2)). \tag{3.12}$$



Rysunek 3.3: Przykład obliczania metryki MC.

3.3 Złożoność czasowa wyznaczania wartości MS i MC

Rozważmy graf dwudzielny $G = (V_1, V_2, E)$ służący do wyznaczania wartości odległości MS dla drzew $T_1, T_2 \in U_L$. Możemy przyjąć, że $|V_1| = |V_2| \leq |L| - 3$, więc $|E| \leq (|L| - 3)^2$. Na złożoność czasową wyznaczania wartości MS składa się zatem koszt związany z wyznaczeniem

wag krawędzi grafu G oraz koszt znalezienia wartości najlżejszego doskonałego skojarzenia w G. Pokażemy, że graf G można skonstruować w czasie $O(|L|^2)$. Najlżejsze doskonałe skojarzenie można w nim znaleźć w czasie $O(|E|\sqrt{|V|}\log(|V|\max_{e\in E} w(e))$ używając algorytmów opisanych w [52, 83]. Ponieważ maksymalna waga krawędzi w G jest $O(|L|)$, otrzymamy wówczas algorytm o złożoności $O(|L|^{2.5}\log|L|)$. W przypadku metryki MC sytuacja jest analogiczna.

Niech $T \in U_L$ będzie dowolnym drzewem różnym od gwiazdy oraz $T' \in R_L$ będzie drzewem powstałym z T w wyniku jego ukorzenienia w dowolnym wierzchołku wewnętrznym. Z każdym rozbiciem s w T możemy skojarzyć wzajemnie jednoznacznie klaster $c \neq L$ w T', taki że $s = c|L \setminus c$. Rozważmy teraz dwa rozbicia $s_1, s_2 \in Splits(L)$, takie że istnieją $c_1, c_2 \subsetneq L$, dla których zachodzi $s_1 = c_1|L \setminus c_1$, $s_2 = c_2|L \setminus c_2$. Otrzymujemy wówczas zależność:

$$\begin{aligned} h_{MS}(s_1, s_2) &= h_{MS}(c_1|L \setminus c_1, c_2|L \setminus c_2) \\ &= \min\{|c_1| + |c_2| - 2|c_1 \cap c_2|, |L| - (|c_1| + |c_2| - 2|c_1 \cap c_2|)\} \\ &= \min\{h_{MC}(c_1, c_2), |L| - h_{MC}(c_1, c_2)\}. \end{aligned} \tag{3.13}$$

W celu wyznaczenia wag krawędzi w grafie G wystarczy zatem skonstruować drzewa ukorzenione $T'_1, T'_2 \in R_L$, odpowiadające drzewom T_1, T_2, następnie wyznaczyć moce klastrów w zbiorach $\sigma(T'_1), \sigma(T'_2)$ oraz macierz I mocy przecięć klastrów z tych zbiorów. Obliczanie wartości h_{MC} sprowadza się do tego samego. Wyznaczenia mocy klastrów z $\sigma(T'_1), \sigma(T'_2)$ można dokonać w czasie $O(|L|)$ odwiedzając wierzchołki drzewa w porządku postorder (tj. w momencie odwiedzania danego wierzchołka, odwiedzone zostały już wszystkie wierzchołki w jego poddrzewach) i korzystając z uprzednio wyznaczonych wartości dla wszystkich dzieci danego wierzchołka. Wyznaczenia macierzy I można dokonać w czasie $O(|L|^2)$ wykorzystując algorytm 1 przedstawiony w [8]. Wyznaczenie mocy przecięcia dwóch jednoelementowych klastrów jest oczywiste. W przypadku większych klastrów $c_1 \in \sigma(T'_1)$, $c_2 \in \sigma(T'_2)$ korzystamy ze znanych już wartości mocy przecięć dla wierzchołków będących dziećmi c_1 lub c_2, które zostały wyznaczone w poprzednich krokach algorytmu. Wykorzystanie wcześniej

obliczonych wartości jest możliwe dzięki specjalnej kolejności odwiedzania wierzchołków drzew. Dla danego wierzchołka $v_1 \in V(T_1)$ wybranego z T_1 według porządku postorder odwiedzane są następnie kolejno wszystkie wierzchołki w T_2 również według tej kolejności. Zatem złożoność opisanego algorytmu obliczania metryk MS i MC wynosi $O(|L|^{2.5} \log |L|)$.

Algorytm 1 Wyznaczanie mocy przecięć pomiędzy klastrami z T_1 i T_2

Wejście: $T_1, T_2 \in R_L$
Wyjście: tablica I o wymiarach $k \times l$, $k = |\sigma(T_1)|$, $l = |\sigma(T_2)|$

$Q_1 \leftarrow$ wierzchołki drzewa T_1 w porządku postorder
$Q_2 \leftarrow$ wierzchołki drzewa T_2 w porządku postorder
for $i \leftarrow 1$ **to** $|Q_1|$ **do**
  $u \leftarrow Q_1[i]$
  for $j \leftarrow 1$ **to** $|Q_2|$ **do**
    $v \leftarrow Q_2[j]$
    if $|c(u)| = 1$ **and** $|c(v)| = 1$ **then**
      /* wierzchołki u i v są liśćmi */
      if $c(u) = c(v)$ **then**
        $I[i][j] \leftarrow 1$
      else
        $I[i][j] \leftarrow 0$
      end if
    else if $|c(u)| = 1$ **and** $|c(v)| > 1$ **then**
      /* wierzchołek u jest liściem, lecz v nie jest */
      /* $Ch(x)$ oznacza zbiór dzieci wierzchołka x */
      /* $Q_2^{ind}(x)$ zwraca indeks x w Q_2 */
      $I[i][j] \leftarrow \sum_{x \in Ch(v)} I[i][Q_2^{ind}(x)]$
    else /* wierzchołek u nie jest liściem */
      /* $Q_1^{ind}(x)$ zwraca indeks x w Q_1 */
      $I[i][j] \leftarrow \sum_{x \in Ch(u)} I[Q_1^{ind}(x)][j]$
    end if
  end for
end for

4 Struktura przestrzeni metrycznej MS

W niniejszym rozdziale gruntownie przebadamy przestrzenie metryczne drzew nieukorzenionych binarnych i niebinarnych. Zaprezentowane zostaną podstawowe własności metryki MS, m. in. rozmiar sąsiedztwa i oszacowanie średnicy przestrzeni. Szczegółowo przeanalizowany zostanie również związek odległości MS z metryką RF.

4.1 Podstawowe własności odległości MS

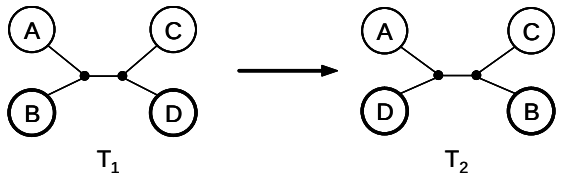
Przyjrzyjmy się teraz podstawowym własnościom MS przyjmując za punkt odniesienia metrykę RF.

Lemat 4.1 ([17, 19])**.** *Niech $T_1, T_2 \in U_L^B$, gdzie $|L| = n$, wówczas jeśli $d_{RF}(T_1, T_2) = 1$, to*

$$2 \leq d_{MS}(T_1, T_2) \leq \left\lfloor \frac{n}{2} \right\rfloor$$

oraz oba ograniczenia w pewnych przypadkach są osiągalne.

Dowód. Zauważmy, że $d_{RF}(T_1, T_2) = 1$ wtedy i tylko wtedy, gdy drzewa T_1 i T_2 dzieli dokładnie jedna operacja uNNI [28]. Drzewa te zostały schematycznie przedstawione na rysunku 4.1. Mamy zatem $d_{MS}(T_1, T_2) = \min\{|B|+|D|, |A|+|C|\}$. Łatwo zauważyć, że maksymalna wartość odległości MS pomiędzy T_1 i T_2 wynosi $\left\lfloor \frac{n}{2} \right\rfloor$, minimalna zaś 2 (gdy $|B| = |D| = 1$ lub $|A| = |C| = 1$). □



RYSUNEK 4.1: Drzewo T_2 powstaje z drzewa T_1 w wyniku pojedynczej operacji uNNI przestawiającej poddrzewa B i D. Okręgi reprezentują poddrzewa binarne nad zbiorami liści A, B, C i D.

Twierdzenie 4.2 ([19]). *Niech $T_1, T_2 \in U_L$, gdzie $|L| = n$, wówczas zachodzi następująca nierówność:*

$$d_{RF}(T_1, T_2) \leq d_{MS}(T_1, T_2) \leq n\ d_{RF}(T_1, T_2). \tag{4.1}$$

Dowód. Pełne grafy dwudzielne występujące w definicji 3.1 i pojawiające się przy obliczaniu wartości $d_{RF}(T_1, T_2)$ oraz $d_{MS}(T_1, T_2)$ różnią się wyłącznie wagami na krawędziach (por. wniosek 3.3). Wagi te są wyznaczone przez odpowiednie metryki na $Splits(L) \cup \{O\}$, które spełniają następujące zależności: $h_{RF}(s_1, s_2) \leq h_{MS}(s_1, s_2) \leq \left\lfloor \frac{n}{2} \right\rfloor h_{RF}(s_1, s_2) \leq n\ h_{RF}(s_1, s_2)$ dla $s_1, s_2 \in Splits(L)$ oraz $\frac{1}{2} = h_{RF}(s, O) \leq h_{MS}(s, O) \leq \left\lfloor \frac{n}{2} \right\rfloor \leq n\ h_{RF}(s, O)$ dla $s \in Splits(L)$. Analogiczna relacja zachodzi więc również pomiędzy odpowiednimi wagami krawędzi, zatem jest ona prawdziwa także dla wag najlżejszych doskonałych skojarzeń w rozważanych grafach. □

Dla drzew binarnych można podać dokładniejsze oszacowanie.

Twierdzenie 4.3 ([19]). *Niech $T_1 \neq T_2 \in U_L^B$, gdzie $|L| = n$, wówczas:*

$$d_{RF}(T_1, T_2) + 1 \leq d_{MS}(T_1, T_2) \leq \left\lfloor \frac{n}{2} \right\rfloor d_{RF}(T_1, T_2). \tag{4.2}$$

Dowód. Używając metody zastosowanej w dowodzie twierdzenia 4.2, z tą tylko różnicą, że dla drzew binarnych T_1, T_2 nie musimy rozważać odległości do elementu O, mamy $h_{MS}(s_1, s_2) \leq \left\lfloor \frac{n}{2} \right\rfloor h_{RF}(s_1, s_2)$ dla $s_1, s_2 \in$

$Splits(L)$. Zatem ta sama relacja występuje dla odpowiednich odległości między drzewami T_1 i T_2.

W celu pokazania poprawności ograniczenia dolnego przeprowadzimy dowód nie wprost. Załóżmy przeciwnie, że zachodzi sytuacja, w której

$$d_{MS}(T_1, T_2) = d_{RF}(T_1, T_2) > 0 \tag{4.3}$$

dla $T_1, T_2 \in U_n^B$ oraz $n > 4$ jest najmniejsze możliwe. Jeśli istnieją dwa liście (gatunki) x, y, które wraz z wierzchołkiem będącym ich wspólnym sąsiadem tworzą tzw. *wiśnię* w obu drzewach (czyli takie, że $xy|L \setminus xy \in \beta(T_1) \cap \beta(T_2)$), to zastępujemy tę wiśnię nowym liściem $z \notin L$ zarówno w T_1, jak i w T_2. Zauważmy, że operacja ta nie zmienia wartości RF oraz nie zwiększa odległości MS (waga krawędzi w grafie z definicji 3.1 może się tylko zmniejszyć lub pozostać bez zmian), stąd (4.3) jest nadal prawdzie, lecz rozmiar drzew zmalał do $n-1$ liści. Ponieważ n było minimalne, to nie może istnieć rozbicie s o $\min(s) = 2$ w zbiorze $\beta(T_1) \cap \beta(T_2)$. Z (4.3) wynika również, iż istnieje takie sparowanie wszystkich rozbić z $\beta(T_1) \setminus \beta(T_2)$ z rozbiciami z $\beta(T_2) \setminus \beta(T_1)$, że dla każdej pary odległość h_{MS} wynosi 1. Niech zatem $s_1 = ab|L \setminus ab \in \beta(T_1)$ będzie sparowane z pewnym rozbiciem $t_1 \in \beta(T_2)$, gdzie $t_1 = abd|L \setminus abd \notin \beta(T_1)$. Stąd jedno z rozbić $t_2 = ad|L \setminus ad$ lub $t_2 = bd|L \setminus bd$ musi należeć do $\beta(T_2)$. Rozważmy pierwszy przypadek, drugi jest analogiczny. Załóżmy, że rozbicie $t_2 = ad|L \setminus ad$ jest sparowane z $s_2 = acd|L \setminus acd \in \beta(T_1)$, $c \neq b$. W ten sposób otrzymujemy sprzeczność (taka sytuacja nie jest możliwa), ponieważ s_2 nie jest kompatybilne z s_1 (por. def. 2.4 oraz tw. 2.1). □

Przykładem osiągania równości z dolnego ograniczenia są dwie gąsienice $T_1, T_2 \in U_n^B$, takie że T_2 powstaje z T_1 w wyniku operacji przeniesienia wybranego liścia z jednego końca drzewa na drugi (por. rysunek 4.8). Górne ograniczenie jest spełnione dla drzew przedstawionych na rysunku 4.1, jeśli np. $|A| = |B| = |C| = |D| = \frac{|L|}{4}$.

Konsekwencją twierdzenia 4.3 jest następujący fakt.

Wniosek 4.4. *Niech $T_1, T_2 \in U_L^B$. Jeśli $d_{MS}(T_1, T_2) = 2$, wówczas drzewa T_1 i T_2 różnią się dokładnie jednym rozbiciem, czyli $d_{uNNI}(T_1, T_2) = d_{RF}(T_1, T_2) = 1$.*

4.2 Rozmiar sąsiedztwa

Rozmiar sąsiedztwa może służyć jako pewna miara regularności przestrzeni. W pierwszej kolejności przedstawimy rozważania dotyczące przestrzeni drzew binarnych.

Zgodnie z definicją drzewa $T_1, T_2 \in U_L^B$ są sąsiadami w metryce RF (ściślej, w przestrzeni z metryką RF), jeśli $d_{RF}(T_1, T_2) = 1$, natomiast są sąsiadami w metryce MS, jeśli $d_{MS}(T_1, T_2) = 2$.

Własności sąsiedztwa w przypadku metryk opartych na operacjach edycyjnych podsumowuje poniższe twierdzenie.

Twierdzenie 4.5 ([94, 6])**.** *Rozmiar sąsiedztwa drzewa $T \in U_L^B$, $|L| = n$ wynosi:*

1. $2n - 6$ *dla metryki uNNI* [94]*,*
2. $2(n-3)(2n-7)$ *dla metryki uSPR* [6]*,*
3. *co najwyżej* $(2n-3)(n-3)^2$ *zależnie od topologii T dla uTBR* [6]*.*

Dla sąsiedztwa w metryce uTBR dokładniejsze oszacowanie zostało niedawno (2010 rok) podane w pracy [63].

Twierdzenie 4.6 ([63])**.** *Rozmiar sąsiedztwa drzewa $T \in U_L^B$, $|L| = n$ w metryce uTBR spełnia warunki:*

$$2n^2 - 8n + 2 + 2\sum_{i=1}^{n-4} l(i) \leq |N_{uTBR}(T)| \leq \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2,$$

gdzie $l(1) = 0$ oraz $l(i) = 2 + 4\sum_{j=2}^{i-1} \lfloor \log_2 j \rfloor$.

W przypadku metryki RF każde drzewo $T \in U_L^B$ podobnie jak dla uNNI posiada dokładnie $2n - 6$ sąsiadów. Wynika to z faktu, że drzewo sąsiednie w RF powstaje wskutek operacji uNNI, a dla każdej krawędzi wewnętrznej w T możliwe są dwie takie operacje. Niestety w literaturze trudno jest odnaleźć analogiczne informacje odnośnie pozostałych metryk wielomianowych, tj. PD, ND i QT.

Przechodząc do rozważań dotyczących sąsiedztwa w metryce MS zauważmy, że na podstawie wniosku 4.4 oraz dowodu lematu 4.1 łatwo opisać operację edycyjną, w wyniku której powstają drzewa sąsiednie w U_L^B. Operacja ta jest specyficznym typem transformacji uNNI, w której przemieszczane poddrzewa są pojedynczymi liśćmi.

Twierdzenie 4.7 ([19])**.** *Rozmiar sąsiedztwa drzewa $T \in U_L^B$, $|L| = n$ w metryce MS wynosi co najwyżej $n-1$ i jest zależny od topologii T.*

Dowód. Łatwo zweryfikować poprawność twierdzenia dla $n = 4, 5, 6$. Dla $n \geq 7$ rozważmy cztery możliwe sposoby rozmieszczenia liści względem krawędzi wewnętrznej przedstawione na rysunku 4.2. Dla krawędzi wewnętrznej e znajdującej się w konfiguracji a) lub b) nie można wykonać operacji pozwalającej na utworzenie drzewa sąsiedniego. W sytuacji c) można wykonać jedną taką operację, natomiast w konfiguracji d) dwie. Niech m_i, $i = 1, 2, 3, 4$ odpowiada ilości krawędzi w $T \in U_n^B$ znajdujących się odpowiednio w konfiguracjach a), b), c) i d). Mamy zatem

$$|N_{MS}(T)| = m_3 + 2m_4. \tag{4.4}$$

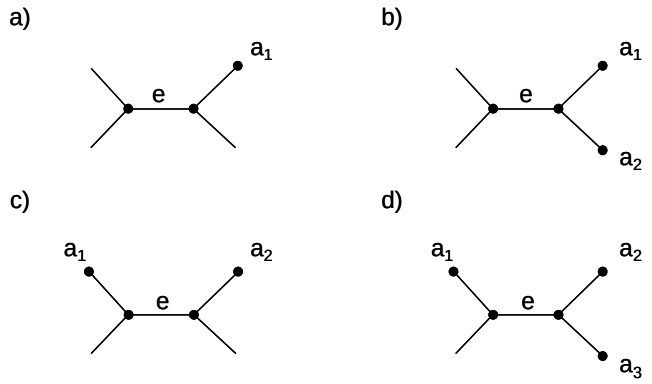
Zliczając liście w T w następujący sposób: liście a_1, a_2 w konfiguracji b) oraz a_2, a_3 w sytuacji d) liczone są pojedynczo, gdyż nie są incydentne z żadną inną krawędzią wewnętrzną w T, natomiast pozostałe liście uwzględniane są z wagą $\frac{1}{2}$, ponieważ sąsiadują one z dokładnie dwiema krawędziami wewnętrznymi, otrzymujemy

$$\frac{1}{2}m_1 + 2m_2 + m_3 + \frac{5}{2}m_4 = n. \tag{4.5}$$

Wyznaczając z (4.5) m_4 i podstawiając do (4.4) dostajemy $|N_{MS}(T)| = \frac{1}{5}(4n + m_3) - \frac{1}{5}(2m_1 + 8m_2)$. Ponieważ $m_3 \leq n - 5$, stąd ostatecznie $|N_{MS}(T)| \leq n - 1$.

Drzewo o maksymalnej liczbie sąsiadów powinno zatem zawierać możliwie dużo liści w konfiguracji c). Takim drzewem jest gąsienica $T_1 \in U_n^B$ przedstawiona na rysunku 4.3, gdzie mamy możliwość wykonania jednej operacji tworzącej drzewo sąsiednie dla każdej z krawędzi $e_2, \ldots, e_{n-4}$ oraz

dwóch takich operacji dla krawędzi e_1 i e_{n-3}, stąd T_1 posiada $n-1$ sąsiadów w U_n^B. Z drugiej strony zauważmy, że istnieją drzewa nieposiadające sąsiadów w MS, np. 8-listne drzewo T_2 przedstawione na rysunku 4.3. Drzewa o tej własności mogą być skonstruowane w podobny sposób dla dowolnej liczby liści większej od 8 oraz dla $n = 6$. □

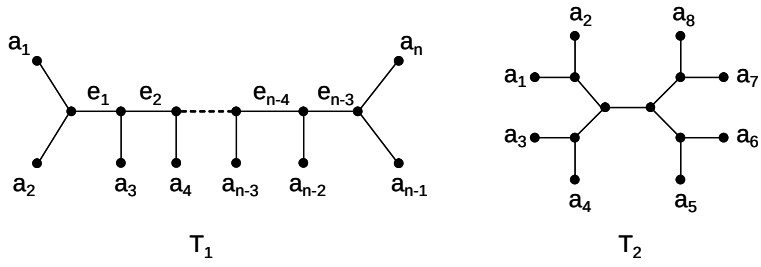


Rysunek 4.2: Możliwe sposoby rozmieszczenia liści względem krawędzi wewnętrznej w drzewach ze zbioru U_n^B, gdzie $n \geq 7$.

W dalszej części pracy zostanie wykazane, że pomimo faktu, iż istnieją w MS drzewa bez sąsiadów, przestrzeń ta nie posiada *izolowanych wysp*, czyli obszarów znacznie oddalonych od pozostałych punktów. Dokładna analiza tej własności wraz z formalną definicją jest umieszczona w podrozdziale 4.5.

Twierdzenie 4.7 ilustruje jedną z kolejnych różnic między metrykami MS i RF. W przypadku odległości RF bowiem, rozmiar sąsiedztwa nie zależy od topologii drzewa.

Rozważymy teraz parametry analogiczne do opisanych powyżej dla zbioru wszystkich drzew U_L, a nie tylko binarnych. Jedna z podstawowych różnic w stosunku do sytuacji w zbiorze U_L^B przejawia się w tym, że drzewa T_1 i T_2 są sąsiadami w U_L, gdy $d_{MS}(T_1, T_2) = 1$. Podstawowe własności dotyczące sąsiedztwa w zbiorze U_L zostały podsumowane w poniższym wniosku.



RYSUNEK 4.3: Gąsienica T_1 ma najwięcej sąsiadów w metryce MS. Drzewo T_2 jest przykładem drzewa bez sąsiadów w MS.

Wniosek 4.8.

1. *Jeśli* $d_{MS}(T_1, T_2) = 1$*, wówczas oba drzewa nie są binarne oraz jedno z nich powstaje z drugiego w wyniku odczepienia liścia połączonego z pewnym wierzchołkiem wewnętrznym* v *i przyłączeniu go do wierzchołka będącego sąsiadem* v*. Przykład takiej operacji został przedstawiony na rysunku* 4.4*.*

2. *Jeśli* $T_1 \in U_L^B$*, wówczas nie istnieje* $T_2 \in U_L$*, takie że* $d_{MS}(T_1, T_2) = 1$ *(tj. drzewa binarne nie mają sąsiadów w* U_L*).*

3. *Niech* $T_1 \in U_L$*. Liczba drzew* $T_2 \in U_L$*, dla których zachodzi równość* $d_{MS}(T_1, T_2) = 1$ *jest* $O(|L|^2)$*.*

4. *Równość* $d_{RF}(T_1, T_2) = d_{MS}(T_1, T_2)$ *może zachodzić dla dowolnie dużych wartości* $d_{RF}(T_1, T_2)$*.*

Dowód.

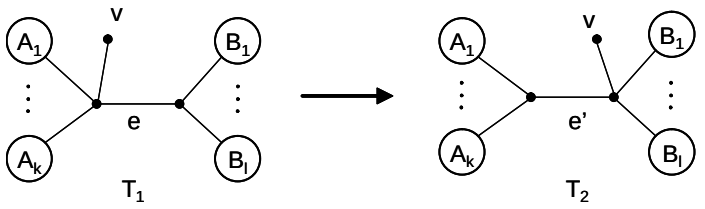
1-2. Zauważmy, że dla drzew $T_1, T_2 \in U_L$ równość $d_{MS}(T_1, T_2) = 1$ może zachodzić wyłącznie gdy $|\beta_*(T_1)| = |\beta_*(T_2)|$, w przeciwnym bowiem razie w wyrażeniu opisującym wartość odległości MS miedzy tymi drzewami występowałby składnik $h_{MS}(O, s) \geq 2$ dla pewnego nietrywialnego rozbicia $s \in \beta_*(T_1) \cup \beta_*(T_2)$. Ponieważ, na podstawie

twierdzenia 4.3, minimalna odległość różnych drzew binarnych wynosi 2 oraz z uwagi na wymaganą równoliczność zbiorów nietrywialnych klastrów, żadne z drzew nie może być drzewem binarnym (stąd otrzymujemy punkt 2). Drzewa te zatem muszą różnić się dokładnie jedną parą rozbić $s_1 \in \beta_*(T_1)$, $s_2 \in \beta_*(T_2)$, $s_1 \neq s_2$, a rozbicia s_1,s_2 z kolei muszą się różnić położeniem dokładnie jednego elementu z L.

3. Zauważmy, że liczba możliwych operacji opisanych w punkcie pierwszym dla danego drzewa $T_1 \in U_L$ jest nie większa niż liczba par (x, u), gdzie $x \in L$ oraz u jest sąsiadem sąsiada liścia x, czyli $O(|L|^2)$. Przykładem sytuacji, gdy moc sąsiedztwa jest $\Omega(|L|^2)$, jest rodzina drzew T_3 na rysunku 4.5.

4. Usunięcie każdego z następujących liści $a_1, \ldots, a_k$ drzewa T_3 przedstawionego na rysunku 4.5 i przyłączenie go do odpowiedniego wierzchołka wewnętrznego $v_1, \ldots, v_k$ powoduje utworzenie drzewa T_3' będącego w odległości $d_{RF}(T_3, T_3') = d_{MS}(T_3, T_3') = k$.

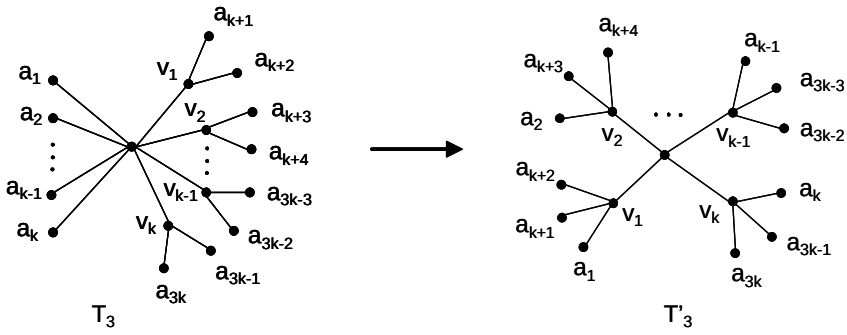
□



Rysunek 4.4: Niebinarne drzewa T_1 i T_2 są sąsiadami w U_L z metryką MS.

4.3 Lokalne modyfikacje drzewa

Rozważane w niniejszym podrozdziale przekształcenia polegają na wprowadzeniu niewielkich zmian w porównywanych drzewach, tzn. na utworze-



RYSUNEK 4.5: Dla $3k$-listnego drzewa T_3 istnieje k^2 drzew oddalonych od T_3 o 1 (w metryce MS). Odległość MS między drzewami T_3 i T_3' wynosi k.

niu lub ściągnięciu pojedynczej krawędzi oraz dołączeniu nowego wierzchołka. Ponieważ analizowane modyfikacje dotyczą pojedynczych elementów, nie powinny znacząco wpływać na zmianę odległości powstałych w ten sposób drzew.

Twierdzenie 4.9 ([19])**.** *Dane są dwa drzewa $T, T_e \in U_L$, $|L| = n$, takie że T_e powstaje z T w wyniku ściągnięcia krawędzi wewnętrznej e odpowiadającej rozbiciu $A|B$, tj. $\beta(T_e) = \beta(T_1) \setminus \{A|B\}$, wówczas:*

$$d_{MS}(T, T_e) = h_{MS}(A|B, O) = \min(A|B) \leq \left\lfloor \frac{n}{2} \right\rfloor .$$

Dowód. Podana nierówność wynika bezpośrednio z lematu 3.1 oraz faktu 3.5, gdyż drzewa T i T_e różnią się wyłącznie jednym rozbiciem związanym z krawędzią e. □

W dalszej części podrozdziału rozważymy wpływ dołączenia nowego liścia do porównywanych drzew na zmianę ich odległości w metryce MS.

Twierdzenie 4.10 ([19])**.** *Niech $T_1, T_2 \in U_L$, gdzie $|L| = n \geq 5$ oraz zbiór $A \subsetneq L$ będzie o jeden element mniejszy od L, czyli $|A| = n - 1$. Zachodzą*

wówczas poniższe zależności:

$$\begin{aligned} d_{MS}(T_1,T_2) &\leq d_{MS}(T_{1|A},T_{2|A}) + n - 4 + \left\lfloor \frac{n}{2} \right\rfloor, \\ d_{MS}(T_1,T_2) &\geq d_{MS}(T_{1|A},T_{2|A}) - \left\lfloor \frac{n-1}{2} \right\rfloor. \end{aligned}$$

Dowód. Niech $L = A \cup \{x\}$. Rozbicia $s \in Splits(A)$ oraz $s' \in Splits(L)$ nazywamy *odpowiadającymi* (wówczas s *odpowiada* s' oraz s' *odpowiada* s), jeśli rozbicie $s = C|D \in Splits(A)$ jest nietrywialne oraz zachodzi jedna z następujących równości: $s' = C \cup \{x\}|D$ lub $s' = C|D \cup \{x\}$. Zauważmy, że w przypadku dołączenia krawędzi wiszącej z nowym liściem do środka istniejącej krawędzi w $T_{1|A}$ powstaje *nowe* rozbicie $s'_{new} \in \beta_*(T_1)$. W zależności od sposobu dołączenia liścia x możliwe są dwie sytuacje. Liść x jest dołączony do krawędzi wiszącej związanej z liściem y. Powstaje wówczas rozbicie $xy|L \setminus xy$, które nie posiada rozbicia odpowiadającego w $\beta_*(T_{1|A})$. W drugim przypadku liść x jest dołączony do krawędzi wewnętrznej związanej z rozbiciem $B|C \in \beta_*(T_1)$, wówczas dla $B|C$ pojawią się dwa rozbicia odpowiadające $B \cup \{x\}|C, B|C \cup \{x\} \in \beta(T_1)$. Uwzględniając obie te sytuacje zdefiniujmy formalnie pojęcie *nowego* rozbicia $s'_{new} \in \beta_*(T_i)$, gdzie $i = 1, 2$, jako rozbicia wybranego tak, że dla każdego $s \in \beta_*(T_{i|A})$ istnieje dokładnie jedno rozbicie odpowiadające w zbiorze $\beta_*(T_i) \setminus \{s'_i\}$. Przyjmijmy również, że dodatkowy element O *odpowiada* sobie samemu. Jeśli elementy $s', t' \in Splits(L) \cup \{O\}$ odpowiadają elementom $s, t \in Splits(A) \cup \{O\}$, to $h_{MS}(s,t) \leq h_{MS}(s',t') \leq h_{MS}(s,t) + 1$.

Rozważmy następujące sparowanie $P = \{(s_i, t_i) : i = 1, \ldots, k\}$, gdzie $s_i \in \beta_*(T_{1|A}) \cup \{O\}$, $t_i \in \beta_*(T_{2|A}) \cup \{O\}$, $k = \max_{i=1,2} |\beta_*(T_{i|A})| \leq n-4$, powstałe w wyniku wyznaczenia wartości $d_{MS}(T_{1|A}, T_{2|A}) = \sum_{i=1}^{k} h_{MS}(s_i, t_i)$. Konstruujemy sparowanie składające się z par elementów $s'_i \in \beta_*(T_1) \cup \{O\}$ i $t'_i \in \beta_*(T_2) \cup \{O\}$ odpowiadających elementom sparowanym w P oraz (jeśli to konieczne) pary rozbić (s_{new}, t_{new}), jeśli oba zbiory $\beta_*(T_1)$ i $\beta_*(T_2)$ zawierają nowe rozbicia lub z pary (s_{new}, O) ewentualnie (O, t_{new}), jeśli dokładnie jeden ze zbiorów $\beta_*(T_1)$, $\beta_*(T_2)$ zawiera nowe rozbicie. Ponieważ koszt skonstruowanego sparowania wyznacza górne ograniczenie dla

odległości $d_{MS}(T_1, T_2)$, otrzymujemy stąd $d_{MS}(T_1, T_2) \leq d_{MS}(T_{1|A}, T_{2|A}) + n - 4 + \left\lfloor \frac{n}{2} \right\rfloor$.

Rozważmy sparowanie M elementów $\beta_*(T_1) \cup \{O\}$ z elementami $\beta_*(T_2) \cup \{O\}$ analogiczne do najlżejszego doskonałego skojarzenia definiującego odległość MS między drzewami T_1 i T_2. Przekształcamy M w sparowanie M' pomiędzy elementami zbiorów $\beta_*(T_{1|A}) \cup \{O\}$ i $\beta_*(T_{2|A}) \cup \{O\}$, zastępując elementy we wszystkich parach, w których nie występują nowe rozbicia, elementami im odpowiadającymi (pary zawierające nowe rozbicia są pomijane przy transformacji, stąd nie występują one w M'). Możliwe są następujące cztery sytuacje.

Przypadek 1. Istnieją dwa nowe rozbicia $s_{new} \in \beta_*(T_1)$, $t_{new} \in \beta_*(T_2)$, które tworzą parę $(s_{new}, t_{new}) \in M$. Wówczas zachodzi $d_{MS}(T_1, T_2) = \sum_{(s,t)\in M} h_{MS}(s,t) \geq \sum_{(s,t)\in M'} h_{MS}(s,t) \geq d_{MS}(T_{1|A}, T_{2|A})$.

Przypadek 2. Nowe rozbicia nie występują ani w $\beta_*(T_1)$, ani $\beta_*(T_2)$, wówczas $d_{MS}(T_1, T_2) \geq d_{MS}(T_{1|A}, T_{2|A})$ podobnie jak w powyższym przypadku.

Przypadek 3. Występują dwa nowe rozbicia $s_{new} \in \sigma_*(T_1)$, $t_{new} \in \sigma_*(T_2)$, lecz nie tworzą one pary w M, tj. $(s_{new}, x_2), (x_1, t_{new}) \in M$. Wówczas rozszerzamy M' o parę (x'_1, x'_2), gdzie x'_i jest elementem odpowiadającym x_i, stąd $d_{MS}(T_1, T_2) = \sum_{(s,t)\in M} h_{MS}(s,t) \geq \sum_{(s,t)\in M'} h_{MS}(s,t) - h_{MS}(x'_1, x'_2) \geq d_{MS}(T_{1|A}, T_{2|A}) - \left\lfloor \frac{n-1}{2} \right\rfloor$.

Przypadek 4. Istnieje dokładnie jedno nowe rozbicie. Bez straty ogólności możemy założyć, że występuje ono w drzewie T_1, tj. $s_{new} \in \sigma_*(T_1)$ oraz $(s_{new}, x) \in M$. Wówczas rozszerzamy M' o parę (O, x'), gdzie x' jest elementem odpowiadającym x, stąd $d_{MS}(T_1, T_2) = \sum_{(s,t)\in M} h_{MS}(s,t) \geq \sum_{(s,t)\in M'} h_{MS}(s,t) - h_{MS}(x', O) \geq d_{MS}(T_{1|A}, T_{2|A}) - \left\lfloor \frac{n-1}{2} \right\rfloor$. □

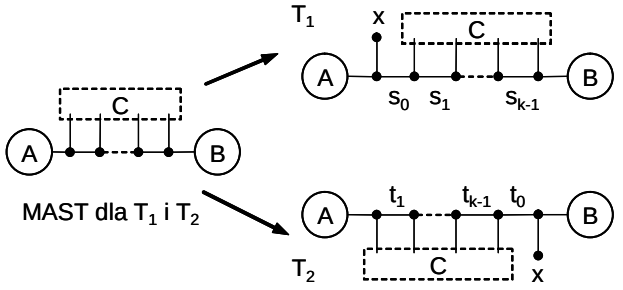
W przypadku gdy oba drzewa T_1 oraz T_2 powstają ze wspólnego drzewa binarnego $T_{1|A} = T_{2|A}$ w wyniku operacji dołączenia liścia, górne ograniczenie może być poprawione.

Twierdzenie 4.11 ([19]). *Jeśli* $T_1, T_2 \in U_L^B$, $|L| = n$ *i* $MAST(T_1, T_2) = n - 1$, *wówczas*

$$d_{MS}(T_1, T_2) \leq n - 2 \tag{4.6}$$

oraz wartość ograniczenia $n-2$ może być osiągalna.

Dowód. Rozważmy drzewa $T_1, T_2 \in U_L^B$ przedstawione schematycznie na rysunku 4.6 i różniące się wyłącznie umiejscowieniem jednego liścia x. Zauważmy, że drzewa te różnią się wyłącznie rozbiciami s_i, t_i dla $i = 0, \ldots, k-1$. Ponadto mamy $h_{MS}(s_i, t_i) \leq 1$ dla $i \neq 0$ oraz $h_{MS}(s_0, t_0) = \min\{|C|+1, n-|C|-1\}$. Zatem $d_{MS}(T_1, T_2) \leq \sum_{i=1}^{k-1} h_S(s_i, t_i) + h_S(s_0, t_0) \leq n-2+k-|C| \leq n-2$. Przykładem osiągania przez (4.6) wartości $n-2$ są gąsienice przedstawione na rysunku 4.8. □



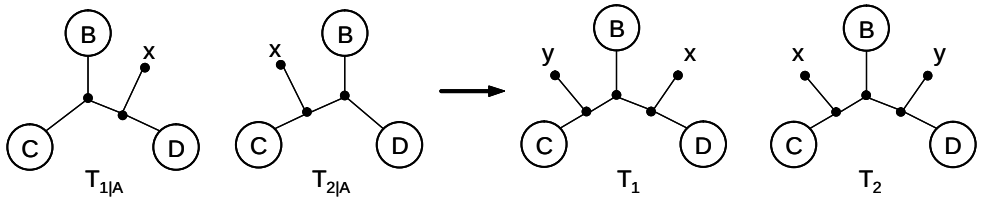
Rysunek 4.6: Okręgi reprezentują dowolne binarne poddrzewa nad zbiorami liści A i B. Zbiór C odpowiada zbiorowi liści w rodzinie poddrzew znajdujących się w wyróżnionym obszarze.

Można postawić hipotezę, że górne ograniczenie z tw. 4.10 dla drzew binarnych może być poprawione do następującej postaci:

$$d_{MS}(T_1, T_2) \leq d_{MS}(T_{1|A}, T_{2|A}) + n - 2.$$

Zauważmy, że dołączenie nowego liścia do drzew może zmniejszyć ich odległość w metryce MS. Przykład takiej sytuacji jest przedstawiony na rysunku 4.7. Drzewa $T_{1|A}$ oraz $T_{2|A}$ różnią się dokładnie jednym rozbiciem, zatem łatwo możemy wyznaczyć ich odległość $d_{MS}(T_{1|A}, T_{2|A}) = h_{MS}(Dx|CB, Cx|BD) = \min\{|C|+|D|, |B|+1\}$. Dystans drzew T_1 i T_2 natomiast możemy oszacować jako $d_{MS}(T_1, T_2) \leq h_{MS}(Cy|BDx, Cx|BDy) + h_{MS}(Dx|BCy, Dy|BCx) \leq 4$. Jeśli zatem przyjmiemy, że $|B| = 8$, $|C| = |D| = 3$, to otrzymamy zmniejszenie odległości z 6 do 4.

Jest to kolejna własność, która odróżnia metrykę MS od RF, gdyż przypadku tej ostatniej mamy zawsze $d_{RF}(T_{1|A}, T_{2|A}) \leq d_{RF}(T_1, T_2)$. Zauważmy jednak, że możliwość zmniejszenia odległości występuje również dla innych metryk, czego przykładem może być odległość węzłowa (ND) lub ścieżkowa (PD). Dla drzew na rysunku 4.7 mamy: $d_{ND}(T_{1|A}, T_{2|A}) = (|B|+1)(|C|+|D|)$, $d_{PD}(T_{1|A}, T_{2|A}) = \sqrt{(|B|+1)(|C|+|D|)}$, $d_{ND}(T_1, T_2) = 4(|C|+|D|)$, $d_{PD}(T_1, T_2) = \sqrt{8(|C|+|D|)}$. W przypadku rozmiarów poddrzew określonych tak samo jak powyżej otrzymujemy: $d_{ND}(T_{1|A}, T_{2|A}) = 54 > 24 = d_{ND}(T_1, T_2)$ oraz $d_{PD}(T_{1|A}, T_{2|A}) = \sqrt{54} > \sqrt{48} = d_{PD}(T_1, T_2)$.

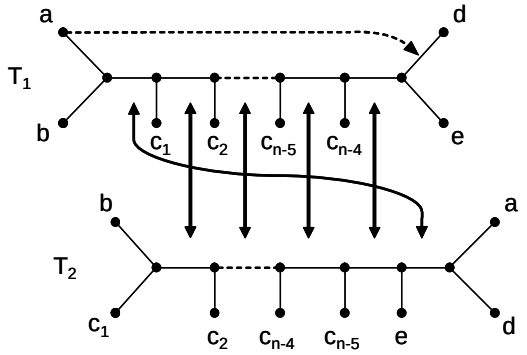


Rysunek 4.7: Przykład konstrukcji drzew, dla których dodanie liścia powoduje zmniejszenie odległości MS. Okręgi reprezentują poddrzewa o zbiorach liści B, C i D.

Oszacowania przedstawione w twierdzeniach 4.9-4.11 ilustrują kluczową cechę metryki MS. Zauważmy, że dla dowolnej stałej $k \in \mathbb{Z}_{>0}$, używając k razy tw. 4.9 lub tw. 4.10, otrzymujemy, że przemieszczenie dowolnej ustalonej liczby k liści lub ściągnięcie i wprowadzenie k krawędzi może zmienić odległość MS o wartość rzędu $O(n)$ i jest to asymptotycznie niewiele w porównaniu do maksymalnej odległości w metryce MS osiąganej w zbiorze U_n, która jest rzędu $\Theta(n^2)$ (dowód tego faktu jest przedstawiony w podrozdziale 4.4). Zatem w przypadku porównywania dużych drzew filogenetycznych przemieszczanie ograniczonej (niewielkiej w stosunku do rozmiaru drzewa) liczby liści nie powoduje istotnych zmian wartości MS.

Własność ta stanowi jedną z podstawowych zalet MS w odniesieniu do metryki RF. W celu zilustrowania tego faktu rozważmy drzewa przedstawione na rysunku 4.8. Drzewo T_2 powstaje z T_1 w wyniku usunięcia liścia a i dołączenia go do krawędzi związanej z liściem d. Jest to znany przykład (pojawia się np. w [85]) prezentujący nieintuicyjne zachowanie się odległo-

ści RF. W wyniku opisanej operacji, która zmienia pozycję tylko jednego liścia, pozostawiając względne relacje pokrewieństwa dla znacznej większości gatunków bez zmian, otrzymujemy bowiem dwa drzewa maksymalnie różne od siebie, tj. $d_{RF}(T_1, T_2) = n - 3$ (metryka RF osiąga maksymalną możliwą wartość dla drzew z U_n^B). Natomiast w przypadku metryki MS mamy $d_{MS}(T_1, T_2) = n - 2$, co jest asymptotycznie dużo mniejszą wartością w porównaniu do maksymalnej odległości wynoszącej $\Theta(n^2)$. Już np. dla $n = 8$ mamy $d_{MS}(T_1, T_2) = 6$, podczas gdy maksymalna wartość odległości MS wynosi 16.



Rysunek 4.8: Przykład obrazujący jedną z podstawowych zalet metryki MS w stosunku do RF. Strzałki narysowane linią ciągłą wskazują sparowanie rozbić wyznaczające wartość odległości MS.

Zbadajmy teraz jak na zmianę odległości w MS wpływają podstawowe operacje edycyjne. Niech $T, T' \in U_n^B$. Jeśli $d_{uNNI}(T, T') = 1$, wówczas na podstawie lematu 4.1 mamy, że $d_{MS}(T, T') \leq \left\lfloor \frac{n}{2} \right\rfloor$. Pozostałe dwie operacje mogą spowodować większą zmianę.

Fakt 4.12. *Istnieją drzewa $T, T' \in U_L^B$, $|L| = n$, takie że:*

1. $d_{uSPR}(T, T') = 1$ *oraz* $d_{MS}(T, T') = \Theta(n^2)$,

2. $d_{uTBR}(T, T') = 1$ *oraz* $d_{MS}(T, T') = \Theta(n^2)$.

Dowód. Zauważmy, że na mocy twierdzenia 4.10 wystarczy pokazać konstrukcję takich drzew dla $n = 3k$, $k \in \mathbb{Z}_{\geq 2}$. W celu pokazania prawdziwości punktu pierwszego rozważmy drzewa $T_1, T_2 \in U_L^B$, $|L| = 3k$ przedstawione na rysunku 4.9. Drzewo T_2 powstaje z T_1 w wyniku pojedynczej operacji uSPR, polegającej na odczepieniu poddrzewa zawierającego liście $A = \{a_1, \ldots, a_k\}$ i przyłączeniu go w środek krawędzi e. Niech $B_i = \{b_1, b_2, \ldots, b_i\}$ oraz $s_i = (A \cup B_i)|L \setminus (A \cup B_i)$ i $t_i = B_{i+1}|L \setminus B_{i+1}$ dla $i = 1, \ldots, k-1$. Mamy zatem $\beta_*(T_1) \setminus \beta_*(T_2) = \{s_i\}_{i=1,\ldots,k-1} = X_1$ oraz $\beta_*(T_2) \setminus \beta_*(T_1) = \{t_i\}_{i=1,\ldots,k-1} = X_2$. Na podstawie lematu 3.1 odległość $d_{MS}(T_1, T_2)$ jest równa sparowaniu o najmniejszej wadze rozbić ze zbiorów X_1 i X_2. Zauważmy, że $h_{MS}(s_i, t_j) = \min\{|A| + |B_i \oplus B_{j+1}|, |L| - |A| - |B_i \oplus B_{j+1}|\}$ dla $1 \leq i, j \leq k-1$. Ponieważ dla $1 \leq i, j \leq k-1$ zachodzi $|B_i \oplus B_{j+1}| \leq k$, stąd $h_{MS}(s_i, t_j) \geq k$. Otrzymaliśmy zatem, że $d_{MS}(T_1, T_2) \geq k(k-1) = \Theta(n^2)$.

Punkt drugi wynika z pierwszego oraz z faktu, że operacja uSPR jest szczególnym przypadkiem uTBR (por. lemat 2.3). □

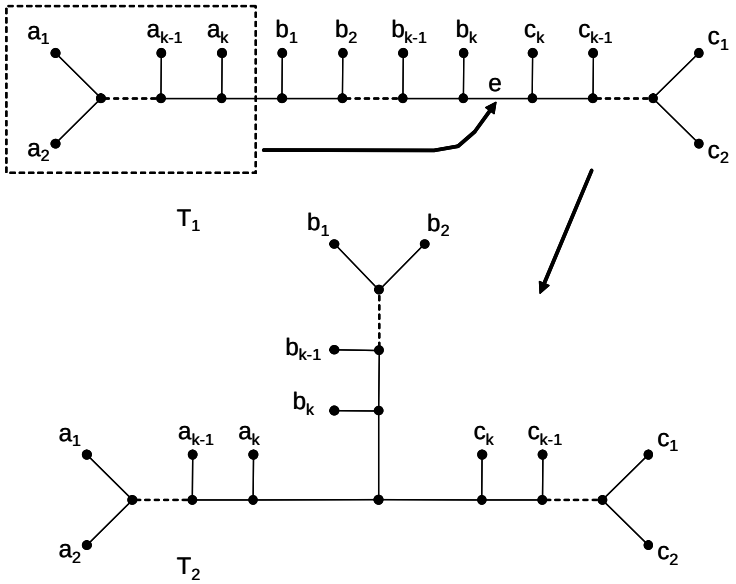
4.4 Średnica przestrzeni z metryką MS

Średnice zbiorów U_n, U_n^B w metrykach filogenetycznych wyznaczają zakres, w jakim dana metryka mierzy odległości między drzewami. W przypadku metryk filogenetycznych wartości te rzadko są znane dokładnie, wprawdzie $\Delta_{d_{RF}}(U_n^B) = n-3$, lecz np. dla metryk opartych na operacjach edycyjnych znamy już tylko asymptotyczne oszacowania.

Twierdzenie 4.13 ([71, 6])**.**

1. $\frac{n}{4}\log_2 n - o(n \log n) \leq \Delta_{d_{uNNI}}(U_n^B) \leq n \log_2 n + O(n)$ [71],
2. $\frac{n}{2} - o(n) \leq \Delta_{d_{uSPR}}(U_n^B) \leq n - 3$ [6],
3. $\frac{n}{4} - o(n) \leq \Delta_{d_{uTBR}}(U_n^B) \leq n - 3$ [6].

W niniejszym podrozdziale zostanie wykazane, że średnica zbiorów U_n i U_n^B w metryce MS wynosi $\frac{3}{8}n^2 + O(n)$. W przeprowadzeniu dowodu tego faktu przydatne będą poniższe lematy.



Rysunek 4.9: Drzewo $T_2 \in U_L^B$, gdzie $|L| = 3k$, powstaje z drzewa $T_1 \in U_L^B$ w wyniku pojedynczej operacji uSPR.

Lemat 4.14 ([19])**.** *W dowolnym ukorzenionym drzewie binarnym $T \in R_n^B$ liczba wierzchołków wewnętrznych posiadających klastry o mocy nie większej niż $k \in \{1, \ldots, n\}$ wynosi co najmniej $k - 1$.*

Dowód. Łatwo zweryfikować poprawność lematu dla $k = n$. Niech $k < n$ oraz u będzie wierzchołkiem wewnętrznym, takim że u nie spełnia warunku wymaganego w lemacie (tj. $|c(u)| > k$), lecz wszystkie wierzchołki wewnętrzne będące potomkami u spełniają ten warunek. Zauważmy, że zawsze taki wierzchołek istnieje, gdyż wystarczy jako u przyjąć dowolny wierzchołek wewnętrzny posiadający klaster o mocy $\min_{v \in V(T)}\{|c(v)| : |c(v)| > k\}$. Wierzchołek u wraz ze swoimi potomkami indukuje podgraf postaci ukorzenionego poddrzewa w T posiadający co najmniej $k + 1$ liści. Poddrzewo to ma zatem co najmniej k wierzchołków wewnętrznych, z których $k - 1$ spełnia warunki lematu. □

Dla $n \in \mathbb{Z}_{>1}$, niech $G_n \in R_n^B$ będzie ukorzenioną binarną gąsienicą. W szczególności przyjmujemy, że drzewa ukorzenione binarne posiadające

2 lub 3 liście są również gąsienicami. Niech $S(n) = \sum_{i=2}^{n} i = \frac{n^2+n}{2} - 1$ dla $n \geq 2$ oraz $S(1) = 0$.

Lemat 4.15 ([19])**.**

1. *Suma mocy klastrów odpowiadających wierzchołkom wewnętrznym w gąsienicy G_a wynosi $S(a)$.*

2. *Dla $a \leq b$ zachodzi $S(a) + S(b) < S(a-1) + S(b+1)$.*

3. *Suma mocy klastrów związanych z wierzchołkami wewnętrznymi w binarnym ukorzenionym drzewie n-listnym osiąga maksymalną wartość dla G_n.*

Dowód. Fakty 1, 2 oraz 3 dla $n < 4$ są prostymi obserwacjami. Niech $T \in R_n^B$ będzie dowolnym ukorzenionym drzewem binarnym niebędącym gąsienicą. W celu uzasadnienia prawdziwości faktu 3 dla $n \geq 4$ pokażemy, że T można przekształcić w drzewo o topologii gąsienicy G_n za pomocą operacji przenoszenia pojedynczego liścia, tak by nie zmniejszyć sumy mocy jego klastrów. Proces transformacji dotyczy ukorzenionych poddrzew i odbywa się w porządku „od dołu do góry" (bottom-up). Zauważmy, że jeżeli korzeń takiego poddrzewa jest połączony z dwoma ukorzenionymi gąsienicami posiadającymi odpowiednio n_1 oraz n_2 liści ($1 < n_1 \leq n_2$), to zgodnie z punktem 2 wartość analizowanej sumy wskutek przeniesienia liścia z pierwszego poddrzewa do drugiego nie zmaleje. Opisaną operację powtarzamy do momentu uzyskania ukorzenionej gąsienicy. □

Lemat 4.16 ([19])**.** *Dla n podzielnego przez 4 liczba krawędzi wewnętrznych w drzewie $T \in U_n^B$, którym odpowiadają rozbicia $A|B$, takie że zachodzi* $\min(A|B) \leq \frac{n}{4}$*, wynosi co najmniej* $\frac{n}{2} - 2$.

Dowód. Wprowadzamy orientację krawędzi drzewa T w kierunku wyznaczonym przez położenie w drzewie mniejszej składowej rozbicia związanego z daną krawędzią. W ten sposób co najwyżej jedna krawędź może nie otrzymać orientacji (taka, która odpowiada rozbiciu $A|B$, gdzie $|A| = |B|$).

Przez stopień *wejściowy* (*wyjściowy*) wierzchołka rozumiemy liczbę krawędzi incydentnych z danym wierzchołkiem i skierowanych do (od) niego. Po wprowadzeniu wspomnianej orientacji dla każdego wierzchołka w T jego stopień wejściowy może wynosić 0 lub 1. Możliwe są dwie sytuacje.

1. Dokładnie jedna krawędź $\{u_1, u_2\}$ została nieskierowana. Drzewo T może być zatem rozważane jako dwa drzewa binarne T_1 oraz T_2 ukorzenione odpowiednio w u_1 i u_2, gdzie $|L(T_1)| = |L(T_2)| = \frac{n}{2}$. Na podstawie lematu 4.14, w T_i $(i = 1, 2)$ istnieje co najmniej $\frac{n}{4} - 1$ wierzchołków wewnętrznych u posiadających klastry o mocy $|c(u)| \leq \frac{n}{4}$.

2. Wszystkie krawędzie zostały skierowane. Istnieje zatem jeden wierzchołek u posiadający stopień wejściowy 0 i T może być rozważane jako suma trzech drzew binarnych ukorzenionych w wierzchołkach sąsiadujących z u. Co najwyżej jedno z tych drzew może posiadać mniej niż $\frac{n}{4}$ liści, stąd pozostałe dwa drzewa T_1 oraz T_2 spełniają $|L(T_i)| \geq \frac{n}{4}$ dla $i = 1, 2$, a to z kolei pozwala na powtórzenie dla nich rozumowania z punktu 1.

□

Lemat 4.17 ([19])**.** *Niech n będzie podzielne przez 4 oraz $T \in U_n^B$. Wszystkie rozbicia $A|B$ z $\beta_*(T)$ sortujemy w porządku niemalejącym według wartości* $\min(A|B)$*. Suma wartości* $\min(A|B)$ *pierwszych $\frac{n}{2}$ rozbić w tym ciągu jest mniejsza niż $\frac{n^2}{16} + O(n)$.*

Dowód. Na podstawie lematu 4.16 wszystkie wybrane rozbicia $A|B$, z wyjątkiem co najwyżej dwóch, których nie będziemy dalej rozważać, spełniają warunek $\min(A|B) \leq \frac{n}{4}$. Wprowadzamy orientację krawędzi drzewa T w kierunku wyznaczonym przez położenie w drzewie mniejszej składowej rozbicia związanego z daną krawędzią. Następnie usuwamy wszystkie wierzchołki, do których nie wchodzi żadna z krawędzi odpowiadających rozważanym rozbiciom. Usuwamy również wszystkie wierzchołki izolowane. W rezultacie T rozpada się na sumę x rozłącznych ukorzenionych drzew binarnych posiadających n_i wierzchołków wewnętrznych $(i = 1, \ldots, x)$. Suma szacowana w niniejszym lemacie, oznaczona dalej jako S, odpowiada

sumie mocy klastrów związanych z wierzchołkami wewnętrznymi otrzymanych w ten sposób drzew. Mamy więc $\sum_{i=1}^{x} n_i \leq \frac{n}{2}$ (każdy wierzchołek wewnętrzny odpowiada pewnemu z wybranych rozbić) oraz $n_i \leq \frac{n}{4} - 1$ dla $i = 1, \ldots, x$. Na podstawie lematu 4.15 otrzymujemy ostatecznie, że $S \leq \sum_{i=1}^{x} S(n_i + 1) \leq 2S(\frac{n}{4} + 1)$. □

Twierdzenie 4.18 ([19])**.** *Średnice zbiorów n-listnych nieukorzenionych drzew filogenetycznych w metryce MS wynoszą:*

$$\Delta_{d_{MS}}(U_n^B) = \frac{3}{8}n^2 \pm O(n),$$
$$\Delta_{d_{MS}}(U_n) = \frac{3}{8}n^2 \pm O(n).$$

Dowód. Dla każdego n mamy $n = 8k + l$, $l = 0, \ldots, 7$, gdzie $k \in \mathbb{Z}_{\geq 0}$. Na podstawie twierdzenia 4.10 możemy zatem ograniczyć rozważania do przypadku $n = 8k$, ponieważ ewentualna zmiana odległości spowodowana dodaniem l liści jest nie większa niż $(\left\lfloor \frac{n}{2} \right\rfloor + n - 4)l$. Na podstawie lematu 3.1 wartość $d_{MS}(T_1, T_2)$ dla $T_1, T_2 \in U_n^B$ jest równa najlżejszemu doskonałemu skojarzeniu $n - 3$ rozbić z $\beta_*(T_1)$ z $n - 3$ rozbiciami z $\beta_*(T_2)$. W celu otrzymania ograniczenia górnego sortujemy elementy zbiorów $\beta_*(T_1)$ i $\beta_*(T_2)$ (osobno elementy z każdego ze zbiorów) w porządku niemalejącym względem rozmiaru mniejszej partycji rozbicia, tworząc dwie sekwencje. Następnie, łącząc elementy występujące na zgodnych pozycjach w obu sekwencjach, tworzymy pewne sparowanie. Waga odpowiadającego przedstawionemu sparowaniu skojarzenia na podstawie lematu 4.17 może być oszacowana przez $2(\frac{1}{16}n^2 + O(n)) + (\frac{n}{2} - 3)\left\lfloor \frac{n}{2} \right\rfloor = \frac{3}{8}n^2 + O(n)$, ponieważ dla dowolnych rozbić s i t zbioru n-elementowego zachodzi $h_{MS}(s, t) \leq \min\left\{\min(s) + \min(t), \left\lfloor \frac{n}{2} \right\rfloor\right\}$.

Pokażemy teraz, że oszacowanie to jest również prawdziwe dla dowolnych (niekoniecznie binarnych) drzew $T_1, T_2 \in U_n$. Na podstawie drzew T_1, T_2 poprzez *rozwiązywanie multifurkacji* (czyli wprowadzanie nowych krawędzi rozdzielających wierzchołki o stopniu większym niż 3) w dowolny sposób konstruujemy odpowiednie drzewa binarne $T_1', T_2' \in U_n^B$. Mamy

zatem $\beta_*(T_i) \subseteq \beta_*(T'_i)$, dla $i = 1, 2$. Sortujemy elementy zbiorów $\beta_*(T'_1)$ i $\beta_*(T'_2)$ podobnie jak powyżej, tj. w porządku niemalejącym względem rozmiaru mniejszej partycji rozbicia. Niech $M' = \{(s'_i, t'_i) : i = 1, \ldots, n-3\}$, gdzie $s'_i \in \beta_*(T'_1)$, $t'_i \in \beta_*(T'_2)$ będzie sparowaniem rozbić występujących na tych samych pozycjach w posortowanych sekwencjach. Na podstawie M' skonstruujemy teraz sparowanie $M = \{(s_i, t_i) : i = 1, \ldots, n-3\}$, gdzie $s_i \in \beta_*(T_1) \cup \{O\}$, $t_i \in \beta_*(T_2) \cup \{O\}$, zastępując rozbicia niewystępujące w zbiorach $\beta_*(T_1)$ i $\beta_*(T_2)$ elementem O. Zauważmy, że każdy z elementów zbiorów $\beta_*(T_1)$ i $\beta_*(T_2)$ występuje w M dokładnie raz. Przyjmując $\min(O) = 0$, na podstawie lematu 4.17 otrzymujemy:

$$\begin{aligned} d_{MS}(T_1, T_2) &\leq \sum_{(s,t)\in M} h_{MS}(s,t) \leq \sum_{i=1}^{\frac{n}{2}} (\min(s_i) + \min(t_i)) + \left(\frac{n}{2} - 3\right) \left\lfloor \frac{n}{2} \right\rfloor \\ &\leq \sum_{i=1}^{\frac{n}{2}} \min(s'_i) + \sum_{i=1}^{\frac{n}{2}} \min(t'_i) + \left(\frac{n}{2} - 3\right) \left\lfloor \frac{n}{2} \right\rfloor \leq \frac{3}{8} n^2 + O(n). \end{aligned}$$

Ograniczenie dolne otrzymamy konstruując dwa drzewa $T_1, T_2 \in U_n^B$, takie że $d_{MS}(T_1, T_2) \geq \frac{3}{8} n^2 - O(n)$. Niech $S(T) = \{s \in \beta_*(T) : \min(s) \leq \frac{n}{4}\}$ będzie zbiorem małych nietrywialnych rozbić w T i $D(T) = \beta_*(T) \setminus S(T)$ będzie zbiorem dużych rozbić w T. Konstrukcja wspomnianych drzew została przedstawiona na rysunku 4.10. Zauważmy, że dla rozbić $s \in \beta_*(T_1)$, $t \in \beta_*(T_2)$ zachodzą następujące relacje:

$$h_{MS}(s,t) = \min(s) + \min(t), \text{ gdzie } s \in S(T_1) \text{ i } t \in S(T_2), \tag{4.7}$$

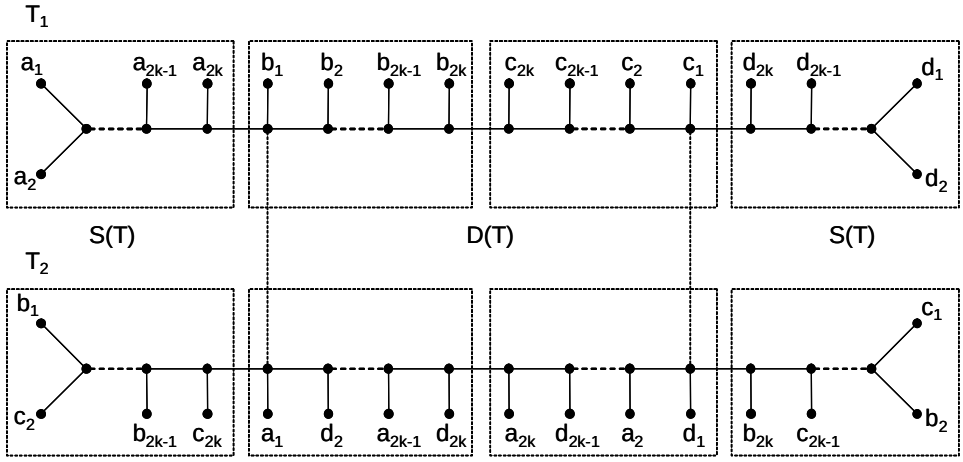
$$h_{MS}(s,t) \geq \frac{n}{2} - 2, \text{ gdzie } s \in D(T_1) \text{ i } t \in D(T_2), \tag{4.8}$$

$$h_{MS}(s,t) \geq \frac{n}{4} + \min(t) - 1, \text{ gdzie } s \in D(T_1) \text{ i } t \in S(T_2), \tag{4.9}$$

$$h_{MS}(s,t) \geq \frac{n}{4} + \min(s) - 1, \text{ gdzie } s \in S(T_1) \text{ i } t \in D(T_2). \tag{4.10}$$

Definiujemy funkcję f na zbiorze nietrywialnych rozbić w następujący sposób: $f(s) = \min(s)$ dla $s \in S(T_1) \cup S(T_2)$ oraz $f(s) = \frac{n}{4} - 1$ dla $s \in D(T_1) \cup D(T_2)$. Wówczas z (4.7) - (4.10) otrzymujemy $h_S(s_1, s_2) \geq f(s_1) + f(s_2)$ dla każdego rozbicia $s_i \in \beta_*(T_i)$, gdzie $i = 1, 2$. Ponieważ w doskonałym skojarzeniu musi znaleźć się każde z rozbić z $\beta_*(T_1) \cup \beta_*(T_2)$,

jego waga (zatem również i wartość $d_{MS}(T_1, T_2)$) jest nie mniejsza niż $\sum_{s\in\beta_*(T_1)\cup\beta_*(T_2)} f(s) = 4\sum_{i=2}^{\frac{n}{4}} i + 2(\frac{n}{4}-1)(\frac{n}{2}-1) = \frac{3}{8}n^2 - n - 2$. □



Rysunek 4.10: Konstrukcja odległych drzew w metryce MS.

4.5 Regularność przestrzeni z metryką MS

W tym podrozdziale rozważymy problem występowania izolowanych wysp w zbiorach U_L^B i U_L z metryką MS. Wystąpienie izolowanej wyspy ma miejsce wtedy, gdy poniższa własność nie jest spełniona:

Definicja 4.1. Niech dla każdego $n \geq 4$ funkcje $d_n^T : X_n \times X_n \to \mathbb{R}_{\geq 0}$ będą metrykami w zbiorach n-listnych drzew filogenetycznych X_n. *Metryki d_n^T nie tworzą izolowanych wysp*, jeśli istnieje taka stała $\delta \in \mathbb{R}_{\geq 0}$ niezależna od n, że dla każdej pary $T_a, T_b \in X_n$ istnieje ciąg drzew $T_a = T_1, T_2, \ldots, T_{k-1}, T_k = T_b$, $T_i \in X_n$ dla $1 \leq i \leq k$, taki że $d_n^T(T_j, T_{j+1}) \leq \delta$ dla $1 \leq j \leq k-1$.

Izolowanych wysp, w sensie powyższej definicji w zbiorach U_L oraz U_L^B, nie posiada oczywiście metryka RF. Dla U_L^B wystarczy przyjąć $\delta = 1$ i wy-

konywać operacje uNNI aby zaleźć sekwencję bliskich binarnych drzew pośrednich. W przypadku drzew dowolnych $T_1, T_2 \in U_L$ własność ta wynika z faktu, że wartość $2d_{RF}(T_1, T_2)$ określa minimalną liczbę operacji ściągnięcia lub wprowadzenia krawędzi niezbędną do przekształcenia drzewa T_1 w T_2 [95]. Szukana sekwencja drzew powstaje zatem w wyniku tych właśnie operacji.

Dla metryk uNNI, uSPR i uTBR w zbiorze U_L^B własność braku izolowanych wysp wynika wprost z ich definicji. Wspomniane wyspy nie występują również w przypadku metryki MS. Fakt ten jest konsekwencją poniższego twierdzenia.

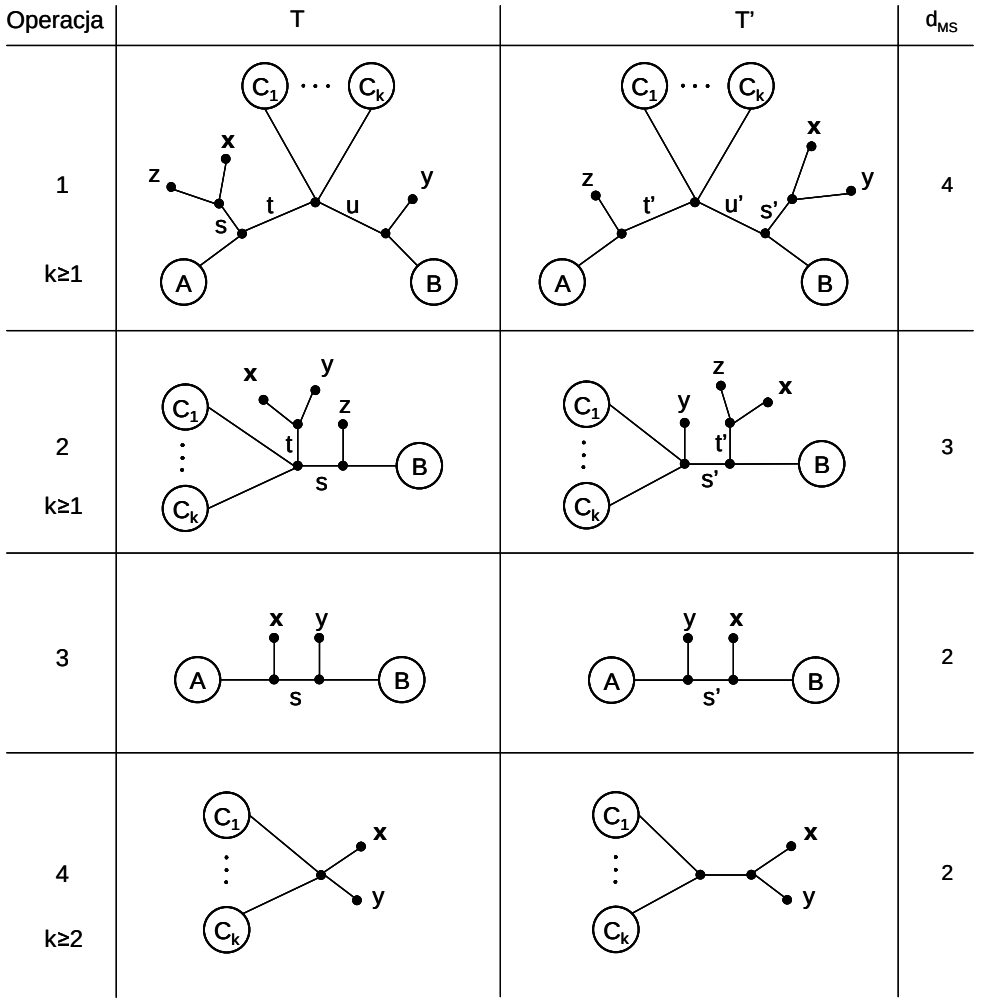
Twierdzenie 4.19 ([19])**.** *Dla dowolnych $T_a, T_b \in U_L$ istnieje sekwencja drzew $T_a = T_1, T_2, \ldots, T_{k-1}, T_k = T_b$, $T_i \in U_L$ dla $i = 1, \ldots, k$, taka że $d_{MS}(T_j, T_{j+1}) \leq 4$, gdzie $j = 1, \ldots, k-1$. Ponadto jeśli dodatkowo $T_a, T_b \in U_L^B$, wówczas istnieje sekwencja o opisanych powyżej własnościach składająca się wyłącznie z drzew binarnych.*

Dowód. Sekwencja drzew pośrednich występująca w twierdzeniu powstaje w wyniku wykonywania operacji przedstawionych na rysunku 4.11, za pomocą których można przekształcić dowolne drzewo $T_a \in U_L$ w dowolne drzewo $T_b \in U_L$. W celu wyznaczenia kolejności wykonywania operacji wybieramy dowolny liść z T_a, tymczasowo go usuwamy i następnie ukorzeniamy w wierzchołku, który sąsiadował z usuniętym liściem. Proces transformacji dotyczy ukorzenionych poddrzew i odbywa się w porządku „od dołu do góry" (bottom-up). Usunięcie wybranego liścia ma na celu wyłącznie ustalenie kolejności wykonywania operacji, tj. ustalenie porządku „od dołu do góry", który możemy zapamiętać, np. wprowadzając orientację krawędzi w kierunku wierzchołka pełniącego rolę korzenia. W procesie transformacji uczestniczy więc nieukorzenione drzewo T_a.

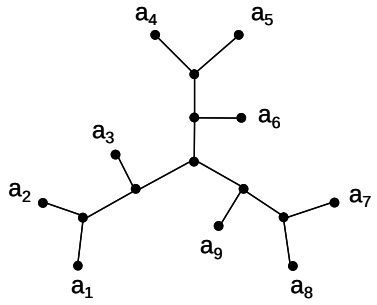
Proces ten polega na przenoszeniu liścia w obrębie poddrzewa, tak by przekształcić je w binarną gąsienicę, co może być wykonane za pomocą operacji 1 oraz 2. Jeśli w trakcie transformacji pojawi się w danym drzewie wierzchołek wewnętrzny o stopniu większym niż 3 połączony z co najmniej dwoma liśćmi, wykonujemy wówczas operację 4. Dla drzew binarnych T_a, T_b operacja 4 jest zbędna. Podobnie w binarną gąsienicę transformujemy

drzewo T_b. Sekwencję drzew pośrednich łączącą powstałe gąsienice otrzymujemy poprzez sortowanie liści jednej z nich za pomocą operacji 3. □

Mimo zmiennej liczby sąsiadów, wynoszącej w niektórych przypadkach nawet 0, przestrzeń z metryką MS zarówno dla U_L^B, jak i U_L posiada więc pewne cechy regularności.



Rysunek 4.11: Lokalne modyfikacje w metryce MS. Literami s', t', u' oznaczono rozbicia, których całkowity koszt sparowania odpowiednio z s, t i u definiuje odległość MS między danymi drzewami.



Rysunek 4.12: Przykład drzewa, dla którego nie istnieje drzewo binarne w odległości 3 w metryce MS.

Tabela 4.1: Liczby drzew $|N_d(T,\delta)|$ ze zbioru U_9^B leżących w pobliżu drzewa T przedstawionego na rysunku 4.12 dla pięciu najmniejszych osiągalnych wartości δ.

Nr	δ	$\lvert N_{MS}(T,\delta)\rvert$	δ	$\lvert N_{RF}(T,\delta)\rvert$	δ	$\lvert N_{PD}(T,\delta)\rvert$	δ	$\lvert N_{QT}(T,\delta)\rvert$
1	2	6	1	12	$\sqrt{14}$	6	6	6
2	4	45	2	96	$\sqrt{20}$	6	12	12
3	5	72	3	724	$\sqrt{28}$	12	18	14
4	6	241	4	5082	$\sqrt{32}$	9	24	36
5	7	432	5	26904	$\sqrt{34}$	24	28	12

Odległość MS wynosząca 4 jest jednocześnie najmniejszą wartością dla U_n^B, $n \geq 4$, która wyznacza sąsiedztwo, takie że zawsze można znaleźć w nim drzewa pośrednie występujące w def. 4.1. W podrozdziale 4.1 wykazane zostało, że w U_n^B nie istnieją drzewa odlegle o 1 oraz istnieją drzewa, które nie mają sąsiadów (czyli drzew w odległości 2). Na rysunku 4.12 przedstawione zostało drzewo $T \in U_9^B$, dla którego nie istnieje w U_9^B drzewo w odległości 3. W tabeli 4.1 przedstawiono moce zbiorów $N_d(T,\delta)$ dla drzewa T dla pięciu najmniejszych wartości δ osiągalnych w poszczególnych metrykach.

Warto w tym miejscu zaznaczyć, że w przypadku innych powszechnie

stosowanych metryk wielomianowych, tj. metryki kwartetowej, ścieżkowej (PD) lub węzłowej (ND), izolowane wyspy występują. W przypadku metryki kwartetowej i zbioru U_n^B wynika to wprost z faktu, że już minimalna odległość w QT miedzy różnymi drzewami rośnie proporcjonalnie do liczby liści n i wynosi $n-3$ [106]. W przypadku pozostałych dwóch metryk PD, ND i zbioru U_n wystarczy zauważyć, że odległość dowolnego drzewa od gwiazdy $T \in U_n$ rośnie wyraz ze wzrostem n, gdyż wprowadzanie każdej nowej krawędzi wewnętrznej wydłuża wzajemne odległości o 1 między co najmniej $2(n-2)$ parami wierzchołków. Zatem dla dowolnego drzewa $T' \in U_n$, $T' \neq T$ mamy $d_{ND}(T,T') \geq 2(n-2)$ oraz $d_{PD}(T,T') \geq \sqrt{2(n-2)}$.

4.6 Podsumowanie

Własności metryki MS dla drzew binarnych:

1. oszacowanie przez RF, tw. 4.3:

$$d_{RF}(T_1,T_2)+1 \leq d_{MS}(T_1,T_2) \leq \left\lfloor \frac{n}{2} \right\rfloor d_{RF}(T_1,T_2), \text{ dla } T_1 \neq T_2,$$

2. minimalna dodatnia odległość, tw. 4.3

$$\min_{T_1 \neq T_2 \in U_n^B} d_{MS}(T_1,T_2) = 2,$$

3. rozmiar sąsiedztwa, tw. 4.7: co najwyżej $n-1$, zależny od topologii drzewa,

4. średnica, tw. 4.18:

$$\max_{T_1,T_2 \in U_n^B} d_{MS}(T_1,T_2) = \frac{3}{8}n^2 \pm O(n).$$

Własności metryki MS dla drzew dowolnych:

1. oszacowanie przez RF, tw. 4.2:

$$d_{RF}(T_1,T_2) \leq d_{MS}(T_1,T_2) \leq n\ d_{RF}(T_1,T_2),$$

2. minimalna dodatnia odległość, wniosek 4.8:

$$\min_{T_1 \neq T_2 \in U_n} d_{MS}(T_1, T_2) = 1,$$

3. rozmiar sąsiedztwa, wniosek 4.8: $O(n^2)$, zależny od topologii drzewa,

4. średnica, tw. 4.18:

$$\max_{T_1, T_2 \in U_n} d_{MS}(T_1, T_2) = \frac{3}{8} n^2 \pm O(n).$$

Wybór sposobu porównywania rozbić za pomocą funkcji h_{MS} jest dość naturalny. Dla drzew niebinarnych zachodzi jednak potrzeba zdefiniowania odległości $h(O, s)$ rozbicia s do elementu O. Przyjęta w rozważaniach definicja $h_{MS}(s, O) = \min(s)$ powoduje, że duża część własności prawdziwych dla drzew binarnych przenosi się również na drzewa dowolne, np. oszacowanie średnicy, reakcja na dodanie nowego wierzchołka, czy własność braku izolowanych wysp. Warto zauważyć, że dość łatwo jest podać inne sposoby definiowania tej odległości gwarantujące metryczność h, np. dla $s \in Splits(L)$ wartość tą możemy zdefiniować jako $h(s, O) = \left\lceil \left\lfloor \frac{|L|}{2} \right\rfloor / 2 \right\rceil$ [18]. Otrzymujemy wówczas jednak pewne osłabienie własności w zbiorze U_L w stosunku do drzew binarnych, objawiające się m.in. powstaniem izolowanych wysp, których przykładem jest gwiazda.

Trzeba zaznaczyć, że metoda mierzenia odległości o zbliżonej naturze do MS pojawia się w pracy Nye i in. [82]. Zaproponowany tam algorytm jest przedstawiony jako miara podobieństwa drzew filogenetycznych (w szczególności nie spełnia aksjomatów przestrzeni metrycznej), w związku z czym opiera się on na wadze najcięższego skojarzenia w grafie dwudzielnym, którego partycje podobnie jak w MS, odpowiadają zbiorom nietrywialnych rozbić w porównywanych drzewach. Metoda ta wykorzystuje jednak inną niż w przypadku MS miarę podobieństwa rozbić, zdefiniowaną jako $s(A|B, C|D) = \max\left\{\min\left\{\frac{|A\cap C|}{|A\cup C|}, \frac{|B\cap D|}{|B\cup D|}\right\}, \min\left\{\frac{|A\cap D|}{|A\cup D|}, \frac{|B\cap C|}{|B\cup C|}\right\}\right\}$. Autorzy pracy [82] wspominają także, że ich algorytm może być stosowany również dla drzew o różnej liczbie krawędzi wewnętrznych, jednak jasno nie

precyzują jak w takim przypadku powinno odbywać się obliczanie wartości podobieństwa. Praca [82] nie zawiera żadnych rozważań teoretycznych ani statystycznych dotyczących własności prezentowanej tam metody.

5 Przestrzeń metryczna MC dla drzew z korzeniem

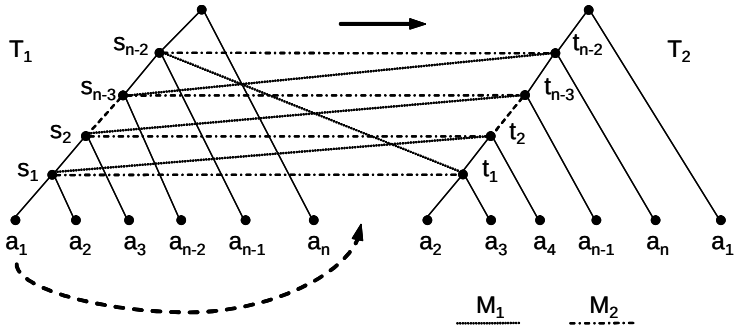
5.1 Dopasowanie wierzchołków drzew za pomocą metryki MC

Dopasowanie wierzchołków porównywanych drzew jest istotnym zagadnieniem w analizie zgodności historii ewolucji [79]. Oprócz ilościowego określenia stopnia podobieństwa realizowanego przez metryki, dopasowanie wierzchołków pozwala na strukturalne zobrazowanie kształtu i położenia najbardziej zbliżonych do siebie obszarów drzew. Jedną z najbardziej popularnych metod umożliwiających skonstruowanie i zobrazowanie opisanego dopasowania jest podejście polegające na wyszukiwaniu wierzchołków wewnętrznych najbardziej zbliżonych do danego, według określonej miary podobieństwa, wraz z interaktywną wizualizacją otrzymanych wyników [79]. Funkcjonalność ta jest dostępna np. w aplikacji TreeJuxtaposer [79].

Warto zauważyć, że pewnego rodzaju dopasowanie, wyznaczone przez odpowiadające najlżejszemu skojarzeniu sparowanie wierzchołków drzew $T_1, T_2 \in R_L$, pojawia się jako element dodatkowy przy obliczaniu odległości MC między drzewami T_1 i T_2. Oczywiście takie sparowanie występuje dla dowolnej metryki skonstruowanej za pomocą definicji 3.1, lecz w przypadku odległości MC posiada ono dodatkowe własności, które zostaną omówione w niniejszym podrozdziale.

Opisane dopasowanie odpowiadające najlżejszemu doskonałemu skojarzeniu nie musi być unikatowe, czego przykładem są sparowania $M_1 = \{\{s_i, t_{i+1}\}\}_{i=1,\ldots,n-3} \cup \{\{s_{n-2}, t_1\}\}$ oraz $M_2 = \{\{s_i, t_i\}\}_{i=1,\ldots,n-2}$ o minimal-

nej wadze równej $2n-4$ przedstawione na rysunku 5.1. Niemniej jednak dla MC zawsze możliwe jest znalezienie takiego skojarzenia, które zachowuje relacje miedzy przodkami i potomkami. Rozważmy dwa dowolne drzewa $T_1, T_2 \in R_L$. Niech $a \in V(T_1)$, $b \in V(T_2)$ oraz wierzchołki $c \in V(T_1)$, $d \in V(T_2)$ będą dowolnymi przodkami odpowiednio a i b, tj. $a \leq_{T_1} c$, $b \leq_{T_2} d$. Zachowanie wspomnianej relacji polega na tym, że zawsze możemy znaleźć takie najlżejsze skojarzenie definiujące wartość MC, w którym nie będzie „skrzyżowanego" sparowania wierzchołków, czyli potomka z przodkiem i przodka z potomkiem, tj. a z d oraz b z c (np. sparowanie M_2 na rysunku 5.1). Warto zauważyć, że analogiczna własność dla metryki MS nie jest znana.



Rysunek 5.1: Drzewo T_2 powstaje z T_1 w wyniku przeniesienia liścia a_1 ponad liść a_n i połączenia go z korzeniem. Sparowania klastrów drzew T_1 i T_2, zaznaczone liniami przerywanymi i oznaczone jako M_1 oraz M_2, posiadają minimalną wagę równą $2n-4$.

Twierdzenie 5.1 (o zachowaniu relacji przodek-potomek)**.** *Niech $T_1, T_2 \in R_L$, $|V(T_1)| \leq |V(T_2)|$ oraz $\dot{V}_i = V(T_i) \setminus (L \cup \{r(T_i)\})$ dla $i = 1, 2$ oznacza zbiór wierzchołków wewnętrznych bez korzenia w T_i. Istnieje iniekcja $f : \dot{V}_1 \to \dot{V}_2$ o całkowitym koszcie $\sum_{v \in \dot{V}_1} |c(v) \oplus c(f(v))| + \sum_{u \in \dot{V}_2 \setminus f(\dot{V}_1)} |c(u)| = d_{MC}(T_1, T_2)$, taka że dla wszystkich wierzchołków wewnętrznych (z wyjątkiem korzenia) $a, b \in \dot{V}(T_1)$ spełnione są następujące zależności:*

1. *jeśli $a \leq_{T_1} b$, wówczas wierzchołki sparowane z a i b w T_2 spełniają $f(a) \leq_{T_2} f(b)$ lub są $\leq_{T_2}$-nieporównywalne,*

2. *jeśli* $f(a) \leq_{T_2} f(b)$*, wówczas wierzchołki* a*,* b *spełniają* $a \leq_{T_1} b$ *lub są* $\leq_{T_1}$*-nieporównywalne.*

Dowód. Niech $a \leq_{T_1} b$, $a \neq b$ oraz $b' = f(b) \leq_{T_2} a' = f(a)$, gdzie f oznacza dowolne sparowanie klastrów, które definiuje wartość $d_{MC}(T_1, T_2)$. Wówczas $A = c(a) \subsetneq c(b) = B = X \oplus A$ oraz $B' = c(b') \subsetneq c(a') = A' = B' \oplus X'$ dla pewnych zbiorów X, X'. Mamy zatem $|A \oplus A'| + |B \oplus B'| = |A \oplus B'| + |B \oplus A'| + 2|X \cap X'|$, stąd po wykonaniu modyfikacji w f, tak aby $f(a) := b'$, $f(b) := a'$ równość $\sum_{v \in \dot{V}_1} |c(v) \oplus c(f(v))| + \sum_{u \in \dot{V}_2 \setminus f(\dot{V}_1)} |c(u)| = d_{MC}(T_1, T_2)$ nadal jest prawdziwa, lecz $|A||B'| + |B||A'| = |A||A'| + |B||B'| + |X||X'|$. Zatem po uwzględnieniu opisanej modyfikacji wartość parametru zdefiniowanego jako $\sum_{v \in \dot{V}_1} |c(v)||c(f(v))|$ wzrosła. Otrzymujemy więc, że co najwyżej $O(|L|^3)$ kolejnych operacji tego typu jest możliwych do wykonania, po uwzględnieniu których otrzymamy f spełniające pierwszy punkt twierdzenia. Drugi punkt jest wnioskiem z pierwszego. □

Niech T_v dla v będącego wierzchołkiem wewnętrznym w $T \in R_L$ oznacza poddrzewo w T ukorzenione w v oraz $I(T) = V(T) \setminus L(T)$ będzie zbiorem wierzchołków wewnętrznych T. Poniższy wniosek ilustruje pewną własność pozwalającą zwiększyć efektywność wyznaczania odległości MC dla drzew silnie niezbalansowanych, np. gąsienic.

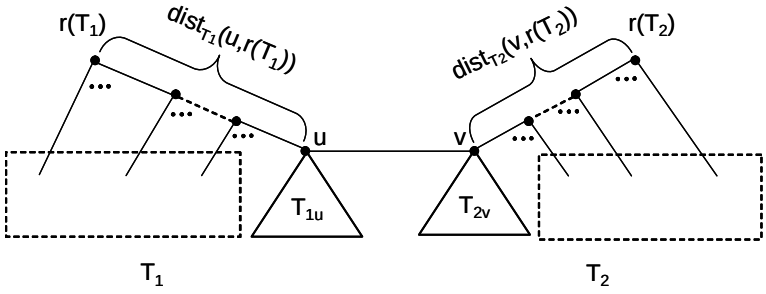
Wniosek 5.2. *Niech* $T_1, T_2 \in R_L$ *oraz* $|\sigma_*(T_1)| \geq |\sigma_*(T_2)|$*. Ponadto niech* $G(V_1, V_2, E)$*,* $|V_1| = |V_2| = |\sigma_*(T_1)|$ *będzie grafem umożliwiającym wyznaczenie odległości* $d_{MC}(T_1, T_2)$ *z lematu 3.2, a wierzchołki* $\dot{u} \in V_1$*,* $\dot{v} \in V_2$ *odpowiadają wierzchołkom wewnętrznym* u*,* v *różnym od korzenia w drzewach odpowiednio* T_1 *i* T_2*. Zbiór krawędzi w* G*, których jednym z końców jest wierzchołek odpowiadający elementowi pomocniczemu* O *oznaczymy jako* E_O*.*

Konstruujemy podgraf G *postaci* $G' = (V_1, V_2, E')$*, gdzie* $E' = E'' \cup E_O \subseteq E$*,* $E'' \cap E_O = \emptyset$*, taki że dla każdej krawędzi* $\{\dot{u}, \dot{v}\} \in E''$*, gdzie* u*,* v *nie są korzeniami* T_1*,* T_2*, spełnione są następujące warunki:*

$$|I(T_{1u})| \leq |\sigma_*(T_1)| - \mathrm{dist}_{T_2}(v, r(T_2)) + 1,$$
$$|I(T_{2v})| \leq |\sigma_*(T_1)| - \mathrm{dist}_{T_1}(u, r(T_1)) + 1,$$

wówczas waga najlżejszego doskonałego skojarzenia w G' również wynosi $d_{MC}(T_1, T_2)$. Ponadto jedyną krawędzią w E' zawierającą wierzchołek odpowiadający korzeniowi T_1 lub T_2 jest $\{\dot{r}(T_1), \dot{r}(T_2)\}$.

Dowód. Niech $M \subseteq E$ będzie skojarzeniem w grafie G spełniającym twierdzenie 5.1. Pokażemy, że wszystkie krawędzie należące do M spełniają podane zależności. Załóżmy, że w M istnieje krawędź $e = \{\dot{u}, \dot{v}\}$ odpowiadająca sparowaniu wierzchołków wewnętrznych $u \in I(T_1) \setminus \{r(T_1)\}$ oraz $v \in I(T_2) \setminus \{r(T_2)\}$ jak na rysunku 5.2. Ponieważ M spełnia warunki twierdzenia 5.1 żaden z wierzchołków wewnętrznych $x \in I(T_{1u}) \setminus \{u\}$ nie może być sparowany z $\mathrm{dist}_{T_2}(v, r(T_2)) - 1$ wierzchołkami leżącymi na ścieżce pomiędzy v oraz $r(T_2)$. Zatem, aby w M mogła istnieć wspomniana krawędź musi być spełniona nierówność $I(T_{1v}) - 1 \leq |\sigma_*(T_1)| - 1 - (\mathrm{dist}_{T_2}(v, r(T_2)) - 1)$. Analogiczna sytuacja musi zachodzi dla wierzchołków wewnętrznych znajdujących się w poddrzewie T_{2v}. Ponieważ takie skojarzenie M istnieje zawsze, możemy z grafu G usunąć te krawędzie o końcach odpowiadających wierzchołkom wewnętrznym drzew T_1, T_2, które nie spełniają przynajmniej jednej z podanych zależności. □



Rysunek 5.2: Ilustracja do wniosku 5.2.

Zauważmy, że jeśli $T \in R_n^B$ jest gąsienicą, to dla każdego $x \in I(T)$ zachodzi $|I(T_x)| + \mathrm{dist}_T(x, r(T)) = n - 1$. Zatem w przypadku, gdy oba drzewa $T_1, T_2 \in R_n^B$ są gąsienicami, na mocy wniosku 5.2 mamy $|I(T_{1u})| \leq |I(T_{1v})|$ oraz $|I(T_{1v})| \leq |I(T_{1u})|$, czyli $I|(T_{1v})| = |I(T_{1u})|$. W celu wyznaczenia odległości MC między T_1 i T_2 wystarczy więc skonstruować graf G' posiada-

jący tylko $n-2$ krawędzi, które jednocześnie tworzą najlżejsze doskonałe skojarzenie w G'.

Niestety użyteczność przedstawionej własności do redukcji ilości krawędzi w grafie konstruowanym przy wyznaczeniu metryki MC maleje wraz ze wzrostem zbalansowania porównywanych drzew.

5.2 Podstawowe własności metryki MC

Poniższy lemat przedstawia prostą zależność między odległościami drzew ukorzenionych i drzew bez korzenia.

Lemat 5.3. *Niech $T_1, T_2 \in R_L$, $T \in R_{L'}$, gdzie $L \cap L' = \emptyset$ i $|L'| \geq |L|$ oraz drzewo $T'_i \in U_{L\cup L'}$, $i = 1, 2$ powstaje w wyniku połączenia korzenia drzewa T_i z korzeniem T. Zachodzą następujące równości:*

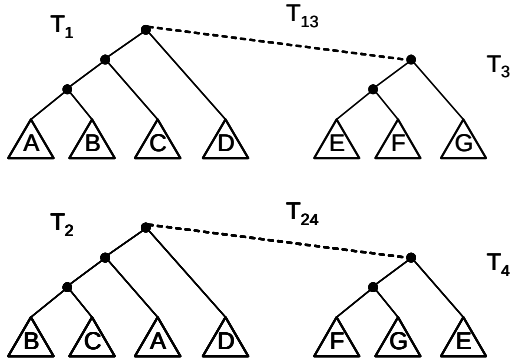
$$d_{RFC}(T_1, T_2) = d_{RF}(T'_1, T'_2),$$
$$d_{MC}(T_1, T_2) = d_{MS}(T'_1, T'_2).$$

Dowód. Definiujemy iniekcję $\iota_1 : \sigma(T_1) \to \beta(T'_1)$, taką że klaster $A \subseteq L$ przechodzi na $\iota(A) = A|L' \cup (L \setminus A)$. W analogiczny sposób wprowadzamy iniekcję $\iota_2 : \sigma(T_2) \to \beta(T'_2)$. Pozostałe rozbicia są identyczne w obu drzewach T'_1, T'_2. Stąd, $d_{RF}(T_1, T_2) = d_{RF}(T'_1, T'_2)$. Skoro $|L'| \geq |L|$, więc dla $A \in \sigma(T_1)$, $B \in \sigma(T_2)$ mamy $h_S(\iota_1(A), \iota_2(B)) = \min\{|A \oplus B|, |L| + |L'| - |A \oplus B|\} = h_{MC}(A, B)$ oraz $h_{MS}(\iota_1(A), O) = |A| = h_{MC}(A, O)$, zatem druga równość zachodzi na mocy lematu 3.1. □

Uogólnienie lematu nie jest prawdziwe, jeśli zamiast wspólnego drzewa $T \in R_{L'}$ użylibyśmy różnych drzew z $R_{L'}$. Jako kontrprzykład rozważmy drzewa przedstawione na rysunku 5.3, gdzie $L = A \cup B \cup C \cup D$, $L' = E \cup F \cup G$, $|L| = |L'| = 14$, $T_1, T_2 \in R^B_L$, $T_3, T_4 \in R^B_{L'}$, $T_{13}, T_{24} \in U^B_{L\cup L'}$. Zachodzi wówczas $d_{MC}(T_1, T_2) + d_{MC}(T_3, T_4) = 10 > d_{MS}(T_{13}, T_{24}) = 8$.

Twierdzenie 5.4. *Niech $T_1, T_2 \in R_L$, gdzie $|L| = n$, wówczas:*

$$d_{RFC}(T_1, T_2) \leq d_{MC}(T_1, T_2) \leq 2(n-1)\, d_{RFC}(T_1, T_2). \tag{5.1}$$



Rysunek 5.3: Litery reprezentują zbiory liści w odpowiednich poddrzewach o następujących liczbach elementów: $|A| = 3$, $|B| = |C| = |E| = |F| = 1$, $|D| = 2$, $|G| = 5$.

Dowód. Zauważmy, że pełne grafy dwudzielne służące do wyznaczania wartości $d_{RF}(T_1, T_2)$ oraz $d_{MC}(T_1, T_2)$ różnią się jedynie wagami na krawędziach. Dla klastrów $c_1, c_2 \subsetneq L$ zachodzi $h_{RFC}(c_1, c_2) \leq h_{MC}(c_1, c_2) \leq n \cdot h_{RFC}(c_1, c_2) \leq 2(n-1)h_{RFC}(c_1, c_2)$, podobnie $\frac{1}{2} = h_{RFC}(c, O) < h_{MC}(c, O) = |c| \leq 2(n-1)h_{RFC}(c, O)$ dla klastra $c \subsetneq L$. □

Przykładem osiągania równości z dolnego ograniczenia są drzewa T_3 i T_3' przedstawione na rysunku 5.7. Górne ograniczenie natomiast staje się równością dla drzew $T, T' \in R_L$, gdzie T jest gwiazdą, a T' posiada dokładnie jeden nietrywialny klaster o mocy $|L| - 1$.

Dla drzew binarnych oszacowanie można poprawić.

Twierdzenie 5.5. *Niech $T_1 \neq T_2 \in R_L^B$, gdzie $|L| = n$, wówczas:*

$$d_{RFC}(T_1, T_2) + 1 \leq d_{MC}(T_1, T_2) \leq (n-1)d_{RFC}(T_1, T_2). \tag{5.2}$$

Dowód. W przypadku dowodu dolnego ograniczenia wykorzystamy twierdzenie 4.3 dla metryki MS. Niech $T \in R_{L'}^B$ będzie dowolnym drzewem,

takim że $L \cap L' = \emptyset$, $|L'| = |L|$. Na podstawie T_1, T_2 oraz T poprzez połączenie krawędzią korzeni drzew T_1, T_2 z wierzchołkiem $r(T)$ konstruujemy dwa drzewa nieukorzenione $T_1', T_2' \in U^B_{L \cup L'}$. Na mocy twierdzenia 4.3 otrzymujmy $d_{RFC}(T_1', T_2') + 1 \leq d_{MS}(T_1', T_2')$. Stąd posługując się lematem 5.3 dostajemy $d_{RF}(T_1, T_2) + 1 \leq d_{MC}(T_1, T_2)$.

Odnośnie górnego ograniczenia zauważmy, że klastry c_1, c_2 tego samego drzewa spełniają warunek kompatybilności (por. tw. 2.2), tj. $c_1 \subseteq c_2$ lub $c_2 \subseteq c_1$ lub $c_1 \cap c_2 = \emptyset$. Niech klastry $A_1, A_2, \ldots, A_{d_{RFC}(T_1,T_2)}$ tworzą zbiór $\sigma_*(T_1) \setminus \sigma_*(T_2)$ oraz $|A_i| \leq |A_j|$ dla $i < j$. W ten sam sposób sortujemy elementy $B_1, B_2, \ldots, B_{d_{RFC}(T_1,T_2)}$ tworzące zbiór $\sigma_*(T_2) \setminus \sigma_*(T_1)$. Pokażemy, że $\sum_{i=1}^{d_{RFC}(T_1,T_2)} |A_i \oplus B_i| \leq (|L|-1) d_{RF}(T_1, T_2)$. Jeśli dla każdego i zachodzi $A_i \neq L \setminus B_i$, to $|A_i \oplus B_i| < |L|$ i teza jest spełniona. W przeciwnym przepadku niech i będzie najmniejszym indeksem, takim że $A_i = L \setminus B_i$. Bez straty ogólności możemy założyć, że $|B_i| \geq |L|/2$. Rozważymy dwa przypadki.

Przypadek 1. $i < d_{RFC}(T_1, T_2)$, wówczas dla $j > i$ na mocy warunku kompatybilności mamy $B_i \subsetneq B_j$ lub $B_i \cap B_j = \emptyset$. Jednak $B_i \cap B_j \neq \emptyset$, gdyż w przeciwnym przypadku $|B_j| \geq |B_i| \geq |L|/2$, a stąd musiałaby zachodzić równość $|B_j| = |L|/2$, co nie jest możliwe, ponieważ wówczas $B_j = L \setminus B_i = A_i$ — sprzeczność z definicją B_j i A_i. Stąd otrzymujemy, że $B_i \subsetneq B_j$ oraz zachodzi $A_i \subsetneq A_j$ albo $A_i \cap A_j = \emptyset$. Zauważmy, że $A_i \cap A_j = \emptyset$ implikuje $A_j \subsetneq B_i \subsetneq B_j$. Dla każdej z obu tych sytuacji, tj. $A_i \subsetneq A_j$ lub $A_j \subsetneq B_i \subsetneq B_j$, mamy $|A_j \oplus B_j| \leq |L| - 2$. Ponieważ dla $j < i$ zachodzi $|A_j \oplus B_j| \leq |L| - 1$, więc otrzymujemy tezę.

Przypadek 2. $i = d_{RFC}(T_1, T_2)$. Załóżmy, że istnieje zbiór $B \in \sigma_*(T_2)$, taki że $B_i \subsetneq B$. Wówczas $B \in \sigma_*(T_1)$, lecz B nie jest kompatybilne z A_i. Zatem taki zbiór B nie istnieje, co oznacza, że B_i musi być klastrem o maksymalnej mocy spośród klastrów z $\sigma_*(T_2)$. Ponieważ T_2 jest drzewem binarnym, dla takiego B_i musi zachodzić $L \setminus B_i = A_i \in \sigma_*(T_2)$. Otrzymujemy sprzeczność z definicją zbioru A_i. Możliwy zatem jest wyłącznie rozważony powyżej przypadek 1. □

Sytuację, w której osiągane jest górne ograniczenie ilustruje pierwszy punkt poniższego wniosku.

Wniosek 5.6.

1. *Istnieją drzewa* $T_1, T_2 \in R_n^B$*, dla których* $d_{RFC}(T_1, T_2) = 1$ *oraz* $d_{MC}(T_1, T_2) = n - 1$.

2. *Jeśli* $T_1, T_2 \in R_n^B$ *oraz* $d_{MC}(T_1, T_2) = 2$*, wówczas* $d_{RFC}(T_1, T_2) = 1$.

Dowód.

1. Konstrukcja takich drzew jest przedstawiona na rysunku 5.4.

2. Zależność ta wynika bezpośrednio z twierdzenia 5.5.

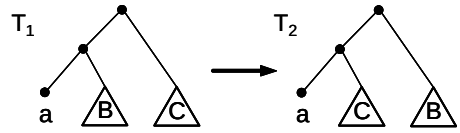
□

Rozważmy teraz dokładniej własności przestrzeni metrycznej generowanej przez metrykę MC, skupiając się w pierwszej kolejności na zbiorze drzew binarnych. Bezpośrednią konsekwencją twierdzenia 5.5 jest definicja sąsiadów w metryce MC. Mianowicie, sąsiadami w zbiorach R_L^B i R_L będą drzewa odległe odpowiednio o 2 i 1. Podobnie jak w przypadku metryki MS, drzewa sąsiednie w MC w zbiorze R_L^B powstają w wyniku operacji zamiany liści znajdujących się po przeciwnych końcach wspólnej krawędzi wewnętrznej (np. liście a_2 oraz a_3 drzewa T_1 z rysunku 5.5). Drzewa sąsiednie w zbiorze R_L^B są zatem przykładem osiągania równości dla dolnego ograniczenia z tw. 5.5.

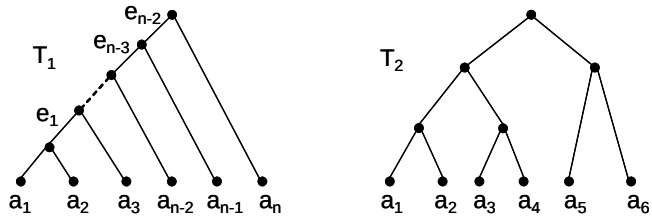
Twierdzenie 5.7. *Dla drzewa* $T_1 \in R_n^B$ *liczba drzew* $T_2 \in R_n^B$*, takich że* $d_{MC}(T_1, T_2) = 2$*, wynosi co najwyżej* $n - 1$*. Dla* $n \geq 4$ *istnieje* $T_1 \in R_n^B$*, takie że dla dowolnego* $T_2 \in R_n^B$ *zachodzi* $d_{MC}(T_1, T_2) > 2$.

Dowód. Pokażemy, że podobnie jak w przypadku metryki MS drzewo binarne o topologii gąsienicy posiada największą możliwą liczbę sąsiadów w MC. Rozważmy drzewa przedstawione na rysunku 5.5. Dla drzewa $T_1 \in R_n^B$ dla każdej krawędzi $e_2, \ldots, e_{n-2}$ możemy wykonać jedną operację zamiany miejscami incydentnych z nią liści, pozwalającą na utworzenie drzewa sąsiedniego. W przypadku krawędzi e_1 możliwe są dwie takie operacje, zatem dla T_1 istnieje $n - 1$ drzew sąsiednich w R_n^B. Z drugiej

strony zauważmy, że T_2 jest przykładem drzewa, które nie posiada żadnej z powyższych konfiguracji liści, stąd nie posiada ono sąsiadów w R_6^B. Konstrukcja drzew o większej liczbie liści przebiega w analogiczny sposób. Dodatkowo jeśli usuniemy liść a_6 oraz a_6 i a_5, otrzymamy drzewa 5- i 4-listne o tej samej własności. □



Rysunek 5.4: Drzewa T_1 i T_2 różnią się jednym klastrem. Odległość w metryce MC między tymi drzewami wynosi $|L| - 1$.



Rysunek 5.5: Gąsienica posiada największą liczbę sąsiadów w R_L^B.

Rozważmy teraz zbliżone własności dla zbioru wszystkich drzew R_L. Własności te podsumowane są w poniższym twierdzeniu.

Twierdzenie 5.8.

1. *Jeśli $d_{MC}(T_1, T_2) = 1$, wówczas oba drzewa nie są binarne oraz jedno z nich powstaje z drugiego w wyniku usunięcia liścia połączonego z pewnym wierzchołkiem wewnętrznym v i przyłączenia go do wierzchołka będącego sąsiadem v. Przykład takiej sytuacji został przedstawiany na rysunku 5.6.*

2. *Jeśli $T_1 \in R_L^B$, wówczas nie istnieje $T_2 \in R_L$, takie że $d_{MC}(T_1, T_2) = 1$.*

3. *Niech $T_1 \in R_L$. Liczba drzew $T_2 \in R_L$, dla których zachodzi równość $d_{MC}(T_1, T_2) = 1$ jest $O(|L|^2)$.*

4. *Równość $d_{RFC}(T_1, T_2) = d_{MC}(T_1, T_2)$ może zachodzić dla dowolnie dużej wartości $d_{RFC}(T_1, T_2)$.*

Dowód.

1. Przedstawiona konfiguracja jest jedyną możliwą sytuacją, w której zbiory $\sigma_*(T_1)$ i $\sigma_*(T_2)$ różnią się dokładnie jedną parą klastrów $c_1 \in \sigma_*(T_1)$, $c_2 \in \sigma_*(T_2)$, $c_1 \neq c_2$ oraz dodatkowo klastry c_1 i c_2 różnią się tylko jednym elementem z L.

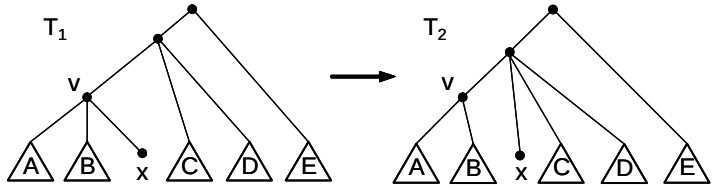
2. Zauważmy, że na podstawie punktu pierwszego, aby opisana sytuacja była możliwa w T_1 musi istnieć wierzchołek o stopniu co najmniej 4 (lub co najmniej 3, w przypadku gdyby wierzchołek ten był korzeniem). Ponieważ drzewo T_1 jest binarne, nie można dla niego przeprowadzić opisanej transformacji.

3. Liczba możliwych operacji opisanych w punkcie pierwszym dla danego drzewa $T_1 \in R_L$ jest nie większa niż liczba par (x, u), gdzie $x \in L$ oraz u jest sąsiadem sąsiada liścia x, stąd oszacowanie $O(|L|^2)$. Przykładem opisanej sytuacji jest drzewo T_3 na rysunku 5.7.

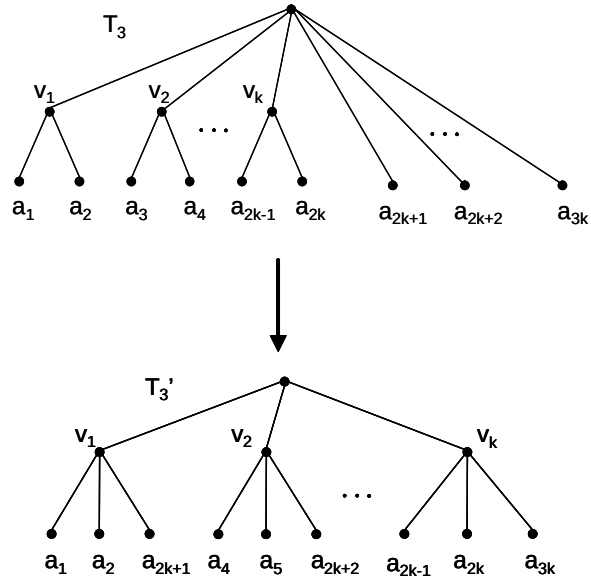
4. Usunięcie każdego z następujących liści $a_{2k+1}, \ldots, a_{3k}$ drzewa T_3 (rysunek 5.7) i przyłączenie go do odpowiedniego wierzchołka wewnętrznego $v_1, \ldots, v_k$ powoduje utworzenie drzewa T_3' znajdującego się w odległości $d_{RFC}(T_3, T_3') = d_{MS}(T_3, T_3') = k$.

□

Podsumowując, podobnie jak w przypadku odległości MS liczba drzew w sąsiedztwie wybranego drzewa w metryce MC zależy od jego topologii. Może się ona wahać w granicach od 0 do $\Theta(n^2)$ w przypadku ogólnym oraz pomiędzy 0 i $\Theta(n)$ dla binarnych drzew n-listnych. Rozmiar sąsiedztwa jest zależny od topologii również dla odległości rSPR [103] (w przeciwieństwie



Rysunek 5.6: Odległość w metryce MC między drzewami T_1 i T_2 wynosi 1.



Rysunek 5.7: W odległości MC wynoszącej 1 od $3k$-listnego drzewa T_3 znajduje się k^2 innych drzew. Odległości MC i RF między drzewami T_3 i T_3' wynoszą k.

do uSPR, por. tw. 4.5). Inaczej jest w przypadku metryki RFC, gdzie w zbiorze R_n^B rozmiar sąsiedztwa nie zależy od topologii drzewa i wynosi zawsze $2n-4$. Jednak w zbiorze drzew dowolnych R_n, podobnie jak dla MC, również i w RFC pojawia się związek między topologią drzewa a liczbą jego sąsiadów. Zauważmy, że sąsiadem gwiazdy $T \in R_n$ w metryce RFC jest każde drzewo posiadające dokładnie jeden nietrywialny klaster, stąd liczba jej sąsiadów wynosi $\sum_{i=2}^{n-1} \binom{n}{i} = 2^n - n - 2$ i rośnie wykładniczo względem n. W metryce MC natomiast gwiazda jest przykładem drzewa

nieposiadającego sąsiadów. Niemniej jednak, podobnie jak w przypadku MS, w przestrzeniach R_L i R_L^B nie występują izolowane wyspy. Własność ta zostanie dokładniej omówiona w podrozdziale 5.4.

5.3 Nieznaczne modyfikacje drzewa a średnica przestrzeni MC

Twierdzenie 5.9. *Niech $T \in R_L$, $|L| = n$, e będzie krawędzią wewnętrzną w T oraz $T_e \in R_L$ będzie drzewem powstałym z T w wyniku ściągnięcia krawędzi e, wówczas $d_{MC}(T, T_e) \leq n - 1$.*

Dowód. Zauważmy, że T_e posiada jeden klaster mniej niż T. Niech $c \in \sigma_*(T) \setminus \sigma_*(T_e)$. Na podstawie lematu 3.1 otrzymujemy zatem $d_{MC}(T, T_e) = h_{MC}(c, O) = |c| \leq n - 1$. □

Twierdzenie 5.10. *Niech $T_1, T_2 \in R_L$, $|L| = n$, $A \subsetneq L$ oraz $|A| = n - 1$, wówczas zachodzą zależności:*

$$d_{MC}(T_1, T_2) \leq d_{MC}(T_{1|A}, T_{2|A}) + 2n - 4,$$
$$d_{MC}(T_1, T_2) \geq d_{MC}(T_{1|A}, T_{2|A}) - n + 1.$$

Dowód. Niech $L = A \cup \{x\}$. Rozważmy zmiany w zbiorze nietrywialnych klastrów pojawiające się po dołączeniu nowego liścia x do drzewa $T_{1|A}$. Klaster $s \in \sigma_*(T_{1|A})$ *przekształca* się w klaster $s' \in \sigma_*(T_1)$, gdy $s' = s$ lub $s' = s \cup \{x\}$, w przypadku jeśli $s \notin \sigma_*(T_1)$. Klastry s i s' będziemy nazywać *odpowiadającymi*. Ponadto dodatkowy element $O = \emptyset$ odpowiada sam sobie. Jeśli x jest dołączony w środku pewnej krawędzi drzewa $T_{1|A}$, wówczas pojawia się jeden *nowy* klaster $s_{new} \in \sigma_*(T_1)$, który nie jest klasterem odpowiadającym. Transformacja drzewa $T_{2|A}$ w T_2 odbywa się w analogiczny sposób. Stąd, jeśli $s, t \subseteq A$ oraz $s', t' \subseteq L$ są klastrami odpowiadającymi s, t w T_1 i T_2, to $h_{MC}(s, t) \leq h_{MC}(s', t') \leq h_{MC}(s, t) + 1$.

Rozważmy sparowanie $P = \{(s_i, t_i) : i = 1, \ldots, k\}$, gdzie $s_i \in \sigma_*(T_{1|A}) \cup \{O\}$, $t_i \in \sigma_*(T_{2|A}) \cup \{O\}$, $k = \max_{i=1,2} |\sigma_*(T_{i|A})| \leq n - 3$, powstające przy wyznaczaniu dystansu $d_{MC}(T_{1|A}, T_{2|A}) = \sum_{i=1}^{k} h_{MC}(s_i, t_i)$. Konstruujemy sparowanie składające się z par elementów $s'_i \in \sigma_*(T_1) \cup \{O\}$ i $t'_i \in \sigma_*(T_2) \cup \{O\}$ odpowiadających elementom sparowanym w P oraz, jeśli to konieczne, z pary klastrów (s_{new}, t_{new}), jeśli oba zbiory $\sigma_*(T_1)$ i $\sigma_*(T_2)$ zawierają nowe klastry, lub z pary (s_{new}, O) ewentualnie (O, t_{new}), jeśli dokładnie jeden ze zbiorów $\sigma_*(T_1)$, $\sigma_*(T_2)$ zawiera nowy klaster. Zauważmy, że jeśli zachodzi sytuacja, w której występuje para klastrów (s_{new}, t_{new}), wówczas $x \in s_{new} \cap t_{new}$, stąd $h_{MC}(s_{new}, t_{new}) \leq n - 1$. Podobnie jest w przypadkach, gdy występuje jeden nowy klaster: $h_{MC}(s_{new}, O) \leq n - 1$, $h_{MC}(t_{new}, O) \leq n - 1$. Zatem $d_{MC}(T_1, T_2) \leq d_{MC}(T_{1|A}, T_{2|A}) + 2n - 4$.

Rozważmy teraz sparowanie M elementów $\sigma_*(T_1) \cup \{O\}$ z elementami $\sigma_*(T_2) \cup \{O\}$ analogiczne do najlżejszego doskonałego skojarzenia definiującego odległość MC między drzewami T_1 i T_2. Przekształcamy M w sparowanie M' pomiędzy elementami zbiorów $\sigma_*(T_{1|A}) \cup \{O\}$ i $\sigma_*(T_{2|A}) \cup \{O\}$, zastępując elementy we wszystkich parach, w których nie występują nowe klastry, elementami im odpowiadającymi (pary zawierające nowe klastry są pomijane przy transformacji, stąd nie występują one w M'). Możliwe są następujące cztery sytuacje.

Przypadek 1. Istnieją dwa nowe klastry $s_{new} \in \sigma_*(T_1)$, $t_{new} \in \sigma_*(T_2)$, które tworzą parę $(s_{new}, t_{new}) \in M$. Wówczas zachodzi $d_{MC}(T_1, T_2) = \sum_{(s,t)\in M} h_{MC}(s, t) \geq \sum_{(s,t)\in M'} h_{MC}(s, t) \geq d_{MC}(T_{1|A}, T_{2|A})$.

Przypadek 2. Nowe klastry nie występują ani w $\sigma_*(T_1)$, ani $\sigma_*(T_2)$, wówczas $d_{MC}(T_1, T_2) \geq d_{MC}(T_{1|A}, T_{2|A})$ podobnie jak w powyższym przypadku.

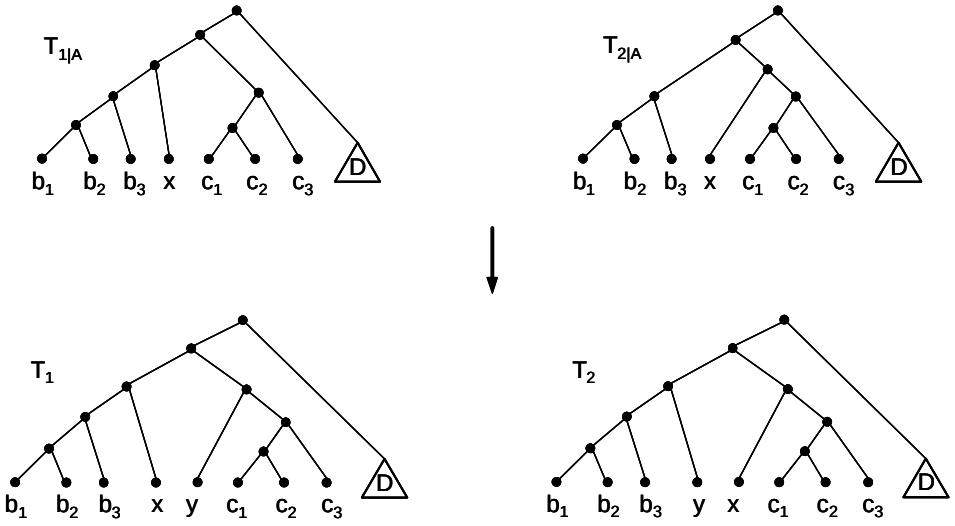
Przypadek 3. Występują dwa nowe klastry $s_{new} \in \sigma_*(T_1)$, $t_{new} \in \sigma_*(T_2)$, lecz nie tworzą one pary w M, tj. $(s_{new}, u_2), (u_1, t_{new}) \in M$. Wówczas rozszerzamy M' o parę (u'_1, u'_2), gdzie u'_i jest elementem odpowiadającym u_i, stąd $d_{MC}(T_1, T_2) = \sum_{(s,t)\in M} h_{MC}(s, t) \geq \sum_{(s,t)\in M'} h_{MC}(s, t) - h_{MC}(u'_1, u'_2) \geq d_{MC}(T_{1|A}, T_{2|A}) - n + 1$.

Przypadek 4. Istnieje dokładnie jeden nowy klaster. Bez straty ogólności możemy założyć, że występuje on w drzewie T_1, tj. $s_{new} \in \sigma_*(T_1)$

oraz $(s_{new}, u) \in M$. Wówczas rozszerzamy M' o parę (O, u'), gdzie u' jest elementem odpowiadającym u, stąd $d_{MC}(T_1, T_2) = \sum_{(s,t)\in M} h_{MC}(s,t) \geq \sum_{(s,t)\in M'} h_{MC}(s,t) - h_{MC}(u', O) \geq d_{MC}(T_{1|A}, T_{2|A}) - n + 2$. □

Podobnie jak w przypadku metryki MS, także dla MC obserwujemy możliwość zmniejszenia dystansu między drzewami po dołączeniu nowego liścia. Sytuacja taka nie ma miejsca w przypadku metryki RFC, lecz występuje ona jednak dla odległości węzłowych dla drzew ukorzenionych.

W celu zilustrowania tej własności rozważmy drzewa przedstawione na rysunku 5.8. Odległość RFC wzrosła $d_{RFC}(T_{1|A}, T_{2|A}) = 1$, $d_{RFC}(T_1, T_2) = 2$, odległość MC zmalała niezależnie od liczby liści w poddrzewie D, tj. $d_{MC}(T_{1|A}, T_{2|A}) = 6$, lecz $d_{MC}(T_1, T_2) = 4$. Wartość odległości SN dla drzew $T_{1|A}$ i $T_{2|A}$ zależy od liczby liści w poddrzewie D, $d^2_{SN}(T_{1|A}, T_{2|A}) = \sqrt{6|D| + 24}$. W przypadku drzew T_1 i T_2 natomiast otrzymujemy wartość niezależną od $|D|$ wynoszącą $\sqrt{24}$.



Rysunek 5.8: Drzewa, dla których dodanie liścia powoduje zmniejszenie odległości MC.

Lemat 5.11. *Rozważmy sekwencję nietrywialnych klastrów drzewa* $T \in R_n^B$ *posortowaną niemalejąco według ich mocy. Suma mocy pierwszych* $\left\lfloor \frac{n}{2} \right\rfloor$

klastrów w tym ciągu wynosi co najwyżej $\frac{n^2}{8} + O(n)$.

Dowód. Na podstawie lematu 4.16 wszystkie wybrane (tj. pierwsze $\left\lfloor \frac{n}{2} \right\rfloor$ w ciągu) klastry A, z wyjątkiem co najwyżej jednego, którego nie będziemy dalej rozważać, spełniają warunek $|A| \leq \left\lfloor \frac{n}{2} \right\rfloor$. Usuwamy z T wszystkie wierzchołki wewnętrzne, których klastry nie będą rozważane. Usuwamy również wszystkie wierzchołki izolowane. W rezultacie T rozpada się na sumę x rozłącznych ukorzenionych drzew binarnych posiadających n_i wierzchołków wewnętrznych ($i = 1, \ldots, x$). Suma szacowana w niniejszym lemacie, oznaczona dalej jako S, odpowiada sumie mocy klastrów związanych z wierzchołkami wewnętrznymi otrzymanych w ten sposób drzew. Mamy więc $\sum_{i=1}^{x} n_i \leq \left\lfloor \frac{n}{2} \right\rfloor$ oraz $n_i \leq \left\lfloor \frac{n}{2} \right\rfloor - 1$ dla $i = 1, \ldots, x$. Na podstawie lematu 4.15 otrzymujemy ostatecznie, że $S \leq \sum_{i=1}^{x} S(n_i + 1) \leq S(\left\lfloor \frac{n}{2} \right\rfloor + 1) = \frac{n^2}{8} + O(n)$. □

Twierdzenie 5.12. *Maksymalna odległość w metryce MC spełnia poniższe zależności:*

$$\frac{n^2 - 4 - (n \mod 2)}{2} \leq \Delta_{d_{MC}}(R_n^B) \leq \Delta_{d_{MC}}(R_n) \leq \frac{3}{4}n^2 + O(n).$$

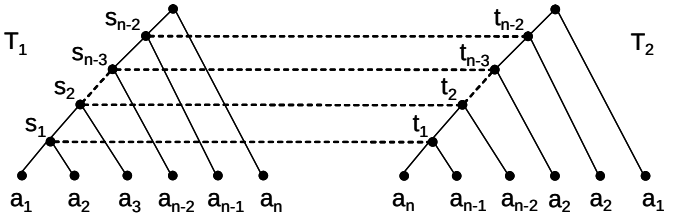
Dowód. Dolne ograniczenie jest realizowane przez dwie ukorzenione binarne gąsienice $T_1, T_2 \in R_n^B$ otrzymane z tej samej nieukorzenionej gąsienicy $T \in U_n^B$ w wyniku jej ukorzenienia na najbardziej odległych od siebie krawędziach, tj. na krawędziach tworzących dwie wiśnie w T (rysunek 5.9). Zauważmy, że w tym przypadku na mocy wniosku 5.2 istnieje tylko jedno sparowanie klastrów $M = \{\{s_i, t_i\}\}_{i=1,\ldots,n-2}$ spełniające warunki twierdzenia 5.1, stąd zaś łatwo możemy wyznaczyć wartość $d_{MC}(T_1, T_2) \geq 2\sum_{i=1}^{\lfloor n/2 \rfloor - 1}(i+1) + 2\sum_{i=1}^{\lceil n/2 \rceil - 1} i \geq (n^2 - 4 - (n \mod 2))/2$.

Niech $T_3, T_4 \in R_n^B$ będą dowolnymi ukorzenionymi drzewami binarnymi. W celu otrzymania ograniczenia górnego sortujemy elementy zbiorów $\sigma_*(T_3)$ i $\sigma_*(T_4)$ w porządku niemalejącym względem ich mocy, tworząc dwie sekwencje. Następnie, łącząc elementy występujące na zgodnych pozycjach w obu sekwencjach, tworzymy pewne sparowanie. Waga odpowiadającego przedstawionemu sparowaniu skojarzenia na podstawie lematu 5.11 może być oszacowana przez $2(\frac{n^2}{8} + O(n)) + (n - 2 - \left\lfloor \frac{n}{2} \right\rfloor)n = \frac{3}{4}n^2 + O(n)$.

Podobnie jak w przypadku twierdzenia 4.18 pokażemy teraz, że górne ograniczenie jest również prawdziwe dla dowolnych drzew $T_5, T_6 \in R_n$. Na podstawie T_5, T_6, poprzez rozwiązywanie multifurkacji w dowolny sposób, konstruujemy odpowiednie drzewa binarne $T_5', T_6' \in R_n^B$. Mamy zatem $\sigma_*(T_i) \subseteq \sigma_*(T_i')$ dla $i = 5, 6$. Sortujemy elementy zbiorów $\sigma_*(T_5')$ i $\sigma_*(T_6')$ w porządku niemalejącym względem ich mocy. Niech $M' = \{(s_i', t_i') : i = 1, \ldots, n-2\}$, gdzie $s_i' \in \sigma_*(T_5')$, $t_i' \in \sigma_*(T_6')$ będzie sparowaniem klastrów występujących na tych samych pozycjach w posortowanych sekwencjach. Na podstawie M' skonstruujemy teraz sparowanie $M = \{(s_i, t_i) : i = 1, \ldots, n-2\}$, gdzie $s_i \in \sigma_*(T_5) \cup \{\emptyset\}$, $t_i \in \sigma_*(T_6) \cup \{\emptyset\}$, zastępując klastry niewystępujące w zbiorach $\sigma_*(T_5)$ i $\sigma_*(T_6)$ zbiorem pustym. Zatem

$$\begin{aligned} d_{MC}(T_5, T_6) &\leq \sum_{(s,t)\in M} h_{MC}(s,t) \leq \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (|s_i| + |t_i|) + \left(n - 2 - \left\lfloor \frac{n}{2} \right\rfloor\right) n \\ &\leq \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} |s_i'| + \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} |t_i'| + \left(n - 2 - \left\lfloor \frac{n}{2} \right\rfloor\right) n \leq \frac{3}{4} n^2 + O(n). \end{aligned}$$

□



Rysunek 5.9: Przykład odległych drzew w metryce MC.

Zauważmy, że średnica zbioru n-listnych drzew ukorzenionych w metryce MC jest większa niż średnica zbioru drzew nieukorzenionych w MS, która wynosi $\frac{3}{8}n^2 \pm O(n)$ (por. tw. 4.18). Można postawić hipotezę, że zachodzi równość $\max_{T_1,T_2 \in R_n} d_{MC}(T_1, T_2) = \frac{1}{2}n^2 \pm O(n)$.

Metryka MC wykazuje podobne zalety jak MS, tj. zmiana odległości spowodowana niewielkimi modyfikacjami drzewa, polegającymi np. na dodaniu liścia czy ściągnięciu krawędzi, jest mała w porównaniu do średnicy.

Zachowanie to obrazuje jednocześnie istotną zaletę metryki MC w stosunku do RFC, w której, podobnie jak dla RF, nawet dołączenie jednego liścia może spowodować wzrost odległości do poziomu wartości maksymalnej. Przykładem opisanej sytuacji są drzewa przedstawione na rysunku 5.1, gdzie $d_{MC}(T_1, T_2) = 2n - 4$, podczas gdy $d_{RFC}(T_1, T_2) = \Delta_{RFC}(R_n^B) = n - 2$.

5.4 Regularność przestrzeni z metryką MC

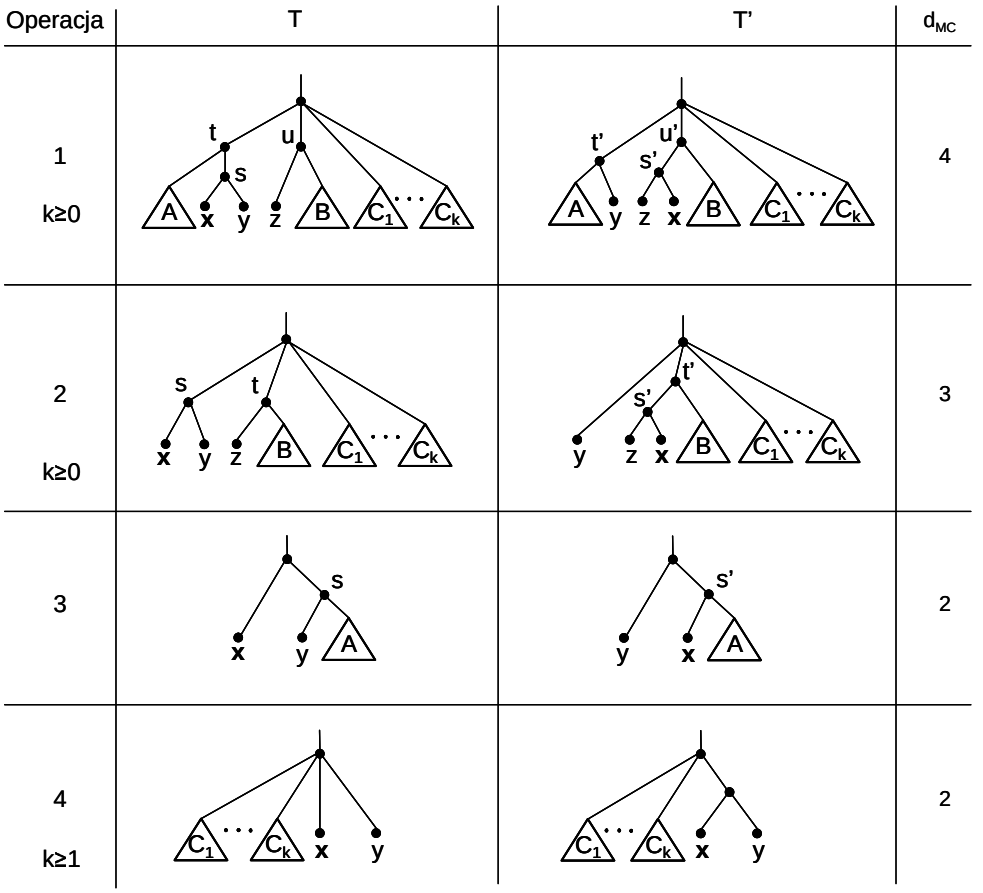
W niniejszym podrozdziale rozważymy problem dotyczący istnienia izolowanych wysp dla metryki MC. Podobnie jak dla odległości MS, w metryce MC izolowane wyspy nie występują.

Twierdzenie 5.13. *Dla dowolnych $T_a, T_b \in R_L$ istnieje sekwencja drzew $T_a = T_1, T_2, \ldots, T_{k-1}, T_k = T_b$, $T_i \in U_L$, $i = 1, \ldots, k$, taka że zachodzi $d_{MC}(T_j, T_{j+1}) \leq 4$, gdzie $j = 1, \ldots, k-1$. Ponadto, jeśli dodatkowo $T_a, T_b \in R_L^B$, wówczas istnieje sekwencja o opisanych powyżej własnościach, składająca się wyłącznie z drzew binarnych.*

Dowód. Dowód przebiega się analogicznie jak dla twierdzenia 4.19 z tą różnicą, że nie ma potrzeby wprowadzania do drzew korzenia (drzewa są już ukorzenione). Transformacja odbywa się w porządku „od dołu do góry" za pomocą operacji zaprezentowanych na rysunku 5.10. W przypadku drzew binarnych $T_a, T_b \in R_L^B$ operacja 4 nie jest wykorzystywana. □

Izolowane wyspy nie występują również w przypadku metryki RFC oraz dla metryk opartych na operacjach edycyjnych (z tych samych względów, dla których nie występują one dla nieukorzenionych odpowiedników podanych odległości, por. podrozdział 4.5).

Wyspy występują jednak w przypadku metryki tripletowej oraz odległości SN. Przykładem wyspy w obu tych metrykach jest ukorzenione drzewo $T \in R_n$ o topologii gwiazdy. Zauważmy, iż wprowadzenie wierzchołka wewnętrznego powoduje, że odległość między co najmniej $2(n-2)$ parami liści wzrośnie o 1. Stąd zaś mamy, że odległość w metryce SN między T a dowolnym innym drzewem $T' \in R_n$ różnym od T spełnia



Rysunek 5.10: Lokalne modyfikacje w metryce MC.

$d^2_{SN}(T,T') \geq \sqrt{2(n-2)}$ i rośnie wraz ze wzrostem n. Ta sama gwiazda jest również wyspą w przypadku metryki tripletowej, ponieważ wskutek wprowadzenia wierzchołka wewnętrznego co najmniej $n-2$ nierozwiązanych tripletów staje się rozwiązanymi, stąd $d_{TT}(T,T') \geq n-2$.

5.5 Związek metryki MC z MS

Twierdzenie 5.14. *Dane są drzewa $T_1, T_2 \in U_L^B$. Niech $T_1', T_2' \in R_L^B$ będą drzewami otrzymanymi odpowiednio z T_1 i T_2 w wyniku operacji ukorzeniania, wówczas zachodzi zależność:*

$$d_{MC}(T_1', T_2') \geq d_{MS}(T_1, T_2).$$

Dowód. Rozważmy drzewa $T_1, T_2 \in U_L^B$ oraz ich ukorzenienia w środku krawędzi e_1, e_2 odpowiadających kolejno rozbiciom $A|B$ i $C|D$. Zauważmy, że każdej krawędzi $e \in E(T_1) \setminus \{e_1\}$ odpowiada dokładnie jeden klaster ze zbioru $\sigma(T_1') \setminus \{A, B, L\}$, gdzie elementami odpowiadającymi są klaster X i rozbicie $X|L \setminus X$. Analogiczna relacja może być wprowadzona dla klastrów $Y \in \sigma(T_2') \setminus \{C, D, L\}$ i rozbić $Y|L \setminus Y \in \beta(T_2) \setminus \{C|D\}$. Ponadto zachodzi $|X \oplus Y| \geq h_{MS}(X|L \setminus X, Y|L \setminus Y)$.

Rozważmy sparowanie M' pomiędzy klastrami ze zbiorów $\sigma(T_1') \setminus \{L\}$ oraz $\sigma(T_2') \setminus \{L\}$, które odpowiada najlżejszemu doskonałemu skojarzeniu definiującemu odległość $d_{MC}(T_1', T_2') = \sum_{(c,c') \in M'} |c \oplus c'|$ oraz spełnia warunki twierdzenia 5.1. Celem jest skonstruowanie sparowania M rozbić ze zbioru $\beta(T_1)$ z rozbiciami z $\beta(T_2)$ o wadze $\sum_{(s,s') \in M} h_{MS}(s, s') \leq d_{MC}(T_1', T_2')$. W pierwszym kroku dla każdej pary $(X_1, X_2) \in M'$, takiej że $X_1 \notin \{A, B\}$ oraz $X_2 \notin \{C, D\}$, dołączmy do M parę $(X_1|L \setminus X_1, X_2|L \setminus X_2)$ składającą się z rozbić odpowiadających wybranym klastrom. W zależności od pozostałych elementów w M', tzn. par zawierających klastry A, B, C, D w M', możliwe są następujące trzy sytuacje.

Przypadek 1. Każdy z klastrów A, B jest sparowany z pewnym klastrem ze zbioru $\{C, D\}$. Powiększamy M o parę $(A|B, C|D)$, co kończy dowód.

Przypadek 2. Tylko jeden klaster z $\{A, B\}$ jest sparowany z klastrem z $\{C, D\}$, np. $(A, C) \in M'$. Stąd mamy $(B, Y), (X, D) \in M'$, gdzie $X \in \sigma(T_1)$, $Y \in \sigma(T_2)$. W tej sytuacji dołączamy do M dwie pary: $(A|B, Y|L \setminus Y)$ i $(X|L \setminus X, C|D)$.

Przypadek 3. Żaden z klastrów A, B nie jest sparowany z klastrami z $\{C, D\}$. Ponieważ M' spełnia warunki twierdzenia 5.1, możliwe są wyłącznie następujące dwa sparowania: $m' = \{(A, c), (a, D), (B, d), (b, C)\} \subseteq M'$ lub wariant symetryczny $\{(A, d), (a, C), (B, c), (b, D)\} \subseteq M'$, gdzie $a \subsetneq A$, $b \subsetneq B$, $c \subsetneq C$ oraz $d \subsetneq D$. Rozważmy pierwszy przypadek, drugi jest analogiczny. Konstruujemy dwa sparowania rozbić:

$$m_1 = \{(a|L \setminus a, c|L \setminus c), (b|L \setminus b, C|D), (A|B, d|L \setminus d)\},$$
$$m_2 = \{(a|L \setminus a, C|D), (b|L \setminus b, d|L \setminus d), (A|B, c|L \setminus c)\}$$

oraz dwa sparowania klastrów:

$$m'_1 = \{(a, c), (b, C), (B, d)\},$$
$$m'_2 = \{(a, D), (b, d), (A, c)\}.$$

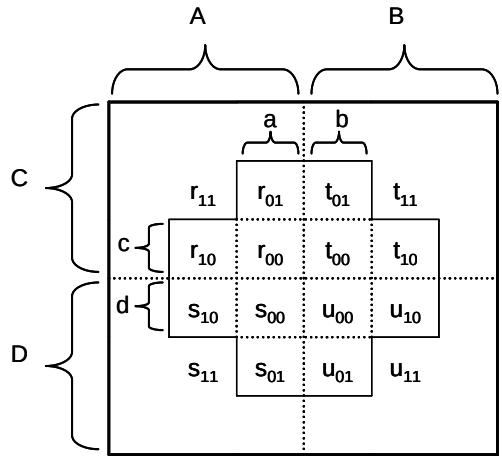
Dla $i = 1, 2$ zachodzi:

$$\sum_{(s,s') \in m_i} h_{MS}(s, s') \leq \sum_{(c,c') \in m'_i} h_{MC}(c, c').$$

Na podstawie obliczeń całkowitych wag sparowań m', m'_1, m'_2 umieszczonych w tabeli 5.1 (por. również diagram Venna na rysunku 5.11) otrzymujemy, że zawsze przynajmniej jedna z nierówności $\sum_{(c,c') \in m'_i} |c \oplus c'| \leq \sum_{(c,c') \in m'} |c \oplus c'|$, $i = 1, 2$ jest prawdziwa. Wynika to z faktu, że w kolumnie $\Delta m'_1$ jest tylko jedna dodatnia wartość, która znajduje się w wierszu r_{10}, zaś jedyna dodatnia wartość w kolumnie $\Delta m'_2$ jest umieszczona w innym wierszu: u_{10}. Zatem możemy rozszerzyć M' o lżejsze spośród dwóch sparowań m_1 i m_2. □

Nierówność z twierdzenia 5.14 może być ostra. Przykładem takiej sytuacji są drzewa przedstawione na rysunku 5.12.

Zauważmy, że zależność ta pozwala na wprowadzenie interesującej interpretacji odległości $d_{MC}(T'_1, T'_2)$ dla drzew binarnych, której wartość możemy przedstawić jako sumę dwóch składników. Pierwszy składnik równy $d_{MS}(T_1, T_2)$, gdzie T_1, T_2 są drzewami powstałymi z T'_1, T'_2 w wyniku usunięcia korzenia i utożsamienia krawędzi z nim incydentnych, mierzy różnicę w topologii nieukorzenionych odpowiedników porównywanych drzew,



Rysunek 5.11: Diagram Venna zbioru L obrazujący możliwe przecięcia zbiorów: A, B, C, D, a, b, c, d.

drugi zaś $d_{MC}(T_1', T_2') - d_{MS}(T_1, T_2)$ wprowadza dodatkową wartość związaną z różnicą w kierunku przepływu czasu wzdłuż krawędzi, wprowadzoną w drzewa T_1', T_2' podczas operacji ukorzeniania. Drugi składnik może przyjmować duże wartości, zbliżone nawet do wartości średnicy, np. dla dwóch identycznych nieukorzenionych gąsienic (będących zatem w odległości 0 w MS) po wprowadzeniu korzenia na skrajnych krawędziach wiszących otrzymamy odległość MC rzędu $\sim \frac{1}{2}n^2$ (por. dowód twierdzenia 5.12).

Z drugiej strony, dla danych drzew $T_1, T_2 \in U_L^B$, możliwe jest na ogół ustalenie kierunku przepływu czasu, tj. umiejscowienie korzenia w taki sposób, że składnik $d_{MC}(T_1', T_2') - d_{MS}(T_1, T_2)$ przyjmuje niewielką wartość. Eksperymenty komputerowe przy użyciu losowych drzew nieukorzenionych (każde drzewo ma jednakowe prawdopodobieństwo wylosowania) wykazują, że wartość $\Delta_{MC}(T_1, T_2) = \min d_{MC}(T_1', T_2') - d_{MS}(T_1, T_2)$, gdzie minimum jest brane po wszystkich możliwych miejscach ukorzenienia (korzeń jako nowy wierzchołek jest wprowadzany na krawędziach T_1 i T_2), jest mała w porównaniu do wartości średniej odległości MC tych drzew. Wyniki eksperymentu zostały przedstawione w tabeli 5.2. W przypadku drzew o rozmiarze nieprzekraczającym 8 liści (pierwsza cześć tabeli) prezentowane wartości są wyznaczone na podstawie porównań wszystkich możliwych

Tabela 5.1: Całkowity koszt sparowań m', m'_1, m'_2. Liczby w wierszach odpowiadają krotności, z którą moc danego zbioru występuje w całkowitym koszcie analizowanych sparowań, np. wartość 2 w wierszu r_{00}, w kolumnie m' oznacza, że składnik $2|r_{00}|$ występuje w wyrażeniu opisującym wagę m'. Kolumny $\Delta m'_1 = m'_1 - m'$ i $\Delta m'_2 = m'_2 - m'$ zawierają różnice pomiędzy odpowiednimi wierszami w kolumnie m' oraz odpowiednio m'_1 i m'_2.

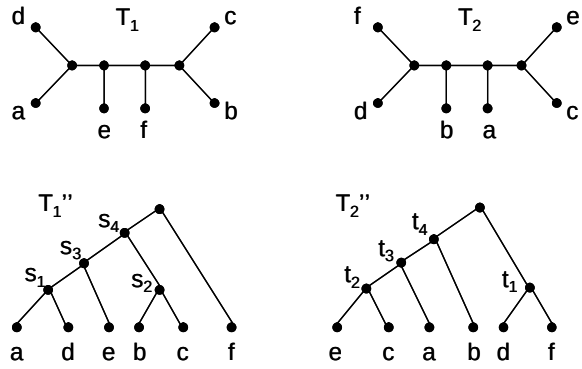
Zbiór	Definicja zbioru	m'	m'_1	m'_2	$\Delta m'_1$	$\Delta m'_2$
r_{00}	$a \cap c$	2	1	1	-1	-1
r_{01}	$(a \cap C) \setminus (a \cap c)$	3	2	2	-1	-1
r_{10}	$(A \cap c) \setminus (a \cap c)$	1	2	0	**1**	-1
r_{11}	$(A \cap C) \setminus (a \cup c)$	2	1	1	-1	-1
s_{00}	$a \cap d$	2	2	2	0	0
s_{01}	$(a \cap D) \setminus (a \cap d)$	1	1	1	0	0
s_{10}	$(A \cap d) \setminus (a \cap d)$	3	1	3	-2	0
s_{11}	$(A \cap D) \setminus (a \cup d)$	2	0	2	-2	0
t_{00}	$b \cap c$	2	2	2	0	0
t_{01}	$(b \cap C) \setminus (b \cap c)$	1	1	1	0	0
t_{10}	$(B \cap c) \setminus (b \cap c)$	3	3	1	0	-2
t_{11}	$(B \cap C) \setminus (b \cup c)$	2	2	0	0	-2
u_{00}	$b \cap d$	2	1	1	-1	-1
u_{01}	$(b \cap D) \setminus (b \cap d)$	3	2	2	-1	-1
u_{10}	$(B \cap d) \setminus (b \cap d)$	1	0	2	-1	**1**
u_{11}	$(B \cap D) \setminus (b \cup d)$	2	1	1	-1	-1

par $T_1, T_2 \in U_n^B$. W przypadku większych drzew (druga cześć tabeli) wyniki zamieszczone w każdym z wierszy pochodzą z analizy 10000 par drzew losowych.

Twierdzenie 5.14 nie jest jednak prawdziwe, gdy zrezygnujemy z ograniczania się do drzew binarnych. Kontrprzykładem w takiej sytuacji są drzewa przedstawiane na rysunku 5.13, gdzie $d_{MS}(T_1, T_2) = 2$, podczas gdy $d_{MC}(T'_1, T'_2) = 1$.

Tabela 5.2: Wartości parametru $\Delta_{MC}(T_1, T_2)$.

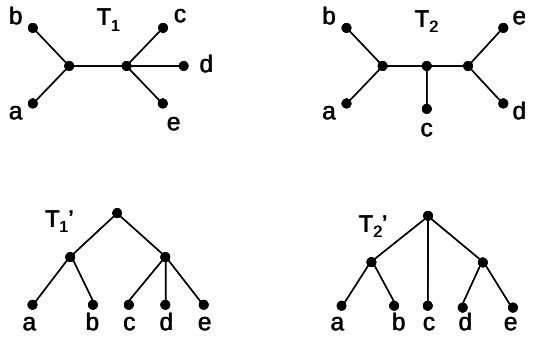
n	średnie $\Delta_{MC}(T_1, T_2)$	max $\Delta_{MC}(T_1, T_2)$	średnie $d_{MC}(T_1', T_2')$
4	0	0	3.360
5	0	0	5.971
6	0.032	1	9.105
7	0.054	1	12.822
8	0.029	2	16.922
10	0.058	2	26.570
20	0.333	4	97.205
30	0.691	4	198.785
40	0.965	5	325.445
50	1.173	5	474.145



Rysunek 5.12: Para najmniejszych drzew T_1, T_2, dla których nierówność z tw. 5.14 jest ostra: $d_{MS}(T_1, T_2) = 6$, ale $\min d_{MC}(T_1', T_2')$ wzięte po wszystkich możliwych ukorzenieniach T_1 i T_2 wynosi $d_{MC}(T_1'', T_2'') = 7$.

Zaprezentowana własność (w odniesieniu do drzew binarnych) jest bardzo interesująca i motywuje do szukania odpowiedzi na pytanie o istnienie innych przykładów par metryk posiadających podobną cechę. Kolejne pytanie może dotyczyć istnienia metryk, dla których własność ta jest praw-

dziwa również w przypadku drzew dowolnych (niekoniecznie binarnych) lub takich par, dla których zawsze istnieje takie ukorzenienie, że wartości odległości miedzy drzewami nieukorzenionymi i pewnymi ich ukorzenionymi odpowiednikami są równe.



Rysunek 5.13: Przykład drzew niebinarnych, dla których zachodzi relacja $d_{MS}(T_1, T_2) > d_{MC}(T_1', T_2')$.

5.6 Podsumowanie własności przestrzeni metrycznej MC

Z uwagi na zbliżony charakter definicji oraz podobne zachowanie w przeanalizowanych sytuacjach, odległość MC może być traktowana jako ukorzeniony odpowiednik metryki MS.

Własności metryki MC dla drzew binarnych:

1. oszacowanie przez RFC, tw. 5.5:

$$d_{RFC}(T_1, T_2) + 1 \leq d_{MC}(T_1, T_2) \leq (n-1) d_{RFC}(T_1, T_2), \text{ dla } T_1 \neq T_2,$$

2. minimalna dodatnia odległość, tw. 5.5:

$$\min_{T_1 \neq T_2 \in R_n^B} d_{MC}(T_1, T_2) = 2,$$

3. rozmiar sąsiedztwa, tw. 5.7: co najwyżej $n-1$, zależny od topologii drzewa,

4. średnica, tw. 5.12:

$$\frac{n^2-4-(n \mod 2)}{2} \leq \max_{T_1,T_2 \in R_n^B} d_{MC}(T_1,T_2) \leq \frac{3}{4}n^2 + O(n).$$

Własności metryki MC dla drzew dowolnych:

1. oszacowanie przez RFC, tw. 5.4:

$$d_{RFC}(T_1,T_2) \leq d_{MC}(T_1,T_2) \leq 2(n-1)\, d_{RFC}(T_1,T_2),$$

2. minimalna dodatnia odległość, tw. 5.8:

$$\min_{T_1 \neq T_2 \in R_n} d_{MC}(T_1,T_2) = 1,$$

3. rozmiar sąsiedztwa, wniosek 5.8: $O(n^2)$, zależny od topologii drzewa,

4. średnica, tw. 5.12:

$$\frac{n^2-4-(n \mod 2)}{2} \leq \max_{T_1,T_2 \in R_n} d_{MC}(T_1,T_2) \leq \frac{3}{4}n^2 + O(n).$$

Metryka MC w zbiorze drzew binarnych wykazuje ciekawe powiązanie z odległością MS, tj. odległość MC dowolnych drzew z R_L^B jest co najmniej równa odległości ich nieukorzenionych odpowiedników (tw. 5.14). Dodatkowo, w odróżnieniu od MS, posiada ona bardzo interesującą własność: sparowanie wierzchołków w porównywanych drzewach wyznaczone przez metrykę MC charakteryzuje się bardzo intuicyjnymi cechami przedstawionymi szczegółowo w tw. 5.1.

W literaturze (praca [22] Boormana i Oliviera) obecna jest definicja dystansu zbliżona do MC. Obie te metryki są tożsame w zbiorze ukorzenionych drzew binarnych, lecz różnią się istotnie w sposobie porównywania drzew niebinarnych. Definicja zaproponowana w pracy [22] nie korzysta z koncepcji elementu dodatkowego, w miejsce której pojawia się podejście

polegające na powielaniu klastra związanego z danym węzłem v posiadającym k dzieci $k-1$ razy.

Pierwsze wyniki i publikacje odnośnie metryk skojarzeniowych (wykonane w ramach pracy nad niniejszą rozprawą) ukazały się jeszcze przed pozyskaniem informacji o tematyce i zawartości pozycji [22]. Metryki w pracy Boormana i Oliviera pojawiają się głównie w kontekście związanym z hierarchicznymi metodami klasteryzacji. Zastosowanie tych metod dla drzew filogenetycznych nie zostało tam wspomniane. Proponowane w [22] metryki dotyczą wyłącznie drzew ukorzenionych. Jedyną częścią wspólną analizy tam zawartej z wynikami prezentowanymi w niniejszej pracy (poza definicją odległości MC dla drzew binarnych) jest odpowiednik lematu 3.1 dla MC.

5.7 Problem mediany dla metryki MC

Jak już zostało wspomniane w pierwszym rozdziale pracy, istnieje wiele metod tworzenia drzew filogenetycznych i często zdarza się, że różne algorytmy generują różne drzewa dla tych samych danych wejściowych. Z drugiej strony już niektóre metody rekonstrukcji, np. metoda maksymalnej parsymonii, generują wyniki w postaci nie jednego, lecz zbioru drzew. Jedną z metod analizy i interpretacji wyników zawartych w zbiorze drzew filogenetycznych jest konstruowanie jednego drzewa, które, wg ustalonego kryterium, najlepiej reprezentuje informacje zgromadzone w drzewach wejściowych. Problem ten, wraz z metodą wyznaczania drzewa o podanych własnościach, został po raz pierwszy sformułowany przez Edwarda N. Adamsa III w pracy [1]. Podążając za [1] drzewo posiadające opisaną cechę nazywamy *drzewem konsensusu* lub krócej *konsensusem*.

W literaturze zaproponowano wiele metod wyznaczania konsensusu, których wyczerpujący i usystematyzowany opis można znaleźć w [27]. Obecnie pojawiają się także różne modyfikacje tego zagadnienia, polegające na wyborze pewnego niewielkiego podzbioru drzew najlepiej reprezentującego informacje zgromadzone w większym zbiorze, wykorzystując do tego celu np. algorytmy kolorowania grafów [21] lub metody klastery-

zacji [107].

Poniżej przedstawione zostaną trzy względnie proste i podstawowe podejścia do standardowego zagadnienia konsensusu oraz ich relacje z problemem mediany w metryce MC. Ogół drzew będących wejściem dla algorytmu wyznaczania konsensusu określa się mianem *profilu* $P = \{T_1, \ldots, T_k\}$, gdzie $T_i \in R_L$ dla $i = 1, \ldots, k$. Drzewa tworzące profil P nie muszą być różne, stąd P nie jest zbiorem lecz multizbiorem.

Definicja 5.1 (Konsensus ścisły)**.** Dany jest profil P drzew z R_L. Drzewo $T \in R_L$ jest *konsensusem ścisłym* (ang. *Strict Consenus Tree*) dla P, jeśli zbiór $\sigma(T)$ zawiera dokładnie wszystkie te klastry, które są wspólne dla drzew z P.

Zatem wszystkie drzewa profilu P rozszerzają konsensus ścisły dla P.

Definicja 5.2 (Konsensus większościowy)**.** Dany jest profil P drzew z R_L. Niech $\sigma_{cons}(P)$ będzie zbiorem wszystkich tych klastrów, które występują w więcej niż połowie drzew z P. Drzewo $T \in R_L$ jest *konsensusem większościowym* (ang. *Majority Rule Tree*) dla P, jeśli $\sigma(T) = \sigma_{cons}(P)$.

Definicja 5.3. Dany jest profil $P = \{T_1, \ldots, T_k\}$, gdzie $T_i \in R_L$, dla $i = 1, \ldots, k$. Niech $d : R_L \times R_L \to \mathbb{R}_{\geq 0}$ będzie metryką w R_L. Drzewo $T \in R_L$ jest *medianą dla P w metryce d*, jeśli T minimalizuje sumę

$$D_d(T, P) = \sum_{i=1}^{k} d(T, T_i).$$

Zbiór $M_d(P) = \{T : D_d(T, P) = \min_{T' \in R_L} D_d(T', P)\}$ jest zbiorem *median dla P w metryce d*.

Definicja 5.4 (Mediana standardowa)**.** Dany jest profil P drzew z R_L. Drzewo $T \in R_L$ jest *medianą* (ang. *Median Consensus Tree*) dla P, jeśli T jest medianą dla P w metryce RFC.

Mediana zdefiniowana za pomocą metryki RFC jest ściśle związana z konsensusem większościowym. Jeśli liczba drzew k w profilu P jest nieparzysta, wówczas drzewo konsensusu większościowego dla P jest równocześnie jego medianą. Jeśli zaś k jest parzyste, wtedy drzewo konsensusu

większościowego T jest nadal medianą oraz medianą jest także każde inne drzewo, które zawiera klastry ze zbioru $\sigma(T)$ oraz dowolne klastry występujące w dokładnie połowie drzew z P [10].

Jeśli profil P składa się z co najmniej trzech drzew binarnych oraz wymagamy by mediana była również drzewem binarnym (tj. szukamy mediany w zbiorze R_L^B a nie R_L), problem ten staje się NP-trudny [74]. Istnieje jednak algorytm wielomianowy, jeśli P posiada dokładnie dwa elementy [26].

W dalszej części podrozdziału zaprezentowane zostaną wstępne rozważania dotyczące własności mediany w metryce MC.

Twierdzenie 5.15. *Dla $|L| \geq 4$ istnieją profile P składające się z drzew z R_L, takie że drzewo $T \in M_{MC}(P)$ zawiera klastry niewystępujące w drzewach z P.*

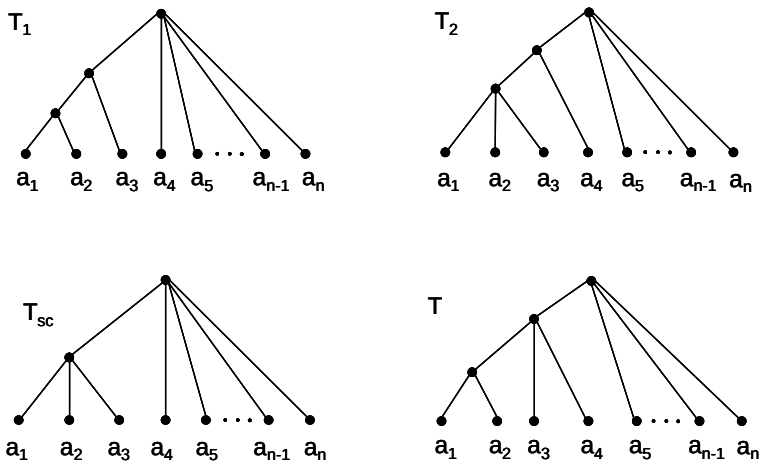
Dowód. Rozważmy zbiór $A = \{a_1, \ldots, a_k\} \subsetneq L$, taki że $3 \leq |A| \leq |L| - 1$ oraz k zbiorów $A_i = A \setminus \{a_i\}$. Konstruujemy k drzew T_i, z których każde posiada dokładnie jeden nietrywialny klaster A_i oraz drzewo T, takie że $\sigma_*(T) = \{A\}$. Niech $P = \{T_1 \ldots, T_k\}$. Wykażemy, że jedynym elementem zbioru $M_{MC}(P)$ jest T. Zauważmy, że $d_{MC}(T, T_i) = 1$ dla $i = 1, \ldots, k$. Mamy więc $D_{MC}(T, P) = k$ oraz $D_{MC}(T_i, P) = 2k - 2$ dla $i = 1, \ldots, k$. Zatem jedynym drzewem będącym medianą dla P w metryce MC jest T. □

Problemem otwartym pozostaje rozstrzygnięcie prawdziwości powyższego twierdzenia w przypadku, gdy ograniczymy P wyłącznie do drzew binarnych. Kolejne interesujące i powiązane pytanie otwarte brzmi następująco: czy mediana w MC dla profilu składającego się z drzew binarnych musi być również drzewem binarnym?

Poniższe twierdzenia ilustrują pewne podstawowe cechy mediany w metryce MC. W szczególności tw. 5.17 dowodzi faktu, że mediana w tej metryce może agregować informacje zawarte w profilu w sposób odmienny niż standardowe metody konstrukcji konsensusu.

Twierdzenie 5.16. *Dla $|L| \geq 4$ istnieją profile P składające się z drzew z R_L, takie że $|M_{MC}(P)| = \Omega(2^{|L|})$.*

Dowód. Rozważmy zbiór $A = \{a_1, \ldots, a_x\} \subseteq L$, taki że $|A| \geq 4$ i jego rozbicie na co najmniej dwuelementowe zbiory B, C. Niech $D \subsetneq A$, $|D| \geq 2$, wówczas $D \setminus C = D \cap B$ oraz $D \setminus B = D \cap C$, zatem zbiory $B \oplus D$ i $C \oplus D$ również tworzą rozbicie A. Stąd $h_{MC}(B, D) + h_{MC}(C, D) = x$. Konstruujemy dwa drzewa: $T_1 \in R_L$ posiadające dokładnie jeden nietrywialny klaster B oraz podobnie $T_2 \in R_L$ z jednym nietrywialnym klastrem C. Zauważmy, że każde drzewo $T \in R_L$ posiadające dokładnie jeden nietrywialny klaster $D \subsetneq A$ jest medianą dla $P = \{T_1, T_2\}$, gdyż $d_{MC}(T, T_1) + d_{MC}(T, T_2) = x$, a skoro zachodzi $d_{MC}(T_1, T_2) = x$, to x jest możliwie najmniejsze. □



Rysunek 5.14: Drzewo T_{sc} jest ścisłym konsensusem dla $P = \{T_1, T_2\}$; T jest medianą dla P w metryce MC.

Twierdzenie 5.17. *Rozważmy profile składające się z drzew z R_L, gdzie $|L| \geq 5$, wówczas:*

1. *istnieją profile P, takie że mediana dla P w metryce MC nie musi być rozszerzeniem ścisłego konsensusu dla P,*

2. *istnieją profile P, takie że dla każdego $T \in M_{MC}(P)$, T nie jest rozszerzeniem konsensusu większościowego dla P.*

Dowód.

1. Definiujemy zbiory $A_i = \{a_1, a_2, \ldots, a_i\} \subsetneq L$ dla $i = 2, 3, 4$ oraz dwa drzewa $T_1 \in R_L$ posiadające dokładnie dwa nietrywialne klastry A_2 i A_3 oraz $T_2 \in R_L$, gdzie $\sigma_*(T_2) = \{A_3, A_4\}$ (rysunek 5.14, $|L| = n$). Drzewo $T_{sc} \in R_L$ posiada dokładnie jeden nietrywialny klaster A_3 i jest konsensusem ścisłym dla $P = \{T_1, T_2\}$. Drzewo $T \in R_L$, takie że $\sigma_*(T) = \{A_2, A_4\}$ jest medianą dla P w MC, gdyż $d_{MC}(T_1, T_2) = d_{MC}(T, T_1) + d_{MC}(T, T_2) = 2$, lecz nie jest rozszerzeniem T_{sc}.

2. Dla $k \geq 2$ rozważmy profil P złożony z $2k+1$ drzew zdefiniowanych w poprzednim punkcie, zawierający k kopii drzewa T_1, k kopii drzewa T_2 oraz drzewo T. Jedyną medianą w MC dla P jest drzewo T, gdyż minimalizuje ono sumę odległości do drzew T_1, T_2 oraz jako jedyne jest w odległości 0 od T. Drzewo T nie zawiera jednak klastra A_3 obecnego we wszystkich, z wyjątkiem jednego, drzewach z P, stąd nie jest ono rozszerzeniem konsensusu większościowego dla P.

□

Drzewa T_1 i T_2 użyte w dowodzie punktu drugiego są również medianami dla $P = \{T_1, T_2\}$ w MC, lecz oba zawierają klaster A_3, zatem są rozszerzeniami drzewa ścisłego konsensusu T_{sc}. Fakt ten motywuje do postawienia pytania o istnienie takiego zestawu danych, dla którego żadne z drzew będących mediną w MC nie rozszerzałoby ścisłego konsensusu.

Przedstawione konstrukcje mogą być łatwo przeniesione i zastosowane dla metryki MS. W takim przypadku budowa odpowiednich drzew nieukorzenionych polega w głównej mierze na dołączaniu jednego lub wielu nowych liści do korzenia analizowanych w tym podrozdziale przykładów. Skonstruowany w opisany sposób odpowiednik twierdzenia 5.15 dla metryki MS można znaleźć w pracy [18]. Istotnym problemem otwartym pozostaje określenie statusu (tj. wielomianowości lub NP-trudności) zagadnienia wyznaczania mediany w metryce MC.

6 Własności metryk MC i MS dla drzew losowych

6.1 Modele losowe drzew filogenetycznych

Modele losowe generacji drzew filogenetycznych są istotną częścią analizy filogenetycznej. Znajomość rozkładu odległości w danej metryce dla drzew losowych wygenerowanych według konkretnego modelu umożliwia lepszą interpretację dystansu między badaną parą drzew. W szczególności często istotne jest czy odległość między danymi drzewami jest mniejsza, zbliżona, czy też większa niż średnia odległość dla drzew losowych. Przydatna jest także znajomość prawdopodobieństwa wygenerowania dwóch losowych drzew będących w odległości mniejszej lub równej zadanej wartości, gdyż pozwala to określić na ile otrzymana wartość dystansu wskazuje na wyższe niż tylko przypadkowe podobieństwo.

W analizie filogenetycznej najczęściej spotyka się dwa modele generowania drzew binarnych [73]: model drzew jednakowo prawdopodobnych (ang. *the Uniform Model* — UM, określany też w biologii jako model PDA) oraz model Yule'a (ang. *the Yule Model* — YM), zwany również modelem Yule'a-Hardinga [98]. W modelu drzew jednakowo prawdopodobnych (UM) prawdopodobieństwo wylosowania drzewa nie zależy od jego topologii, tj. dla $T \in U_n^B$, $T' \in R_n^B$ wynosi ono odpowiednio $\Pr(X = T) = \frac{1}{|U_n^B|}$ oraz $\Pr(X = T') = \frac{1}{|R_n^B|}$. W modelu Yule'a (YM) natomiast prawdopodobieństwa wylosowania różnych drzew mogą być inne.

Generacja n-listnego ($n \geq 2$) binarnego drzewa ukorzenionego w modelu Yule'a przebiega następująco. Zakładamy, że etykietami liści generowanego drzewa są liczby ze zbioru $\{1, \ldots, n\}$. Proces generacji rozpoczyna

się od utworzenia ukorzenionego drzewa 2-listnego, którego liście posiadają różne etykiety wybrane losowo (z jednakowym prawdopodobieństwem) z $\{1, \ldots, n\}$. Do momentu otrzymania drzewa n-listnego, powtarzana jest następująca procedura: w skonstruowanym do tej pory drzewie losujemy (z jednakowym prawdopodobieństwem) krawędź wiszącą, następnie do nowo utworzonego wierzchołka stopnia dwa na tej krawędzi dołączamy kolejną krawędzią liść o etykiecie wylosowanej (z jednakowym prawdopodobieństwem) z $\{1, \ldots, n\}$ niewystępującej jeszcze w drzewie. W przypadku generacji drzewa nieukorzenionego postępujemy w ten sam sposób z tą tylko różnicą, że w ostatnim kroku w utworzonym drzewie ukorzenionym usuwamy (ściągamy) korzeń.

Sposób generacji drzew jednakowo prawdopodobnych wygodniej jest opisać na przykładzie generacji binarnych drzew nieukorzenionych. Procedura generacji przebiega podobnie jak w modelu Yule'a. Rozpoczynamy od drzewa trzylistnego o losowo wybranych niepowtarzających się etykietach z $\{1, \ldots, n\}$. Do momentu uzyskania drzewa n-listnego, w każdym kroku dołączamy nową wiszącą krawędź do losowo wybranej (z jednakowym prawdopodobieństwem) krawędzi spośród wszystkich (a nie tylko wiszących, jak to było w YM) istniejących krawędzi w drzewie. Etykieta nowego liścia jest wybierana ze zbioru $\{1, \ldots, n\}$ pomniejszonego o elementy już występujące w drzewie.

W celu wygenerowania n-listnego drzewa ukorzenionego, wygodnie jest wygenerować $(n + 1)$-listne drzewo nieukorzenione, następnie ukorzenić drzewo w wierzchołku wewnętrznym sąsiadującym z liściem o etykiecie $n+1$ i ostatecznie usunąć ten liść wraz z krawędzią łączącą go z korzeniem.

Ustalanie etykiet kolejnych liści w modelu Yule'a powinno odbywać się w porządku losowym. W przypadku modelu UM, kolejność ta może być zarówno losowa, jak i ustalona arbitralnie [73].

Drzewa wygenerowane według modelu Yule'a są bardziej zbalansowane niż drzewa w modelu UM. Ściślej (na podstawie [73]), średnio $\sim \frac{2}{3}n$ liści w drzewach wygenerowanych wg modelu YM znajduje się w konfiguracji tworzącej tzw. wiśnię (dwa liście sąsiadują ze wspólnym wierzchołkiem wewnętrznym), podczas gdy w drzewach wygenerowanych zgodnie z mo-

delem UM wartość ta wynosi średnio tylko $\sim \frac{n}{2}$; przy czym najmniej liści w opisanej konfiguracji posiada gąsienica (4 liście), najwięcej zaś drzewo doskonale zrównoważone (n liści, dla n parzystego).

Definicja 6.1 ([106])**.** Rozkład prawdopodobieństwa w zbiorze A n-listnych drzew filogenetycznych opisany przez zmienną losową X jest *niezależny od permutacji etykiet*, jeśli dla każdego drzewa $T \in A$ zachodzi

$$\Pr(X = T) = \Pr(X = T'),$$

gdzie $T' \in A$ jest dowolnym drzewem o tej samej topologii co T.

Zarówno model Yule'a, jak i UM generuje drzewa z rozkładem niezależnym od permutacji etykiet [106].

Wiele analiz, np. [50, 3, 76], ujawnia pewną prawidłowość dotyczącą kształtu rzeczywistych drzew filogenetycznych, mianowicie ich topologia na ogół jest bardziej zbalansowana (symetryczna) niż dla drzew wygenerowanych według modelu UM, lecz mniej niż w przypadku modelu Yule'a.

W kolejnych podrozdziałach przedstawione zostaną własności metryk MS i MC dla drzew losowych w obu opisanych powyżej modelach. Aby lepiej zobrazować i porównać omawiane cechy, w rozważaniach zostanie dodatkowo uwzględnionych 6 klasycznych metryk. Dla drzew nieukorzenionych będzie to metryka Robinsona Foulds'a (dla rozbić) — RF, metryka ścieżkowa — PD, metryka kwartetowa — QT, zaś dla drzew ukorzenionych: metryka Robinsona Foulds'a (dla klastrów) — RFC, metryka węzłowa z normą L^2 — SN oraz metryka tripletowa — TT.

6.2 Odległości drzew nieukorzenionych

6.2.1 Rozkłady odległości

W tym podrozdziale zaprezentowane zostaną rozkłady odległości w poszczególnych metrykach w dwóch opisanych powyżej podstawowych modelach generacji drzew losowych. W celu wyznaczenia rozkładów odległości przedstawionych na rysunkach 6.1 i 6.2 losowo wygenerowano 10000 par drzew o 50 liściach (dla każdego z modeli osobno).

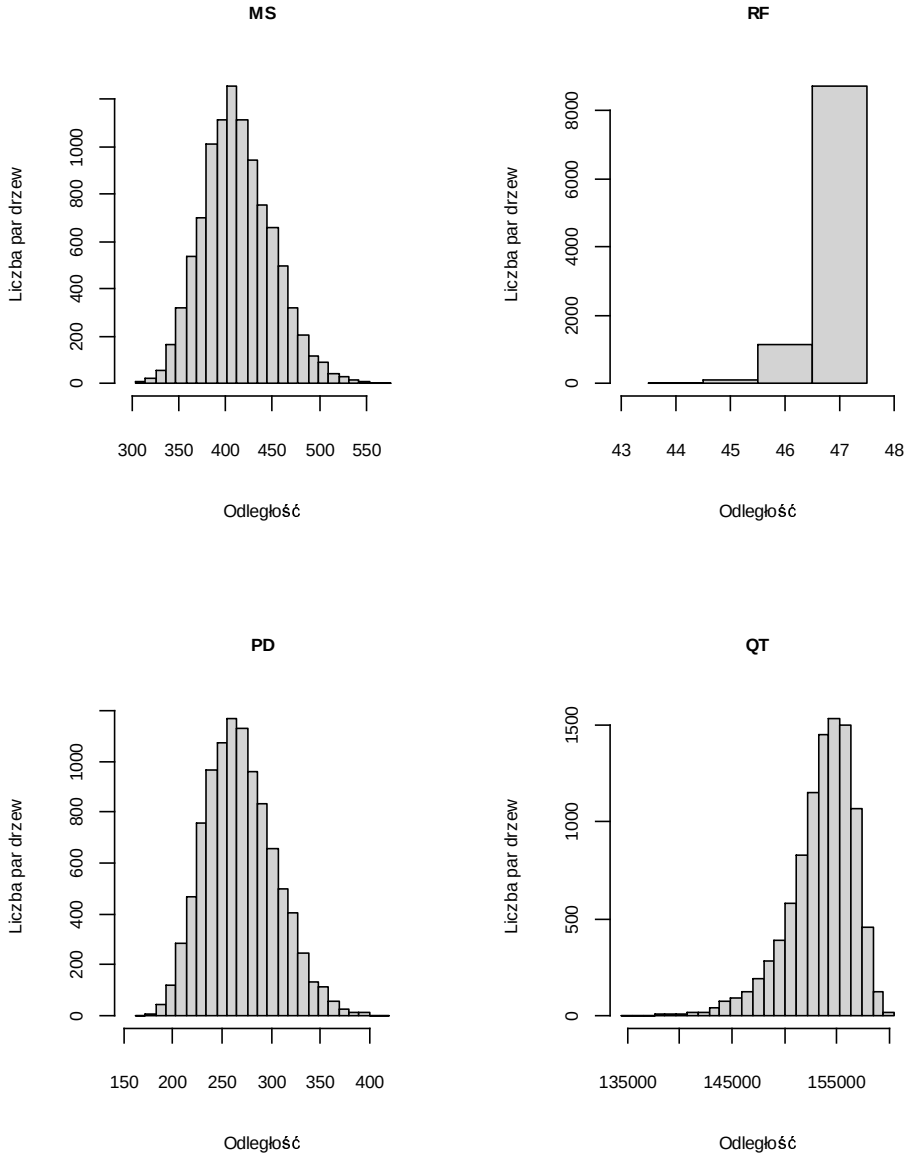
W przypadku metryk MS, PD, QT słupki na histogramach reprezentują przedziały o szerokości odpowiadającej 1/25 różnicy pomiędzy wartością maksymalną i minimalną. W przypadku odległości RF każdy słupek odpowiada jednej wartości metryki. Parametry charakterystyczne dla poszczególnych rozkładów, takie jak wartość średnia, odchylenie standardowe, liczba różnych wartości oraz kwartyle, zostały umieszczone w tabelach 6.1 i 6.2.

Tabela 6.1: Parametry rozkładów odległości w modelu UM.

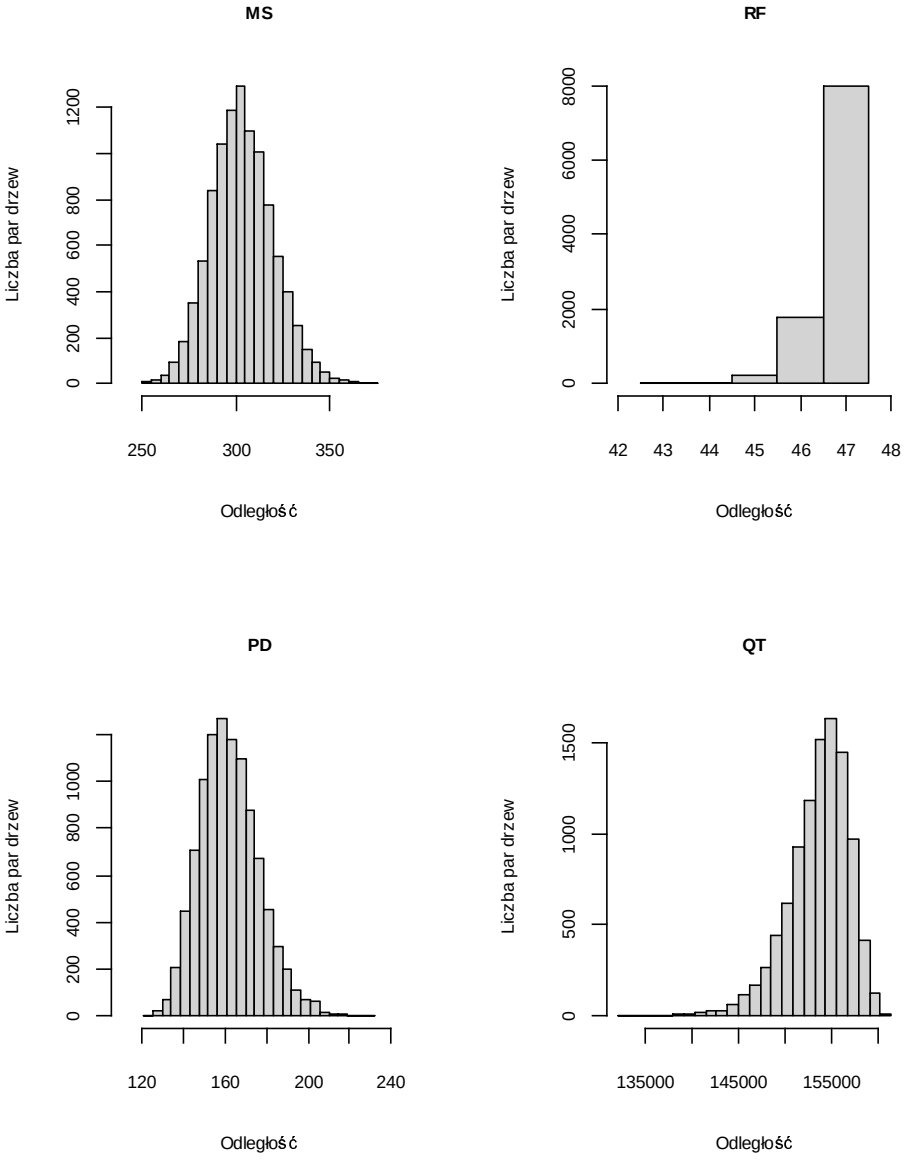
	MS	RF	PD	QT
Liczba wartości	236	4	8670	6455
Średnia	413.327	46.864	269.076	153538.838
Odchylenie std.	37.075	0.373	35.795	3170.604
Min	304	44	162.136	134429
Kwartyl 1	387	47	243.302	151987
Mediana	411	47	266.349	154085.5
Kwartyl 3	437	47	292.076	155764.75
Max	575	47	420.373	160544

Tabela 6.2: Parametry rozkładów odległości w modelu YM.

	MS	RF	PD	QT
Liczba wartości	121	5	5862	6580
Średnia	304.012	46.772	162.100	153550.788
Odchylenie std.	16.383	0.485	14.289	3245.354
Min	250	43	120.449	132014
Kwartyl 1	293	47	151.938	151744.5
Mediana	303	47	161	154002.5
Kwartyl 3	315	47	171.045	155833
Max	375	47	232.433	161362



Rysunek 6.1: Rozkłady odległości między losowymi drzewami nieukorzenionymi o 50 liściach w modelu UM.



Rysunek 6.2: Rozkłady odległości między losowymi drzewami nieukorzenionymi o 50 liściach w modelu YM.

Analizując kształt prezentowanych histogramów można łatwo zauważyć, że najbardziej stromy rozkład posiada metryka RF. Dla pary drzew losowych, niezależnie od modelu, przyjmuje ona na ogół swoją maksymalną możliwą wartość, czyli 47 (por. mediana i wartość średnia w tablach 6.1 i 6.2). Rozkład RF jako jedyny z przedstawionych w znacznym stopniu skupiony jest przy swojej wartości maksymalnej.

Metryka MS podobnie jak RF bazuje na porównywaniu rozbić, lecz jej rozkład znacznie różni się od rozkładu RF. Jest on zdecydowanie bardziej zbalansowany. Warto również zauważyć, że mimo iż kształt rozkładów odległości MS w obu modelach jest zbliżony, to ich parametry wykazują istotne różnice w zależności od modelu, np. wartość średnia w YM (304.012) jest znacznie mniejsza od analogicznej wartości w modelu UM (413.327). Podobna zależność występuje również dla metryki PD, natomiast nie występuje dla odległości RF i QT, w przypadku których parametry rozkładów w obu modelach są do siebie zbliżone. Cecha ta zostanie dokładniej przedstawiona w następnym podrozdziale.

Zaprezentowane wyniki potwierdzają teoretyczne rozważania przedstawione w [106], gdzie wykazano, że rozkłady odległości RF w obu modelach zbliżają się asymptotycznie do rozkładu Poissona, a wartość oczekiwana zmierza do $n-3$.

Twierdzenie 6.1 ([106])**.** *Dla wartości oczekiwanej odległości RF drzew nieukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z U_n^B z dowolnym rozkładem niezależnym od permutacji etykiet zachodzi*

$$\lim_{n\to\infty} \frac{\mathbb{E}[d_{RF}(T'_{1_n}, T'_{2_n})]}{n-3} = 1.$$

Twierdzenie 6.2 ([106])**.** *Niech $\mu_{c,n}$ oraz $\sigma_{c,n}$ oznaczają odpowiednio wartość oczekiwaną i odchylenie standardowe liczby wiśni w drzewie wylosowanym z U_n^B z dowolnym rozkładem Y niezależnym od permutacji etykiet. Wówczas rozkład odległości RF drzew nieukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z U_n^B zgodnie z Y, pod warunkiem, że $\lim_{n\to\infty} \frac{\mu_{c,n}}{n} \to c > 0$ oraz $\lim_{n\to\infty} \frac{\sigma_{c,n}}{n} \to 0$, jest asymptotycznie opisany rozkładem Poissona.*

Dla metryki QT wartość średnia również nie zależy od modelu [106].

Twierdzenie 6.3 ([106]). *Wartość oczekiwana odległości QT dla drzew nieukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z U_n^B z dowolnym rozkładem niezależnym od permutacji etykiet wynosi*

$$\mathbb{E}[d_{QT}(T'_{1_n}, T'_{2_n})] = \frac{2}{3}\binom{n}{4}.$$

Wartości otrzymane eksperymentalne wynoszą: 153538.838 (UM) oraz 153550.788 (YM) i są bardzo zbliżone do teoretycznej wartości równej ≈ 153533.34.

Dla metryki PD w modelu UM teoretyczna wartość średnia wynosi ≈ 271.58; wartość otrzymana eksperymentalnie jest również zbliżona i wynosi 269.076. Dokładny wzór wyrażający wartość średnią odległości PD w modelu UM jest dość skomplikowany.

Twierdzenie 6.4 ([106]). *Wartość oczekiwana odległości PD dla drzew nieukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z rozkładem równomiernym z U_n^B, tj. według modelu UM, wynosi*

$$\mathbb{E}[d_{PD}(T'_{1_n}, T'_{2_n})] = \sqrt{2\binom{n}{2}\left(4n - 6 - \frac{2^{2(n-2)}}{\binom{2(n-2)}{n-2}} - \left(\frac{2^{2(n-2)}}{\binom{2(n-2)}{n-2}}\right)^2\right)}$$
$$\sim \sqrt{n(n-1)\left((4-\pi)n - \sqrt{\pi n}\right)}.$$

Na podstawie powyższych rozważań można przypuszczać, że sposób określania odległości między drzewami filogenetycznymi wyrażony przez metrykę MS jest najbardziej zbliżony do odległości PD. Wskazuje na to choćby podobny kształt rozkładów odległości oraz zależność parametrów rozkładu od modelu generowania drzew. W celu weryfikacji tego przypuszczenia możemy posłużyć się współczynnikami korelacji zamieszczonymi w tabeli 6.3.

Współczynnik korelacji rangowej Spearmana [104] określa stopień monotonicznej zależności dwóch ciągów wartości. Przyjmuje on wartości z przedziału $[-1, 1]$ o następującej interpretacji: wartość 1 oznacza doskonałą korelację dodatnią (oba ciągi są jednakowo monotoniczne), -1 oznacza

Tabela 6.3: Współczynniki korelacji rangowej Spearmana analizowanych metryk dla drzew nieukorzenionych.

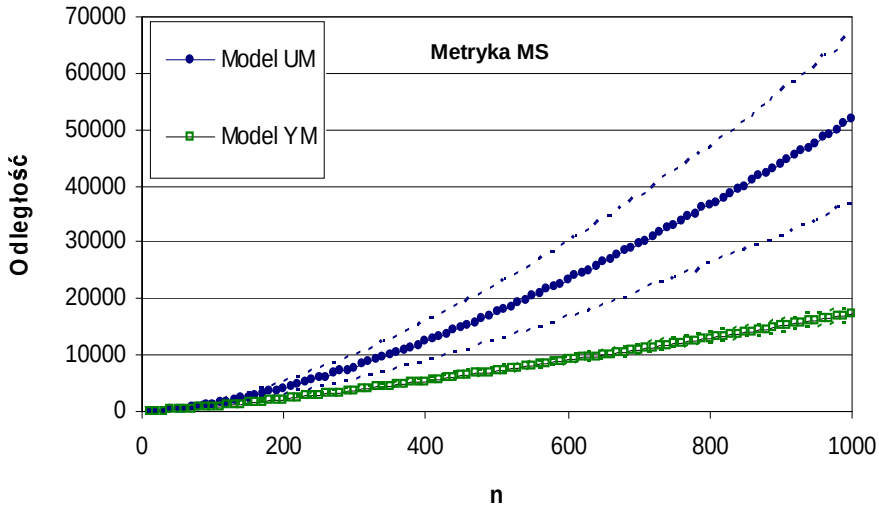
	Model UM	Model YM
MS — RF	0.0485	0.1267
MS — PD	0.9351	0.8395
MS — QT	0.2856	0.4903
RF — PD	0.0100	0.0304
RF — QT	0.1019	0.1026
PD — QT	0.0465	0.1037

doskonałą korelację ujemną (oba ciągi są przeciwnie monotoniczne), im wartości są bliższe zeru tym zależność monotoniczna jest mniejsza. Współczynnik Spearmana operuje nie na bezwzględnych wartościach występujących w wejściowych ciągach, lecz na ich względnych pozycjach, na których pojawiają się one w porządku rosnącym w ramach jednego ciągu, np. dla wartości niepowtarzających się najmniejsza z nich otrzymuje rangę 1, druga z kolei 2 itd. W przypadku występowania identycznych wartości każde z wystąpień ma przypisaną tę samą rangę równą średniej arytmetycznej pozycji w porządku rosnącym. Z uwagi na ten fakt współczynnik Spearmana wydaje się być bardziej odpowiedni do analizy podobieństwa metod mierzenia odległości niż np. współczynnik korelacji Pearsona, operujący wprost na wartościach występujących w sekwencjach.

Analiza wyników przedstawionych w tabeli 6.3 wskazuje na duże podobieństwo metryk MS i PD (0.9351, 0.8395), potwierdzając tym samym wysunięte przypuszczenie. Widzimy też, że podobieństwo MS do metryki QT (0.2856, 0.4903) jest większe niż podobieństwo do QT odległości PD (0.0465, 0.1037). Metryka RF może być natomiast określona jako najbardziej „wyróżniająca" się spośród badanych odległości, gdyż wartości jej korelacji z innymi metrykami są niewielkie.

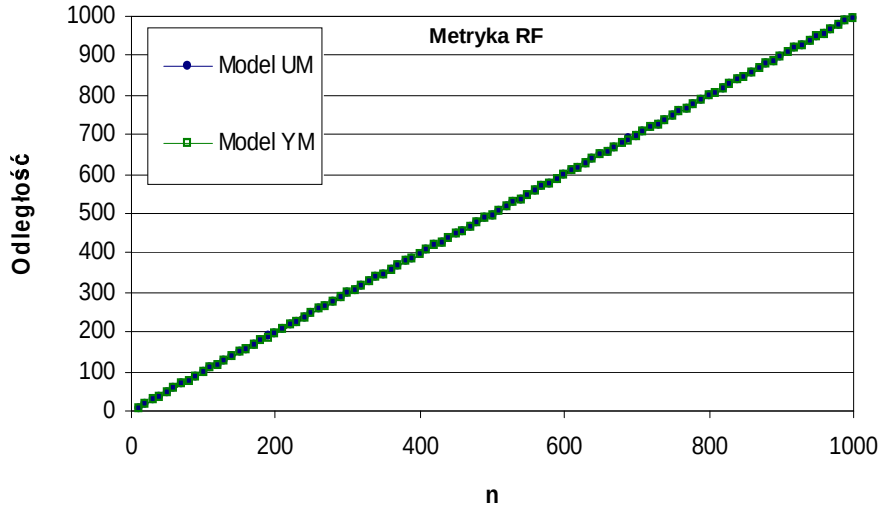
6.2.2 Wartość średnia i odchylenie standardowe

Wartość średnia odległości między drzewami losowymi może być traktowana jako punkt odniesienia w ocenie stopnia podobieństwa porównywanych drzew. Na rysunkach 6.3-6.6 przedstawione zostały wykresy wartości średniej odległości w danej metryce w obu rozpatrywanych modelach. Linie przerywane odpowiadają wartościom $\bar{x} + 3\sigma$, gdzie $\bar{x}$ oznacza średnią z wyników otrzymanych w eksperymencie dla ustalonej liczby liści n, natomiast σ jest odchyleniem standardowym od tej średniej. Prezentowane wyniki pochodzą z analizy 1000 losowych par drzew n-listnych (dla każdego z modeli), dla n przyjmującego wartości $10, 20, \ldots, 1000$.

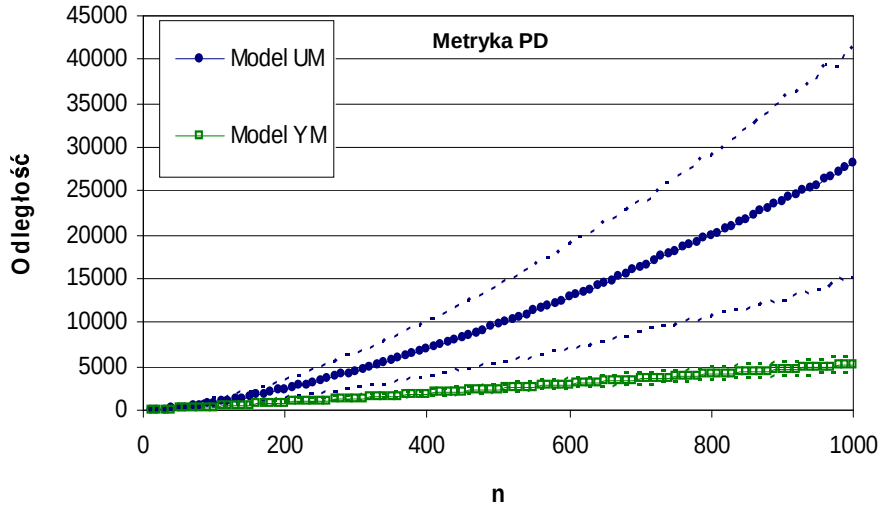


Rysunek 6.3: Wartość średnia i odchylnie standardowe w metryce MS.

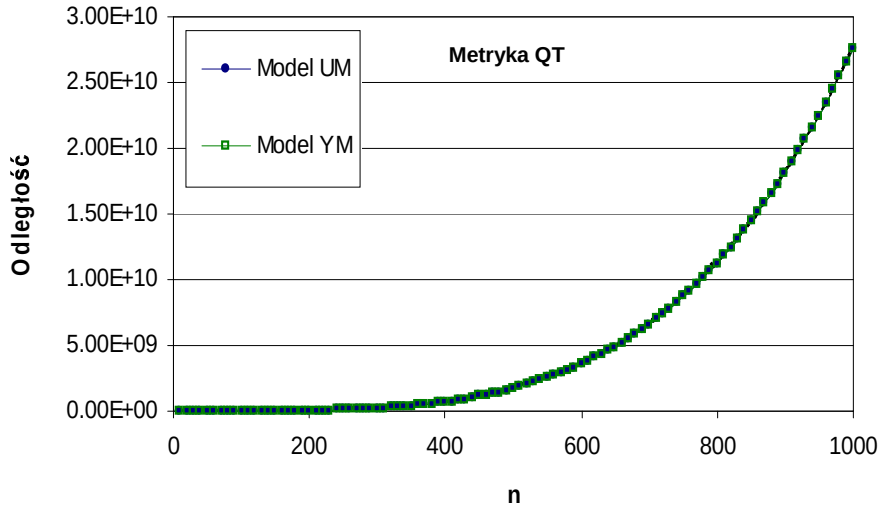
Analiza przedstawionych wykresów potwierdza wnioski zasygnalizowane w poprzednim podrozdziale dotyczące silnej zależności wartości metryk MS i PD od modelu generowania drzew oraz braku tej zależności w przypadku pozostałych dwóch metryk. Ponadto wartość średnia w modelu Yule'a, jak również i odchylenie standardowe dla MS i PD, są znacznie mniejsze niż w modelu UM. Jednocześnie warto zauważyć, że wartość od-



Rysunek 6.4: Wartość średnia i odchylnie standardowe w metryce RF.



Rysunek 6.5: Wartość średnia i odchylnie standardowe w metryce PD.



Rysunek 6.6: Wartość średnia i odchylnie standardowe w metryce QT.

chylenia standardowego w stosunku do wartości średniej, niezależnie od modelu dla RF i QT, jest bardzo mała (nie jest widoczna na wykresach). Fakt ten wskazuje na duże skoncentrowanie wartości tych metryk w pobliżu średniej. Największy rozrzut wartości ma miejsce w przypadku metryki PD w modelu UM. Metryka MS pod względem tego kryterium zajmuje drugie miejsce.

6.3 Odległości drzew ukorzenionych

W niniejszym podrozdziale zaprezentowane zostaną rozkłady odległości dla drzew ukorzenionych w metrykach RFC, MC, SN oraz TT. Podobnie jak w przypadku drzew nieukorzenionych, w eksperymentach zostały uwzględnione dwa najbardziej popularne modele generacji drzew losowych: model drzew jednakowo prawdopodobnych (UM) i model Yule'a (YM). W celu wyznaczenia rozkładów odległości przedstawionych na rysunkach 6.7 i 6.8 losowo wygenerowane zostało 10000 par drzew ukorzenionych

binarnych o 50 liściach (dla każdego z modeli osobno).

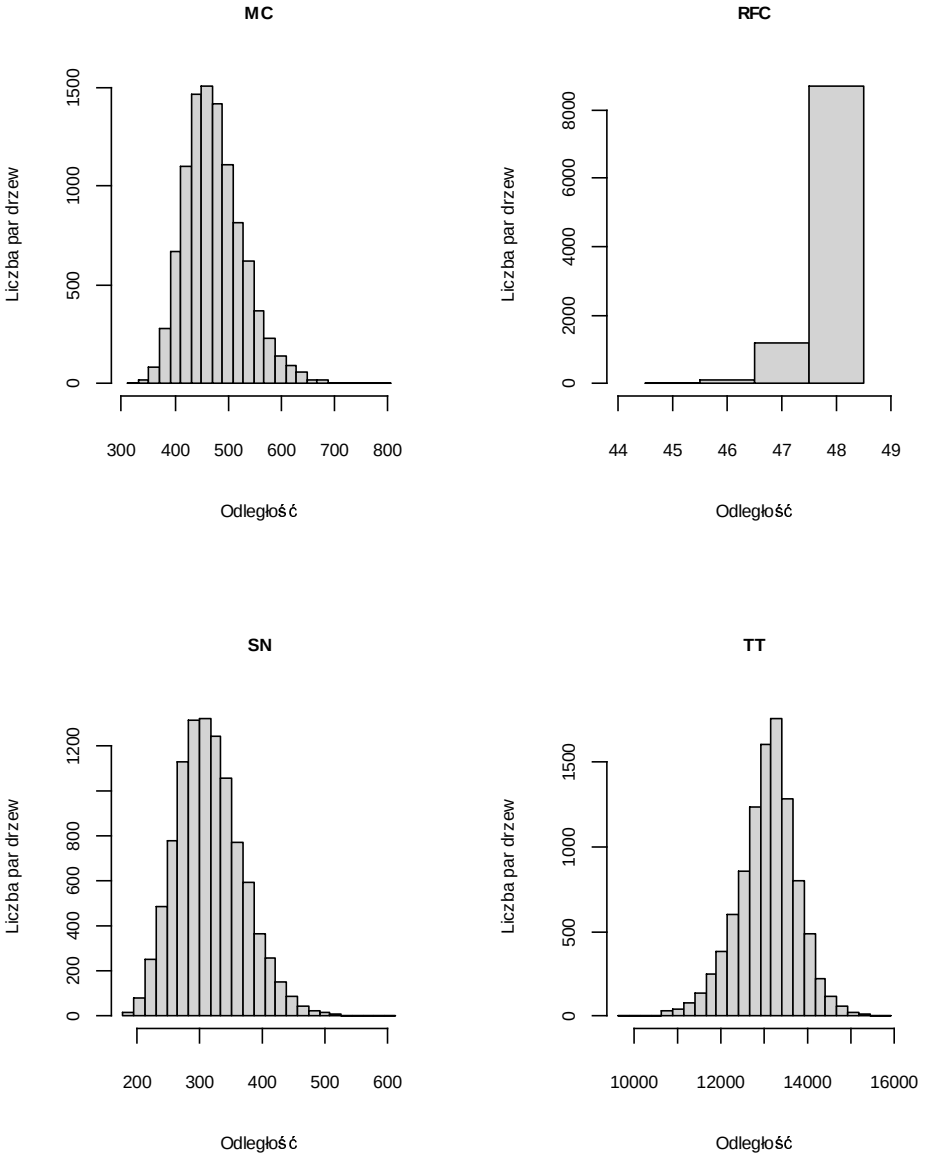
6.3.1 Rozkłady odległości

Podobnie jak w przypadku metryk dla drzew nieukorzenionych, słupki na histogramach dla odległości MC, SN oraz TT reprezentują przedziały o szerokości odpowiadającej 1/25 różnicy pomiędzy wartością maksymalną i minimalną. W przypadku odległości RFC każdy słupek odpowiada dokładnie jednej wartości metryki. Parametry charakterystyczne dla poszczególnych rozkładów, takie jak wartość średnia, odchylenie standardowe, liczba różnych wartości oraz kwartyle, zostały umieszczone w tabelach 6.4 oraz 6.5.

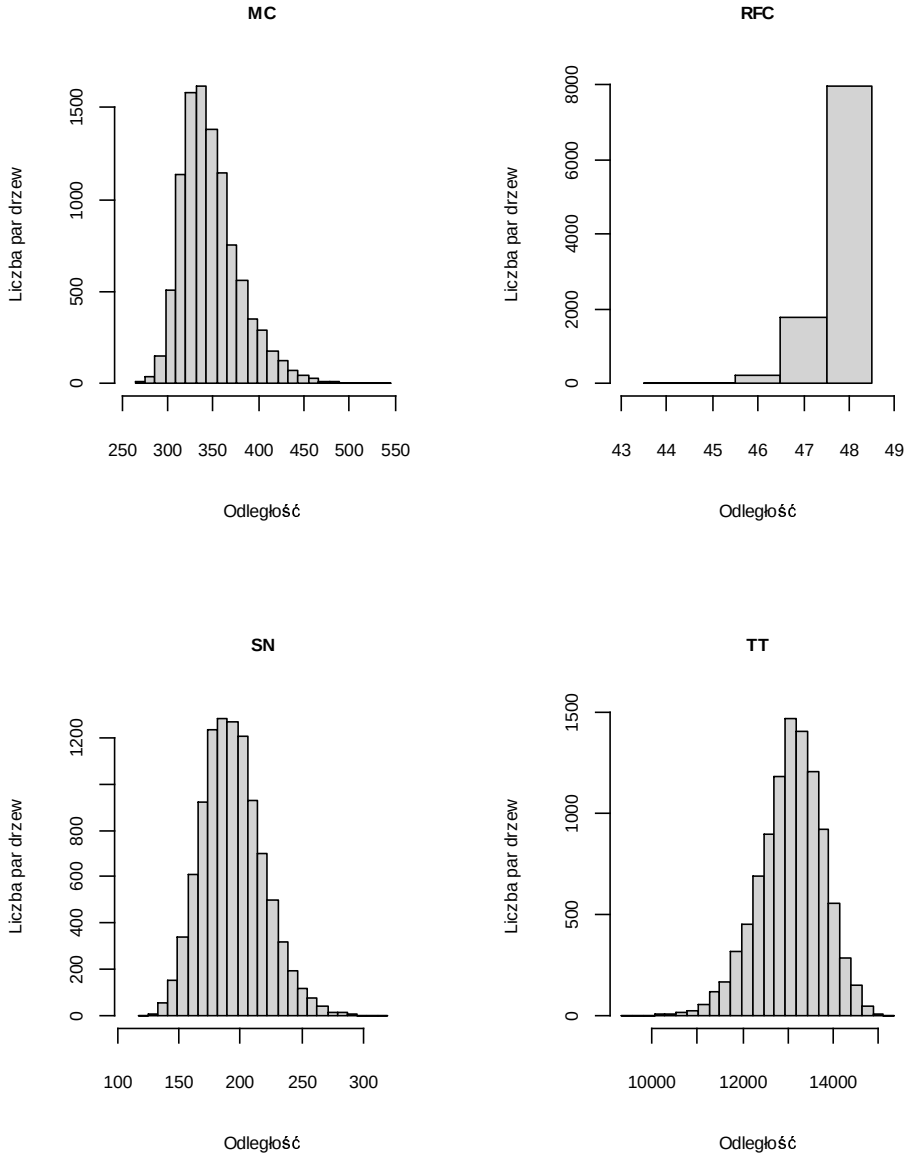
Tabela 6.4: Parametry rozkładów odległości w modelu UM.

	MC	RFC	SN	TT
Liczba wartości	339	4	9571	2852
Średnia	474.245	47.863	317.021	13081.914
Odchylenie std.	54.660	0.368	52.382	681.575
Min	312	45	179.053	9654
Kwartyl 1	436	48	279.812	12694
Mediana	468	48	312.977	13137
Kwartyl 3	507	48	349.391	13511
Max	805	48	612.949	15938

Łatwo zauważyć, że histogram metryki RFC jest najbardziej stromy i odległość ta przyjmuje tylko 4 lub 5 różnych wartości (w zależności od modelu). Najwięcej różnych wartości przyjmuje metryka SN i w odróżnieniu od blisko spokrewnionej metryki PD dla drzew nieukorzenionych liczba ta jest największa w każdym z rozpatrywanych modeli. Za najbardziej zrównoważony można uznać rozkład wartości odległości TT, którego kształt nie jest zależny od modelu i asymptotycznie może być opisany rozkładem normalnym [38].



Rysunek 6.7: Rozkłady odległości między losowymi drzewami ukorzenionymi o 50 liściach w modelu UM.



Rysunek 6.8: Rozkłady odległości między losowymi drzewami ukorzenionymi o 50 liściach w modelu Yule'a.

Tabela 6.5: Parametry rozkładów odległości w modelu YM.

	MC	RFC	SN	TT
Liczba wartości	215	5	8612	2900
Średnia	347.639	47.773	193.837	13067.694
Odchylenie std.	32.002	0.480	24.862	711.337
Min	264	44	116.722	9336
Kwartyl 1	325	48	176.376	12642
Mediana	342	48	192.171	13114
Kwartyl 3	365	48	209.439	13557.25
Max	545	48	319.620	15349

Twierdzenie 6.5 ([38])**.** *Wartość oczekiwana odległości TT dla drzew ukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z R_n^B z dowolnym rozkładem niezależnym od permutacji etykiet wynosi*

$$\mathbb{E}[d_{TT}(T'_{1_n}, T'_{2_n})] = \frac{2}{3}\binom{n}{3}.$$

Twierdzenie 6.6 ([38])**.** *Rozkład standaryzowanej zmiennej losowej opisującej odległość TT dla drzew ukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z R_n^B z dowolnym rozkładem niezależnym od permutacji etykiet dla $n \to \infty$ zbiega (względem rozkładu) do rozkładu normalnego.*

Od modelu generacji drzew nie zależy również kształt rozkładu RFC. Średnie wartości odległości dla TT wyznaczone eksperymentalnie są bardzo zbliżone do teoretycznej wartości wynoszącej ≈ 13066.67.

Przyjrzyjmy się teraz bliżej wzajemnym relacjom zachodzącym pomiędzy analizowanymi metrykami (tabela 6.6). Metryka MC zarówno dla drzew generowanych według modelu UM, jak i YM jest najbardziej zbliżona do SN, lecz odpowiednie współczynniki korelacji przyjmujące wartości 0.7608 i 0.4998 są znacznie niższe niż w przypadku pary metryk MS — PD (0.9351, 0.8395). Zauważmy również, że metryka MC posiada największe wartości współczynników korelacji w stosunku do dowolnej innej metryki

(w każdym z modeli), tzn. jest „najbardziej podobna" do każdej z pozostałych metryk. Można powiedzieć, że stanowi pewnego rodzaju konsensus pomiędzy nimi. Fakt ten jest tym bardziej interesujący, że opisana własność dla odległości MS występuje tylko w modelu Yule'a; w modelu UM jej współczynnik korelacji względem RF wynosi 0.0485, podczas gdy dla pary metryk QT, RF przyjmuje on wartość większą, równą 0.1019.

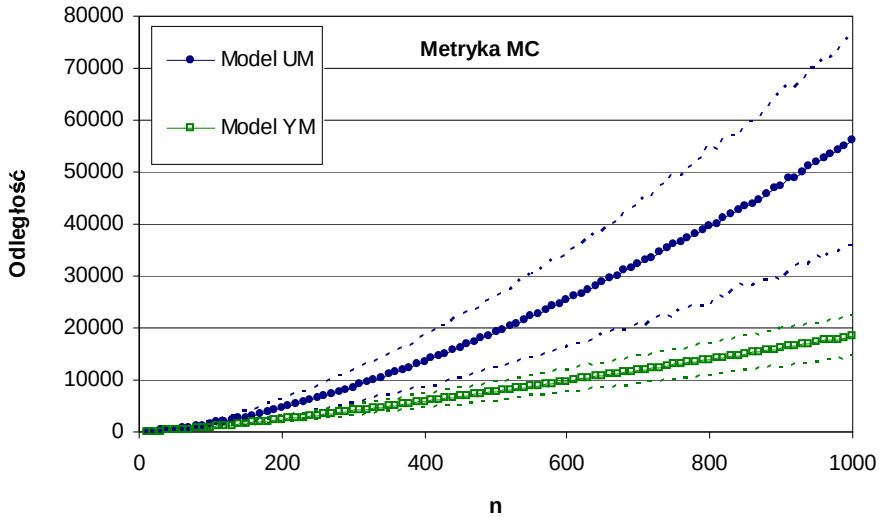
Tabela 6.6: Współczynniki korelacji rangowej Spearmana analizowanych metryk dla drzew ukorzenionych.

	Model UM	Model YM
MC — RFC	0.0730	0.0644
MC — SN	0.7608	0.4998
MC — TT	0.2575	0.2807
RFC — SN	0.0417	0.0080
RFC — TT	0.0515	0.0132
SN — TT	0.1895	0.2488

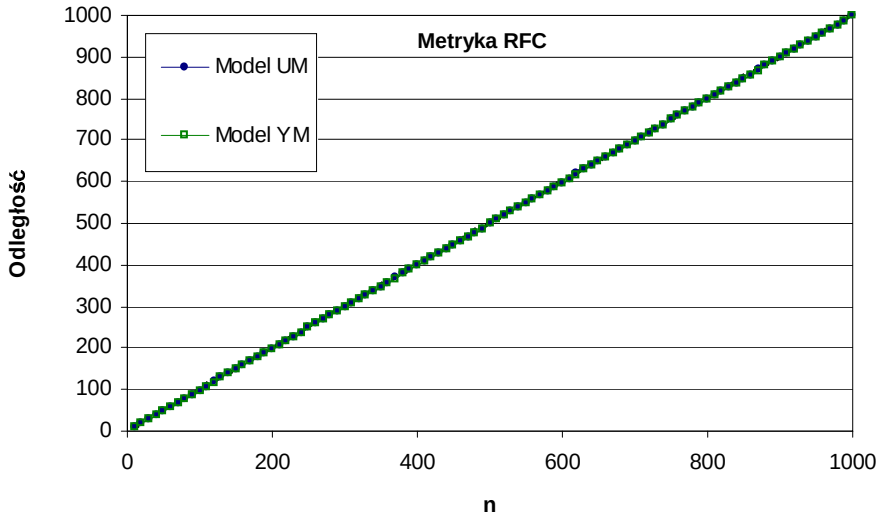
6.3.2 Wartość średnia i odchylenie standardowe

Na rysunkach 6.9-6.12 przedstawione zostały wykresy wartości średniej odległości w danej metryce w obu rozpatrywanych modelach. Podobnie jak w przypadku metryk dla drzew nieukorzenionych, linie przerywane odpowiadają wartościom $\bar{x} + 3\sigma$. Zbiory danych testowych zostały skonstruowane analogicznie jak w podrozdziale 6.2.2.

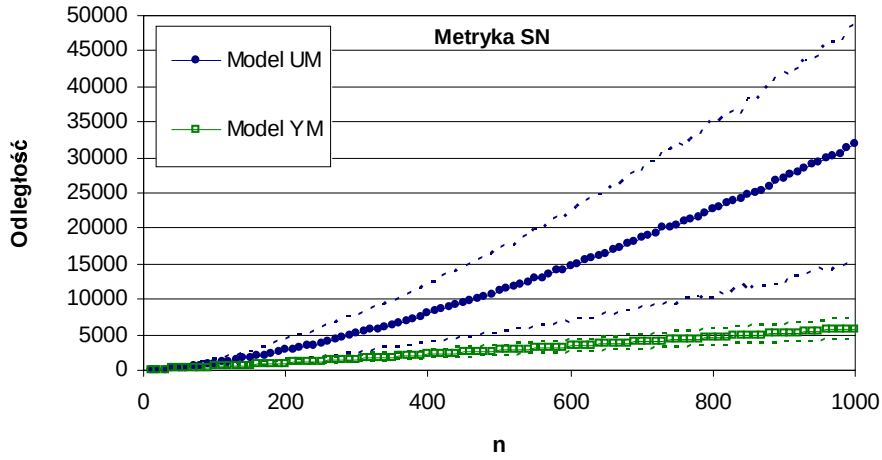
Przedstawione wykresy są bardzo zbliżone do wykresów dla odpowiednich metryk dla drzew nieukorzenionych. Największy stosunek odchylenia standardowego do wartości średniej obserwujemy dla metryki SN, w następnej kolejności cecha ta uwidacznia się dla odległości MC. W obu tych przypadkach zarówno wartość średnia, jak i odchylenie standardowe mocno uzależnione są od modelu generacji drzew losowych. W rezultacie, wartości obu tych parametrów są znacznie mniejsze w modelu Yule'a. W przypadku pozostałych metryk opisana zależność nie występuje.



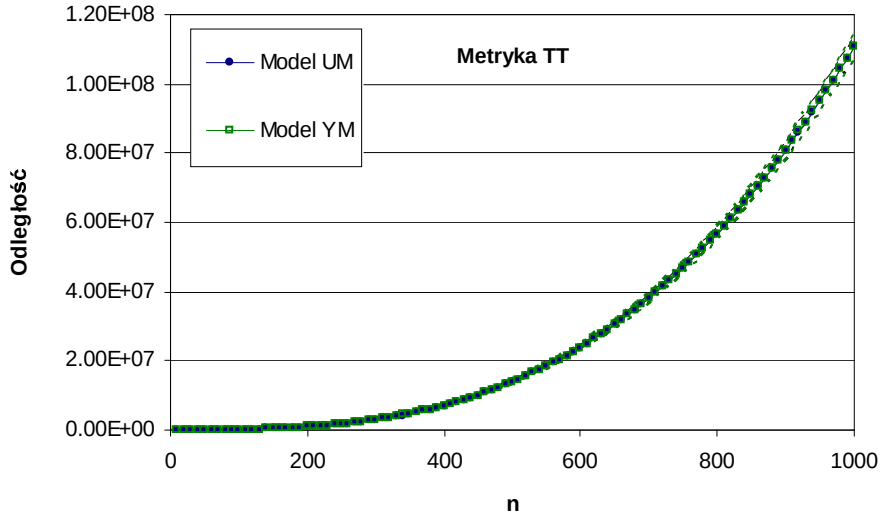
Rysunek 6.9: Wartość średnia i odchylnie standardowe w metryce MC.



Rysunek 6.10: Wartość średnia i odchylnie standardowe w metryce RFC.



Rysunek 6.11: Wartość średnia i odchylnie standardowe w metryce SN.



Rysunek 6.12: Wartość średnia i odchylnie standardowe w metryce TT.

6.4 Asymptotyka wartości oczekiwanej odległości w MS i MC

Niech $S(T)$ będzie zdefiniowane jako $\sum_{c\in\sigma(T)}|c|$, gdzie $T \in R_L^B$, oraz jako $\sum_{s\in\beta(T)}\min(s)$, jeśli $T \in U_L^B$. Definicja sumy $S(T)$, dla T będącego drzewem ukorzenionym, jest bardzo zbliżona do formuły opisującej wartość *indeksu Sackina* $S_{ind}(T)$ traktowanego jako miara zbalansowania drzewa [96, 99]. W szczególności dla $T \in R_n^B$ zachodzi $S(T) = S_{ind}(T)+n$. W celu uzyskania asymptotycznego oszacowania tempa wzrostu wartości średniej odległości między drzewami w metrykach MS i MC w modelu UM wykorzystamy następujące dwa silne fakty podane przez Blum'a i in. w [14].

Twierdzenie 6.7 ([14])**.** *Niech T_n będzie drzewem wybranym losowo z rozkładem równomiernym z R_n^B.*

1. *Dystrybuanta zmiennej losowej $\frac{S(T_n)}{n^{3/2}}$ jest punktowo zbieżna do dystrybuanty zmiennej losowej rozkładu Airy ($\mathcal{A}$).*

2. *Zachodzi równość*
$$\lim_{n\to\infty}\frac{\mathbb{E}[S(T_n)]}{n^{3/2}} = \sqrt{\pi}.$$

Poniższe twierdzenie odpowiada na pytanie o asymptotyczne tempo wzrostu wartości średniej odległości w metryce MS postawione w [19].

Twierdzenie 6.8.

1. *Wartość oczekiwana odległości MC dla drzew ukorzenionych T_{1_n}, T_{2_n} wylosowanych niezależnie z rozkładem równomiernym z R_n^B spełnia zależność:*
$$\mathbb{E}[d_{MC}(T_{1_n}, T_{2_n})] = \Theta(n^{3/2}).$$

2. *Wartość oczekiwana odległości MS dla drzew nieukorzenionych T'_{1_n}, T'_{2_n} wylosowanych niezależnie z rozkładem równomiernym z U_n^B spełnia zależność:*
$$\mathbb{E}[d_{MS}(T'_{1_n}, T'_{2_n})] = \Theta(n^{3/2}).$$

Dowód. Niech drzewa T_1, T_2 będą wybrane niezależnie z rozkładem równomiernym z R_n^B oraz M będzie sparowaniem ich klastrów (włącznie z klastrami trywialnymi związanymi z liśćmi i korzeniem), takim że zachodzi $\sum_{(A,B)\in M} |A \oplus B| = d_{MC}(T_1, T_2)$. Wówczas drzewa T_1', T_2', utworzone na podstawie T_1 oraz T_2 poprzez dołączenie nowego liścia $n+1$ do ich korzeni, są drzewami wylosowanymi z rozkładem równomiernym z U_{n+1}^B (por. [98], wniosek 2.2.3). Definiujemy dwa sparowania rozbić drzew T_1', T_2': $M' = \{(A|\{1,\ldots,n+1\}\setminus A, B|\{1,\ldots,n+1\}\setminus B) : (A,B) \in M\}$ oraz sparowanie M'' jako takie, które wyznacza wartość $d_{MS}(T_1', T_2')$. Korzystając z (3.9) mamy zatem

$$\begin{aligned}
|S(T_1') - S(T_2')| &\leq \sum_{(s_1,s_2)\in M''} |\min(s_1) - \min(s_2)| \\
&\leq \sum_{(s_1,s_2)\in M''} h_{MS}(s_1,s_2) = d_{MS}(T_1', T_2') \\
&\leq \sum_{(s_1,s_2)\in M'} h_{MS}(s_1,s_2) \leq d_{MC}(T_1, T_2) \\
&\leq \sum_{(A,B)\in M} (|A| + |B|) = S(T_1) + S(T_2). \qquad (6.1)
\end{aligned}$$

Stąd, na podstawie twierdzenia 6.7, otrzymujemy górne oszacowanie dla wartości oczekiwanej w obu przypadkach.

Pozostaje wykazać, że $\mathbb{E}[|S(T_1') - S(T_2')|] = \Omega(n^{3/2})$. Zauważmy, że każdemu klastrowi $A \in \sigma(T_1)$ odpowiada rozbicie $A|\{1,\ldots,n+1\}\setminus A \in \beta(T_1')$, stąd

$$S(T_1') \leq S(T_1). \qquad (6.2)$$

Niech $T \in U_L^B$ będzie n-listnym nieukorzenionym drzewem binarnym. Wprowadzamy orientację krawędzi T w kierunku mniej licznej partycji rozbicia związanego z daną krawędzią. W ten sposób co najwyżej jedna z krawędzi nie otrzyma orientacji (tj. taka, która jest związana z rozbiciem o równolicznych partycjach). Wejściowy stopień każdego z wierzchołków drzewa wynosi zatem 0 lub 1. Możliwe są dwie sytuacje.

Przypadek 1. Dokładnie jedna krawędź $\{u_1, u_2\}$ została nieskierowana. Drzewo T może być w tym przypadku traktowane jako suma dwóch drzew binarnych T_{u_1} oraz T_{u_2}, ukorzenionych odpowiednio w u_1 i u_2, gdzie

$|L(T_{u_1})| = |L(T_{u_2})| = \frac{n}{2}$. W tej sytuacji określamy T jako *związane z rozbiciem* $\{L(T_{u_1}), L(T_{u_2})\}$ zbioru L.

Przypadek 2. Wszystkie krawędzie zostały skierowane. Wówczas istnieje jeden wierzchołek u posiadający stopień wejściowy równy 0. Drzewo T może być w tym przypadku traktowane jako suma trzech drzew binarnych T_{u_1}, T_{u_3}, T_{u_3}, ukorzenionych odpowiednio w wierzchołkach u_1, u_2, u_3. W tej sytuacji będziemy określać T jako *związane z* 3*-rozbiciem* $\{L(T_{u_1}), L(T_{u_2}), L(T_{u_3})\}$ zbioru L, gdzie rozbicie $\{L(T_{u_1}), L(T_{u_2}), L(T_{u_3})\}$ jest rozumiane jako nieuporządkowana trójka.

Niech T będzie drzewem wybranym z U_n^B z rozkładem równomiernym. Na podstawie (6.2) mamy $\lim\limits_{n\to\infty} \mathbb{E}\left[\frac{S(T)}{n^{3/2}}\right] \leq \sqrt{\pi}$. Stąd, dla wystarczająco dużych wartości n (takich, że $\mathbb{E}\left[\frac{S(T)}{n^{3/2}}\right] < 2$), na podstawie nierówności Markowa otrzymujemy:

$$\Pr\left(S(T) \leq 4n^{3/2}\right) \geq 1 - \Pr\left(S(T) \geq 2\mathbb{E}[S(T)]\right) \geq \frac{1}{2}. \tag{6.3}$$

Niech $p > 0$ będzie mniejsze niż prawdopodobieństwo, że zmienna losowa Airy $\mathcal{A}$ przyjmie wartość większą niż $5 \cdot 3^{3/2}$. Wybrane losowo drzewo T jest związane z pewnym 2- lub 3-rozbiciem zbioru L, którego największa z partycji $A \subsetneq L$ spełnia nierówność $|A| \geq n/3$. Niech $T_A \in R_A^B$ będzie ukorzenionym poddrzewem T związanym z A. Dla wystarczająco dużych wartości n ukorzenione drzewo T_A, gdzie $|A| \geq n/3$, spełnia $\frac{S(T_A)}{|A|^{3/2}} \geq 5 \cdot 3^{3/2}$ z prawdopodobieństwem większym niż p. Wówczas $S(T_A) \geq 5n^{3/2}$ i ostatecznie:

$$\Pr\left(S(T) \geq 5n^{3/2}\right) \geq p. \tag{6.4}$$

Łącząc (6.3) z (6.4) otrzymujemy, że dla niezależnie wylosowanych drzew T_1' i T_2' z prawdopodobieństwem co najmniej p jedno z nich spełnia warunek (6.3), podczas gdy drugie spełnia (6.4). Stąd zaś mamy $\mathbb{E}[|S(T_1') - S(T_2')|] = \Omega(n^{3/2})$. □

Z twierdzenia 6.8 wynika kolejna pożądana własność metryk MS i MC, dla których wprawdzie średnia odległość w modelu UM jest asymptotycznie mniejsza od średnicy, lecz zarazem jest ona większa niż zmiana, którą

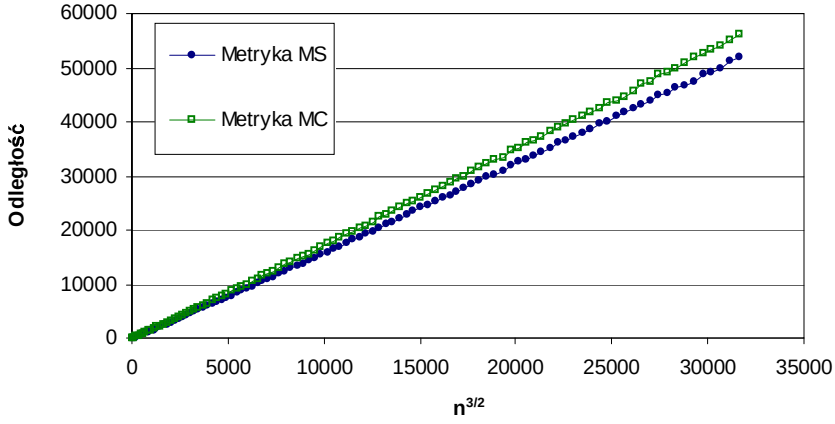
może spowodować niewielka modyfikacja drzewa (np. przeniesienie pojedynczego liścia). Fakt ten odróżnia metryki MS, MC od RF i RFC, gdzie zarówno wartość oczekiwana, jak i średnica jest $\Theta(n)$, a ponadto również przemieszczenie liścia może powodować zmiany odległości tego samego rzędu.

Ilustracja oszacowania z twierdzenia 6.8 jest przedstawiona na rysunkach 6.13 oraz 6.14. Wartość dla ustalonego n jest średnią wyznaczoną dla 1000 losowych par drzew. Łatwo zauważyć, że jeśli na osi odciętych umieścimy wartości $n^{3/2}$, to otrzymamy kształt wykresu bardzo zbliżony do odcinka prostej. Aproksymacja równania tej prostej metodą najmniejszych kwadratów prowadzi do następujących relacji:

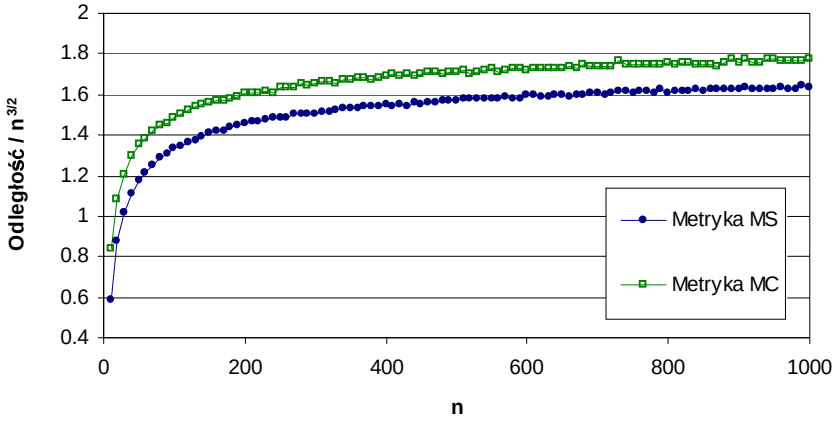
$$\text{MS: } y(x) \approx 1.6414x - 574.6,\ R^2 \approx 0.9998,$$
$$\text{MC: } y(x) \approx 1.7747x - 523.41,\ R^2 \approx 0.9998,$$

gdzie R^2 oznacza kwadrat współczynnika korelacji liniowej Pearsona.

Jeśli natomiast wykreślimy wartość średnią odległości przeskalowaną przez $n^{3/2}$ (rysunek 6.14), wówczas wyraźnie uwidacznia się rosnący charakter tego ilorazu, zmierzający do wartości około 1.8 w przypadku MC i około 1.65 dla MS.



Rysunek 6.13: Wartość średnia metryk MS i MC w modelu UM dla liczby liści $n = 10, 20, \ldots, 1000$. Na osi X wykresu znajdują się wartości odpowiadające $n^{3/2}$.



Rysunek 6.14: Wartość średnia metryk MS i MC w modelu UM. Na osi Y wykresu znajduje się iloraz wartości średniej odległości przez $n^{3/2}$.

7 Część eksperymentalna

Jednym z podstawowych zastosowań metryk filogenetycznych jest ilościowe określanie jakości metod konstrukcji drzew. W tej części przedstawiona zostanie analiza jakości 10 popularnych metod rekonstrukcji. Ponieważ na ogół zrekonstruowane drzewo jest nieukorzenione (do ustalania miejsca korzenia stosowane są inne metody), w analizie jakości zostaną wykorzystane metryki RF, PD, QT i MS. Wszystkie te odległości, wraz z czterema ich odpowiednikami dla drzew ukorzenionych, zostały zaimplementowane w opisanej poniżej aplikacji TreeCmp [118], wykonanej w ramach pracy nad niniejszą rozprawą.

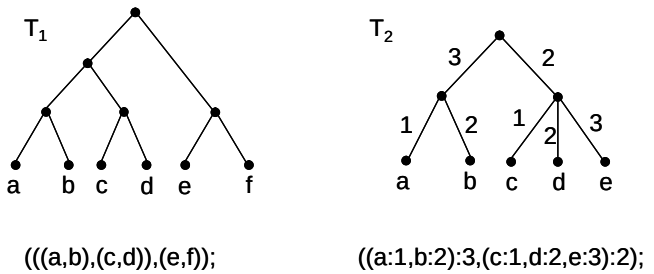
7.1 Aplikacja TreeCmp

Aplikacja TreeCmp została wykorzystana w eksperymentach opisanych w dalszej części rozdziału oraz w badaniach dotyczących drzew losowych zamieszczonych w rozdziale 6. Program ten, napisany w języku Java, umożliwia wygodne przeprowadzanie analiz podobieństwa dowolnych (niekoniecznie binarnych) drzew filogenetycznych za pomocą wszystkich ośmiu metryk rozważanych w niniejszej pracy.

Danymi wejściowymi dla TreeCmp są pliki tekstowe zawierające definicje drzew filogenetycznych w formacie Newick (określanym też jako New Hampshire). Formalny opis składni Newick jest dostępny na stronach internetowych Uniwersytetu w Waszyngtonie:

- `http://evolution.genetics.washington.edu/phylip/newicktree.html`,
- `http://evolution.genetics.washington.edu/phylip/newick_doc.html`.

Newick jest powszechnie stosowany w aplikacjach operujących na drzewach filogenetycznych. Idea tego formatu opiera się na wykorzystaniu notacji nawiasowej, za pomocą której hierarchicznie grupuje się węzły będące potomkami danego wierzchołka (por. rysunek 7.1). Ciąg znaków kodujących pojedyncze drzewo jest zakończony średnikiem. Newick pozwala również na zapisanie długości krawędzi. W tym przypadku ich wartości umieszczane są po znaku „:”.



Rysunek 7.1: Przykłady zapisu drzew filogenetycznych w formacie Newick.

Przetwarzanie notacji tekstowej drzewa na obiekt w pamięci, na którym przeprowadzane są dalsze operacje w TreeCmp, odbywa się za pomocą biblioteki PAL (Phylogenetic Analysis Library) [44].

Wyniki generowane przez aplikację są zapisywane w plikach tekstowych o strukturze tabelarycznej (TSV, ang. Tab Separated Values). Format ten umożliwia wygodny eksport rezultatów obliczeń do innych programów (np. MS Excel, R) w celu dalszego ich przetwarzania. Aplikacja pozwala również na wyznaczenie optymalnego dopasowania drzew, zwracając listę składającą się z par elementów, tj. rozbić, klastrów lub elementu O, tworzących najlżejsze doskonałe skojarzenie (por. definicja 3.1 oraz podrozdział 5.1).

Jedną z głównych zalet TreeCmp jest różnorodność dostępnych metod przy równoczesnej efektywnej ich implementacji. Stąd też aplikacja ta pozwala na analizowanie dużych drzew filogenetycznych, posiadających np. 5000 liści. Wykorzystanie TreeCmp możliwe jest również, w przypadku gdy porównywane drzewa posiadają różne zbiory liści $L_1 \neq L_2$. W takiej

sytuacji przed wyznaczeniem wartości danej metryki drzewa te zostają zastąpione poddrzewami indukowanymi nad wspólnym zbiorem liści $L_1 \cap L_2$.

Oprócz raportowania bezwzględnych wartości odległości, TreeCmp prezentuje również odległości znormalizowane. Normalizacja ta polega na wyznaczeniu ilorazu wartości dystansu obliczonego w danej metryce przez empiryczną wartość średnią tej odległości pomiędzy drzewami losowymi. W celu dokonania opisanej operacji aplikacja wykorzystuje uprzednio wyliczone i dołączone do programu dane, w skład których wchodzą podstawowe parametry rozkładów dla każdej z metryk w obu, rozważanych w niniejszej pracy, modelach generacji drzew losowych. Obliczenia te zostały przeprowadzone dla drzew o liczbie liści n, dla każdego n z zakresu $4, \ldots, 1000$, na podstawie 1000 par drzew losowych. Wybrane początkowe wartości dostępnych paramentów dla metryk MS i MC znajdują się w tabelach 7.9-7.12. Szerszy zestaw wyników statystycznych można znaleźć na stronie internetowej [118].

Przy wyznaczaniu wartości odległości RF i RFC w aplikacji TreeCmp zastosowano technikę hashingu. Pozwoliło to na optymalizację procesu wyszukiwania identycznych rozbić (lub klastrów) reprezentowanych jako obiekty klasy BitSet w języku Java. Mimo że otrzymany w ten sposób algorytm posiada gorszą niż liniowa złożoność obliczeniową, bardzo dobrze sprawdza się w praktyce (por. wykres 7.2).

Metryki SN i PD są wyznaczane algorytmem o złożoności kwadratowej. W przypadku metryki QT w TreeCmp zaadaptowano i usprawniono implementację algorytmu dostępną w aplikacji QuartetDist [36]. Czas obliczania metryki QT zależy od stopni wierzchołków wewnętrznych porównywanych drzew (jest kwadratowy dla drzew binarnych — przypadek optymistyczny, zaś sześcienny w najgorszym przypadku) [36].

W celu wyznaczania wartości metryki TT aplikacja używa dwóch różnych algorytmów o złożoności kwadratowej: stosunkowo prostej metody dla drzew binarnych [38] i dużo bardziej skomplikowanej nowszej metody dla drzew niebinarnych [8].

Wyszukiwanie najlżejszego doskonałego skojarzenia w grafach dwudzielnych, będące najbardziej kosztowną obliczeniowo fazą wyznaczania war-

tości metryk MS i MC, odbywa się za pomocą algorytmu podanego przez Jonkera i Volgenant'a [64] (określanego często jako LAPJV). Według analiz porównawczych przeprowadzonych w pracach [42] oraz [30] efektywność algorytmu LAPJV w porównaniu do innych dostępnych implementacji jest wysoka, dzięki czemu dobrze nadaje się on do praktycznych zastosowań.

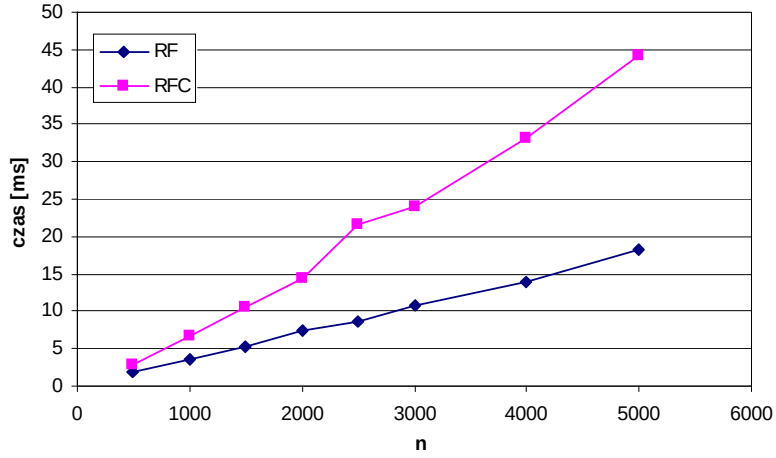
Gdy porównywane drzewa są do siebie podobne, np. w sytuacji opisanej w dalszej części rozdziału, gdzie porównujemy drzewo wzorcowe z drzewem zrekonstruowanym pewną metodą, współdzielą one na ogół dużą część swoich rozbić. Korzystając z tego faktu i posługując się formułą (3.1) możemy znacznie przyspieszyć wyznaczanie metryk skojarzeniowych w takich przypadkach. Dla drzew o 5000 liściach otrzymujemy wówczas nawet dziesięciokrotne zmniejszenie czasu obliczeń (por. tabela 7.1).

Tabela 7.1: Średni czas jednego porównania dla metryk MS i MC w przypadku drzew podobnych (kolumna sim) oraz losowych (kolumna rand). Wartości dla drzew losowych są wyznaczone na podstawie 100 porównań. W przypadku drzew podobnych wartości te są średnimi z odpowiednio: 3080 porównań dla drzew o 250 liściach (na podstawie odległości dla 308 przypadków testowych i 10 metod rekonstrukcji), 784 porównań dla 1250 liści (92 przypadki testowe, 8 metod rekonstrukcji), oraz 56 porównań dla 5000 liści (7 zestawów testowych, 8 metod rekonstrukcji).

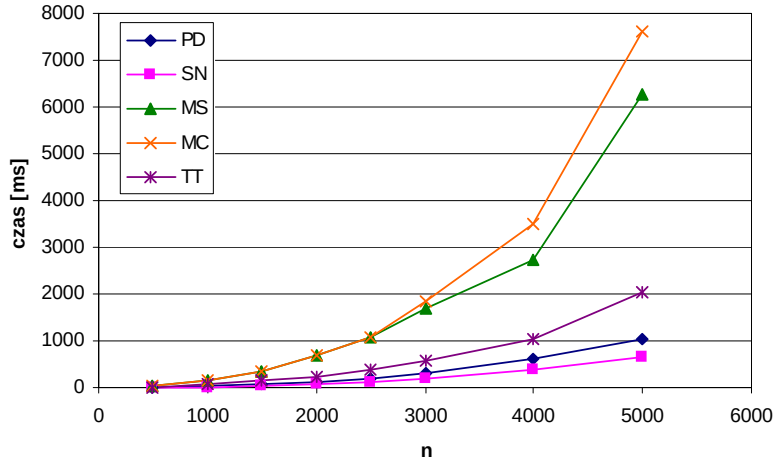
Liczba liści	Metryka MS		Metryka MC	
	rand [ms]	sim [ms]	rand [ms]	sim [ms]
250	12.2	2.16	19.66	2.9
1250	248.37	22.4	286.27	28.3
5000	6287.87	565.1	8644.62	549.04

Średni czas obliczeń dla poszczególnych metryk, wyznaczony na podstawie 100 porównań, jest przedstawiony na rysunkach 7.2, 7.3 i 7.4.

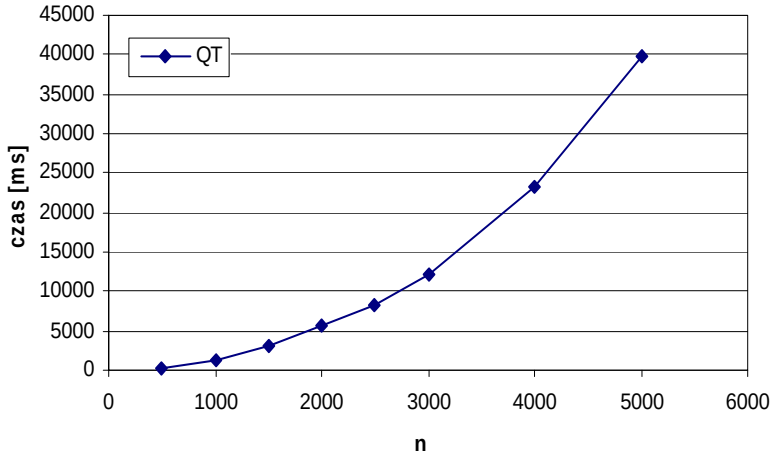
Warto zauważyć, że dostępne są również inne aplikacje pozwalające wyznaczać wartości odległości drzew filogenetycznych w klasycznych metrykach. Implementują one na ogół podzbiór metod rozważanych w niniejszej



Rysunek 7.2: Średni czas jednego porównania w metrykach RF i RFC.



Rysunek 7.3: Średni czas jednego porównania w metrykach MS, MC, PD, SN i TT.



Rysunek 7.4: Średni czas jednego porównania w metryce QT.

pracy, np. aplikacja COMPONENT 2.0 [101] pozwala na wyznaczanie metryk RF, TT, QT. Niestety dla poprawnego działania programu COMPONENT rozmiar drzew wejściowych nie może przekraczać 100 liści.

Kolejnym programem o zbliżonej funkcjonalności, umożliwiającym obliczanie odległości RF, PD, QT oraz TT, jest TOPD/FMTS [90]. Aplikacja ta jest jednak istotnie wolniejsza od TreeCmp, np. pojedyncze porównanie pary drzew o 5000 liściach przy użyciu metryki RF w TOPD/FMTS trwa ok. 2 min, w TreeCmp natomiast wynik otrzymujemy w czasie poniżej 1s; podobnie jest dla pozostałych metryk (por. tabela 7.2). Testy efektywności zostały przeprowadzone na serwerze o następujących parametrach: procesor Intel Core i7 920, 2.66GHz, pamięć 12GB RAM, system operacyjny Ubuntu 10.10.

7.2 Opis eksperymentu

Motywacja do przeprowadzenia poniższej analizy pochodzi z pracy [89], w której opisano nową heurystykę dla problemu rekonstrukcji drzew fi-

Tabela 7.2: Porównanie efektywności aplikacji TOPD/FMTS i TreeCmp.

Metryka	Liczba liści	TOPD/FMTS	TreeCmp
RF	5000	$\approx$ 2min	$<$ 1s
PD	1250	$>$ 1h	$<$ 1s
TT	100	$>$ 30min	$<$ 1s
QT	100	$>$ 20h	$<$ 1s

logenetycznych. Skuteczność zaproponowanej tam metody, zaimplementowanej w aplikacji FastTree 2, została potwierdzona eksperymentalnie przy użyciu miary jakości odpowiadającej ilorazowi poprawnie odtworzonych rozbić do ich całkowitej liczby w danym drzewie. Użyta metoda pomiaru wykorzystuje więc to samo podejście, które pojawia się w definicji metryki RF. Ponieważ metryka ta posiada pewne słabe punkty, wykazane w poprzednich rozdziałach, zasadne jest przeprowadzenie pełniejszej analizy wykorzystującej również inne dostępne metody mierzenia podobieństwa drzew filogenetycznych (w szczególności metrykę MS).

Wśród porównywanych metod rekonstrukcji znajdują się następujące algorytmy:

- RAxML 7 [105] z operacją SPR,
- PhyML 3 [55], [56] — dwa warianty, z których jeden używa operacji SPR (oznaczony dalej jako PhyML-SPR), natomiast drugi nie,
- BIONJ [54], gdzie odległości sekwencji zostały wyznaczone za pomocą aplikacji PROTDIST, która jest częścią pakietu PHYLIP [48],
- FastTree 2 [89] — dwa warianty, z których jeden używa transformacji NNI do poprawy drzewa (oznaczony dalej jako FT-NNI), w drugim przypadku zaś stosowane są wyłącznie operacje SPR (FT),
- FastME 2.06 [43],
- metoda maksymalnej parsymonii (MP) zaimplementowana w aplikacji RAxML 7.2.5,

- metoda przyłączania sąsiada (NJ) [97],

- bardziej efektywny wariant powyższej metody zaimplementowany w aplikacji Clearcut [46].

Zauważmy, że aby ocenić jakość metod rekonstrukcji musimy znać drzewo wzorcowe. Powszechnie stosowaną metodą tworzenia takich zestawów jest symulowanie ewolucji sekwencji przebiegającej wzdłuż znanego drzewa. Istnieje wiele aplikacji pozwalających na wykonanie takich symulacji, np. aplikacja Evolver z pakietu PAML [114], czy aplikacja Rose [108] użyta do wygenerowania danych opisywanych poniżej.

Zbiory sekwencji, drzew testowych i wzorcowych wykorzystane w przeprowadzonych eksperymentach pochodzą z pracy opisującej nową heurystykę FastTree [88]. Te same dane zostały również użyte w testowaniu udoskonalonej wersji tej aplikacji: FastTree 2 [89]. Danymi wejściowymi dla analizowanych metod są symulowane sekwencje aminokwasów, w których skład wchodzi: 308 różnych zestawień po 250 sekwencji, 92 zestawień zawierających 1250 sekwencji oraz 7 zestawiań mających 5000 sekwencji.

7.3 Metody pomiaru

W celu określenia dokładności badanych metod rekonstrukcji i wykazania zalet heurystyki FastTree 2, w pracy [89] wykorzystano stosunek liczby rozbić poprawnie zrekonstruowanych (tj. takich, które są obecne w drzewie wzorcowym) do ich całkowitej liczby. Podejście to bazuje na idei zaczerpniętej z konstrukcji metryki RF i gwarantuje, że otrzymany iloraz będzie w przedziale od 0 do 1. W przypadku pozostałych metryk wybór metody normalizacji jest mniej oczywisty. Za punkt odniesienia przyjmijmy więc średnią odległość w danej metryce dla drzew wygenerowanych według modelu Yule'a [115, 58]. Model ten wg [14] jest najbardziej popularnym modelem generacji drzew losowych używanym w literaturze.

Niech $T_* \in U_n$ będzie drzewem wzorcowym, zaś $T_r \in U_n$ drzewem zrekonstruowanym pewną metodą. Zdefiniujmy parametr będący *współczyn-*

nikiem podobieństwa drzew T_*, T_r następująco:

$$TA_d(T_r, T_*) = \frac{\text{avg}_d(n) - d(T_r, T_*)}{\text{avg}_d(n)}, \tag{7.1}$$

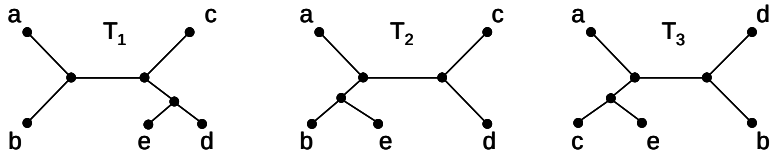
gdzie $\text{avg}_d(n)$ oznacza empiryczną średnią odległość między n-listnymi drzewami w modelu Yule'a. Parametr $TA_d(T_r, T_*)$ może być traktowany jako znormalizowana miara jakości danej metody rekonstrukcji. W tabeli 7.3 umieszczone zostały wartości $\text{avg}_d(n)$ dla $n = 250, 1250, 5000$, wyznaczone na podstawie odległości dla 10000 par drzew losowych. Zauważmy, że z faktu, iż wartość średnia metryki RF bardzo szybko zmierza do średnicy [106] wynika, że wartości parametru TA_{RF} są bardzo zbliżone do miary podobieństwa opartej na względnej liczbie rozbić odtworzonych poprawnie, zastosowanej w [89].

Tabela 7.3: Wartości średnie odległości w modelu Yule'a.

Liczba liści	MS	RF	PD	QT
250	2939.20	246.78	1112.62	105924513.79
1250	22606.81	1246.78	6608.38	67492177209.29
5000	118474.06	4996.78	29197.47	17340286310252.20

Zanim przystąpimy do analizy jakości wymienionych metod rekonstrukcji, przyjrzyjmy się bliżej dość skomplikowanemu zagadnieniu jakim jest ilościowa ocena podobieństwa dwóch drzew. Rozważmy dla przykładu drzewa przedstawione na rysunku 7.5. Możemy postawić pytanie: które z drzew T_1 czy T_3 jest bardziej podobne do drzewa T_2?

Posługując się metryką RF otrzymujemy, że stopień podobieństwa obu drzew do T_2 jest identyczny. Jednak według każdej z pozostałych metryk podobieństwo T_1 do T_2 jest większe niż w przypadku drzew T_2 i T_3. Druga odpowiedź, sugerowana przez trzy metryki, jest bardziej intuicyjna, gdyż w przypadku drzew T_1 i T_2 wystarczy usunąć tylko jeden liść e, aby otrzymać drzewa identyczne, podczas gdy podobna operacja nie ma zastosowania dla pary T_2 i T_3.



$$d_{RF}(T_1,T_2) = 2 \quad = \quad d_{RF}(T_2,T_3) = 2$$
$$d_{MS}(T_1,T_2) = 3 \quad < \quad d_{MS}(T_2,T_3) = 4$$
$$d_{PD}(T_1,T_2) \approx 3.7417 \quad < \quad d_{PD}(T_2,T_3) = 4$$
$$d_{QT}(T_1,T_2) = 4 \quad < \quad d_{QT}(T_2,T_3) = 5$$

Rysunek 7.5: Przykłady drzew wraz z ich wzajemnymi odległościami w rozpatrywanych metrykach.

Przyjrzyjmy się jeszcze raz przypadkowi gąsienic zasygnalizowanemu w podrozdziale 4.3, z których jedna powstaje z drugiej w wyniku przeniesienia skrajnego liścia a_1 na przeciwną stronę drzewa (por. rysunek 4.8). Współczynniki podobieństwa tych drzew zostały przedstawione w tabeli 7.4. Otrzymane wyniki możemy podzielić na trzy kategorie:

1. wartość TA_{PD} znacznie poniżej 0 wskazuje, że badane drzewa różnią się bardziej niż drzewa wygenerowane losowo,

2. wartość TA_{RF} w okolicy 0 wskazuje, że badane drzewa różnią się w podobnym stopniu co drzewa wygenerowane losowo,

3. wartości parametrów TA_{MS} i TA_{QT} znajdują się w zakresie powyżej 90%, co możemy interpretować jako stwierdzenie, że analizowane drzewa są do siebie bardzo podobne (tzn. znacznie bardziej niż drzewa losowe).

Zauważamy, że w obu przypadkach najbardziej intuicyjną interpretację podobieństwa otrzymujemy za pomocą metryk MS i QT. Powyższe różnice potwierdzają zasadność stosowania w analizach wielu metod pomiaru.

Tabela 7.4: Wartości parametru TA_d dla dwóch gąsienic, z których jedna powstaje z drugiej w wyniku przeniesienia skrajnego liścia na drugi koniec (rysunek 4.8).

TA_d	$n = 250$	$n = 1250$	$n = 5000$
TA_{RF}	-0.09%	-0.02%	0.00%
TA_{PD}	-103.88%	-285.64%	-598.91%
TA_{MS}	91.56%	94.48%	95.78%
TA_{QT}	97.60%	99.52%	99.88%

7.4 Wyniki analizy

Wyniki eksperymentu przedstawione są w tabelach 7.5-7.8. Wiersze są posortowane względem wartości średniej odległości w odpowiedniej metryce dla drzew o 250 liściach. Z powodu niskiej efektywności aplikacji PhyML przeprowadzanie obliczeń przy jej wykorzystaniu było możliwe tylko dla najmniejszych drzew.

Tabela 7.5: Średnie odległości i wartości współczynnika TA_d (w %) dla metryki MS.

Nr	Metoda	$n = 250$		$n = 1250$		$n = 5000$	
		d_{MS}	TA_{MS}	d_{MS}	TA_{MS}	d_{MS}	TA_{MS}
1	RAxML	222.98	92.41	2256.15	90.02	12410.0	89.53
2	PhyML-SPR	234.31	92.03	-	-	-	-
3	FT-NNI	293.95	90.00	3325.75	85.29	19830.4	83.26
4	PhyML	313.62	89.33	-	-	-	-
5	BIONJ	482.04	83.60	4418.58	80.45	20555.4	82.65
6	FT	499.15	83.02	4402.59	80.53	23219.1	80.40
7	MP	510.11	82.64	4418.33	80.46	27135.7	77.10
8	FastME	510.89	82.62	4501.62	80.09	23796.0	79.91
9	NJ	556.27	81.07	5086.80	77.50	25391.6	78.57
10	Clearcut	610.25	79.24	5183.72	77.07	25276.9	78.66

Tabela 7.6: Średnie odległości i wartości współczynnika TA_d (w %) dla metryki RF.

Nr	Metoda	$n = 250$		$n = 1250$		$n = 5000$	
		d_{RF}	TA_{RF}	d_{RF}	TA_{RF}	d_{RF}	TA_{RF}
1	RAxML	23.55	90.46	145.03	88.37	577.4	88.44
2	PhyML-SPR	24.85	89.93	-	-	-	-
3	FT-NNI	32.28	86.92	203.48	83.68	786.0	84.27
4	PhyML	34.55	86.00	-	-	-	-
5	FastME	48.06	80.52	264.09	78.82	1148.0	77.03
6	FT	48.36	80.41	270.71	78.29	1168.1	76.62
7	MP	52.33	78.80	268.85	78.44	1429.4	71.39
8	BIONJ	55.19	77.63	328.57	73.65	1343.4	73.11
9	NJ	59.23	76.00	341.9	72.58	1420.7	71.57
10	Clearcut	60.57	75.45	346.08	72.24	1423.4	71.51

Tabela 7.7: Średnie odległości i wartości współczynnika TA_d (w %) dla metryki PD.

Nr	Metoda	$n = 250$		$n = 1250$		$n = 5000$	
		d_{PD}	TA_{PD}	d_{PD}	TA_{PD}	d_{PD}	TA_{PD}
1	RAxML	245.42	77.94	1982.16	70.01	10844.4	62.86
2	PhyML-SPR	254.38	77.14	-	-	-	-
3	FT-NNI	303.92	72.68	2603.88	60.60	17695.3	39.39
4	PhyML	325.41	70.75	-	-	-	-
5	BIONJ	438.70	60.57	3326.51	49.66	14568.2	50.10
6	FT	440.38	60.42	3243.49	50.92	14911.7	48.93
7	FastME	449.37	59.61	3420.10	48.25	15564.0	46.69
8	MP	452.24	59.35	3284.08	50.30	18126.6	37.92
9	NJ	483.03	56.59	3749.12	43.27	16207.2	44.49
10	Clearcut	541.15	51.36	3927.25	40.57	15842.3	45.74

Tabela 7.8: Średnie odległości i wartości współczynnika TA_d (w %) dla metryki QT.

Nr	Metoda	$n = 250$		$n = 1250$		$n = 5000$	
		d_{QT}	TA_{QT}	d_{QT}	TA_{QT}	d_{QT}	TA_{QT}
1	RAxML	$4.35 \cdot 10^6$	95.89	$5.67 \cdot 10^9$	91.60	$1.75 \cdot 10^{12}$	89.90
2	PhyML-SPR	$4.76 \cdot 10^6$	95.50	-	-	-	-
3	FT-NNI	$8.16 \cdot 10^6$	92.29	$1.01 \cdot 10^{10}$	85.01	$2.87 \cdot 10^{12}$	83.44
4	PhyML	$8.86 \cdot 10^6$	91.63	-	-	-	-
5	BIONJ	$1.40 \cdot 10^7$	86.77	$1.30 \cdot 10^{10}$	80.74	$2.89 \cdot 10^{12}$	83.34
6	MP	$1.57 \cdot 10^7$	85.21	$1.39 \cdot 10^{10}$	79.36	$5.47 \cdot 10^{12}$	68.46
7	FT	$1.64 \cdot 10^7$	84.56	$1.46 \cdot 10^{10}$	78.39	$4.48 \cdot 10^{12}$	74.16
8	FastME	$1.70 \cdot 10^7$	83.99	$1.51 \cdot 10^{10}$	77.68	$3.94 \cdot 10^{12}$	77.28
9	NJ	$1.81 \cdot 10^7$	82.93	$1.64 \cdot 10^{10}$	75.71	$4.94 \cdot 10^{12}$	71.51
10	Clearcut	$1.95 \cdot 10^7$	81.63	$1.64 \cdot 10^{10}$	75.69	$4.94 \cdot 10^{12}$	71.53

Zauważmy, że w większości przypadków klasyfikacja jakości rozpatrywanych metod rekonstrukcji według użytych metryk jest zgodna. Otrzymujemy bowiem tę samą kolejność na pozycjach 1-4 oraz 9-10. Różnice pojawiają się w klasyfikacji czterech pozostałych metod: BIONJ, FastTree 2.0.0, FastME oraz maksymalnej parysmonii (MP). Po pierwsze według każdej z metryk metoda maksymalnej parsymonii dla bardzo dużych drzew (5000 liści) daje najgorsze rezultaty. Po drugie, na podstawie odległości MS, QT i PD (z wyjątkiem przypadku 1250 liści) spośród wspomnianych czterech metod, BIONJ rekonstruuje drzewa najlepiej. Pomijając przypadek QT dla 5000 liści, wszystkie metryki z wyjątkiem RF wskazują, że dokładność rekonstrukcji FastTree 2.0.0 jest wyższa niż FastME.

Kolejna obserwacja dotyczy wartości parametru TA_d. W przypadku metryki PD wartości te są znacznie niższe niż dla pozostałych trzech odległości. Znajdują się one w przedziale 37.92 – 77.94%, podczas gdy dla odległości RF mamy 71.39 – 90.46%, zaś dla metryk MS i QT odpowiednio 77.10 – 92.41% oraz 68.46 – 95.89%.

Tabela 7.9: Parametry rozkładów odległości w metryce MS w modelu UM, gdzie $\bar{x}$ — wartość średnia, σ — odchylenie standardowe.

n	$\bar{x}$	σ	Centyle												
			0.02	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.97
4	1.4	0.9	0	0	0	0	0	2	2	2	2	2	2	2	2
5	2.6	0.9	0	0	2	2	2	3	3	3	3	3	4	4	4
6	5.1	1.4	2	3	3	4	4	5	6	6	6	6	6	7	7
7	7.5	1.6	3	4	5	6	7	7	8	8	9	9	9	10	10
8	11.0	2.0	6	7	8	9	10	11	11	12	12	13	13	14	14
9	14.4	2.3	9	10	11	13	13	14	15	15	16	16	17	18	18
10	18.7	2.8	12	14	15	16	17	18	19	20	20	21	22	22	23
11	22.8	3.1	16	17	19	20	21	22	23	24	25	25	26	27	28
12	27.9	3.5	19	22	24	25	26	27	28	29	30	31	32	33	34
13	33.1	3.9	25	26	28	30	31	32	33	34	35	36	38	39	40
14	38.7	4.5	29	31	33	35	36	38	39	40	41	42	44	46	47
15	44.5	4.8	34	37	38	41	42	44	45	46	47	48	51	52	53
16	50.7	5.6	39	42	44	46	48	50	51	52	54	55	58	60	61
17	57.3	6.2	45	47	49	52	54	56	57	59	60	63	65	67	69
18	64.4	6.9	50	53	56	59	61	62	65	66	68	70	73	75	77
19	71.3	7.1	56	60	63	66	67	69	71	73	75	77	80	82	85
20	78.8	7.9	62	66	69	72	74	77	79	81	83	85	89	92	94
21	86.7	8.5	70	73	76	80	82	84	86	88	91	94	98	100	103
22	94.9	9.4	76	81	83	87	89	92	94	97	100	103	107	111	113
23	102.5	9.7	82	87	90	94	97	100	102	105	108	111	115	119	120
24	111.3	10.9	88	94	98	102	105	108	111	114	117	121	126	129	131
25	120.4	11.4	96	102	106	111	115	118	120	123	126	130	135	140	142
26	129.1	11.5	106	110	114	119	123	126	128	132	135	139	144	148	151
27	137.9	12.3	114	119	121	127	131	135	138	141	144	148	154	159	162
28	147.7	14.0	120	125	129	136	140	144	147	151	155	159	165	172	176
29	157.9	14.9	128	134	139	145	150	154	158	161	165	170	177	182	187
30	167.7	15.3	139	143	148	154	159	163	167	171	175	181	187	194	198
40	281.1	25.0	234	243	249	259	267	274	280	287	294	302	314	323	330
50	414.9	37.5	344	355	368	383	393	402	413	423	434	448	466	479	488
60	563.3	50.6	470	485	502	520	536	548	559	571	586	603	633	654	669
70	734.8	65.5	610	634	652	679	699	715	731	749	764	789	824	848	863
80	920.2	79.2	787	809	824	853	871	891	912	933	959	989	1021	1063	1081
90	1118.9	102.8	928	958	992	1034	1060	1087	1113	1136	1167	1200	1253	1300	1335
100	1334.2	123.8	1112	1145	1181	1224	1260	1296	1327	1354	1394	1439	1495	1548	1588

Tabela 7.10: Parametry rozkładów odległości w metryce MS w modelu YM, gdzie $\bar{x}$ — wartość średnia, σ — odchylenie standardowe.

n	$\bar{x}$	σ	Centyle												
			0.02	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.97
4	1.4	0.9	0	0	0	0	2	2	2	2	2	2	2	2	2
5	2.6	1.0	0	0	2	2	2	3	3	3	3	3	4	4	4
6	5.0	1.5	2	2	3	4	4	5	6	6	6	6	7	7	7
7	7.4	1.7	3	4	5	6	7	7	8	8	9	9	9	9	10
8	10.7	2.2	6	7	8	9	10	10	11	12	12	12	13	14	14
9	14.0	2.2	9	10	11	12	13	14	14	15	15	16	17	17	17
10	17.7	2.7	12	13	14	15	16	17	18	19	19	20	21	22	22
11	21.7	2.9	15	17	18	19	20	21	22	23	23	24	25	26	27
12	26.4	3.5	18	20	22	24	25	26	26	27	28	29	31	32	32
13	30.4	3.7	22	24	26	27	29	30	31	32	33	34	35	36	36
14	35.6	4.1	27	29	30	32	34	35	36	37	38	39	41	42	43
15	40.0	4.4	30	33	34	36	38	39	40	41	42	44	45	47	48
16	45.5	4.6	36	38	39	42	43	44	46	46	48	49	51	53	54
17	50.8	4.9	41	42	44	47	48	50	51	52	53	55	57	59	60
18	56.7	5.1	47	48	50	52	54	56	57	58	59	61	63	65	66
19	62.3	5.6	50	53	56	58	59	61	62	64	65	67	69	72	73
20	68.3	6.1	56	58	61	63	65	67	68	70	72	73	76	78	80
21	74.3	6.6	60	64	66	69	71	73	75	76	78	80	83	85	87
22	81.0	6.7	66	70	73	75	77	80	81	83	85	86	90	92	93
23	86.9	7.1	73	76	78	81	83	85	87	88	90	93	96	98	100
24	93.7	7.4	78	81	84	87	90	92	94	96	98	100	103	106	108
25	100.4	7.5	85	88	91	94	96	98	100	102	105	107	110	112	114
26	107.7	8.2	90	94	97	101	104	106	108	110	112	114	118	122	123
27	114.3	8.1	98	101	104	108	110	112	114	116	118	121	125	128	130
28	121.2	8.5	103	107	110	114	117	119	121	123	126	128	132	135	137
29	128.4	9.0	111	113	117	121	124	126	128	131	133	136	140	143	146
30	136.2	9.1	118	121	125	129	131	134	136	138	141	143	148	151	154
40	215.4	12.9	190	196	199	205	209	211	214	218	221	226	233	237	242
50	305.1	16.4	274	279	285	291	297	301	304	308	313	319	326	332	338
60	400.3	20.0	361	370	376	384	389	394	399	404	410	416	426	436	442
70	504.3	24.7	460	467	474	484	490	496	502	508	516	525	537	550	556
80	611.0	28.8	558	567	576	587	594	602	609	615	625	635	648	658	668
90	723.7	31.5	665	675	685	696	705	713	722	731	741	751	765	779	787
100	840.2	33.5	775	788	798	811	822	831	840	847	856	869	884	896	904

Tabela 7.11: Parametry rozkładów odległości w metryce MC w modelu UM, gdzie $\bar{x}$ — wartość średnia, σ - odchylenie standardowe.

n	$\bar{x}$	σ	Centyle												
			0.02	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.97
4	3.4	1.3	0	2	2	2	3	3	3	4	4	4	6	6	6
5	5.9	1.7	2	3	4	4	6	6	6	6	6	7	8	8	10
6	9.2	2.3	4	5	6	7	8	9	9	10	10	11	12	13	13
7	12.7	2.9	7	8	9	10	11	12	13	14	14	15	16	18	18
8	17.0	3.1	11	12	13	14	15	16	17	18	18	20	21	22	23
9	21.3	3.9	14	15	16	18	19	20	21	22	23	24	26	28	28
10	26.4	4.5	17	19	21	23	24	25	26	27	28	30	32	34	36
11	32.0	4.9	22	24	26	28	29	31	32	33	34	36	38	40	41
12	37.8	5.7	26	29	31	33	35	36	38	39	41	42	45	47	49
13	44.1	6.4	32	34	36	39	41	42	44	45	47	49	53	56	57
14	50.5	7.0	37	39	42	44	47	48	50	52	54	56	60	62	64
15	57.5	8.0	42	45	48	51	53	55	57	59	61	64	68	71	74
16	64.6	8.5	47	51	54	58	60	62	64	67	69	71	75	79	81
17	72.7	9.5	55	58	61	65	67	70	72	74	77	81	86	90	91
18	80.4	10.0	62	65	68	72	75	77	80	82	85	89	94	98	102
19	88.9	11.5	67	72	75	79	82	85	88	91	94	98	104	109	113
20	97.3	11.9	74	79	83	87	91	94	96	99	103	106	112	118	122
21	105.8	12.6	82	87	91	96	99	102	105	108	112	116	122	127	131
22	115.8	14.5	88	94	99	104	107	111	115	119	122	127	136	141	144
23	123.9	14.8	95	101	106	111	116	120	123	127	131	135	143	150	155
24	135.4	16.6	104	111	115	121	126	130	134	138	143	149	159	165	168
25	144.3	17.5	113	118	123	130	135	138	142	147	152	158	168	175	181
26	156.4	19.3	122	127	134	141	146	150	155	159	164	172	180	192	197
27	164.5	20.0	129	135	141	147	152	158	162	168	174	180	191	201	206
28	175.8	20.3	134	144	150	159	165	170	175	180	185	192	202	210	215
29	186.7	21.2	145	154	162	168	175	181	185	190	196	204	215	222	230
30	198.1	23.2	156	164	170	178	184	190	197	202	209	216	229	239	245
40	327.5	38.8	258	271	281	295	305	315	324	333	344	360	380	397	409
50	478.6	55.1	382	397	412	432	448	460	471	486	504	522	553	582	598
60	643.9	75.3	511	535	555	581	601	619	636	652	675	702	738	784	823
70	829.4	97.6	665	694	718	748	771	794	816	843	869	906	958	1013	1039
80	1034.3	119.9	823	859	889	928	964	998	1029	1055	1092	1128	1186	1233	1281
90	1247.8	148.0	1015	1043	1080	1123	1162	1197	1232	1267	1306	1351	1428	1507	1564
100	1485.6	171.7	1200	1236	1279	1342	1381	1424	1467	1515	1561	1622	1718	1790	1859

Tabela 7.12: Parametry rozkładów odległości w metryce MC w modelu YM, gdzie $\bar{x}$ — wartość średnia, σ — odchylenie standardowe.

n	$\bar{x}$	σ	Centyle												
			0.02	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.97
4	3.3	1.4	0	0	2	2	3	3	3	4	4	4	6	6	6
5	6.0	1.7	2	3	4	4	6	6	6	6	7	7	8	8	10
6	9.0	2.2	4	6	6	7	8	8	9	10	10	11	12	13	13
7	12.6	2.6	7	8	9	10	12	12	13	13	14	14	16	17	18
8	16.4	3.1	10	11	13	14	15	16	16	17	18	19	20	21	22
9	20.6	3.7	12	14	16	18	19	20	21	21	22	24	25	26	28
10	25.4	4.0	17	19	21	22	23	24	25	26	28	29	31	32	33
11	30.4	4.7	20	23	24	26	28	29	30	32	33	34	36	38	39
12	35.5	5.0	25	27	29	31	33	34	36	37	38	39	42	44	45
13	40.8	5.7	29	32	34	36	38	39	41	42	43	45	48	50	52
14	46.7	6.2	34	36	39	42	44	45	47	48	50	52	55	57	59
15	51.8	6.5	39	42	44	46	48	50	51	53	55	57	60	63	65
16	58.4	7.0	45	48	50	52	55	56	58	60	62	64	68	70	72
17	64.6	7.6	50	53	55	58	60	63	64	66	68	71	74	78	80
18	71.0	8.3	55	58	61	64	66	69	71	73	75	78	81	84	88
19	77.5	9.0	60	64	66	70	73	75	77	79	82	84	89	93	96
20	85.2	9.4	67	70	74	78	80	82	85	87	89	93	98	102	104
21	91.6	10.3	73	77	79	83	85	88	91	93	96	100	105	110	113
22	98.8	11.2	78	83	85	89	93	95	98	100	104	108	114	119	122
23	106.0	11.5	85	89	92	97	100	103	105	107	111	115	121	127	129
24	113.3	11.8	91	95	99	104	107	110	112	115	118	122	130	135	138
25	121.2	12.6	98	102	106	111	114	118	120	123	126	131	137	144	147
26	129.2	13.0	106	110	114	118	122	125	128	131	135	139	147	153	157
27	136.5	14.5	110	114	119	125	128	132	135	139	143	148	156	162	166
28	144.3	14.5	119	123	127	132	136	139	143	147	151	155	164	170	174
29	151.7	15.4	125	130	134	139	143	147	150	154	158	163	172	180	185
30	160.9	15.7	132	137	142	148	152	156	159	163	168	173	181	190	195
40	249.5	23.6	211	216	222	230	236	241	247	252	259	268	280	290	303
50	346.0	31.5	295	302	311	321	328	335	341	348	357	369	386	403	417
60	455.0	40.6	393	403	412	421	431	439	447	458	468	485	510	535	549
70	564.7	48.9	492	503	511	524	536	546	556	566	580	601	633	658	676
80	683.7	60.5	598	611	621	636	648	660	670	683	701	726	766	797	831
90	803.6	67.4	706	720	733	750	764	776	789	802	825	852	893	932	960
100	932.8	76.8	819	837	849	872	888	902	917	936	958	986	1037	1079	1119

8 Podsumowanie

W pracy sformułowano ogólną metodę konstrukcji metryk dla dowolnych drzew filogenetycznych, tj. nieukorzenionych, ukorzenionych, binarnych oraz posiadających multifurkacje (rozdział 3). Na szczególną uwagę zasługuje fakt wykorzystania w tej metodzie nowatorskiej koncepcji elementu dodatkowego, która umożliwia łatwe i zarazem elastyczne budowanie metryk dla drzew niebinarnych. Wykorzystując opisaną metodę, zaproponowano dwie metryki MS i MC. Metryka MS jest nową metodą, której definicja pojawiła się po raz pierwszy w pracy [16].

Oryginalnym wkładem jest również szczegółowa analiza własności metryk MS i MC, przedstawiona w rozdziałach 4-7, wraz z aplikacją TreeCmp umożliwiającą wygodne wyznaczanie podobieństwa drzew filogenetycznych za pomocą odległości rozważanych w pracy (tj. MS, MC, RF, RFC, PD, SN, QT, TT). Przeprowadzona analiza wskazuje, że zaproponowane metody posiadają wiele cech, które wydają się być intuicyjne i pożądane przy ocenie podobieństwa drzew filogenetycznych. Co więcej, przedstawione odległości mogą być traktowane jako rozszerzenie klasycznych metryk, tj. RF i RFC, gdyż w swoich definicjach wykorzystują one te same fundamentalne elementy, czyli rozbicia i klastry. W przypadku metryk MS i MC wzajemne podobieństwo wspomnianych elementów jest jednak oceniane bardziej precyzyjnie (tzn. niebinarnie).

Kierunki dalszych badań mogą obejmować rozszerzenie zaproponowanych metod na inne obiekty, takie jak:

- drzewa filogenetyczne z dodatnimi wagami liczbowymi na krawędziach. Takie drzewa ważone przedstawiają pełniejszy, niż drzewa nieobciążone, obraz historii ewolucji, gdyż wagi przyjmują na ogół wartości proporcjonalne do czasu jaki upłynął między poszczególny-

mi specjacjami. Przybliżonych długości krawędzi dostarczają, m.in. metody bayesowskie, NJ lub ML.

- sieci filogenetyczne. Zjawiska takie jak np. ponowne krzyżowanie się osobnych linii gatunkowych lub horyzontalny transfer genów sprawiają, że popularnym staje się opisywanie historii ewolucji za pomocą struktur ogólniejszych niż drzewa, nazywanych *sieciami filogenetycznymi*. Sieci te są skierowanymi ukorzenionymi grafami acyklicznymi o zbiorze liści L, które odpowiadają współczesnym gatunkom. Na strukturę sieci filogenetycznych często nakładane są dodatkowe ograniczenia. Proponowane są także różne sposoby ich porównywania poprzez wprowadzanie struktury przestrzeni metrycznej w rodzinie sieci o danym zbiorze liści L. Metody te w wielu przypadkach opierają się na uogólnieniach znanych metryk filogenetycznych dla drzew (takich jak np. RF czy odległości węzłowe [31], [32]). Warto zwrócić uwagę, że definicja 3.1 (przy wykorzystaniu metryki h_{MC}) określa odległość między dowolnymi rodzinami niepustych podzbiorów L, jest to więc także metryka dla sieci filogenetycznych w każdym modelu, w którym sieć taka jest jednoznacznie wyznaczona przez rodzinę klastrów jej węzłów wewnętrznych, czyli np. dla sieci typu *Tree-Child Time Consistent* [32].

- drugorzędowe struktury RNA. Kwas rybonukleinowy RNA występuje na ogół w postaci jednoniciowej. Jego *pierwszorzędowa struktura*, tj. sekwencja nukleotydów jest przedstawiana jako ciąg liter nad alfabetem $\{A, C, G, U\}$. Natomiast struktura RNA określana jako *drugorzędowa* zawiera dodatkowo informacje odnośnie lokalizacji w tej sekwencji zasad sparowanych. Porównywanie drugorzędowych struktur RNA jest jednym z podstawowych problemów obliczeniowych towarzyszących analizie RNA [5]. Ponieważ struktury te mogą być reprezentowane jako ukorzenione drzewa uporządkowane [78, 5, 51], ocena przydatności oraz ewentualna adaptacja zaproponowanego w pracy podejścia skojarzeniowego do porównywania obiektów tego typu wydają się być interesującymi obszarami dalszych badań.

- drzewa filogenetyczne o różnych zbiorach liści [69]. Niekiedy konieczne jest agregowanie i porównywanie informacji z wyników badań filogenetycznych drzew T_1 i T_2 dotyczących różnych (tylko częściowo pokrywających się) zbiorów gatunków (liści) L_1 i L_2. Tradycyjne podejście polega tu na wykorzystaniu klasycznych metryk filogenetycznych dla drzew indukowanych przez wspólny zbiór liści, tj. dla $T_{1|L_1\cap L_2}$ i $T_{2|L_1\cap L_2}$. W literaturze pojawiają się także podejścia polegające na konstruowaniu wspólnej przestrzeni metrycznej w zbiorze wszystkich drzew, czyli w $\bigcup_{L\in 2^{\mathbb{Z}_{>0}}} R_L$, $1 < |L| < \infty$ dla drzew ukorzenionych i w $\bigcup_{L\in 2^{\mathbb{Z}_{>0}}} U_L$, $2 < |L| < \infty$ dla nieukorzenionych [69]. W przypadku drzew ukorzenionych definicja metryki MC może być przeniesiona bez zmian. Rozszerzenie odległości MS na zbiór wszystkich drzew nieukorzenionych wymaga dalszych badań.

Dane pochodzące z metod filogenetycznych generujących duże zbiory drzew o znacznej wiarygodności (np. metody bayesowskie) mogą być w następnej kolejności przetwarzane za pomocą klasteryzacji, mającej na celu wyznaczenie osobnych skupisk złożonych z podobnych drzew. Skupiska te mają określać alternatywne możliwe rozwiązania [107]. Zastosowanie do tego problemu popularnych metod klasteryzacji zbiorów punktów z $\mathbb{R}^n$ (np. algorytm k-średnich) utrudnia fakt, iż nie jest łatwe określenie „środka ciężkości" czy „średniego drzewa" dla podzbioru niekompatybilnych drzew tworzących jeden klaster. W pracy [107] zaproponowano adaptację algorytmu k-średnich, gdzie drzewa porównywane są przy użyciu metryki RF, zaś funkcję środka rodziny drzew pełni odpowiednio wybrany zbiór rozbić (niekoniecznie kompatybilnych ze sobą) występujących w danej rodzinie. Tę metodę postępowania w naturalny sposób można zastosować dla metryk określonych zgodnie definicją 3.1, która pozwala porównywać dowolne podzbiory 2^L lub $Splits(L)$, również nietworzące drzew. Weryfikacja skuteczności tego podejścia wymaga dalszych badań eksperymentalnych.

Bardzo interesujące wydaje się również zbadanie przydatności zaproponowanych metryk do konstrukcji efektywnych heurystyk dla problemu identyfikacji horyzontalnego transferu genów (HGT), gdyż względnie niedawno (2010 rok) w pracy [15] przedstawiono heurystykę dla tego proble-

mu jawnie korzystającą z informacji dostarczanych przez miary odległości w przestrzeni drzew. Spośród przebadanych w [15] metod (tj. RF, PD, QT, BD) najlepsze wyniki otrzymano dla miary BD (zdefiniowanej w tej samej pracy), opierającej się na metryce w zbiorze rozbić h_{MS}. Niestety, zaproponowana miara BD nie jest metryką, stąd z racji zbliżonego charakteru definicji naturalne wydaje się pytanie o określenie stopnia przydatności metryki MS w tym zagadnieniu.

Bibliografia

[1] E. N. Adams. Consensus techniques and the comparison of taxonomic trees. *Systematic Biology*, 21:390–397, 1972.

[2] R. Alberich, G. Cardona, F. Rosselló, G. Valiente. An algebraic metric for phylogenetic trees. *Applied Mathematics Letters*, 22:1320–1324, 2009.

[3] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Statistical Science*, 16:23–34, 2001.

[4] S. Alizon, V. von Wyl, T. Stadler, R. D. Kouyos, S. Yerly, B. Hirschel, J. Böni, C. Shah, T. Klimkait, H. Furrer, A. Rauch, P. L. Vernazza, E. Bernasconi, M. Battegay, P. Bürgisser, A. Telenti, H. F. Günthard, S. Bonhoeffer, the Swiss HIV Cohort Study. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathogens*, 6:e1001123, 2010.

[5] J. Allali, M.-F. Sagot. A new distance for high level rna secondary structure comparison. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:3–14, 2005.

[6] B. L. Allen, M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–15, 2001.

[7] A. Amir, D. Keselman. Maximum agreement subtree in a set of evolutionary trees. *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, 758–769, 1994.

[8] M. S. Bansal, J. Dong, D. Fernández-Baca. Comparing and aggregating partially resolved trees. *Theoretical Computer Science*, 412:6634 – 6652, 2011.

[9] A. C. Barbrook, C. J. Howe, N. Blake, P. Robinson. The phylogeny of the canterbury tales. *Nature*, 394:839, 1998.

[10] J.-P. Barthélemy, F. R. McMorris. The median procedure for n-trees. *Journal of Classification*, 3:329–334, 1986.

[11] A. D. Baxevanis, B. F. F. Ouellette. *Bioinformatyka podręcznik do analizy genów i białek*. Wydawnictwo Naukowe PWN, 2005.

[12] R. Beiko, N. Hamilton. Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology*, 6:15, 2006.

[13] J. Bluis, D.-G. Shin. Nodal distance algorithm: Calculating a phylogenetic tree comparison metric. *Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*, 87–94, 2003.

[14] M. G. B. Blum, O. François, S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*, 16:2195–2214, 2006.

[15] A. Boc, H. Philippe, V. Makarenkov. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic Biology*, 59:195–211, 2010.

[16] D. Bogdanowicz. Comparing phylogenetic trees using a minimum weight perfect matching. *Information Technology, 2008, 1st International Conference on*, 451–454, 2008.

[17] D. Bogdanowicz. Analyzing sets of phylogenetic trees using metrics. *Applicationes Mathematicae*, 8:1–16, 2011.

[18] D. Bogdanowicz, K. Giaro. Comparing arbitrary unrooted phylogenetic trees using generalized matching split distance. *Information*

Technology (ICIT), 2010 2nd International Conference on, 259 –262, 2010.

[19] D. Bogdanowicz, K. Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9:150–160, 2012.

[20] L. Bolikowski, A. Gambin. New metrics for phylogenies. *Fundamenta Informaticae*, 78:199–216, 2007.

[21] C. Bonnard, V. Berry, N. Lartillot. Multipolar consensus for phylogenetic trees. *Systematic Biology*, 55:837–843, 2006.

[22] S. A. Boorman, D. C. Olivier. Metrics on spaces of finite trees. *Journal of Mathematical Psychology*, 10:26–59, 1973.

[23] M. Bordewich, C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8:409–423, 2005.

[24] L. M. Boykin, L. S. Kubatko, T. K. Lowrey. Comparison of methods for rooting phylogenetic trees: A case study using orcuttieae (poaceae: Chloridoideae). *Molecular Phylogenetics and Evolution*, 54:687–700, 2010.

[25] G. S. Brodal, R. Fagerberg, C. N. Pedersen. Computing the quartet distance between evolutionary trees in time $O(n \log\ n)$. *Algorithmica*, 38:377–395, 2004.

[26] D. Bryant. *Building Trees, Hunting for Trees, and Comparing Trees – Theory and Methods in Phylogenetic Analysis*. Praca doktorska, Department of Mathematics, University of Canterbury, 1997.

[27] D. Bryant. A classification of consensus methods for phylogenetics. *Bioconsensus*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. DIMACS-AMS, 2001.

[28] D. Bryant. The splits in the neighborhood tree. *Annals of Combinatorics*, 8:1–11, 2004.

[29] P. Buneman. The recovery of trees from measures of dissimilarity. *Mathematics the the Archeological and Historical Sciences*, 387–395. Edinburgh University Press, 1971.

[30] R. Burkard, M. Mauro Dell'Amico, S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, 2009.

[31] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente. Metrics for phylogenetic networks I: Generalizations of the Robinson-Foulds metric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:46–61, 2009.

[32] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente. Metrics for phylogenetic networks II: Nodal and triplets metrics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:454–469, 2009.

[33] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente. Nodal distances for rooted phylogenetic trees. *Journal of Mathematical Biology*, 61:253–276, 2010.

[34] D. Chen, O. Eulenstein, D. Fernández-Baca, J. Burleigh. Improved heuristics for minimum-flip supertree construction. *Evolutionary Bioinformatics*, 2:347–356, 2006.

[35] B. Chor, T. Tuller. Finding a maximum likelihood tree is hard. *Journal of the ACM*, 53:722–744, 2006.

[36] C. Christiansen, T. Mailund, C. Pedersen, M. Randers, M. Stissing. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, 1:16, 2006.

[37] R. Cole, M. Farach, R. Hariharan, T. Przytycka, M. Thorup. An $O(n \log n)$ algorithm for the maximum agreement subtree problem for binary trees. *SIAM Journal on Computing*, 30:1385–1404, 2001.

[38] D. E. Critchlow, D. K. Pearl, C. Qian. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45:323–334, 1996.

[39] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, L. Zhang. On distances between phylogenetic trees. *Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, 427–436, 1997.

[40] W. H. E. Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2:7–28, 1985.

[41] W. H. E. Day, D. S. Johnson, D. Sankoff. The computational complexity of inferring rooted phylogenles by parsimony. *Mathematical Biosciences*, 81:33–42, 1986.

[42] M. Dell'Amico, P. Toth. Algorithms and codes for dense assignment problems: the state of the art. *Discrete Applied Mathematics*, 100:17–48, 2000.

[43] R. Desper, O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9:687–705, 2002.

[44] A. Drummond, K. Strimmer. PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, 17:662–663, 2001.

[45] G. F. Estabrook, F. R. McMorris, C. A. Meacham. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Biology*, 34:193–200, 1985.

[46] J. Evans, L. Sheneman, J. Foster. Relaxed neighbor joining: A fast distance-based phylogenetic tree construction method. *Journal of Molecular Evolution*, 62:785–792, 2006.

[47] J. S. Farris. On comparing the shapes of taxonomic trees. *Systematic Biology*, 22:50–54, 1973.

[48] J. Felsenstein. PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[49] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

[50] D. J. Ford. Probabilities on cladogram: Introduction to the alpha model. arXiv:math/0511246v1 — część rozprawy doktorskiej, 2005.

[51] D. Fukagawa, T. Tamura, A. Takasu, E. Tomita, T. Akutsu. A clique-based method for the edit distance between unordered trees and its application to analysis of glycan structures. *BMC Bioinformatics*, 12:S13, 2011.

[52] H. N. Gabow, R. E. Tarjan. Faster scaling algorithms for network problems. *SIAM Journal on Computing*, 18:1013–1036, 1989.

[53] B. Gaschen, J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, B. Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296:2354–2360, 2002.

[54] O. Gascuel. BIONJ: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14:685–695, 1997.

[55] S. Guindon, F. Delsuc, J.-F. Dufayard, O. Gascuel. Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology*, 537:113–137, 2009.

[56] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59:307–321, 2010.

[57] D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21:19–28, 1991.

[58] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3:44–77, 1971.

[59] G. Hickey, F. Dehne, A. Rau-Chaplin, C. Blouin. SPR distance computation for unrooted trees. *Evolutionary Bioinformatics*, 4:17–27, 2008.

[60] P. G. Higgs, T. K. Attwood. *Bioinformatyka i ewolucja molekularna*. Wydawnictwo Naukowe PWN, 2008.

[61] T. Hill, K. Nordström, M. Thollesson, T. Säfström, A. Vernersson, R. Fredriksson, H. Schiöth. Sprit: Identifying horizontal gene transfer in rooted phylogenetic trees. *BMC Evolutionary Biology*, 10:42, 2010.

[62] D. M. Hillis, T. A. Heath, K. S. John. Analysis and visualization of tree space. *Systematic Biology*, 54:471–482, 2005.

[63] P. Humphries. Bounds on the size of the tbr unit-neighbourhood. *Annals of Combinatorics*, 14:479–485, 2010.

[64] R. Jonker, A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38:325–340, 1987.

[65] M.-Y. Kao, T.-W. Lam, W.-K. Sung, H.-F. Ting. A decomposition theorem for maximum weight bipartite matchings with applications to evolutionary trees. *Lecture Notes in Computer Science*, 1643:694–694, 1999.

[66] M.-Y. Kao, T.-W. Lam, W.-K. Sung, H.-F. Ting. An even faster and more unifying algorithm for comparing trees via unbalanced bipartite matchings. *Journal of Algorithms*, 40:212–233, 2001.

[67] M. Karim, A. Walenstein, A. Lakhotia, L. Parida. Malware phylogeny generation using permutations of code. *Journal in Computer Virology*, 1:13–23, 2005.

[68] V. Knoop. The mitochondrial dna of land plants: peculiarities in phylogenetic perspective. *Current Genetics*, 46:123–139, 2004.

[69] J. Koperwas, K. Walczak. Tree edit distance for leaf-labelled trees on free leafset and its comparison with frequent subsplit dissimilarity and popular distance measures. *BMC Bioinformatics*, 12:204, 2011.

[70] M. K. Kuhner, J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11:459–68, 1994.

[71] M. Li, J. Tromp, L. Zhang. On the nearest neighbour interchange distance between evolutionary trees. *Journal of Theoretical Biology*, 182:463–467, 1996.

[72] D. R. Maddison. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Biology*, 40:315–328, 1991.

[73] A. McKenzie, M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164:81–92, 2000.

[74] F. R. McMorris, M. Steel. The complexity of the median procedure for binary trees. *Proceedings of the 4th Conference of the International Federation of Classification Societies*, 1993.

[75] M. Meegaskumbura, F. Bossuyt, R. Pethiyagoda, K. Manamendra-Arachchi, M. Bahir, M. C. Milinkovitch, C. J. Schneider. Sri lanka: An amphibian hot spot. *Science*, 298:379, 2002.

[76] A. O. Mooers, S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72:31–54, 1997.

[77] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38:114–117, 1965.

[78] V. Moulton, M. Zuker, M. Steel, R. Pointon, D. Penny. Metrics on RNA secondary structures. *Journal of computational biology*, 7:277–292, 2000.

[79] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, Y. Zhou. Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. *ACM Transactions on Graphics*, 22:453–462, 2003.

[80] D. C. Nickle, M. A. Jensen, G. S. Gottlieb, D. Shriner, G. H. Learn, A. G. Rodrigo, J. I. Mullins. Consensus and ancestral state HIV vaccines. *Science*, 299:1515–1518, 2003.

[81] J. Nielsen, A. Kristensen, T. Mailund, C. Pedersen. A sub-cubic time algorithm for computing the quartet distance between two general trees. *Algorithms for Molecular Biology*, 6:15, 2011.

[82] T. M. Nye, P. Liò, W. R. Gilks. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22:117–119, 2006.

[83] J. B. Orlin, R. K. Ahuja. New scaling algorithms for the assignment and minimum mean cycle problems. *Mathematical Programming*, 54:41–56, 1992.

[84] D. Penny, L. R. Foulds, M. D. Hendy. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*, 297:197–200, 1982.

[85] D. Penny, M. D. Hendy. The use of tree comparison metrics. *Systematic Zoology*, 34:75–82, 1985.

[86] D. Penny, E. E. Watson, M. A. Steel. Trees from languages and genes are very similar. *Systematic Biology*, 42:382–384, 1993.

[87] S. Pompei, V. Loreto, F. Tria. On the accuracy of language trees. *PLoS ONE*, 6:e20109, 2011.

[88] M. N. Price, P. S. Dehal, A. P. Arkin. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26:1641–1650, 2009.

[89] M. N. Price, P. S. Dehal, A. P. Arkin. Fasttree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5:e9490, 2010.

[90] P. Puigbò, S. Garcia-Vallvé, J. O. McInerney. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, 23:1556–1558, 2007.

[91] A. Rambaut, D. Posada, K. A. Crandall, E. C. Holmes. The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5:52–61, 2004.

[92] H. Rasiowa. *Wstęp do matematyki współczesnej*. Wydawnictwo Naukowe PWN, 2009.

[93] G. Restrepo, M. Héber, E. J. Llanos. Three dissimilarity measures to contrast dendrograms. *Journal of Chemical Information and Modeling*, 47:761–770, 2007.

[94] D. F. Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory*, 11:105–119, 1971.

[95] D. F. Robinson, L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[96] M. J. Sackin. ”Good” and ”bad” phenograms. *Systematic Zoology*, 21:225–226, 1972.

[97] N. Saitou, M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[98] C. Semple, M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[99] K.-T. Shao, R. R. Sokal. Tree balance. *Systematic Zoology*, 39:266–276, 1990.

[100] J. Shen, J. Ma, Q. Wang. Evolutionary trends of A(H1N1) influenza virus hemagglutinin since 1918. *PLoS ONE*, 4:e7789, 2009.

[101] J. B. Slowinski. Review of the computer program component. *Cladistics*, 9:351–353, 1993.

[102] Y. Smolenskii. A method for the linear recording of graphs. *USSR Computational Mathematics and Mathematical Physics*, 2:396 – 397, 1963.

[103] Y. S. Song. On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, 7:365–379, 2003.

[104] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101, 1904.

[105] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22:2688–2690, 2006.

[106] M. A. Steel, D. Penny. Distributions of tree comparison metrics – some new results. *Systematic Biology*, 42:126–141, 1993.

[107] C. Stockham, L.-S. Wang, T. Warnow. Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics*, 18:S285–S293, 2002.

[108] J. Stoye, D. Evers, F. Meyer. Rose: generating sequence families. *Bioinformatics*, 14:157–163, 1998.

[109] S.-J. Sul, S. Matthews, T. Williams. Using tree diversity to compare phylogenetic heuristics. *BMC Bioinformatics*, 10:S3, 2009.

[110] G. Valiente. A fast algorithmic technique for comparing large phylogenetic trees. *Lecture Notes in Computer Science*, 3772:370–375, 2005.

[111] J. T. L. Wang, H. Shan, D. Shasha, W. H. Piel. Fast structural search in phylogenetic databases. *Evolutionary Bioinformatics Online*, 1:37–46, 2005.

[112] R. J. Wilson. *Introduction to Graph Theory*. Academic Press, New York, 1972.

[113] Y. Wu. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 25:190–196, 2009.

[114] Z. Yang. Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24:1586–1591, 2007.

[115] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.

[116] K. A. Zaretskii. Constructing trees from the set of distances between pendant vertices. *Uspekhi Matematicheskikh Nauka*, 20:90–92, 1965.

[117] E. Zuckerkandl, L. Pauling. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, 97–166. Academic Press, 1965.

[118] `http://www.kaims.pl/~dambo/treecmp`.