Wykład X: Maksymalne zgodne poddrzewa i lasy

Anna Gambin
12.12.2002

1 Poddrzewa zgodne

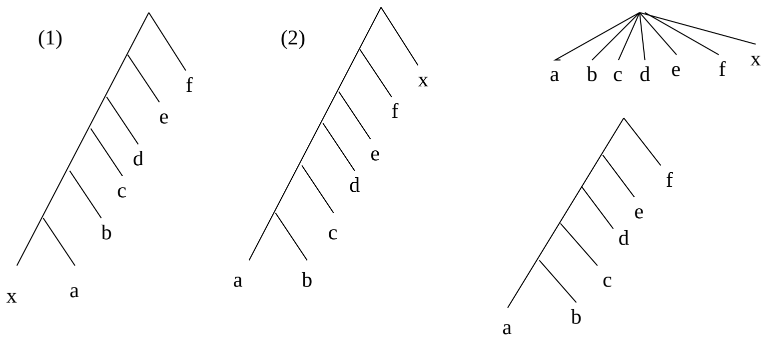
Rozważane przez nas do tej pory metody konsensusu dla rodziny drzew zwracały drzewo o *tym samym zbiorze liści.* Ich uogólnieniem są metody budujące superdrzewo, które nie wymagają, aby wszystkie drzewa w rodzinie miały ten sam zbiór liści. Natomiast wymaga się, aby w konstruowanym przez nie superdrzewie były reprezentowane wszystkie liście występujące w rodzinie drzew.

Okazuje się, że nie zawsze jest to najlepszy sposób na reprezentowanie wspólnej informacji zawartej w drzewach. Rozważmy dwa drzewa na rysunku poniżej. Różnią się one jedynie położeniem liścia x. Niestety ich drzewo dokładnego konsensusu nie zawiera żadnej informacji (jest gwiazdą). Dużo bardziej informatywne drzewo otrzymamy po prostu usuwając liść x. Otrzymamy w ten sposób poddrzewo zgodne, a w tym przypadku jest to również **maksymalne poddrzewo zgodne** (ang. maximum agreement subtree)(**MAST**).

W ogólnym przypadku, dla rodziny drzew $\mathcal{T} = \{T_1, T_2, \ldots T_k\}$ drzewo T o liściach ze zbioru X jest **poddrzewem zgodnym** jeżeli T jest indukowanym poddrzewem wszystkich drzew T_i, czyli:

$$T = T_{1|X} = T_{2|X} = \ldots = T_{k|X}.$$

Maksymalne poddrzewo zgodne (MAST) to poddrzewo zgodne zawierające maksymalna liczbę liści. Dla danej rodziny drzew może istnieć wiele maksymalnych wspólnych poddrzew (nawet wykładniczo wiele). Pierwsze zaproponowane algorytmy dla tego problemu działały w czasie powyżej wykładniczym, wielomianowy algorytm został zaproponowany m.in. w [3], gdzie zakłada się, że przynajmniej jedno drzewo ma ograniczone stopnie wierzchołków. Bez tego ograniczenia znalezienie drzewa MAST dla więcej niż dwóch drzew jest problemem NP-trudnym [1]; udowodniono też, że trudno jest nawet aproksymować rozmiar drzewa MAST [4].



Rysunek 1: Drzewo konsensusu dokładnego dla (1) oraz (2) jest gwiazdą, natomiast drzewo MAST zawiera dużo więcej informacji.

1.1 NP zupełność problemu MAST dla trzech drzew

Sformułujmy problem 3-**MAST**, którego złożoność zbadamy:

> DANE: Trzy drzewa T_1, T_2, T_3 o nieograniczonym stopniu wierzchołka i o liściach należacych do zbioru $S = \{s_1, s_2, \ldots s_n\}$, liczba $i \leq n$ (czyli $\mathcal{L}(T_1) \cup \mathcal{L}(T_2) \cup \mathcal{L}(T_3) = S$).
>
> PYTANIE: Czy istnieje podzbiór $S' \subseteq S$ rozmiaru i, taki że: $T_{1|S'} = T_{2|S'} = T_{3|S'}$?

Twierdzenie 1 *Problem 3-MAST jest NP zupełny.*

Dowód: Problem jest w sposób oczywisty w NP. Zredukujemy problem trójwymiarowego skojarzenia (*ang.* 3-*Dimensional Matching* 3-DM) do problemu 3-MAST. Wejście dla problemu 3-DM stanowi zbiór trójek $M \subseteq W\text{x}X\text{x}Y$, gdzie trzy zbiory W, X, Y są rozłączne, każdy o mocy q. W problemie 3-DM musimy rozstrzygnąć czy istnieje podzbiór $M' \subseteq M$ o mocy q, taki że każde dwie trójki z M' różnia się na wszystkich trzech współrzędnych.

Konstruujemy drzewa T_1, T_2 i T_3 następująco: Drzewo T_i ma korzeń r_i, do którego doczepiamy q synów odpowiadających elementom odpowiednio ze zbiorów W, X i Y. (synowie r_1 to elementy ze zbioru W, etc.) Weźmy $S = M$, dla każdego $e = (w, x, y) \in M$ doczepiamy liść o etykiecie e do wierzchołka w w drzewie T_1, wierzchołka x w T_2 oraz y w T_3. Załóżmy na chwile, że korzenie r_1, r_2, r_3 należą do maksymalnego zgodnego poddrzewa tych trzech drzew. Zauważmy, że wtedy żadne dwa elementy M, które zgadzają się na jednej lub dwóch współrzędnych nie wejdą do maksymalnego zgodnego poddrzewa (tzn. odpowiadające im liście). Ażeby wymusić to, że korzenie należą do MAST(T_1, T_2, T_3) dodajemy do każdego z drzew $q+1$ liści podczepionych bezpośrednio do korzenia etykietowanych różnowartościowo nowymi symbolami: $x_1, x_2, \ldots x_{q+1}$. Łatwo widać, że istnieje zbiór M' rozmiaru q wttw istnieje poddrzewo zgodne rozważanych trzech drzew o rozmiarze $2q+1$.

1.2 Algorytm aproksymacyjny

Pokażemy algorytm aproksymacyjny, który konstruuje zbiór liści nie występujących w maksymalnym zgodnym poddrzewie. Wykorzystamy następujące twierdzenie dotyczące czwórek:

Twierdzenie 2 *Dwa drzewa filogenetyczne są homeomorficzne wttw wszystkie ich poddrzewa indukowane przez czwórki liści są homeomorficzne.*

Dowód: W jedną strone twierdzenie jest oczywiste. Udowodnimy, że jeśli wszystkie poddrzewa indukowane przez czwórki liści są homeomorficzne to drzewa są również homeomorficzne. Nazwijmy wierzchołek o stopniu większym niż 2: *rozgałęzionym*. O dwóch liściach powiemy, że są *braćmi* jeżeli na ścieżce je łączącej znajduje się co najwyżej jeden *rozgałęziony* wierzchołek. Można pokazać, że każde drzewo zawiera co najmniej jedną parę liści — *braci* (zadanie 3). Udowodnimy twierdzenie indukcyjnie względem liczby liści. Dla $n \leq 4$ twierdzenie jest w trywialny sposób prawdziwe. Dla $n > 4$ rozważmy parę braci x i y w drzewie T_1. Rozgałęziony wierzchołek leżący na ścieżce je łączącej nazwijmy z. Pokażemy, że x i y są braćmi w drzewie T_2. Załóżmy, że tak nie jest czyli mamy na ścieżce $x \rightsquigarrow y$ dwa rozgałęzione wierzchołki z_1 i z_2. Ponieważ sa one rozgałęzione to isnieją liście p oraz q, gdzie p jest w poddrzewie o korzeniu z_1 a q w poddrzewie o korzeniu z_2. Rozważmy teraz poddrzewa indukowane przez czwórke liści x, y, p, q w T_1 oraz przez tę samą czwórkę w T_2. w poddrzewie T_1 na ścieżce łaczącej x i y znajduje się tylko jeden rozgałęziony wierzchołek, podczas gdy w indukowanym poddrzewie T_2 dwa. Czyli poddrzewa indukowane przez tę czwórkę liści nie są homeomorficzne.

Możemy teraz usunąć w obu drzewach ścieżke z z do y. Na podstawie założenia indukcyjnego mniejsze drzewa będą homeomorficzne i pozostaną takie po dodaniu spowrotem tej ścieżki. ∎

Oznaczmy przez S_4 zbiór wszystkich 4 elementowych podzbiorów S, które nie indukują homeomorficznych poddrzew we wszystkich k drzewach z rodziny $\mathcal{T}$. Zbiór S_4 może zostać skonstruowany w czasie $\mathcal{O}(kn^5)$. Podzbiór $S' \subseteq S$ jest *pokryciem* S_4 jeśli:

$$\forall \; \{a,b,c,d\} \in S_4 \;\; \exists \; e \in S' \; : \; e \in \{a,b,c,d,\}.$$

S' jest *minimalnym pokryciem* jeżeli jest pokryciem i każde inne pokrycie ma moc co najmniej taką jak S'.

Z Twierdzenia 2 wynika, że jeśli S' jest minimalnym pokryciem zbioru S_4 to dla $i = 1, 2, \ldots k$ poddrzewo T_i indukowane przez zbiór $\mathcal{L}(T_i) \setminus S'$ jest poddrzewem zgodnym dla rodziny $\mathcal{T}$. Poniższy algorytm aproksymuje minimalne pokrycie S' ze współczynnikiem aproksymacji 4.

Algorithm 1 Pokrycie S_4

$S_a := \emptyset$
Powtarzaj dopóki $S_4 \neq \emptyset$
  wybierz element $\alpha \in S_4$
  $S_a := S_a \cup \alpha$
  usuń z S_4 wszystkie czwórki β, takie że $\alpha \cap \beta \neq \emptyset$

1.3 MAST dla trzech drzew jest MAX SNP trudny

Wzmocnimy teraz wynik mówiący o NP zupełności problemu 3-MAST: pokażemy, że ten problem jest w rzeczywistości MAX SNP trudny, czyli, że nie istnieje dla niego algorytm aproksymacyjny o współczynniku aproksymacji $1+\epsilon$ dla dowolnie małego ϵ o ile P$\neq$ NP. Rozważamy teraz optymalizacyjną wersję problemu czyli chcemy znaleźć maksymalny podzbiór $S' \subseteq S$ taki, że indukuje on poddrzewa zgodne: $T_{1|S'} = T_{2|S'} = T_{3|S'}$.

Lemat 1 *Problem 3-MAST jest MAX SNP trudny.*

Dowód: Musimy pokazać, że istnieje L-redukcja jakiegoś problemu o którym wiadomo, że jest nieaproksymowalny do naszego problemu. Przypomnijmy definicję L-redukcji: dla dwóch problemów aproksymacyjnych Π_1 oraz Π_2 mówimy, że Π_1 *L-redukuje się* do problemu Π_2 jeżeli istnieją dwa wielomianowe algorytmy f i g oraz dwie stałe $\alpha > 0$ i $\beta > 0$ takie, że dla każdej instancji I problemu Π_1, $f(I)$ jest instancją problemu Π_2 oraz:

1. $opt(f(I)) \leq \alpha \cdot opt(I)$; gdzie $opt(I)$ jest rozwiązaniem optymalnym dla I.
2. Mając dane rozwiązanie dla $f(I)$ o wartości s_2, algorytm g znajduje w czasie wielomianowym rozwiązanie dla I o wartości s_1 spełniające:
$$|s_1 - opt(I)| \leq \beta \cdot |s_2 - opt(f(I))|.$$

Pokażemy L-redukcję z MAX SNP trudnego problemu **maksymalnego ograniczonego trójwymiarowego skojarzenia (MAX 3 DM-B)** (ang. Maximum Bounded 3-Dimensional Matching) rozszerzając nieznacznie dowód Twierdzenia 1.

W problemie MAX 3 DM-B szukamy maksymalnego podzbioru $M' \subseteq M \subseteq W\text{x}X\text{x}Y$ takiego, że każde dwie trójki z M' różnią się na wszystkich współrzędnych, przy dodatkowym założeniu, że każdy element $W \cup X \cup Y$ występuje w co najwyżej B trójkach w M.

Dla danej instancji 3 DM-B konstruujemy 3 drzewa w sposób opisany w dowodzie Twierdzenia 1. Pamiętamy, że zbiór M ma trójwymiarowe skojarzenie rozmiaru k wttw drzewa T_1, T_2, T_3 mają poddrzewo zgodne rozmiaru $k + q + 1$. Ponieważ każdy element $W \cup X \cup Y$ występuje w co najwyżej B trójkach w M, możemy zagwarantować, że w M istnieje skojarzenie rozmiaru co najmniej q/B. Mamy więc:

$$opt(T_1, T_2, T_3) = opt(M) + q + 1 \leq opt(M) + 2q \leq$$
$$\leq opt(M) + 2B opt(M) = (2B + 1) opt(M).$$

Dodatkowo mając poddrzewo zgodne dla T_1, T_2, T_3 rozmiaru $s_2 = q + 1 + k$ możemy w czasie wielomianowym skonstruować trójwymiarowe skojarzenie rozmiaru $s_1 = k$. Zauważmy też, że:

$$opt(M) - s_1 = 1 + q + opt(M) - (1 + q + s_1) =$$
$$= 1 + q + opt(M) - s_2 = opt(T_1, T_2, T_3) - s_2.$$

A więc opisana redukcja jest L-redukcją, co kończy dowód.

1.4 MAST dla drzew o ograniczonym stopniu wierzchołka

Omówimy algorytm opisany w [2], który ma taką samą złożoność obliczeniową jak algorytm podany w pracy [3]. Dla uproszczenia założymy dodatkowo, że jedno z drzew w $\mathcal{T}$ jest binarne, czyli oznaczając przez $f(T_i)$ zbiór "nierozwiązanych" trójek (gwiazd) w drzewie T_i, mamy:

$$F = \bigcap_i f(T_i) = \emptyset.$$

Zachodzi następujący fakt:

Fakt 1 *Drzewo T jest zgodnym poddrzewem rodziny $\mathcal{T}$ wttw*

$$r(T) \subseteq R = \bigcap_i r(T_i).$$

Zauważmy, że liczba liści w drzewie MAST rodziny $\mathcal{T}$ jest równa maksimum po wszystkich liściach a i b z następującej wartości:

$$MAST(a,b) = \max\{MAST(a,x) + MAST(b,y) : x \in A, y \in B\}$$

gdzie $A = \{x : ax|b \in R\} \cup \{a\}$ oraz $B = \{y : by|a \in R\} \cup \{b\}$.

Oznaczmy przez $\mathcal{T}(a,b)$ zbiór poddrzew zgodnych T rodziny $\mathcal{T}$, dla których a i b występują w różnych maksymalnych poddrzewach T (czyli korzeń T jest najbliższym wspólnym przodkiem a i b). Zgodnie z tym oznaczeniem $\mathcal{T}(a,a)$ jest pojedynczym wierzchołkiem a.

Poprawność algorytmu PODDRZEWO ZGODNE uzasadnia następujące twierdzenie:

Twierdzenie 3 *(i) Algorytm PODDRZEWO ZGODNE zawsze zwraca drzewo ze zbioru $\mathcal{T}(a,b)$.*
(ii) Algorytm PODDRZEWO ZGODNE może zwrócić dowolne drzewo z tego zbioru.

Dowód (Indukcja względem częsciowego porządku $<$ zdefiniowanego na parach liści: $(c,d) < (a,b)$ jeśli najbliższy wspólny przodek c i d jest właściwym potomkiem najbliższego wspólnego przodka a i b.)

Algorithm 2 PODDRZEWO ZGODNE (a, b)

1. Jeśli $a = b$ to zwróć wierzchołek a.
2. Skonstruuj zbiory:

$$A \leftarrow \{x : ax|b \in R\} \cup \{a\}$$
$$B \leftarrow \{y : by|a \in R\} \cup \{b\}$$

3. Wybierz $x^* \in A$.
4. Wybierz $y^* \in B$.
5. Skonstruuj rekurencyjnie PODDRZEWO ZGODNE (a, x^*) oraz PODDRZEWO ZGODNE (b, y^*).
6. Podłącz otrzymane drzewa do wspólnego korzenia i zwróć to drzewo.

Algorithm 3 MAST(a, b)

1. Jeśli $a = b$ to zwróć 1.
2. Skonstruuj zbiory:

$$A \leftarrow \{x : ax|b \in R\} \cup \{a\}$$
$$B \leftarrow \{y : by|a \in R\} \cup \{b\}$$

3. Wybierz $x^* \in A$, który maksymalizuje MAST(a, x^*).
4. Wybierz $y^* \in B$, który maksymalizuje MAST(b, y^*).
5. Zwróć MAST(a, x^*) + MAST(b, y^*).

Kolejny algorytm znajduje rozmiar drzewa MAST dla $\mathcal{T}(a,b)$. Mając policzony rozmiar drzewa możemy łatwo skonstruować to drzewo.

Ażeby policzyć teraz drzewo MAST dla rodziny $\mathcal{T}$ musimy uruchomić algorytm MAST(a,b) dla każdej pary liści (a,b) i wybrać największą wartość. Złożoność całego algorytmu dla k drzew, z których jedno jest binarne wyniesie $\mathcal{O}(kn^3)$.

2 Lasy zgodne

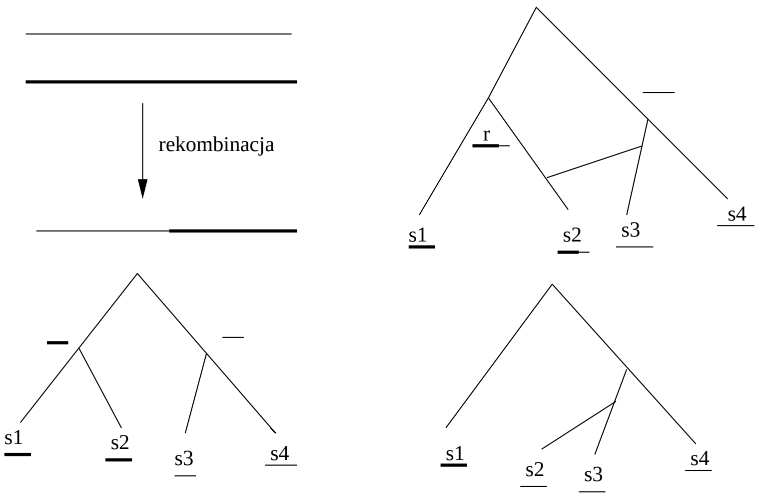
2.1 Rekombinacja, transfer horyzontalny i przemieszczanie poddrzew

Zjawisko *rekombinacji* sekwencji DNA polega na tym, że dwie sekwencje A i B "spotykają" się i powstaje nowa złożona z dwóch kawałków: na lewo od punktu rekombinacji sekwencja jest fragmentem sekwencji A, na prawo od tego punktu nowa sekwencja stanowi fragment sekwencji B. Z filogenetycznego punktu widzenia przed zdarzeniem rekombinacji material genetyczny znajdował się w dwóch sekwencjach. Po rekombinacji ewolucja sekwencji nie może być dalej opisywana przy użyciu jednego drzewa filogenetycznego. Potrzebujemy teraz dwóch drzew: jedno opisuje ewolucję fragmentu dziedziczonego z sekwencji A, a drugie fragmentu dziedziczonego z sekwencji B. Takie dwa drzewa przedstawione są na Rysunku 2.

Okazuje się, że rozważane drzewa filogenetyczne nie mogą się dowolnie różnić: jedno musi być otrzymywalne z drugiego poprzez operację **przemieszczenia poddrzewa (ST)** (ang. subtree transfer). Polega ona na usunięciu pewnej krawędzi i podczepieniu otrzymanego poddrzewa w innym miejscu.

Jeżeli zajdzie więcej zdarzeń rekombinacji, możemy modelować ewolucję sekwencji jako rodzinę drzew, które mogą być przekształcane między sobą ciągiem operacji ST. Bardzo podobna sytuacja zachodzi, jeśli chcemy uwzględniać zjawisko *horyzontalnego transferu.* Odległością ST dla dwóch drzew nazwiemy minimalna liczbę operacji ST pozwalającą przeprowadzić jedno drzewo w drugie. Zastanowimy się nad problemem liczenia odległości ST dla drzew.

Podamy alternatywne sformułowanie tego problemu używające pojęcia zgod-



Rysunek 2: Drzewa modelujące ewolucję sekwencji w przypadku zdarzenia rekombinacji.

nego lasu, bedącego uogólnieniem zgodnego poddrzewa.

Dla dwóch drzew T_1 i T_2 **zgodny las** jest lasem, który może być otrzymany z obydwu drzew przez wycięcie pewnej (tej samej) liczby krawędzi w T_1 i T_2 i wykonaniu kontrakcji wierzchołków stopnia 2 w otrzymanych składowych. Rozmiar zgodnego lasu to liczba drzew, które zawiera. W problemie **maksymalnego zgodnego lasu (MAF)** (ang. Maximum Agreement Forest) szukamy zgodnego lasu o jak *najmniejszym* rozmiarze.

Lemat 2 *Rozmiar maksymalnego zgodnego lasu (MAF) dla T_1 i T_2 jest o jeden większy od ich odległości przemieszczania poddrzew (ST).*

2.2 NP zupełność problemu MAF

Udowodnimy, że problem MAF (czyli również liczenie odległości ST) jest NP zupełny.

Twierdzenie 4 *Znalezienie maksymalnego zgodnego lasu jest NP trudne dla dwóch drzew binarnych.*

Dowód: Zredukujemy problem **dokładnego pokrycia** 3 **zbiorami** do problemu MAF. Formułuje się on następująco:

> DANE: Rodzina C podzbiorów skończonego zbioru S, gdzie każdy $c \in C$ jest trzyelementowy oraz każdy element z S występuje w co najwyżej trzech podzbiorach w C.
>
> PYTANIE: znajdź dokładne pokrycie $C' \subseteq C$, czyli rodzinę wzajemnie rozłącznych zbiorów, których suma wynosi S.

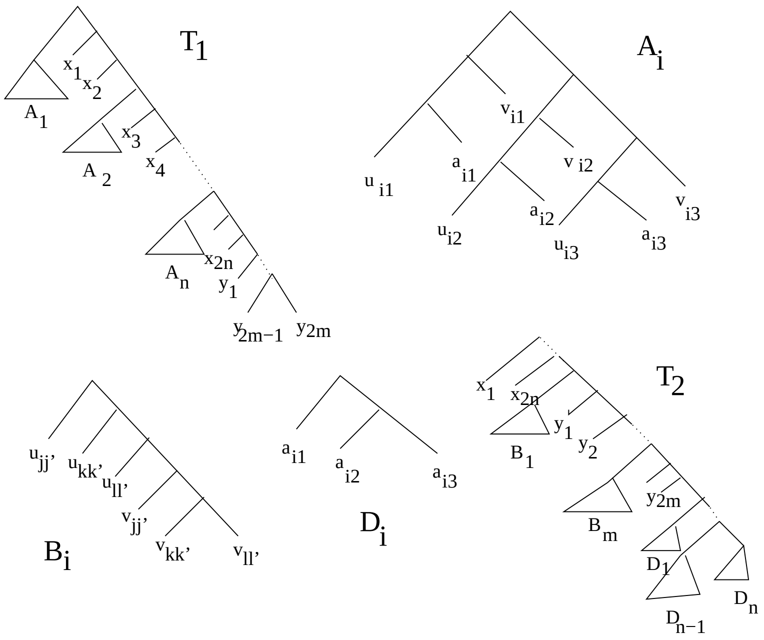
Niech $S = \{s_1, s_2, \ldots s_m\}$ i $C = \{C_1, C_2, \ldots C_n\}$ będzie instancją problemu dokładnego pokrycia, zakładamy że $m = 3q$.

Drzewo T_1 tworzymy doczepiając n poddrzew $A_1, A_2, \ldots A_n$ do łańcucha zawierającego $2n+2m$ liści $x_1, x_2, \ldots x_{2n}, y_1 \ldots y_{2m}$ (Rysunek 3). Każde z poddrzew A_i odpowiada zbiorowi $C_i = \{c_{i1}, c_{i2}, c_{i3}\}$ i ma 9 liści jak na Rysunku 3. Budowa drzewa T_2 też jest przedstawiona na Rysunku 3; dla każdych trzech wystąpień $s_i \in S$ w C: $c_{jj'}, c_{kk'}, c_{ll'}$ mamy poddrzewo B_i, dla każdego C_i mamy poddrzewo D_i. Poddrzewa B_i oraz D_i są podwieszone do łańcucha o liściach $x_1, x_2, \ldots x_{2n}, y_1 \ldots y_{2m}$.

Zauważmy, że każda sąsiednia para poddrzew A_i oraz A_{i+1} w T_1 jest rozdzielona przez łańcuch długości 2 występujący również w T_2. Czyli najlepszą strategią prowadzącą do otrzymania zgodnego lasu jest poobcinać poddrzewa A_i w T_1 i podobnie "wystrzyc" poddrzewa B_i w T_2. Póżniej musimy jeszcze poobcinać $D_i, \ldots D_n$ w T_2. Teraz dla każdego A_i możemy albo poucinać liście $u_{i1}, v_{i1}, u_{i2}, v_{i2}, u_{i3}, v_{i3}$ otrzymując poddrzewo o liściach a_{i1}, a_{i2}, a_{i3}, co prowadzi do 7 składowych lub poobcinać liście a_{i1}, a_{i2}, a_{i3} a następnie porozcinać krawędzie pomiędzy poddrzewami o liściach $\{u_{i1}, vi1\}$, $\{u_{i2}, v_{i2}\}$ oraz $\{u_{i3}, v_{i3}\}$ (ponieważ są one odseparowane w T_2). W tym drugim przypadku najlepsze na co możemy liczyć to 6 składowych (o ile uda nam się utrzymać dwulistne poddrzewa w maksymalnym lesie).

Udowodnimy teraz, że C zawiera dokładne pokrycie S wttw T_1 i T_2 mają zgodny las o rozmiarze $1 + 6q + 7(n - q) = 7n - q + 1$.

($\rightarrow$) Załóżmy, że mamy dane dokładne pokrycie $\{C_{i_1}, C_{i_2}, \ldots, C_{i_q}\} \subseteq C$ zbioru S. Dla każdego $A_{i_l}(l = 1, 2, \ldots q)$ w drzewie T_1 obcinamy a_{i_l1}, a_{i_l2} oraz a_{i_l3} jak również trzy poddrzewa o liściach $\{u_{i_l1}, v_{i_l1}\}$, $\{u_{i_l2}, v_{i_l2}\}$ oraz $\{u_{i_l3}, v_{i_l3}\}$. Odpowiednio obcinając w T_2 otrzymamy też 6 składowych. Zauważmy, że każde B_i w T_2 może wnieść co najwyżej jedno poddrzewo zawierające $\{u_{i_lj}, v_{i_lj}\}$ $(j = 1, 2, 3)$. W ten sposób każde A_{i_l} wniesie 6 składowych



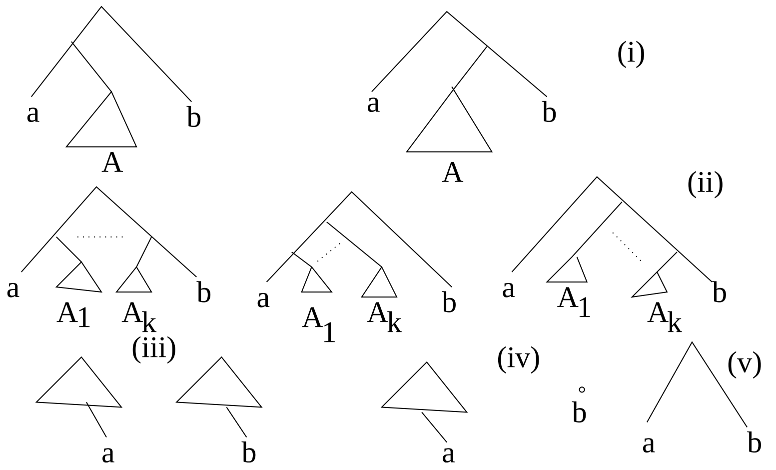
Rysunek 3: Struktura drzew T_1 (góra) i T_2(dół) używanych w dowodzie Twierdzenia 4.

do zgodnego lasu. W pozostałych $n-q$ poddrzewach A_r ($r \in \{1, 2, \ldots, n\} \setminus \{i_1, i_2, \ldots, i_q\}$) możemy poobcinać liście $u_{r1}, v_{r1}, u_{r2}, v_{r2}, u_{r3}, v_{r3}$. Odpowiednio możemy otrzymać 7 składowych strzygąc drzewo T_2. Reasumując całkowita liczba składowych w zgodnym lesie wynosi: $1 + 6q + 7(n-q) = 7n - q + 1$.

($\leftarrow$) Załóżmy, że mamy zgodny las rozmiaru $7n-q+1$, z poprzedniej dyskusji wiemy, że:

- każde poddrzewo A_i w T_1 wnosi co najmniej 6 składowych;
- spośród n poddrzew A_i co najwyżej q może wnieść dokładnie 6 składowych i to są te A_i, które odpowiadają rozłącznym podzbiorom C_i w C.

Reasumując ze zgodnego lasu rozmiaru $7n - q + 1$ potrafimy odtworzyć dokładne pokrycie S.



Rysunek 4: Pięć przypadków rozważanych w algorytmie aproksymującym maksymalny zgodny las.

2.3 Algorytm aproksymacyjny

Rozważmy następujący algorytm aproksymacyjny znajdujący MAF dla dwóch drzew zaproponowany w pracy [4]. Przypadki *(i)* — *(v)* zilustrowane są na Rysunku 4.

Algorithm 4 MAF (T_1, T_2)

---

1. $N := 0$
2. Dla pary sąsiednich liści (liści braci) w T_1 rozważ ich położenie w T_2 i poobcinaj następująco:
   (i): obetnij poddrzewo A w T_2; $N := N + 1$;
   (ii): obetnij liście a i b w obydwu drzewach; $N := N + 2$;
   (iii): obetnij liście a i b w obydwu drzewach, $N := N + 2$;
   (iv): obetnij liść b w T_1;
   (v): zamień rozważaną parę liści na nowy liść etykietowany (a, b) w obydwu drzewach.
3. Jeśli pewna składowa w lesie powstałym z T_1 ma rozmiar większy niż 1 powtórz krok 2.

2.4 Współczynnik aproksymacji ?

Autorzy opisanego przez nas algorytmu aproksymacyjnego utrzymywali, że ma on współczynnik aproksymacji równy 3; w pracy [5] pokazano kontrprzykład, dla którego algorytm ten aproksymuje MAF ze współczynnikiem 4, oraz zaproponowano modyfikację algorytmu tak aby osiągał on współczynnik aproksymacji 3. Otwartym problemem pozostaje znalezienie lepszego algorytmu aproksymacyjnego.

Literatura

[1] Amihood Amir and Dmitry Keselman. Maximum agreement subtree in a set of evolutionary trees: Metrics and efficient algorithms. *SIAM Journal on Computing*, 26(6):1656–1669, 1997.

[2] D. Bryant. Hunting for trees, buiding trees and comparing trees: theory and method in phylogenetic analysis. *phd thesis, Department of Mathematics, University of Canterbury*, 1997.

[3] M. Farach, M.T. Przytycka, and M. Thorup. On the agreement of many trees. *Information Processing Letters*, 55:297–301, 1995.

[4] J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. In Z. Galil and E. Ukkonen, editors, *Proceedings of the 6th Annual Symposium on Combinatorial Pattern Matching*, number 937, pages 177–190, Espoo, Finland, 1995. Springer-Verlag, Berlin.

[5] E.M. Rodrigues, M-F. Sagot, and Y. Wakabayashi. Some approximation results for the maximum agreement forest problem. In *APPROX*, 2001.

3 Zadania

1. Pokaż, że każde drzewo zawiera co najmniej dwa liście połączone scieżką przechodzącą przez co najwyżej jeden wierzchołek stopnia większego niż 2.

2. Uzasadnij poprawność algorytmu aproksymującego minimalne pokrycie zbioru S_4, oczacuj jego złożoność i jakość aproksymacji.

3. Zmodyfikuj algorytm znajdujący poddrzewa zgodne, tak aby nie wymagał założenia, że jedno z drzew w rodzinie $\mathcal{T}$ jest binarne, zamiast tego załóż, że co najmniej jedno drzewo ma stopień wierzchołków ograniczony przez d. Algorytm powinien działać w czasie $\mathcal{O}(kn^3 + n^d)$.

4. Udowodnij Lemat 2 (indukcyjnie po liczbie liści); zauważ, że intuicyjnie lemat pozwala rozbić przemieszczanie poddrzew na dwie fazy: fazę strzyżenia (wycinania poddrzew) i fazę wklejania w nowe miejsca.