# Spis rzeczy

**8 Drzewa ewolucyjne** **118**
8.1 Drzewa ultrametryczne . . . . . 118
8.2 Metody zdobywania ultrametrycznych danych . . . . . 121
8.3 Drzewa addytywne . . . . . 122
8.4 Drzewa filogenetyczne . . . . . 123
8.4.1 Problem doskonałej filogenezy . . . . . 125
8.4.2 Kompatybilność drzew . . . . . 127
8.4.3 Uogólnienie doskonałej filogenezy . . . . . 128
8.5 Problem maksymalnej oszczędności i drzewa Steinera . . . . . 128

# 8 Drzewa ewolucyjne

Zaczniemy od omówienia metod budowania drzew ewolucyjnych opartych na pojęciu odległości.

## 8.1 Drzewa ultrametryczne

Zakładamy, że wszystkie roważane macierze spełniają następujące dwa warunki:

**(m1)** Są kwadratowe i symetryczne.

**(m2)** Na głównej przekątnej stoją same zera, a poza główną przekątną liczby rzeczywiste dodatnie.

Niech $D$ będzie $n \times n$ macierzą. Powiemy, że drzewo $T$ z korzeniem jest *ultrametryczne dla* $D$, gdy $T$ spełnia następujące warunki:

**(U1)** $T$ ma dokładnie $n$ liści, każdy jest etykietowany wierszem macierzy $D$ w sposób różnowartościowy.

**(U2)** Każdy wierzchołek wewnętrzny $T$ ma co najmniej dwóch synów i jest etykietowany liczbą stojącą w $D$.

**(U3)** Etykiety stojące wzdłóż każdej ścieżki od korzenia do liścia tworzą ściśle malejący ciąg.

**(U4)** Dla dowolnych dwóch liści $i, j$, liczba $D(i, j)$ jest etykietą najpóźniejszego wspólnego przodka $i$ oraz $j$.

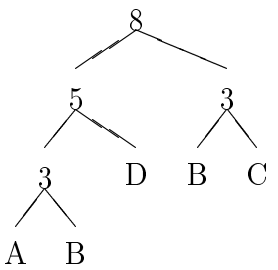
Powiemy, że drzewo $T$ z korzeniem jest *min-ultrametryczne* dla $D$, gdy spełnia ono warunki (U1), (U2), (U4) oraz warunek (U3').

**(U3')** Etykiety stojące wzdłóż każdej ścieżki od korzenia do liścia tworzą ściśle rosnący ciąg.

**Przykład 8.1.1** Niech $D$ będzie następującą macierzą.

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| $A$ | 0 | 8 | 8 | 5 | 3 |
| $B$ | | 0 | 3 | 8 | 8 |
| $C$ | | | 0 | 8 | 8 |
| $D$ | | | | 0 | 5 |
| $E$ | | | | | 0 |

Następujące drzewo jest ultrametryczne dla $D$.




□

Zauważmy, że drzewo ultrametryczne dla $n \times n$ macierzy może mieć co najwyżej $n-1$ wewnętrznych wierzchołków. Zatem macierz $D$ nie może mieć więcej niż $n-1$ różnych wartości.

Intuicja dla drzew ultrametrycznych jest następująca. Każdy wierzchołek $v$ w drzewie $T$ przedstawia moment w historii, kiedy nastąpiło ewolucyjne rozdzielenie dwóch (lub więcej) jednostek taksonomicznych (np. gatunków) $i$ oraz $j$. Wówczas $D(i, j)$ przedstawia czas jaki upłynął od momentu rozdzielenia się $i$ oraz $j$. Dla drzew min-ultrametrycznych jest podobnie, tylko $D(i, j)$ oznacza czas, kiedy nastąpiło rozdzielenie $i$ oraz $j$. Częściej spotykane w zastosowaniach są drzewa ultrametryczne.

Powiemy, że macierz $D$ jest *ultrametryczna*, gdy dla dowolnych $i, j, k$ zbiór elementów maksymalnych w zbiorze $\{D(i, j), D(i, k), D(j, k)\}$ ma co najmniej dwa elementy. Powiemy, że $D$ jest *min-ultrametryczna*, gdy dla dowolnych $i, j, k$ zbiór elementów minimalnych w zbiorze $\{D(i, j), D(i, k), D(j, k)\}$

ma co najmniej dwa elementy. Nietrudno jest zauważyć, że macierz ultrametryczna jest metryką. Podobnie dla macierzy min-ultrametrycznej.

**Twierdzenie 8.1.1** *Dla macierzy $D$ spełniającej (m1) oraz (m2) mamy:*

(i) *Istnieje drzewo ultrametryczne dla $D \Longleftrightarrow D$ jest ultrametryczna.*

(ii) *Istnieje drzewo min-ultrametryczne dla $D \Longleftrightarrow D$ jest min-ultrametryczna.*

(iii) *Drzewo ultrametryczne dla $D$, o ile istnieje, jest dokładnie jedno.*

**Dowód:** Udowodnimy (i). Dla dowodu ($\Longrightarrow$) załóżmy, że $D$ ma drzewo ultrametryczne $T$ i niech $i, j, k$ będą dowolnymi liśćmi $T$. Niech $v_{i,j}$ będzie najpóźniejszym wspólnym przodkiem $i$ oraz $j$. Podobnie definiujemy $v_{i,k}$ oraz $v_{j,k}$. Jeśli $v_{i,j} = v_{i,k}$, to $v_{j,k}$ musi leżeć na tej samej ścieżce co pozostałe dwa wierzchołki, co więcej jeśli $v_{j,k}$ jest różne od pozostałych dwóch wierzchołków, to znajduje się on na ścieżce poniżej $v_{i,j} = v_{i,k}$. Zatem w obu przypadkach mamy

$$D(j,k) \leq D(i,j) = D(i,k).$$

Jeśli natomiast $v_{i,j} \neq v_{i,k}$, to ponieważ $v_{i,j}$ oraz $v_{i,k}$ muszą leżeć na jednej ścieżce, to możemy przyjąć, że $v_{i,j}$ leży dalej od korzenia niż $v_{i,k}$. Wówczas mamy $v_{i,k} = v_{j,k}$ i postępujemy tak samo jak w poprzednim przypadku.

Dowód ($\Longleftarrow$) przeprowadzamy przez indukcję ze względu na rozmiar $n$ macierzy $D$. Dla $n = 2$ teza jest oczywista. Załóżmy, że $n > 2$ i weźmy dowolny wiersz $i$ macierzy $D$. W zbiorze wierszy $X = \{j \mid j \neq i\}$ definiujemy relację równoważności $\sim$ w następujący sposób

$$j_1 \sim j_2 \Longleftrightarrow D(i,j_1) = D(i,j_2).$$

Jeśli $D$ ma mieć drzewo ultrametryczne $T$, to na ścieżce $v_1, \ldots, v_k, i$ w $T$ od korzenia do liścia $i$ musi być następująca sytuacja. Zbiór liści $j \in X$ widocznych z $v_k$ tworzy klasę abstrakcji $[j_k]_\sim$ o najmniejszej wartości $D(i,j_k)$, spośród wszystkich $D(i,j)$ dla $j \in X$. Zbiór liści widocznych z $v_{k-1}$, do których droga nie przechodzi przez $v_k$ musi tworzyć klasę abstrakcji $[j_{k-1}]_\sim$ o wartości $D(i,j_{k-1}) > D(i,j_k)$ drugiej od dołu, itd. Wreszcie, zbiór liści widocznych z korzenia $v_1$, do których droga nie przechodzi przez $v_2$ tworzy klasę abstrakcji $[j_1]_\sim$ o największej wartości $D(i,j_1)$, spośród wszystkich $D(i,j)$ dla $j \in X$. Etykietą wierzchołka $v_l$ musi być $D(i,j_l)$.

Rozważmy klasę abstrakcji $[j]_\sim$. Niech $v$ będzie wierzchołkiem odpowiadającym klasie $[j]$ (tak jak to jest opisane w poprzednim akapicie). Niech $D'$ będzie podmacierzą otrzymaną z $D$ przez wzięcie kolumn i wierszy o obydwu

numerach występujących w $[j]_\sim$. $D'$ jest też ultrametryczna i niech $T'$ będzie drzewem ultrametrycznym znalezionym dla $D'$ na mocy założenia indukcyjnego. Niech $r$ będzie wartością stojącą w korzeniu $T'$. Jeśli $r < D(i,j)$, to drzewo $T'$ doklejamy do wierzchołka $v$ przez dodanie krawędzi łączącej $v$ z korzeniem drzewa $T'$. Jeśli natomiast $r = D(i,j)$, to doklejamy $T'$ do wierzchołka $v$ przez utożsamienie $v$ z korzeniem $T'$. Łatwo jest sprawdzić, że tak otrzymane drzewo jest ultrametryczne dla $D$.

Dowód (ii) jest analogiczny do powyższego dowodu. (iii) wynika z konstrukcji zawartej w dowodzie ($\Longleftarrow$) punktu (i). ∎

## 8.2 Metody zdobywania ultrametrycznych danych

Wiele metod jest opartych na tzw. *teorii zegara molekularnego*, zaproponowanej na początku lat 60. przez E. Zuckerkandla i L. Paulinga. Zakłada ona, że dla każdego białka mutacje akceptowalne jego składu aminokwasów zachodzą ze stałą częstością. Zatem liczba takich mutacji jest proporcjonalna do upływu czasu. Trzeba jednak pamiętać, że częstość mutacji akceptowalnych zależy od rodzaju białka i bywa różna dla różnych białek. Konkurencyjna do niej jest *neutralna teoria molekularnej ewolucji* zaproponowana przez M. Kimurę (1979). Kładzie ona większy nacisk na selekcję naturalną w procesie ewolucji, niż na losowy dryf genetyczny.

Stosowane są dwie metody szacowania odległości ewolucyjnej pomiędzy dwoma jednostkami taksonomicznymi na podstawie liczby mutacji akceptowalnych:

- *Metoda laboratoryjna:* denaturuje się DNA pochodzące z dwóch jednostek taksonomicznych, miesza się je ze sobą, a następnie ponownie doprowadza się do hybrydyzacji. Następnie sprawdza się temperaturę denaturacji. Im jest ona wyższa tym cząsteczki DNA są do siebie bardziej podobne (w sensie liczby mutacji akceptowalnych).

- *Metoda oparta na sekwencjach:* sekwencjonuje się cząsteczki, a następnie oblicza odległość edycyjną.

Problemem wspólnym dla obydwu wyżej wspomnianych metod jest to, że dane zbierane w ten sposób bardzo rzadko spełniają warunek ultrametryczności. W związku z tym wprowadza się szerszą klasę macierzy i odpowiadających im drzew.

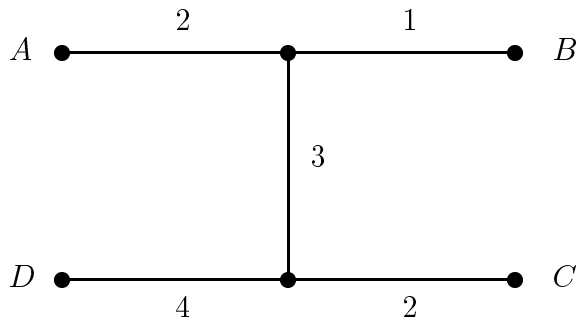
## 8.3 Drzewa addytywne

Niech $D$ będzie $n \times n$ macierzą spełniającą warunki (m1) i (m2). Niech $T$ będzie drzewem o krawędziach etykietowanych liczbami oraz o wyróżnionych $n$ wierzchołkach, którym przypisano w sposób różnowartościowy wiersze macierzy $D$. Powiemy, że $T$ jest drzewem *addytywnym* dla $D$, gdy dla dowolnych wróżnionych wierzchołków $i, j$, koszt drogi w $T$ od $i$ do $j$ wynosi $D(i, j)$. Powiemy, że macierz $D$ jest *addytywna*, gdy istnieje drzewo addytywne dla $D$.

**Przykład 8.3.1** Niech $D$ będzie następującą macierzą

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $A$ | 0 | 3 | 7 | 9 |
| $B$ | | 0 | 6 | 8 |
| $C$ | | | 0 | 6 |
| $D$ | | | | 0 |

Macierz ta jest addytywna, a drzewem addytywnym dla $D$ jest




□

**Zadanie 8.3.1** Dowieść, że jeśli macierz $D$ jest ultrametryczna to jest addytywna.

**Zadanie 8.3.2** Dowieść, że macierz spełniająca warunki (m1), (m2) jest ultrametryczna $\Longleftrightarrow$ $D$ ma drzewo addytywne posiadające wierzchołek $v$ równo odległy od wszystkich wyróżnionych wierzchołków.

Powiemy, że $n \times n$ macierz $D$ jest *addytywna w sposób zwarty*, gdy $D$ ma drzewo addytywne o dokładnie $n$ wierzchołkach.

**Twierdzenie 8.3.1** *Jeśli $D$ ma addytywne drzewo $T$, to $T$ musi być jedynym minimalnym drzewem rozpinającym dla grafu $G(D)$ (pełny graf o $n$ wierzchołkach, krawędź pomiędzy $i$ oraz $j$ ma wagę $D(i,j)$).*

**Dowód:** Jeśli $T$ jest zwartym addytywnym drzewem dla $D$, to oczywiście $T$ jest drzewem rozpinającym dla $G(D)$. Weźmy dowolną krawędź $e = (i,j)$ w $G(D)$, nie należącą do $T$ i załóżmy, że istnieje inne minimalne drzewo rozpinające $T'$, do którego $e$ należy. Jeśli usuniemy $e$ z $T'$ to dostaniemy dwie spójne składowe $S$ oraz $\bar{S}$. Mamy $i \in S$ oraz $j \in \bar{S}$. Rozważmy ścieżkę w $T$ od $i$ do $j$ i niech $e' = (i',j')$ będzie krawędzią na tej ścieżce taką, że $i' \in S$ oraz $j' \in \bar{S}$. Oczywiście $e'$ nie należy do $T'$ oraz $D(i',j') < D(i,j)$ (ponieważ droga w $T$ od $i$ do $j$ zawiera oprócz $e'$ jeszcze inne krawędzie). Zatem drzewo $T''$ otrzymane z $T'$ przez zamianę $e$ na $e'$ ma mniejszą wagę, co przeczy minimalności $T'$. Pokazaliśmy więc, że żadna krawędź nie należąca do $T$ nie może należeć do minimalnego drzewa rozpinającego. To dowodzi, że $T$ jest jedynym minimalnym drzewem rozpinającym. ∎

Zauważmy, że powyższe twierdzenie daje algorytm działający w czasie $O(n^2)$, sprawdzający czy dana macierz jest addytywna w sposób zwarty.

## 8.4 Drzewa filogenetyczne

Obecnie zajmiemy się innym podejściem do rekonstrukcji ewolucji. Podejście to jest oparte na tzw. metodzie *oszczędności* lub *parsymonii*. Mamy dany pewien zbiór atrybutów (cech), które każdy z rozważanych obiektów może posiadać lub nie. Omawiane podejście jest jednym z wielu, a wersja tutaj przedstawiana jest uproszczoną wersją tzw. problemu maksymalnej oszczędności.

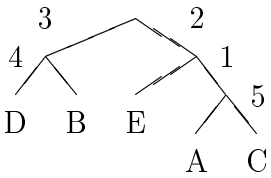
Dana jest $n \times m$ macierz zero-jedynkowa $M$. Wiersze macierzy reprezentują obiekty, a kolumny reprezentują cechy. Tak więc, pozycja $(p,i)$ macierzy $M$ zawiera 1, gdy obiekt $p$ posiada cechę $i$. Powiemy, że drzewo $T$ z korzeniem jest *filogenetyczne* dla $M$, gdy $T$ ma dokładnie $n$ liści oraz spełnia następujące trzy warunki:

- Każdy z $n$ obiektów etykietuje dokładnie jeden liść.
- Każda z $m$ cech etykietuje dokładnie jedną z krawędzi drzewa (drzewo może mieć więcej krawędzi).
- Dla każdego obiektu $p$, wszystkie etykiety krawędzi stojące na drodze od korzenia do $p$ opisują dokładnie wszystkie cechy jakie ten obiekt posiada.

**Przykład 8.4.1** Rozważmy następującą macierz $M$.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $A$ | 1 | 1 | 0 | 0 | 0 |
| $B$ | 0 | 0 | 1 | 0 | 0 |
| $C$ | 1 | 1 | 0 | 0 | 1 |
| $D$ | 0 | 0 | 1 | 1 | 0 |
| $E$ | 0 | 1 | 0 | 0 | 0 |

Poniższe drzewo jest filogenetyczne dla $M$




Natomiast poniższa macierz nie posiada drzewa filogenetycznego.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $A$ | 1 | 1 | 0 | 0 | 0 |
| $B$ | 0 | 0 | 1 | 0 | 1 |
| $C$ | 1 | 1 | 0 | 0 | 1 |
| $D$ | 0 | 0 | 1 | 1 | 0 |
| $E$ | 0 | 1 | 0 | 0 | 0 |

□

Omawiany model ewolucji oparty jest na założeniach, że praprzodek nie posiada żadnej z cech oraz każdy obiekt (obecnie żyjący lub wcześniej), jeśli nabędzie daną cechę, to już jej nie może się pozbyć. Cechy brane do klasyfikacji mają albo morfologiczny charakter (np. posiadanie kręgosłupa) lub mogą być oparte na sekwencji DNA lub na białkach danego gatunku (np. to czy dane białko zawiera pewien charakterystyczny ciąg aminokwasów). Cechy morfologiczne mogą prowadzić do niewłaściwej interpretacji bo mogą się zmieniać pod wpływem presji selekcyjnej (np. pewna cecha może pojawiać i zanikać kilka razy w procesie ewolucji). Przykładowo uważa się, że skrzydła podlegały niezależnie kilka razy zasadniczym zmianom ewolucyjnym. Poniżej podajemy kilka podejść stanowiących podstawę do klasyfikacji gatunków.

- Fragmenty DNA nie mające funkcyjnego charakteru (np. brane spoza genów, sekwencji regulatorowych, itp.). Uważa się, że podlegają one losowym mutacjom. Zatem dla odległych ewolucyjnie gatunków te fragmenty powinny się znacznie różnić.

- Występowanie nukleotydu na ustalonej pozycji. Fragmenty DNA uliniawia się i sprawdza jaki nukleotyd znajduje się na ustalonej pozycji (dla każdego gatunku). Aby cecha ta miała binarny charakter, nukleotydy klasyfikuje się w grupy: puryny (A, G) i pirimidyny (C, T).

- Inną ważną binarną cechą jest to czy ustalone białko jest regulowane przez inne ustalone białko. Np. myszy i ludzie mają bardzo podobne zbiory białek, a to co odróżnia od siebie obydwa gatunki to sposób regulacji białek (zakodowany poza sekwencjami kodującymi geny).

- Ważną cechą jest rozkład przerw w uliniowieniach (wielokrotnych) pewnych białek pochodzących od różnych gatunków. Np. w ten sposób badano grzyby, rośliny i zwierzęta. Okazało się, że biorąc pod uwagę rozkład przerw, grzyby i zwierzęta są sobie bliższe z ewolucyjnego punktu widzenia niż grzyby i rośliny.

### 8.4.1 Problem doskonałej filogenezy

Dana jest $n \times m$ macierz binarna $M$. Problem polega na sprawdzeniu czy $M$ posiada drzewo filogenetyczne.

Każdą kolumnę macierzy $M$ traktujemy jak liczbę w zapisie binarnym (najbardziej znaczący bit znajduje się w pierwszym wierszu) i permutujemy kolumny tak, aby odpowiadające im liczby tworzyły ciąg nierosnący. Niech $\bar{M}$ będzie tak otrzymaną macierzą. Oczywiście $M$ ma drzewo filogenetyczne $\Longleftrightarrow$ $\bar{M}$ ma drzewo filogenetyczne.

Dla każdej kolumny $k$ macierzy $\bar{M}$, niech $O_k$ będzie zbiorem wszystkich obiektów mających cechę $k$, czyli

$$O_k = \{p \mid \bar{M}(p,k) = 1\}.$$

Zauważmy, że jeśli $O_j$ jest właściwym podzbiorem $O_k$, to kolumna $k$ musi być na lewo od kolumny $j$ w macierzy $\bar{M}$.

**Twierdzenie 8.4.1** *Macierz $\bar{M}$ ma drzewo filogenetyczne $\Longleftrightarrow$ dla dowolnych kolumn $i, j$, zbiory $O_i$ oraz $O_j$ są albo rozłączne albo jeden zawiera drugi.*

**Dowód:** ($\Longrightarrow$) Niech $T$ będzie drzewem filogenetycznym dla $\bar{M}$ i niech $e_i$ oraz $e_j$ będą krawędziami w $T$ odpowiadającymi kolumnom $i$ oraz $j$. Jeśli $e_i$ oraz $e_j$ leżą na jednej ścieżce prowadzącej od korzenia (powiedzmy, że $e_i$ leży wyżej niż $e_j$), to wiecej obiektów ma cechę $i$. Zatem $O_j \subseteq O_i$. Jeśli $e_i$ oraz $e_j$ nie leżą na jednej ścieżce, to $O_i \cap O_j = \emptyset$.

($\Longleftarrow$) Dla każdego obiektu $p$ tworzymy słowo $W_p$ składające się z cech, które $p$ posiada, w kolejności ich wystąpienia w $\bar{M}$. Każde ze słów $W_p$

kończymy specjalnym symbolem \$. Zatem żadne słowo nie jest prefiksem innego. Z założenia twierdzenia wynika, że dla dowolnych obiektów $p, q$, słowa $W_p$ oraz $W_q$ muszą być identyczne do pewnego miejsca i od tego miejsca nie zawierają wspólnej cechy. Istotnie, weźmy największe $k$ o tej własności, że $\bar{M}(p,k) = \bar{M}(q,k) = 1$. Jeśli dla pewnego $i < k$ mamy $\bar{M}(p,i) = 1$ oraz $\bar{M}(q,i) = 0$, to ponieważ $O_i \cap O_k \neq \emptyset$, to $O_i$ musiałoby być właściwym podzbiorem $O_k$. Czyli $k < i$, sprzeczność.

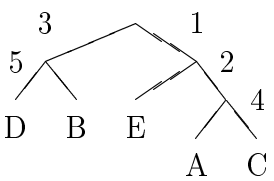
Zatem słowa $\{W_p \mid 1 \leq p \leq n\}$ można ułożyć w drzewo (usuwamy etykietę \$). To drzewo jest filogenetyczne dla $\bar{M}$. ∎

**Przykład 8.4.2** Weźmy macierz z Przykładu 8.4.1. Po uporządkowaniu kolumn dostajemy macierz $\bar{M}$.

| | 2 | 1 | 3 | 5 | 4 |
|---|---|---|---|---|---|
| $A$ | 1 | 1 | 0 | 0 | 0 |
| $B$ | 0 | 0 | 1 | 0 | 0 |
| $C$ | 1 | 1 | 0 | 1 | 0 |
| $D$ | 0 | 0 | 1 | 0 | 1 |
| $E$ | 1 | 0 | 0 | 0 | 0 |

Dostajemy następujące słowa: $W_A = 12\$$, $W_B = 3\$$, $W_C = 124\$$, $W_D = 35\$$, $W_E = 1\$$.

Drzewem filogenetycznym dla $\bar{M}$ jest natępujące drzewo.




□

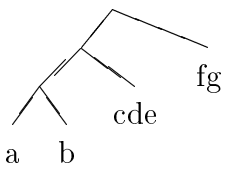
Twierdzenie 8.4.1 sugeruje algorytm, który rozstrzyga problem doskonałej filogenezy w czasie $O(nm^2)$. Używając algorytmu radix-sort i wskaźników można zaprojektować algorytm działający w czasie $O(mn)$.

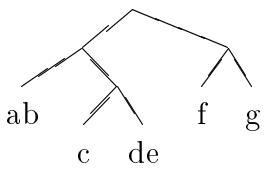
Zarówno problem drzew addytywnych, jak i problem doskonałej filogenezy można zredukować do problemu drzew ultrametrycznych.

### 8.4.2 Kompatybilność drzew

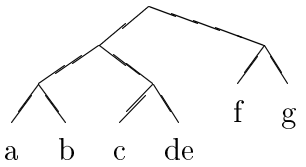
Problem sprawdzania kompatybilności drzew pojawia się wtedy, gdy mamy pewien zbiór drzew przedstawiających fragmenty procesu ewolucji, bądź też drzewa przedstawiające ten sam fragment, ale otrzymane w wyniku zastosowania różnych metod. Powiemy, że drzewo $T'$ jest *ulepszeniem* drzewa $T$, gdy $T$ może być otrzymane z $T'$ w drodze zastosowania pewnego ciągu kontrkcji krawędzi drzewa $T'$.

**Przykład 8.4.3** Niech $T_1$ i $T_2$ będą następującymi drzewami

$T_1 =$ 


$T_2 =$ 

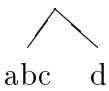

Drzewo $T_3$ jest ulepszeniem drzew $T_1$ i $T_2$.

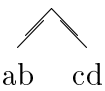
$T_3 =$ 


□

Powiemy, że dwa drzewa $T_1$ i $T_2$ są *kompatybilne*, gdy istnieje drzewo będące ulepszeniem $T_1$ i $T_2$.

**Przykład 8.4.4** Drzewa $T_1$ i $T_2$ z Przykładu 8.4.3 są kompatybilne. Natomiast drzewa




nie są kompatybilne. □

Problem sprawdzania kompatybilności drzew można sprowadzić do problemu doskonałej filogenezy. Mamy dane $T_1$ oraz $T_2$. Budujemy macierz $M_1$ (rozmiar $n \times m_1$, gdzie $n$ jest liczbą obiektów, a $m_1$ jest liczbą wierzchołków drzewa $T_1$). $M_1(i,j) = 1 \iff$ obiekt $i$ jest widoczny z wierzchołka $j$. Podobnie budujemy macierz $M_2$ (rozmiaru $n \times m_2$). Następnie tworzymy macierz $M_3$ rozmiaru $n \times (m_1 + m_2)$ przez zestawienie macierzy obok siebie.

**Twierdzenie 8.4.2** *Drzewa $T_1$ i $T_2$ są kompatybilne $\Longleftrightarrow$ macierz $M_3$ opisana wyżej ma drzewo filogenetyczne. Ponadto drzewo filogenetyczne dla $M_3$ jest ulepszeniem drzew $T_1$ i $T_2$.*

### 8.4.3 Uogólnienie doskonałej filogenezy

Cechy mogą przyjmować więcej niż dwa stany. Tak więc macierz $M$ zawiera dla każdego obiektu i każdej cechy stan, w jakim ten obiekt może się znaleźć. Przyjmujemy, że na początku wszystkie obiekty mają ten sam stan (dla każdej cechy). *Uogólnione drzewo filogenetyczne* dla $M$ jest to drzewo z korzeniem, każdy liść jest etykietowany pewnym obiektem (w sposób różnowartościowy). Pewne krawędzie są etykietowane. Etykiety krawędzi są trójkami postaci $(c, x, y)$, co oznacza, że cecha $c$ zmieniła stan z $x$ na $y$. Dla każdej cechy $c$ i stanu $y$ jest co najwyżej jedna etykieta postaci $(c, x, y)$ (czyli każda cecha zmienia stan co najwyżej jeden raz).

Problem doskonałej filogenezy polega na tym, że mamy daną macierz $M$ (każda cecha może przyjąć co najwyżej $r$ stanów). Należy sprawdzić czy istnieje uogólnione drzewo filogenetyczne dla $M$. Problem ten został postawiony przez P. Bunemena w 1974 i jego złożoność pozostawała otwarta przez 20 lat. Agarwala i Fernandez-Baca (1994) pokazali, że dla każdego ustalonego $r$ problem jest wielomianowy (od rozmiaru macierzy). Jeśli $r$ jest częścią danych to, jak pokazał M. Steel (1992), problem jest NP-zupełny.

## 8.5 Problem maksymalnej oszczędności i drzewa Steinera

Problem maksymalnej oszczędności najwygodniej jest sformułować w języku drzew Steinera. Dany graf $G = \langle V, E\rangle$ z wagami (krawędzie mają wagi). Niech $X \subseteq V$. *Drzewem Steinera* dla $X$ nazwiemy każde spójne poddrzewo grafu $G$ zawierające $X$. *Problem drzew Steinera* polega na znalezieniu drzewa Steinera (dla $X$) o minimalnej wadze.

Niech $d \geq 1$. *$d$-wymiarowa kostka* jest to graf $G = \langle V, E\rangle$ taki, że $V = \{0,1\}^d$ oraz dla słów $U, W$ istnieje krawędź łącząca $U$ z $W \Longleftrightarrow U$ i $W$ różnią się na dokładnie jednej pozycji (tzw. odległość Hamminga jest równa 1). Waga takiej krawędzi wynosi 1. Problem *maksymalnej oszczędności* to problem drzew Steinera dla $d$-wymiarowej kostki, gdzie zbiorem $X$ wyróżnionych wierzchołków są obiekty (czyli wiersze macierzy).

Problem drzew Steinera dla kostek jest NP-trudny (Foulds, Graham'82). Później pokazano, że problem drzew Steinera dla kostek z dowolnymi wagami jest też NP-trudny (Gusfield'84). Znaleziono algorytm aproksymacyjny dla

ogólnego problemu drzew Steinera (dla dowolnego grafu) o ilorazie błędu 11/6 (Zelikovsky'93 oraz Berman, Ramaiyer'94).