Drzewa filogenetyczne

Filogenetyka molekularna zajmuje się odtwarzaniem historii ewolucji żyjących obecnie gatunków na podstawie porównania ich sekwencji biologicznych.

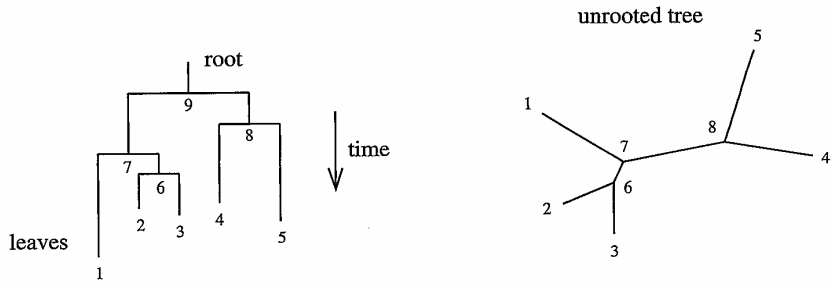
Co to jest drzewo filogenezy?

Drzewo filogenetyczne T jest odpowiednikiem „drzewa genealogicznego" pewnej grupy gatunków.

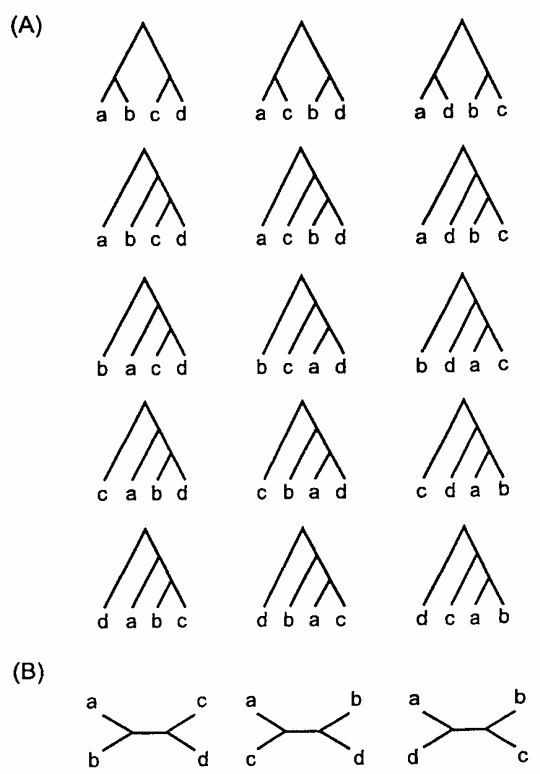
- Zbiór $L(T)$ jego liści (przypomnienie: wierzchołki stopnia 1) to zbiór badanych gatunków żyjących dziś.
- Wierzchołki wewnętrzne $V(T)\backslash L(T)$ to wspólni przodkowie (gatunki wymarłe).
- Jeśli drzewo jest **ukorzenione**, zawiera wówczas jeden wyróżniony wierzchołek r – **korzeń** (często stopnia 2) przedstawiający najstarszego antenata badanej grupy. Wówczas krawędzie można zorientować zgodnie z upływem czasu od korzenia w kierunku liści – przedstawiają one historię rozchodzenia się linii gatunkowych.
- Metody filogenetyki często tworzą drzewa **nieukorzenione**, obrazujące wzajemne relacje bliskości– odległości pokrewieństwa międzygatunkowego bez informacji o położeniu najstarszego przodka. Drzewo można **ukorzenić** wstawiając r w wierzchołku lub „na środku" dowolnej krawędzi. W praktyce często ukorzenia się otrzymane drzewo dołączając do grupy „dalekiego krewnego" (outgroup) – badamy do której krawędzi należy go doczepić wyznaczając w ten sposób ustawienie korzenia.

- W drzewach filogenetycznych nie interesują nas wierzchołki stopnia 2 (ew. z wyjątkiem korzenia) – „ściągamy" je, gdyż nie niosą żadnej informacji.
- Drzewo nazwiemy **binarnym** gdy wszystkie wierzchołki nie będące liśćmi są stopnia 3 (ew. z wyjątkiem korzenia, który ma wtedy stopień 2) – niosą one najwięcej informacji o historii ewolucji.
- Drzewa T i T' o tym samym zbiorze liści $L(T)=L(T')$ (i ew. korzeniu) są **równoważne** (tzn. mają tą samą **topologię**, opisują tą samą historię ewolucji) gdy istnieje izomorfizm tych grafów $f:V(T)\rightarrow V(T')$ będący identycznością dla liści (i ew. korzenia).

Przykład. Ukorzenienie 5–listnego drzewa binarnego.



Przykład. Wszystkie topologie drzew binarnych dla 4 gatunków. A) drzewa ukorzenione. B) nieukorzenione.



Zliczanie nierównoważnych topologii

Niech $n(T)=|L(T)|$ – liczba liści, $w(T)=|V(T)\backslash L(T)|$ – ilość wierzchołków wewnętrznych, $m(T)=|E(T)|$ – liczba krawędzi.

Przypomnienie: w każdym drzewie jest $m=n+w-1$ krawędzi.

Najwięcej krawędzi przy ustalonej liczbie liści n ma drzewo binarne. Z lematu o uściskach dłoni (przypadek nieukorzeniony): $2m=n+3w$, stąd:

$$w=n-2, \quad m=2n-3$$

Drzewo z korzeniem stopnia 2 będzie miało o jedną krawędź i jeden wierzchołek wewnętrzny (korzeń) więcej:

$$w=n-1, \quad m=2n-2$$

Problem. Ile jest możliwych topologii dla drzewa binarnego n–listnego?

Binarne drzewo n–listne bez korzenia można ukorzenić w „środku” każdej z $2n-3$ krawędzi uzyskując n–listne drzewo z korzeniem. W ten sam sposób zamiast korzenia możemy dodać gatunek $n+1$ otrzymując drzewo $(n+1)$–listne bez korzenia. Mamy zatem:

Rozwiązanie. Jest **3*5*...*(2*n*–5)** możliwych topologii drzew binarnych nieukorzenionych i **3*5*...*(2*n*–3)** takich drzew z korzeniem.

Ilość możliwych przebiegów historii ewolucji rośnie superwykładniczo w funkcji rozmiaru badanej grupy gatunków.

Przykład. Liczba drzew bez korzenia

$n=5$	15
$n=10$	$\sim 2\times 10^{6}$
$n=15$	$\sim 7.9\times 10^{12}$
$n=20$	$\sim 2.2\times 10^{20}$

Niegrafowy opis topologii drzewa

Problem Doskonałej Filogenezy. Dany jest zbiór obiektów (gatunków) L i cech binarnych C. Wiemy które cechy mają poszczególne obiekty $v \in L$. Czy istnieje drzewo z korzeniem o zbiorze liści L, którego krawędzie (nie koniecznie wszystkie) są poetykietowane cechami z C, takie że:

- każda cecha $c \in C$ występuje na jednej krawędzi,
- zbiór cech każdego obiektu $v \in L$ to zbiór wszystkich etykiet krawędzi ścieżki łączącej v z korzeniem?

Jest to model cech gatunkowych, **nieodwracalnych** i **unikalnych**, pojawiających się tylko raz. Drzewo ma prezentować ewolucję. Krawędź etykietowana cechą to fragment historii, w którym pewien gatunek nabył tą cechę przekazując ją wszystkim swoim potomkom (tylko oni ją posiadają).

Model ten był stosowany w klasycznej filogenetyce. Jednak cechy morfologiczne, „konstrukcyjne" organizmów często naruszają powyższe warunki:

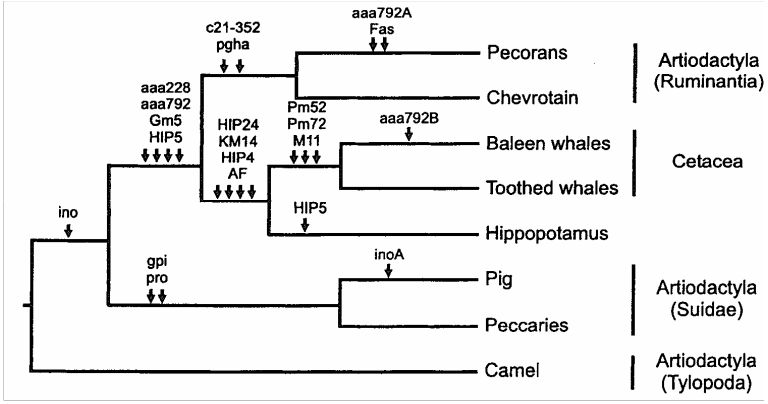
- raz nabyte mogą zanikać w czasie,
- podobne przystosowania często pojawiają się niezależnie (tzw. ewolucja konwergentna).

np. żyworodność, zdolność latania, pływania itp.

Historia genomów zna zdarzenia lepiej pasujące do wymagań nieodwracalności i unikalności (**markery genetyczne**) np.

- wbudowanie się lub powielenie „śmieciowych" sekwencji retrotranspozonów, sekwencji SINE (do kilkuset par) i LINE (do kilku tys.), wstawek retrowirusowych itp.
- rozległe rearanżacje chromosomów,
- pojawienie się nowego intronu w genie.

Przykład. Drzewo filogenetyczne uzyskane na podstawie 21 różnych wstawek typu SINE.



Niech zbiory L i C będą danymi do problemu Doskonałej Filogenezy. Dla cechy $c \in C$ przez $L_c \subseteq L$ oznaczymy podzbiór gatunków posiadających tą cechę.

Twierdzenie (o Doskonałej Filogenezie). Rozwiązanie dla problemu doskonałej filogenezy istnieje wtedy i tylko wtedy, gdy dla każdej pary L_c,L_d (c,$d \in C$) zbiory te zawierają się w sobie ($L_c \subseteq L_d$ lub $L_d \subseteq L_c$) lub są rozłączne ($L_c \cap L_d = \emptyset$). Wówczas odpowiednie drzewo można uzyskać w czasie wielomianowym.

Dowód. $\Rightarrow$ Dla pary cech c,d rozważamy etykietowane nimi krawędzie e_c, e_d. Cechę c mają tylko liście poddrzewa „zawieszonego" na kawędzi e_c, podobnie z e_d. Albo jedna z tych krawędzi leży w poddrzewie drugiej (ew. $e_c = e_d$) i wówczas $L_c \subseteq L_d \vee L_d \subseteq L_c$, albo są to „oddzielne" poddrzewa, czyli $L_c \cap L_d = \emptyset$.

$\Leftarrow$ Bez zmniejszenia ogólności można przyjąć, że $L_c \neq \emptyset$ dla wszystkich $c \in C$, ponadto pewną cechę $l \in C$ mają wszystkie gatunki ($L_l = L$), oraz dla każdego gatunku istnieje cecha posiadana wyłącznie przez niego. Rozważmy diagram Hassego relacji częściowego porządku inkluzji "$\subseteq$" pomiędzy zbiorami L_c dla $c \in C$. Na łukach wychodzących z L_c ustawiamy etykietę c, a w wierzchołkach jednoelementowych $\{x\}$ umieszczamy gatunki $x \in L$. Uzyskujemy drzewo doskonałej filogenezy (z łukami skierowanymi do korzenia), gdyż:

- zbiór L jest elementem największym tej relacji porządku,
- wierzchołki o stopniu wejściowym 0 to zbiory $\{x\}$ dla $x \in L$,
- diagram jest drzewem skierowanym – z żadnego wierzchołka nie wychodzi więcej, niż jeden łuk. Gdyby bowiem istniały łuki $L_c \rightarrow L_d$ i $L_c \rightarrow L_e$ ($L_d \neq L_e$), to $L_c \subseteq L_d \cap L_e$, czyli $L_d \subseteq L_e$ lub $L_e \subseteq L_d$ i jeden z tychże łuków byłby przechodni. ■

Dla skończonego zbioru gatunków L **<u>rodziną klastrów</u>** nazwiemy dowolny taki zestaw jego podzbiorów, że:

- zawiera ona wszystkie klastry jednoelementowe $\{x\}$ dla $x \in L$ oraz trywialny klaster – cały L,
- dla dowolnych klastrów A,B zachodzi $A \subseteq B$ lub $B \subseteq A$ lub $A \cap B = \emptyset$.

Rodzinę klastrów dla L możemy traktować jako **<u>alternatywny opis</u>** drzewa ukorzenionego bez wierzchołków stopnia 2 (być może poza korzeniem) o zbiorze liści L, w sposób jednoznaczny określający jego topologię.

- każdy klaster można uważać za zbiór liści posiadających pewną cechę,
- w drzewie doskonałej filogenezy klastry odpowiadają zbiorowi liści–potomków wierzchołków,
- klastry jednoelementowe odpowiadają liściom,
- w wielomianowym czasie możemy przetłumaczyć drzewo na zbiór klastrów i odwrotnie (por. algorytm z poprzedniego twierdzenia).

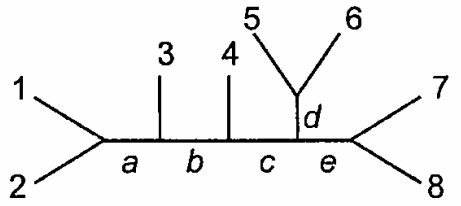
Opis „zbiorowy" topologii drzewa bywa wygodniejszy od „grafowego":

- prostsze struktury danych,
- niezależność od „nazewnictwa" wierzchołków wewnętrznych,
- dostarcza prostej metody sprawdzenia czy drzewa są równoważne (weryfikacja równości rodzin zbiorów).

Podobnie można w sposób jednoznaczny „niegrafowo” określić topologię drzewa nieukorzenionego o liściach L.

- **rozbiciem** zbioru L nazwiemy dowolną nieuporządkowaną parę $\{A,B\}$ jego niepustych podzbiorów, spełniającą $A\cup B=L$, $A\cap B=\emptyset$,
- rozbicia $\{A,B\}\neq\{C,D\}$ dla L są **zgodne** gdy dokładnie jeden zbiór spośród $A\cap C$, $A\cap D$, $B\cap C$, $B\cap D$ jest pusty.
- **Rodziną zgodnych rozbić** $\delta(T)$ dla **drzewa nieukorzenionego** T o zbiorze liści L nazywamy wszystkie rozbicia $\{A,B\}$ jego liści powstające przez usunięcie pewnej krawędzi (co dzieli drzewo na dwie części).
- Odwrotnie: zgodna rodzina rozbić δ zbioru L zawierająca wszystkie **rozbicia trywialne** $\{\{v\},L\setminus\{v\}\}$ dla $v\in L$ jednoznacznie określa topologię nieukorzenionego drzewa T (bez stopni 2) z liśćmi L, takiego że $\delta(T)=\delta$.
- Translacji między obydwoma opisami topologii można dokonać w czasie wielomianowym.

Przykład. Poniższe drzewo zawiera w $\delta(T)$ następujące rozbicia nietrywialne odpowiadające usunięciom pewnych krawędzi:



a – {{1,2},{3,4,5,6,7,8}}, b – {{1,2,3},{4,5,6,7,8}},
c – {{1,2,3,4},{5,6,7,8}}, d – {{1,2,3,4,7,8},{5,6}},
e – {{1,2,3,4,5,6},{7,8}}.

W „języku rozbić”:

- rozbicie $\{A,B\}$ – krawędź (A,B – zbiory liści z obu stron),
- rozbicie trywialne $\{\{v\},L\setminus\{v\}\}$ – krawędź wisząca (z liściem v) ,
- usunięcie rozbicia – ściągnięcie krawędzi,
- dodanie nowego zgodnego rozbicia – „rozdzielenie” wierzchołka wewnętrznego na dwa połączone krawędzią,
- drzewo binarne – $2|L|-3$ rozbicia.

Gorzej, jeśli dysponujemy zafałszowanymi danymi:

Twierdzenie. Problem znalezienia w rodzinie rozbić zbioru L największej możliwej podrodziny zgodnej jest NP–trudny.

Dowód. Redukcja z problemu Największego Zbioru Niezależnego w grafie bez wierzchołków izolowanych i wiszących. Niech $G(V,E)$ – graf, definiujemy rodzinę rozbić R zbioru $E'=E\cup\{S\}$ złożoną z par:

1. wszystkie rozbicia trywialne $\{\{x\},E'\setminus\{x\}\}$ dla $x\in E'$,
2. $\{E_v,E'\setminus E_v\}$ dla $v\in V$, gdzie $E_v\subseteq E$ – krawędzie stykające się z v.

Szukanie największej zgodnej podrodziny z R sprowadza się do poszukiwania największego zbioru parami zgodnych rozbić typu 2, gdyż dodanie nowego rozbicia trywialnego nie może popsuć zgodności. Jednak dla $u\neq v\in V$ wykluczone jest $E_v\subseteq E_u$ i $E_u\subseteq E_v$, więc $\{E_v,E'\setminus E_v\}$, $\{E_u,E'\setminus E_u\}$ są zgodne jedynie gdy $E_v\cap E_u=\emptyset$, czyli u,v nie sąsiadują. Dlatego największy zgodny podzbiór rozbić odpowiada największemu zbiorowi niezależnemu w G. ■

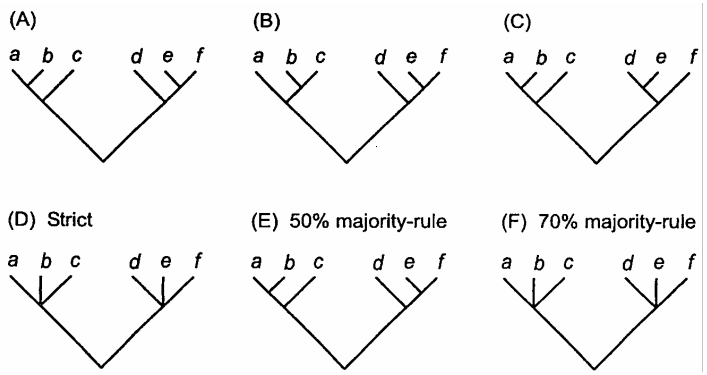
Drzewa konsensusu

Problem. Niektóre badania filogenetyczne dla danego zbioru gatunków L mogą dostarczyć cały zbiór drzew przedstawiających nieco odmienne propozycje historii ich ewolucji. Jak znaleźć drzewo „wyławiające" zgodne informacje, a odrzucające sprzeczne dane?

Drzewo pełnego konsensusu: wybieramy rozbicia zbioru L występujące we wszystkich zaproponowanych drzewach.

Im więcej rozbić (czyli krawędzi) odrzucimy, tym mniej nośny informacyjnie jest wynik. Bardziej „tolerancyjne" jest dla $x \geq 50\%$:

Drzewo konsensusu x%: wybieramy rozbicia obecne w więcej niż x% zaproponowanej populacji drzew.
Przykład. Konsensus o różnym poziomie tolerancji dla drzew A, B, C.



Poprawność konstrukcji. Czy zbiór rozbić do budowy drzewa konsensusu będzie zgodny?

Tak: każde dwa rozbicia występują w >50% drzew, muszą więc spotkać się przynajmniej w jednym drzewie. ■

Kompatybilny zbiór drzew

Dane są dwa nieukorzenione drzewa filogenetyczne T i T' o liściach $L(T) \subseteq L(T')$. Powiemy, że T' jest **rozszerzeniem** T (zapis $T \leq T'$), gdy T można otrzymać z poddrzewa T' indukowanego w nim przez wszystkie najkrótsze ścieżki łączące liście $L(T)$ poprzez ew. ściągnięcie niektórych krawędzi. Zatem drzewo T' zawiera wszystkie informacje o ewolucji z drzewa T i być może jeszcze inne fakty.

Problem. Dla zbioru drzew filogenetycznych $T_1,...,T_k$ sprawdź czy istnieje drzewo T, będące rozszerzeniem każdego z nich (a więc uogólniające ich wszystkie dane). Taki zbiór nazwiemy **kompatybilnym**.

Jest to inne zagadnienie, niż szukanie drzewa konsensusu – zamiast wychwytywania informacji wspólnych, potwierdzonych przez każde drzewo lub znaczny ich procent szukamy drzewa zawierającego wszystkie fakty z ciągu $T_1,...,T_k$.

Fakt. Jeśli $T_1,...,T_k$ mają ten sam zbiór liści, wówczas problem kompatyblilności jest wielomianowy: wystarczy sprawdzić zgodność rodziny rozbić $\delta = \delta(T_1) \cup ... \cup \delta(T_k)$. W takim przypadku δ jest zbiorem rozbić dla drzewa rozszerzającego.

Fakt. Jeśli zbiory liści $L(T_i)$ są różne (szukamy drzewa rozszerzającego z $L(T) \supseteq L(T_1) \cup ... \cup L(T_k)$) – wówczas tak sformułowany problem kompatybilności jest NP–trudny.

Nawet pytanie o kompatybilność rodziny drzew 4–listnych jest NP–zupełne.

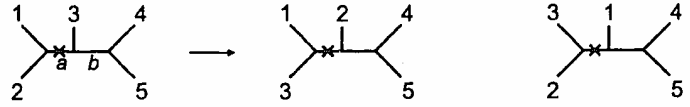
Wiadomo też, że:

- analogiczny problem dla drzew ukorzenionych jest wielomianowy,
- zarówno w przypadku drzew ukorzenionych, jak i bez korzeni pojawia się problem niejednoznaczności rozwiązania (może istnieć wiele różnych drzew rozszerzających dany zbiór).

Sąsiedztwa w przestrzeni drzew

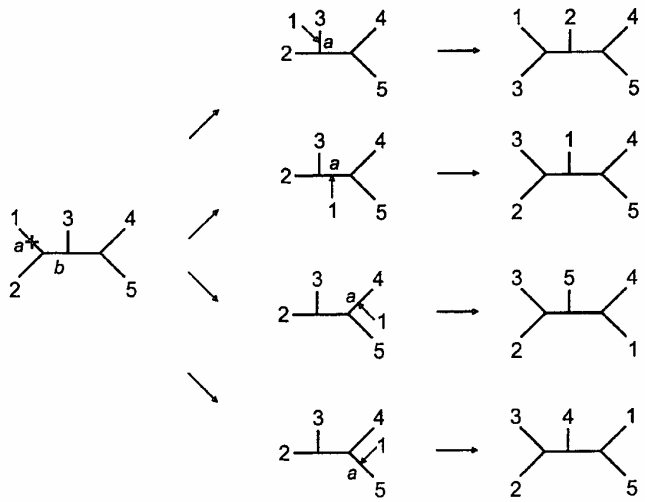
Niektóre metody filogenetyczne wymagają przeszukania olbrzymiej przestrzeni różnych topologii w celu znalezienia jednej, (sub)optymalnej względem pewnej funkcji jakości. Można wówczas stosować metodę gradientową lub bardziej wyrafinowane heurystyki np. Simulated Annealing, Tabu Search. Techniki powyższe wymagają określenia dla każdego elementu przeglądanej przestrzeni pewnego ograniczonego sąsiedztwa „bliskich mu" punktów, między którymi algorytm możne przechodzić w jednym kroku. Podamy trzy sposoby definiowania tworzenia otoczeń (coraz szerszych) w przestrzeni *n*–listych nieukorzenionych drzew binarnych.

- **Nearest Neighbour Interchanges** (**NNI**): sąsiada drzewa *T* tworzymy wybierając dwa poddrzewa wyrastające z przeciwległych końców pewnej krawędzi wewnętrznej i zamieniając je miejscami.



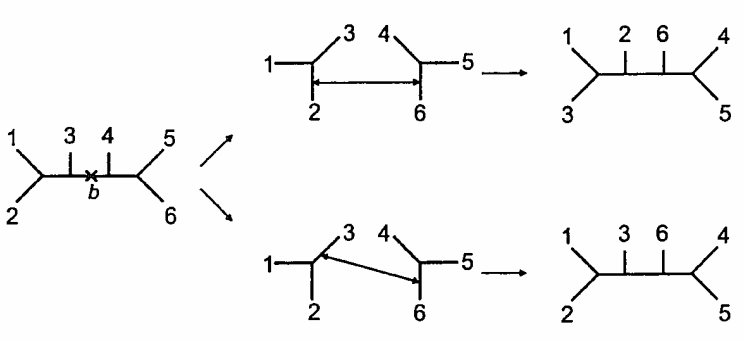
Rozmiar sąsiedztwa: 2*n*–6.

- **Subtree Prunning and Regrafting** (**SPR**): odrywamy krawędź wraz z całym poddrzewem zaczepiając je na środku innej krawędzi (ściągamy też powstały wierzchołek stopnia 2).



Rozmiar sąsiedztwa: 2(*n*–3)(2*n*–7).

- **Tree Bisection–Reconnection** (**TBR**): rozcinamy krawędź drzewa tworząc dwa osobne i scalamy je ponownie nową krawędzią łączącą parę krawędzi z obu drzew (również ściągamy wierzchołki stopnia 2).



Rozmiar sąsiedztwa: $\leq(2n-3)(n-3)^2$.

Miary odległości między topologiami drzew

Problem. Dwa drzewa nieukorzenione T i T' o tym samym zbiorze liści L nie są równoważne. Jak określić „stopień odmienności" ich topologii?

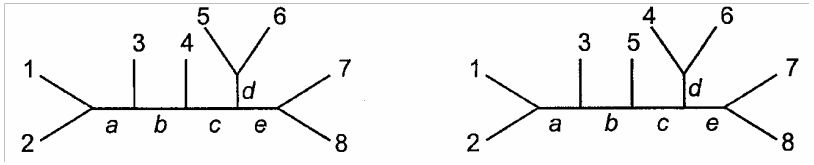
Dobrą metryką „międzydrzewową" jest liczba różnych występujących w nich rozbić:

$$\rho(T,T')=|\delta(T)\setminus\delta(T')|+|\delta(T')\setminus\delta(T)|=|\delta(T)|+|\delta(T')|-2|\delta(T)\cap\delta(T')|$$

W szczególności:

- $\rho(T,T')=0$ wtedy i tylko wtedy, gdy topologie są identyczne,
- maksymalna odległość między drzewami n–listnymi w tej metryce (średnica zbioru drzew) wynosi $2n-6$,
- otoczenie NNI dla binarnego T zawiera wszystkie drzewa binarne T' o odległości topologicznej $\rho(T,T')\leq 2$.

Przykład. Drzewa o odległości topologicznej $\rho(T,T')=4$.



Jeśli T i T' są binarne, wówczas można określać inne metryki: $\rho_{NNI}(T,T')$, $\rho_{SPR}(T,T')$, $\rho_{TBR}(T,T')$ jako minimalną liczbę przekształceń drzew (odpowiedniego typu) transformujących T w T'. Wyznaczenie niektórych z tych wielkości (np. ρ_{NNI}) jest NP–trudne. Zachodzą oszacowania:

$$\rho_{SPR}(T,T')/2 \leq \rho_{TBR}(T,T') \leq \rho_{SPR}(T,T') \leq \rho_{NNI}(T,T') \geq \rho(T,T')/2$$

Średnica zbioru drzew binarnych n–listnych wynosi $\Theta(n \log n)$ dla $\rho_{NNI}(T,T')$ oraz $\Theta(n)$ dla $\rho_{SPR}(T,T')$ i $\rho_{TBR}(T,T')$.