Konstruowanie drzew filogenetycznych – metody odległościowe

Algorytmy te opierają się na dostarczonej wraz ze zbiorem badanych gatunków L **macierzą odległości** (zwykle metryką) $d{:}L^2 \to R^+$ odzwierciedlającą ich „**dystans ewolucyjny**". Na jej podstawie konstruuje się drzewo filogenetyczne o liściach L.

Macierz d można odczytać porównując różne wersje tej samej (tzn. pochodzącej od wspólnego przodka) **sekwencji biologicznej** występującej u gatunków z L. Należy dokonać ich liniowego wielodopasowania, a następnie wyznaczyć między parami wierszy (odpowiadających gatunkom) stosowną „odległość", używając pewnego modelu ewolucji sekwencji (Jukes–Cantor, Kimura itp.).

Zalety:

- prostota i efektywność algorytmów,
- metody te „proponują" gotowe drzewo o konkretnej topologii (inne techniki wymagają przeszukania przestrzeni drzew, dostarczając jedynie sposobu oceny ich „wiarygodności"),
- oprócz drzewa uzyskujemy szacunkowe długości krawędzi tj. okresów samodzielnej ewolucji gatunków między momentami rozdzielenia się linii.

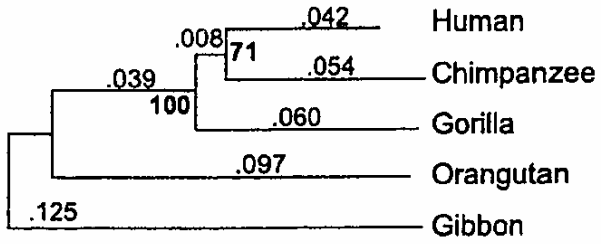
Wady:

- drzewa są wiarygodne tylko dla niewielkiej liczby liści (przy dużym $|L|$ produkt metody odległościowej można traktować jak „pierwsze przybliżenie", którego otoczenie należy dokładniej przebadać),
- trudność doboru właściwego modelu ewolucji i miary dystansu stosownie do rodzaju źródłowej sekwencji (białko, gen jądrowy, gen mitochondrialny itp.) oraz horyzontu czasowego odtwarzanej historii.

Przykład. Odległości (Kimura Distance) między pewnym wspólnym odcinkiem DNA mitochondrialnego dla pięciu gatunków ... małp.

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee	0.095 ± 0.011			
Gorilla	0.113 ± 0.012	0.118 ± 0.013		
Orangutan	0.183 ± 0.016	0.201 ± 0.018	0.195 ± 0.017	
Gibbon	0.212 ± 0.018	0.225 ± 0.019	0.225 ± 0.019	0.222 ± 0.018

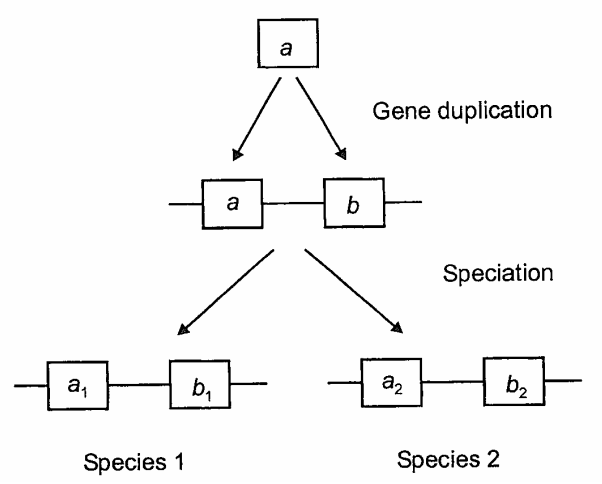
Uzyskane z nich drzewo filogenetyczne (odmiana metody NJ).



Dodatkowym problemem (wspólnym dla technik opartych na analizie „spokrewnionych sekwencji") jest właściwe określenie porównywanych miejsc w genomach różnych gatunków. Podobieństwo sekwencji nie wystarcza, gdyż mogą to być geny:

- **ortologiczne** – pochodzące od wspólnego przodka (właśnie tego szukamy),
- **paralogiczne** – powstałe w wyniku duplikacji genu u przodka,
- **ksenologiczne** – efekt np. międzygatunkowego transferu horyzontalnego materiału genetycznego „w poprzek" genealogii (rzadkość u organizmów wyższych, ale bardzo częste np. u bakterii).

Przykład. Gen *a* uległ duplikacji przed rozdzieleniem się linii gatunkowych. Współczesny badacz rekonstruując filogenezę dysponuje genomami gatunków 1 i 2. Nie wie jednak, czy należy porównywać sekwencję a_1 z a_2 i b_1 z b_2, czy np. a_1 z b_2.



Bootstrap Test

Prosta metoda oceny wiarygodności drzewa T (a także konkretnych jego gałęzi) uzyskanego z wielodopasowania (l–kolumnowego) sekwencji pochodzących od badanych gatunków. Losujemy z powtórzeniami l kolumn z wielodopasowania, uzyskując nowe „dopasowanie" pewnych przypadkowych sekwencji. Na jego podstawie tą samą metodą budujemy drzewo T_1. Procedurę powtarzamy kilkaset razy. Dla każdej krawędzi z T sprawdzamy w ilu procentach drzew T_1, T_2, T_3... pojawia się odpowiadające jej rozbicie. Przyjmuje się, że krawędzie o wartości testu bootstrap powyżej 95% są wiarygodne.

Drzewa ultrametryczne

Wyidealizowany model ewolucji oparty na kontrowersyjnej hipotezie **Zegara Molekularnego** – w osobnych liniach gatunkowych zachodzą zmiany ewolucyjne z jednakową szybkością.
Konsekwencja – dwa gatunki potomne „w jednakowym stopniu" zmieniły się od czasu ich oddzielenia się od wspólnego przodka.

Drzewem ultrametrycznym dla gatunków L i macierzy odległości d nazywamy ukorzenione drzewo filogenetyczne T z $L(T)=L$ o **wierzchołkach wewnętrznych** etykietowanych liczbami $t:V(T)\backslash L(T)\rightarrow R^{+}\backslash\{0\}$, spełniające warunki:

- dla każdej ścieżki prowadzącej od korzenia do liści etykiety wierzchołków wewnętrznych maleją,
- dla każdej pary liści $u\neq v\in L$ ich najbliższy wspólny przodek x ma etykietę $t(x)=d(u,v)$.

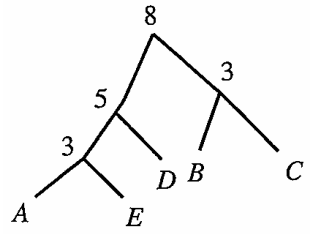
Idea: etykiety wierzchołków–przodków mają przedstawiać dystans jaki minął od chwili rozdzielenia się gatunku na linie potomne.

Macierz ultrametryczna – macierz odległości, dla której istnieje drzewo ultrametryczne.

Przykład. Poniższa macierz odległości pięciu gatunków jest ultrametryczna (korzystając z symetrii wypisano jedynie połowę):

	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

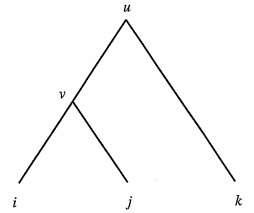
Drzewo ultrametryczne:



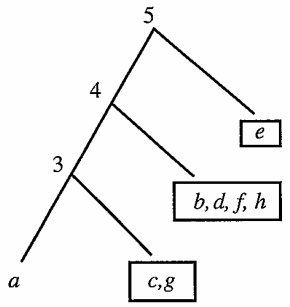
Macierze ultrametryczne występują stosunkowo rzadko.

Twierdzenie. Macierz odległości d jest ultrametryczna wtedy i tylko wtedy, gdy dla każdej trójki gatunków i,j,k wśród liczb $d(i,j)$, $d(i,k)$, $d(j,k)$ wartość największą spośród nich przyjmują co najmniej dwie (tzw. **warunek trzech punktów**).

Dowód. $\Rightarrow$ Wystarczy spojrzeć na fragment drzewa indukowany przez najkrótsze ścieżki łączące te liście:



$\Leftarrow$ Indukcja względem liczby liści. Dla $n=|L|>1$ wybieramy $a\in L$ i kreślimy ścieżkę prowadzącą do korzenia hipotetycznego drzewa. Pozostałe liście $b\in L$ dzielimy na klasy równoważności według rosnącej wartości odległości $d(a,b)$ – klasy „zawieszamy” na kolejnych wierzchołkach–przodkach umieszczanych na ścieżce:



Z założenia indukcyjnego wszystkie te klasy można ułożyć w poddrzewa ultrametryczne zawieszone na wierzchołkach–przodkach – otrzymujemy całe drzewo T.

Z warunku trzech punktów wynika, że:

- etykiety na ścieżkach od korzenia do liści nie rosną,
- drugi warunek def. drzewa ultrametrycznego jest spełniony także dla pary liści z różnych klas.

Ewentualnie ściągając „krawędzie zerowe” uzyskamy drzewo ultrametryczne. ■

Wniosek. Macierz ultrametryczna jednoznacznie wyznacza swoje drzewo ultrametryczne, które można odtworzyć w czasie wielomianowym.

Prosta metoda budowy drzewa ultrametrycznego.

Algorytm UPGMA
(ang. Unweighted Pair Group Method using Arithmetic averages)
Dane: macierz ultrametryczna d dla zbioru L.

Procedura działa w pętli:

- w każdym przebiegu L jest rozbity na sumę rozłącznych klastrów,
- każdemu klastrowi C przypisano drzewo ultrametryczne T_C o $L(T_C)=C$ i etykietach zgodnych z macierzą d obciętą do C,
- algorytm rozpoczyna od $|L|$ klastrów jednoelementowych, a kończy z chwilą uzyskania jednego klastra L,
- w każdym przebiegu dwa klastry C_a,C_b są łączone (i usuwane) w jeden klaster C, a jego drzewo T_c tworzymy wprowadzając nowy korzeń – wspólnego przodka dla T_a i T_b.
  - korzeń T_c otrzymuje etykietę równą średniej odległości pary liści ze scalanych klastrów:

$$d(C_a,C_b)=[\Sigma_{a\in C_a,\, b\in C_b} d(a,b)]/|C_a||C_b|$$

  - strategia wyboru pary scalanych klastrów: weź dwa „najbliższe” tj. o najmniejszym $d(C_a,C_b)$.

Niestety w praktyce rzadko spotykamy się z danymi choćby w przybliżeniu ultrametrycznymi.

Drzewa addytywne

Drzewem addytywnym dla gatunków L i macierzy odległości d nazywamy nieukorzenione drzewo filogenetyczne T z $L(T)=L$ o krawędziach etykietowanych liczbami $D:E(T)\rightarrow R^+$, spełniające warunek:

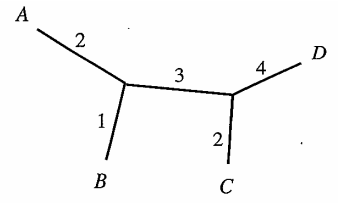
- dla każdej pary liści $u\neq v\in L$ suma etykiet krawędzi łączącej je prostej ścieżki wynosi $d(u,v)$,
- etykietę 0 mogą mieć tylko niesąsiednie krawędzie wiszące.

Idea: etykiety krawędzi przedstawiają „dystans ewolucyjny” przebyty przez gatunek pomiędzy rozwidleniami drzewa filogenetycznego (wierzchołki). Nie przyjmujemy hipotezy zegara molekularnego – odcinki łączące współczesne gatunki ze wspólnym przodkiem mogą mieć różną „ilość zmian”.

Macierz addytywna – macierz odległości, dla której istnieje drzewo addytywne.

Przykład. Poniższa (symetryczna) metryka dla czterech gatunków i jej drzewo addytywne.

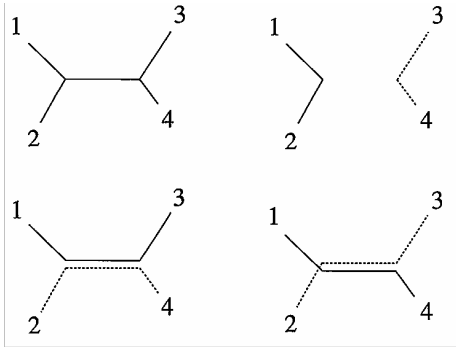
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Jest to model ogólniejszy od poprzedniego: każde drzewo ultrametryczne można zmienić w addytywne (krawędziom przypisujemy połowę odległości etykiet ich końców), lecz odwrotnie – nie zawsze.

Twierdzenie. Macierz odległości *d* jest addytywna wtedy i tylko wtedy, gdy dla każdej czwórki gatunków *i*,*j*,*k*,*l* wśród liczb $d(i,j)+d(k,l)$, $d(i,k)+d(j,l)$, $d(i,l)+d(j,k)$ wartość największą spośród nich przyjmują co najmniej dwie (tzw. **warunek czterech punktów**).

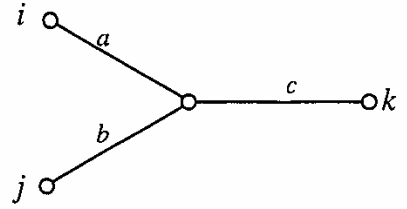
Połowa dowodu. $\Rightarrow$ Rysunek przestawia poddrzewo indukowane ścieżkami łączącymi liście 1,2,3,4 oraz odcinki, których długości są zliczane w wyrażeniach z warunku czterech punktów.



■

Problem. Jak konstruować drzewo addytywne macierzy odległości *d*?

Dla trzech liści rozwiązanie jest jednoznaczne.



$d(i,j)=D(a)+D(b)$, $d(i,k)= D(a)+D(c)$, $d(j,k)=D(b)+D(c)$, więc:

$$D(a)=[d(i,j)+d(i,k)– d(j,k)]/2,$$
$$D(b)=[d(i,j)+d(j,k)– d(i,k)]/2,$$
$$D(c)=[d(i,k)+d(j,k)– d(i,j)]/2,$$

Dla $|L|>3$:

Algorytm Neighbour–Joining (**NJ**):

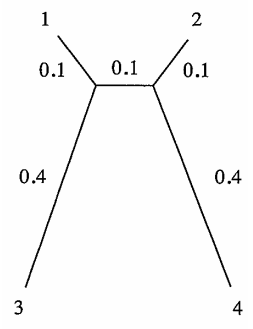
while $|L|>3$ **do begin**

1. znajdź parę $i,j \in L$ **<u>sąsiednich</u>** liści tzn. takich, które muszą mieć wspólnego sąsiada $k \in V(T) \setminus L$ (*jak???*)
2. $L := L \cup \{k\} \setminus \{i,j\}$
3. uaktualnienie d: wyznacz odległości $d(k,l)$ dla $l \in L \setminus \{k\}$ (*jak???*)
4. połącz krawędziami k z i,j
5. wyznacz ich długości $D(\{k,i\})$, $D(\{k,j\})$ (*jak???*)

end;
obsłuż „małe drzewo”;

Ad 1. Metoda „wyboru najbliższych” (jak w UPGMA) jest błędna.

Przykład.



Można wybrać dowolny liść a, a następnie znaleźć parę innych liści i,j maksymalizującą „odległość od rozgałęzienia” $(d(a,i)+d(a,j)-d(i,j))/2$. Wtedy i,j sąsiadują. Czas $O(n^2)$.

Twierdzenie (Studier & Keppler). Dla liści $a \in L$ drzewa addytywnego definiujemy:

$$r_a = [\Sigma_{b \in L}\, d(a,b)]/(|L|-2)$$

Wtedy para liści i,j dla których wartość $d(i,j)-(r_i+r_j)$ jest najmniejsza to liście sąsiednie.

Ad 3. Z definicji drzewa addytywnego musi zachodzić

$$d(k,l) = [d(i,l) + d(j,l) - d(i,j)]/2$$

Ad 5. Posługując się dowolnym innym liściem l uzyskamy:

$$d(i,k) = [d(i,j) + d(i,l) - d(j,l)]/2$$

Dla zwiększenia precyzji zwykle jednak przyjmuje się formułę „uśrednioną” po l:

$$d(i,k) = [d(i,j) + r_i - r_j]/2$$

oraz oczywiście $d(j,k) = d(i,j) - d(i,k)$.

Wniosek. Macierz addytywna jednoznacznie wyznacza swoje drzewo addytywne, które można odtworzyć w czasie wielomianowym.

Metoda „najmniejszych kwadratów”

Rozważamy układ równań liniowych

$$y_1=a_{11}\,x_1+a_{12}\,x_2+\ldots+a_{1n}\,x_n$$
$$y_2=a_{21}\,x_1+a_{22}\,x_2+\ldots+a_{2n}\,x_n$$
$$\ldots$$
$$y_m=a_{m1}\,x_1+a_{m2}\,x_2+\ldots+a_{mn}\,x_n$$

gdzie $m>n$. Macierzowo:

$$Y=AX$$

gdzie Y, X to wektory kolumnowe o kolejnych elementach $y_1,\ldots,y_m$ i $x_1,\ldots,x_n$. Zakładamy, że rząd macierzy współczynników jest maksymalny **rz(A)=n**, czyli kolumny z A rozpinają n–wymiarową podprzestrzeń w R^m, a przekształcenie $X\to Y=AX$ jest różnowartościowe (lecz oczywiście nie jest „na”). Nie dla wszystkich $y_1,\ldots,y_m$ istnieje rozwiązanie układu równań. Możemy jednak szukać rozwiązań przybliżonych próbując zminimalizować błąd.

Problem. Dla wektora – kolumny Y (wymiar m) i macierzy A ($m\times n$) znajdź wektor – kolumnę X (wymiar n), dla którego wektor błędu rozwiązania X:

$$E=Y-AX$$

ma możliwie najmniejszą długość $\|E\|^2$.

Inaczej: nawet jeżeli ścisłe rozwiązanie nie istnieje, ciąg liczb:

$$y_1'=a_{11}\,x_1+a_{12}\,x_2+\ldots+a_{1n}\,x_n$$
$$y_2'=a_{21}\,x_1+a_{22}\,x_2+\ldots+a_{2n}\,x_n$$
$$\ldots$$
$$y_m'=a_{m1}\,x_1+a_{m2}\,x_2+\ldots+a_{mn}\,x_n$$

ma znaleźć się tak blisko $y_1,\ldots,y_m$ jak to tylko możliwe, przy czym błąd określamy wg metryki pitagorejskiej:

$$\|E\|^2=(y_1-y_1')^2+(y_2-y_2')^2+\ldots+(y_m-y_m')^2$$

Rozwiązanie. Błąd minimalizuje wektor $X=(A^TA)^{-1}A^TY$.

Dowód. Istnienie $(A^TA)^{-1}$: gdyby $\det(A^TA)=0$, istniałby wektor niezerowy $v\in R^n$, taki że $A^TAv=0$, stąd $0=v^TA^TAv=(Av)^TAv=\|Av\|^2$, czyli $Av=0$ – sprzeczność gdyż $v\to Av$ jest różnowartościowe.

Dalej rozpiszemy $\|Y-AX'\|^2$ dla argumentu $X'=X-u$.

$$\|Y-AX'\|^2=\|Y-A[(A^TA)^{-1}A^TY-u]\|^2=\|Y-A(A^TA)^{-1}A^TY+Au\|^2=$$
$$=[Y^T-Y^TA(A^TA)^{-1}A^T+u^TA^T][Y-A(A^TA)^{-1}A^TY+Au]=$$
$$=Y^TY-Y^TA(A^TA)^{-1}A^TY+Y^TAu-$$
$$-Y^TA(A^TA)^{-1}A^TY+Y^TA(A^TA)^{-1}A^TA(A^TA)^{-1}A^TY-Y^TA(A^TA)^{-1}A^TAu+$$
$$+u^TA^TY-u^TA^TA(A^TA)^{-1}A^TY+u^TA^TAu=$$
$$=Y^TY-Y^TA(A^TA)^{-1}A^TY+u^TA^TAu=$$
$$=Y^TY-Y^T[A(A^TA)^{-1}A^T]^TA(A^TA)^{-1}A^TY+u^TA^TAu=$$
$$=\|Y\|^2-\|A(A^TA)^{-1}A^TY\|^2+\|Au\|^2.$$

Minimalny $\|Y-AX'\|^2$ wynosi więc $\|Y\|^2-\|A(A^TA)^{-1}A^TY\|^2$ dla $X=(A^TA)^{-1}A^TY$, a ew. zmiana X o $u\neq 0$ zwiększa ten błąd (składnik $\|Au\|^2$). ■

Rozważmy macierz odległości $d_{wzór}$ dla zbioru gatunków L. Zazwyczaj nie spełnia ona ściśle warunku addytywności.

Metoda **Least Squares** stawia sobie za cel znalezienie drzewa addytywnego T_{LS} dla L, którego poetykietowanie krawędzi $D:E(T_{LS})\rightarrow R^+$ minimalizuje **kwadratowy błąd** odtworzenia pierwotnej macierzy odległości $d_{wzór}$:

$$Err^2=\Sigma_{\{u,v\}\subseteq L,\, u\neq v}\,(d_{wzór}(u,v)-d_{T_{LS}}(u,v))^2$$

T_{LS} jest wynikowym drzewem filogenetycznym metody LS.

Pojawiają się dwa zagadnienia:

- Jak wyszukać w przestrzeni różnych topologii drzewo minimalizujące wartość funkcji błędu Err^2? Można zastosować jedną z metaheurystyk, o ile wcześniej umiemy ...
- dla ustalonego drzewa filogenetycznego T znaleźć etykiety (długości) krawędzi $D:E(T)\rightarrow R^+$ minimalizujące kwadratowy błąd odtworzenia metryki $d_{wzór}$ na L.

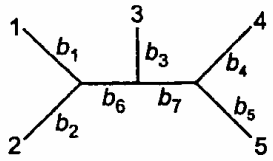
Drugi problem sprowadza się do rozważanej wcześniej minimalizacji błędu kwadratowego dla układu równań liniowych.

- poszukiwane długości krawędzi tworzą $|E(T)|$–elementowy wektor niewiadomych b,
- $|L|(|L|-1)/2$ odległości między różnymi parami liści tworzy znany wektor d („przemieszczone” wartości z górnej połowy macierzy $d_{wzór}$),
- odległości między parami liści w drzewie addytywnym są funkcjami liniowymi (sumami) długości krawędzi; zero–jedynkową macierz przekształcającą

  etykiety krawędzi→odległości liści

  oznaczymy przez A,
- szukamy wektora b, dla którego $d'=Ab$ jest możliwie najbliższy d, wiemy już że jest to $\boldsymbol{b=(A^TA)^{-1}A^Td}$.

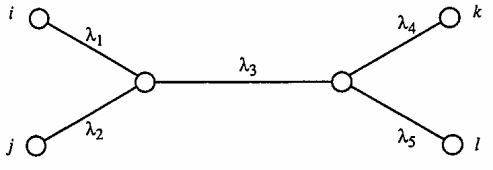
Przykład. Etykiety krawędzi 5–listnego drzewa T ustawiamy w wektor kolumnowy $b^T=(b_1,b_2,b_3,b_4,b_5,b_6,b_7)$



Odległości między liśćmi również kodujemy kolumną $d^T=(d_{wzór}(1,2), d_{wzór}(1,3), d_{wzór}(1,4), d_{wzór}(1,5), d_{wzór}(2,3), d_{wzór}(2,4), d_{wzór}(2,5), d_{wzór}(3,4), d_{wzór}(3,5), d_{wzór}(4,5))$.
Wtedy macierz A z równania $d=Ab$ ma postać:

$$\mathbf{A} = \begin{bmatrix} 1&1&0&0&0&0 \\ 1&0&1&0&0&1&0 \\ 1&0&0&1&0&1&1 \\ 1&0&0&0&1&1&1 \\ 0&1&1&0&0&1&0 \\ 0&1&0&1&0&1&1 \\ 0&1&0&0&1&1&1 \\ 0&0&1&1&0&0&1 \\ 0&0&1&0&1&0&1 \\ 0&0&0&1&1&0&0 \end{bmatrix}$$

Poprawność procedury. Czy macierz A spełnia założenia metody tzn. ma maksymalny możliwy rząd $rz(A)=|E(T)|$?
Tak, gdyż przekształcenie długości krawędzi (wektor b) $\rightarrow$ odległości liści (wektor d) jest odwracalne, znając dystans między liśćmi możemy wyznaczyć etykietę każdej krawędzi wewnętrznej np. tak:



$\lambda_3=[d(i,k)+d(j,l)-d(i,j)-d(k,l)]/2$. ■

Aproksymacja macierzy odległości metrykami addytywnymi

Metoda Least Squares odpowiada szczególnemu przypadkowi ogólniejszej rodziny problemów optymalizacyjnych. Niech dana będzie metryka $d_{wzór}$ dla zbioru gatunków L. Szukamy drzewa addytywnego lub ultrametrycznego T (z $L(T)=L$), takiego by jego metryka d_T (określona jak w definicji drzewa addytywnego) minimalizowała błąd:

- $l^1=\Sigma_{\{u,v\}\subseteq L,\, u\neq v} |d_{wzór}(u,v)-d_T(u,v)|$,

lub

- $l^2=\Sigma_{\{u,v\}\subseteq L,\, u\neq v} (d_{wzór}(u,v)-d_T(u,v))^2$,

lub

- $l^\infty=\max_{\{u,v\}\subseteq L,\, u\neq v} |d_{wzór}(u,v)-d_T(u,v)|$.

Zagadnienie minimalizacji l^∞, w którym T ma być ultrametryczne jest wielomianowe. Wszystkie pozostałe problemy są NP–trudne. Jeżeli topologia drzwa T jest ustalona, a optymalizacji podlegają jedynie nieujemne wagi jego krawędzi, wówczas dla każdego z powyższych problemów znany jest efektywny algorytm dokładny.