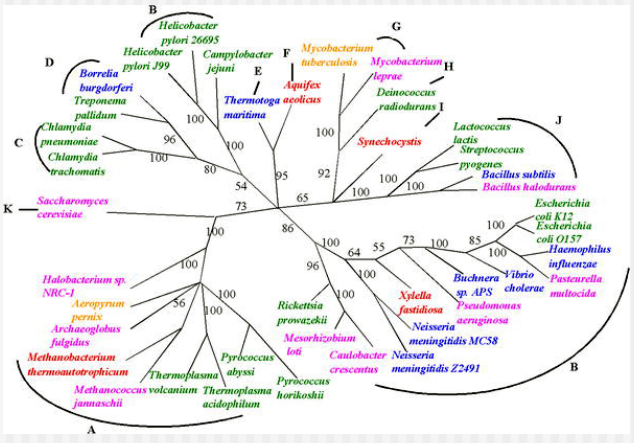
Algorytmy kombinatoryczne w bioinformatyce

wykład 7: drzewa filogenetyczne

prof. dr hab. inż. **Marta Kasprzak**
Instytut Informatyki, Politechnika Poznańska

Drzewa filogenetyczne

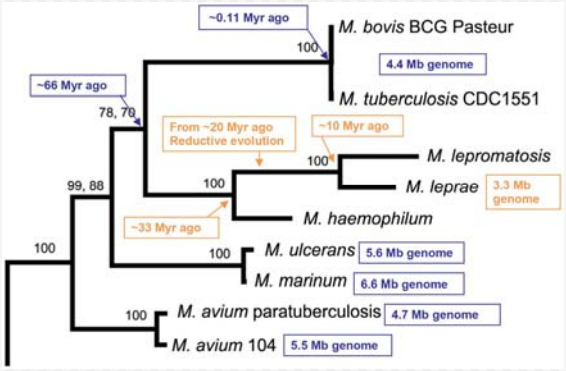
Drzewa filogenetyczne odzwierciedlają powiązania ewolucyjne pomiędzy organizmami reprezentowanymi przez liście



[F. Ge i in., *PLOS Biol.* 3 (2005)]

Drzewa filogenetyczne

Drzewo ukorzenione opisuje hipotetyczną historię ewolucji, z korzeniem odpowiadającym wspólnemu przodkowi



[X. Han i F. Silva, *PLOS Negl. Trop. Dis.* 8 (2014)]

Drzewa dla macierzy znakowych

- Dwa rodzaje danych: macierze znakowe i numeryczne
- Dane znakowe reprezentują cechy porównywanych obiektów o skończonej liczbie stanów (np. kształt płetwy, liczba zębów)

Table 2 Matrix of morphological character states in *Ctenogobiops* species

Species	D	DF	FIL	PEC	GO	LS	CP	LR
C. aurocingulus	11	1st	N	Yellow/white blotch	S	45–58	Spots	3
C. crocineus	11	2nd	N	White streak	L	52–56	Streak	4
C. feroculus	11	1st/2nd	N	Yellow/white dash	S	50–58	Streak	3
C. formosa	11	2nd/3rd	N	Pink/orange blotch	S	45–49	Streak	4
C. maculosus	11	2nd	N	White streak	L	52–56	Streak	4
C. mitodes	11	2nd	Y	White blotch	S	43–52	Streak	4
C. phaeostictus	13	3rd	N	Absent	S	49	Spots	4
C. pomastictus	11	2nd	N	White blotch	S	55–59	Spots	4
C. tangaroai	11	2nd	Y	White streak	L	47–51	Streak	4
C. tongaensis	11	2nd	Y	White blotch	S	50–51	Spots	4

Characters are as described in Lubbock and Polunin (1977) and Randall et al. (2003, 2007)

D dorsal fin ray count, *DF* longest dorsal spine, *FIL* presence of dorsal filament, *PEC* pectoral fin pigment, *GO* gill opening long (*L*) or short (*S*), *LS* longitudinal series scale count, *CP* cheek pigment, specifically whether the anterior portion of the ventralmost cheek pigment row is present as a streak or as distinct spots, *LR* number of lateral rows of blotches

[C. Thacker i in., *Ichthyol. Res.* 57 (2010)]

Drzewa dla macierzy znakowych

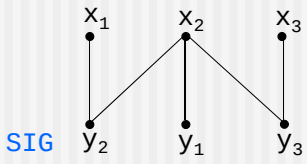
- Macierze znakowe mają *n* wierszy odpowiadających obiektom i *m* kolumn odpowiadających ich cechom. Pola macierzy zawierają stany cech przypisane danym obiektom
- Idealne drzewo filogenetyczne utworzone na podstawie macierzy znakowej powinno być pozbawione przejawów ewolucji równoległej i wstecznej. Innymi słowy, dany stan cechy przypisany jest do węzłów tworzących spójne poddrzewo
- Stany poszczególnych cech mogą być uporządkowane lub nie
- Problem konstrukcji idealnego drzewa filogenetycznego jest w ogólności NP-trudny. Łatwe obliczeniowo są jego warianty z ograniczoną liczbą stanów lub ograniczoną liczbą cech

Drzewa dla macierzy znakowych

[G.F. Estabrook i F.R. McMorris, *Journal of Mathematical Biology* 4 (1977)]

- Problem konstrukcji idealnego drzewa filogenetycznego dla danych znakowych i dwóch rozpatrywanych cech
- Na podstawie macierzy konstruowany jest graf przecięcia stanów (`SIG`)
- Macierz stanów umożliwia konstrukcję drzewa idealnego, jeśli `SIG` jest acykliczny

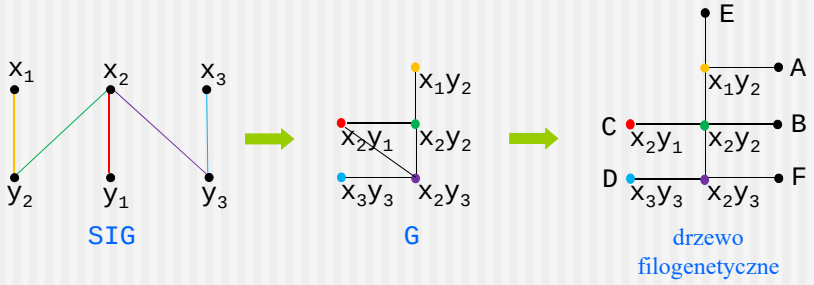
obiekty	cechy x	cechy y
A	x_1	y_2
B	x_2	y_2
C	x_2	y_1
D	x_3	y_3
E	x_1	y_2
F	x_2	y_3



Drzewa dla macierzy znakowych

[G.F. Estabrook i F.R. McMorris, *J. Math. Biol.* 4 (1977)] – cd.

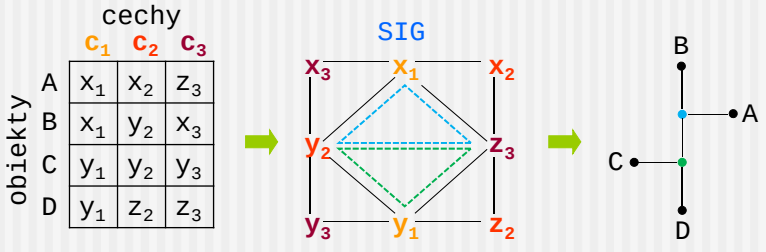
- Algorytm przekształca `SIG` w jego graf liniowy `G`, którego drzewo rozpinające stanowi szkielet wynikowego drzewa filogenetycznego. Pod węzły `G` podpinane są obiekty posiadające odpowiednią parę stanów



Drzewa dla macierzy znakowych

[S. Kannan i T. Warnow, *SIAM J. Discrete Math.* 5 (1992)]

- Problem konstrukcji idealnego drzewa filogenetycznego dla danych znakowych i trzech rozpatrywanych cech
- Macierz stanów umożliwia konstrukcję drzewa idealnego ⇔ graf przecięcia stanów (`SIG`) może być *c-triangulowany*



Drzewa dla macierzy znakowych

- Zazwyczaj macierz opisująca obiekty uniemożliwia skonstruowanie drzewa idealnego. W takich przypadkach dąży się do konstrukcji drzewa optymalnego pod względem pewnego przyjętego kryterium
- Najczęściej stosowane kryteria:
  - oszczędności, gdy minimalizowana jest liczba wystąpień przypadków ewolucji równoległej i wstecznej
  - kompatybilności, gdy minimalizowana jest liczba cech (kolumn) usuwanych z macierzy w celu doprowadzenia do postaci idealnej
- Powyższe problemy optymalizacyjne są trudne obliczeniowo

Drzewa dla macierzy numerycznych

- Macierze numeryczne mają n wierszy i n kolumn odpowiadających obiektom. Ich pola zawierają odległości ewolucyjne pomiędzy obiektami (np. różnice w sekwencjach DNA), które są poddawane operacjom arytmetycznym
- Idealne drzewo filogenetyczne utworzone na podstawie macierzy numerycznej ma wszystkie obiekty w liściach, a krawędzie opisane są takimi odległościami, że suma wag krawędzi ścieżki dla każdej pary obiektów równa się ich odległości z macierzy
- O macierzy umożliwiającej konstrukcję takiego drzewa idealnego mówimy, że jest addytywna. Algorytm konstruujący drzewo dla macierzy addytywnej ma złożoność wielomianową

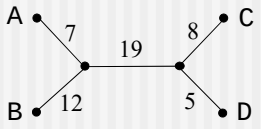
Drzewa dla macierzy numerycznych

- Macierz numeryczna jest *addytywna* wtedy i tylko wtedy, gdy każde cztery jej obiekty można zaetykietować i, j, k, l w taki sposób, że

$$d_{ij} + d_{kl} = d_{ik} + d_{jl} \geq d_{il} + d_{jk}$$

- Przykład macierzy addytywnej i drzewa idealnego

	A	B	C	D
A	0	19	34	31
B	19	0	39	36
C	34	39	0	13
D	31	36	13	0



Drzewa dla macierzy numerycznych

- Podobnie jak w przypadku macierzy znakowej, poszukiwanie rozwiązania jak najbliższego idealnemu dla macierzy numerycznej z błędami jest problemem NP-trudnym
- Niedokładność pomiaru odległości może niekiedy zostać wyrażona jako interwał określający dolne i górne ograniczenie wartości. Problem konstrukcji drzewa idealnego na podstawie dwóch macierzy z dolnymi i górnymi ograniczeniami jest łatwy obliczeniowo

Odległość filogenetyczna

- Odległość filogenetyczna pomiędzy organizmami mierzona jest zwykle wartością dopasowania sekwencji z nich pochodzących
- To podejście może nie dać dobrych rezultatów w przypadkach rearanżacji genomów, innych zaburzeń w ciągłości informacji genetycznej lub gdy sekwencje różnią się znacznie długością
- Jednym z alternatywnych sposobów wyznaczania odległości jest analiza profili zawartości *k*-merów w sekwencjach
- Inna metoda pomija większość obecnej w sekwencjach informacji i opiera się na faktoryzacji metodą LZ (Lempel-Ziv)

Odległość filogenetyczna

[H.H. Otu i K. Sayood, *Bioinformatics* 19 (2003)]

- W tym podejściu zaproponowano nowe miary odległości filogenetycznej wywiedzione z podobieństwa sekwencji po faktoryzacji metodą LZ
- Wykorzystano w nich wartości *złożoności LZ*, zdefiniowanej jako liczba komponentów będących rezultatem faktoryzacji
- Wartość każdej z miar zależy jedynie od złożoności LZ obliczanej dla porównywanych sekwencji oraz dla ich konkatenacji. Nie bierze się tu pod uwagę podobieństwa zbiorów fragmentów po faktoryzacji obu sekwencji. Oparto się na obserwacji mówiącej, że liczba komponentów jest znacznie niższa niż przeciętna, gdy zestawiane sekwencje są podobne

Odległość filogenetyczna

[H.H. Otu i K. Sayood, *Bioinformatics* 19 (2003)] – cd.

- Przykład

S=ATCACAGTA, fact(***S***)=A·T·C·AC·AG·TA, c(***S***)=6

Q=AGACTACT, fact(***Q***)=A·G·AC·T·ACT, c(***Q***)=5

SQ=ATCACAGTAAGACTACT
fact(***SQ***)=A·T·C·AC·AG·TA·AGA·CT·ACT, c(***SQ***)=9

QS=AGACTACTATCACAGTA
fact(***QS***)=A·G·AC·T·ACT·AT·C·ACA·GT·A, c(***QS***)=10

Odległość filogenetyczna

[H.H. Otu i K. Sayood, *Bioinformatics* 19 (2003)] – cd.

- Miara I
  $d_1(S,Q) = \max\{ c(SQ) - c(S), c(QS) - c(Q) \}$
- Miara II
  $d_2(S,Q) = \max\{ c(SQ) - c(S), c(QS) - c(Q) \} / \max\{ c(S), c(Q) \}$
- Miara III
  $d_3(S,Q) = c(SQ) - c(S) + c(QS) - c(Q)$
- Miara IV
  $d_4(S,Q) = (c(SQ) - c(S) + c(QS) - c(Q)) / c(SQ)$
- Miara V
  $d_5(S,Q) = (c(SQ) - c(S) + c(QS) - c(Q)) / \frac{1}{2} (c(SQ) + c(QS))$

Odległość filogenetyczna

[H.H. Otu i K. Sayood, *Bioinformatics* 19 (2003)] – cd.

- Przykładowo, dla trzech sekwencji:
  S=ATCACAGTA, ***Q***=AGACTACT, ***R***=TCACTAC,
  ich odległość mierzona każdą z miar jest następująca:

$d_1(S,Q) = 5$,	$d_1(Q,R) = 3$,	$d_1(R,S) = 4$,
$d_2(S,Q) = 0.83$,	$d_2(Q,R) = 0.6$,	$d_2(R,S) = 0.67$,
$d_3(S,Q) = 8$,	$d_3(Q,R) = 6$,	$d_3(R,S) = 7$,
$d_4(S,Q) = 0.89$,	$d_4(Q,R) = 0.75$,	$d_4(R,S) = 0.78$,
$d_5(S,Q) = 0.84$,	$d_5(Q,R) = 0.75$,	$d_5(R,S) = 0.78$.

  Wszystkie miary dają tutaj taką samą relację w zbiorze.

Drzewa konsensusowe

- Różne metody konstrukcji drzew filogenetycznych mogą zwrócić różne rezultaty dla tej samej macierzy wejściowej. Wybór właściwego drzewa stanowi złożony problem
- Często stosuje się podejście generowania jednej struktury dla wszystkich uzyskanych rezultatów, zgodnej w najwyższym stopniu ze strukturami częściowymi. Problem wyboru optymalnego drzewa konsensusowego jest w stosowanych sformułowaniach trudny obliczeniowo

Drzewa konsensusowe

- Wybrane reguły składania drzew konsensusowych:
  - ścisły konsensus — tylko poddrzewa współdzielone przez wszystkie drzewa na wejściu algorytmu są używane do zbudowania drzewa wyjściowego
  - reguła większościowa — brane pod uwagę są te poddrzewa, które występują w większości drzew wejściowych
  - reguła zachłanna — reguła większościowa uzupełniona o możliwość dodania do drzewa także innych poddrzew, rozważanych w kolejności malejącej częstości ich wystąpienia w zbiorze wejściowym

Drzewa konsensusowe

[C. Bonnard i in., *Systematic Biology* 55 (2006)]

- W przypadku rozbieżnych drzew metody konsensusowe mogą nie dać satysfakcjonującego wyniku. Zaproponowana metoda *multipolar consensus* generuje więcej niż jedno drzewo konsensusowe i różne drzewa składa z fragmentów niedających się pogodzić. Liczba drzew na wyjściu algorytmu jest minimalizowana, a do ich konstrukcji używa się wszystkie liczące się struktury obecne w zbiorze wejściowym (o liczbie wystąpień przekraczającej pewien założony próg odcięcia)
- Autorzy sprowadzają rozwiązywany problem do znanego problemu kolorowania wierzchołkowego grafu. Do jego rozwiązania stosują heurystykę zachłanną

Drzewa konsensusowe

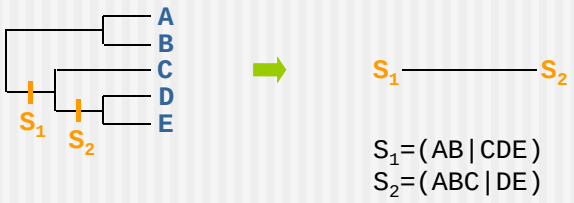
[C. Bonnard i in., *Systematic Biology* 55 (2006)] – cd.

- Każde z ukorzenionych drzew na wejściu algorytmu może zostać podzielone na dwa poddrzewa na wiele sposobów. Każde z tych cięć pociąga za sobą podział zbioru liści drzewa na dwa podzbiory (rozważane są jedynie podzbiory od dwóch liści w górę)
- Takie same podzbiory można uzyskać z podziału różnych drzew wejściowych pomimo różnej topologii uzyskanych poddrzew. Podzbiory występujące częściej otrzymują wyższą wagę, równą liczbie wystąpień
- Podziały o wadze wyższej od założonego progu odcięcia biorą udział w tworzeniu tzw. grafu kompatybilności

Drzewa konsensusowe

[C. Bonnard i in., *Systematic Biology* 55 (2006)] – cd.

- W *grafie kompatybilności* wierzchołki odpowiadają podziałom zbioru obiektów (liści drzewa), nieskierowane krawędzie łączą wierzchołki, jeśli dane podziały nie są sprzeczne



Drzewa konsensusowe

[C. Bonnard i in., *Systematic Biology* 55 (2006)] – cd.

- Każde drzewo wejściowe to klika w utworzonym grafie. Każda klika w grafie to możliwe drzewo konsensusowe na wyjściu
- W zaproponowanej metodzie poszukiwany jest najmniej liczny zbiór klik pokrywający graf kompatybilności
- Wierzchołki połączone ze wszystkimi innymi wierzchołkami tworzą tzw. klikę-rdzeń (ang. *kernel clique*)
- Wierzchołki wchodzące w skład *kernel clique* wystąpią we wszystkich drzewach wyjściowych, pozostałe wierzchołki w dokładnie jednym drzewie

Drzewa konsensusowe

[C. Bonnard i in., *Systematic Biology* 55 (2006)] – cd.

Algorytm *multipolar consensus*:

- Konstrukcja grafu kompatybilności
- Wykrycie *kernel clique* i usunięcie tych wierzchołków z grafu
- Znalezienie zbioru klik pokrywającego pozostałą część grafu:
  - przekształcenie grafu w jego dopełnienie
  - rozwiązanie problemu minimalnego pokolorowania dopełnienia grafu
  - połączenie wierzchołków o tym samym kolorze w klikę
- Dodanie *kernel clique* do wszystkich znalezionych klik
- Utworzenie na ich podstawie drzew konsensusowych

Drzewa konsensusowe

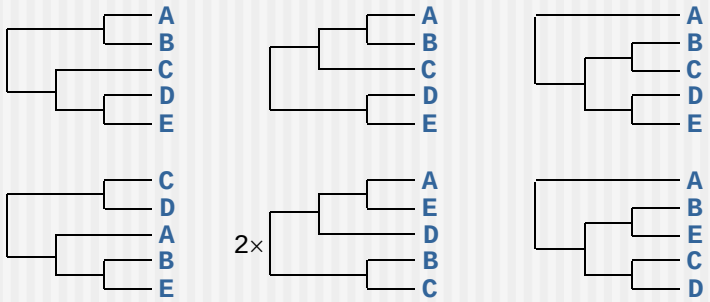
[C. Bonnard i in., *Systematic Biology* 55 (2006)] – cd.

- Problem minimalnego pokolorowania grafu polega na poprawnym pokolorowaniu wierzchołków grafu w taki sposób, że liczba użytych kolorów jest minimalna
- Problem ten, podobnie jak komplementarny do niego problem pokrycia grafu maksymalnymi klikami, jest silnie NP-trudny
- Do rozwiązania problemu użyta została heurystyka zachłanna:
  - wierzchołki grafu porządkowane są w kolejności malejącej wagi (liczby wystąpień danego podziału)
  - po kolei przypisywane są im kolory wg zasady, że indeks koloru ma być jak najniższy przy zachowaniu różnych kolorów wierzchołków sąsiadujących

Drzewa konsensusowe

Przykład działania metody Bonnard i in.

- Zbiór drzew wejściowych:



Drzewa konsensusowe

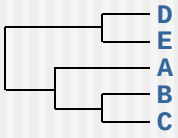
Przykład działania metody Bonnard i in.

- Lista podziałów wraz z liczbą ich wystąpień w zbiorze:

	A	B	C	D	E	#
S_1	0	0	0	1	1	3
S_2	0	1	1	0	0	3
S_3	1	1	0	0	0	2
S_4	0	1	0	0	1	2
S_5	0	0	1	1	0	2
S_6	1	0	0	0	1	2

Graf kompatybilności przy założeniu liczby wystąpień ≥ 3:

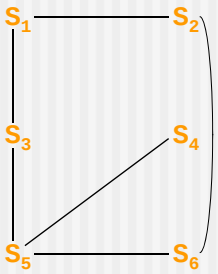
oraz przykładowe drzewo konsensusowe:

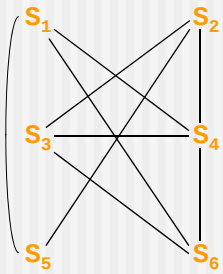


Drzewa konsensusowe

Przykład działania metody Bonnard i in.

- Graf kompatybilności i jego dopełnienie dla liczby wystąpień ≥ 2:

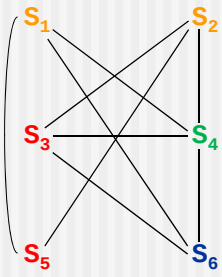




Drzewa konsensusowe

Przykład działania metody Bonnard i in.

- Rozwiązanie znalezione przez heurystykę zachłanną: *kernel clique* = ∅



Cztery kliki przekładające się na cztery drzewa konsensusowe:

S_1, S_2
S_3, S_5
S_4
S_6

Drzewa konsensusowe

- Rozwiązanie optymalne dla tej instancji: