Wykład V: Kompatybilność, wstęp do metod probabilistycznych

Anna Gambin
31.10.2002

1 Kompatybilność

Kolejnym kryterium optymalności rozważanym dla drzew filogenetycznych i pokrewnym parsymonii jest kompatybilność. Do tej pory umawialiśmy się, że w liściach drzewa znajdują się sekwencje molekularne, teraz potraktujemy problem bardziej "klasycznie". Przypomnijmy, że przed erą biologii molekularnej klasyfikacja gatunków opierała się na porównywaniu pewnych cech morfologicznych (takich jak np. budowa kośćca). Cechą nazwiemy dowolny skończony uporządkowany zbiór wartości (każdy rozważany gatunek posiada pewną wartość danej cechy). Dobierając odpowiedni zbiór cech mamy szansę odtworzyć historię ewolucji rozważanych gatunków. Podobnie jak w przypadku parsymonii podejście oparte o cechy pozwala nam na poznanie topologii drzewa, nie możemy natomiast powiedzieć nic sensownego o długości krawędzi (chyba że mamy dodatkowe dane pochodzące z wykopalisk). Rozważane w rekonstrukcji drzew cechy mogą również pochodzić z sekwencji genetycznych. Przykładem doskonałej cechy binarnej może być np. występowanie ustalonego słowa w DNA badanych gatunków. Zazwyczaj lepiej nadają się do tego niefunkcjonalne fragmenty DNA, o których zakłada się, że mutują w sposób losowy. Innym przykładem może być występowanie danego nukleotydu na danej pozycji w sekwencji DNA (nie jest to cecha binarna ale stanie się nią kiedy pogrupujemy puryny i pirymidyny).

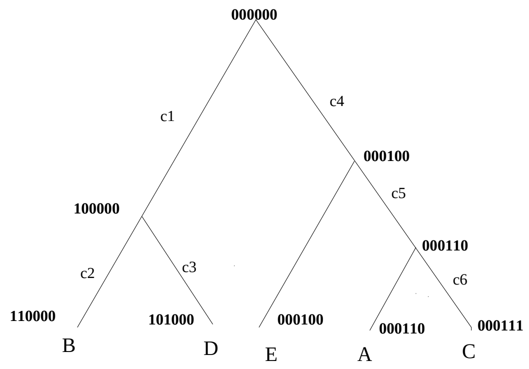
Zdefiniujemy teraz pojęcie kompatybilności cechy z drzewem. Będziemy poszukiwać drzew, które są kompatybilne z jak największą liczbą cech.

Definicja 1 *Cecha binarna jest kompatybilna z drzewem T jeśli istnieje takie etykietowanie wierzchołków wewnętrznych T wartościami cechy, że tylko jedna krawędź T łączy wierzchołki o różnych etykietach (mówimy, że mutacja nastąpiła tylko po tej krawędzi).*

Zauważmy, że liście rozważanego przez nas drzewa są już poetykietowane wartościami cechy. Z definicji wynika, że etykietowanie wierzchołków wewnętrznych drzewa musi grupować wierzchołki o tej samej etykiecie w pojedyncze spójne składowe drzewa. Taka sama własność musi zachodzić jeśli rozważymy cechę o więcej niż dwu wartościach.

Definicja 2 *k wartościowa cecha c jest kompatybilna z drzewem T o etykietowanych liściach jeśli istnieje etykietowanie wierzchołków wewnętrznych drzewa T wartościami c, takie że sumaryczna liczba zmian po krawędziach wynosi $k-1$.*

	c1	c2	c3	c4	c5	c6
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0



Rysunek 1: Kompatybilne drzewo dla 6 binarnych cech.

Dla zbioru m cech (przedstawianego jako macierz o wymiarach n x m gdzie n jest liczbą gatunków) interesuje nas pytanie: czy istnieje drzewo T o n etykietowanych liściach, takie że każda spośród m cech jest z nim kompatybilna (Rysunek 1 przedstawia drzewo kompatybilne ze zbiorem 6-ciu cech binarnych). Okazuje się, że problem ten jest w ogólności NP trudny, natomiast istnieje efektywny algorytm dla cech binarnych.

2 Algorytm dla cech binarnych

Omówimy algorytm działający w dwóch fazach: w pierwszej decyduje czy wejściowa macierz M (zbiór cech) ma kompatybilne drzewo i jeśli tak konstruuje je w drugiej fazie.

Przyjmiemy pewne upraszczające założenie (można go uniknąć — zastanów się w jaki sposób?) a mianowicie wartość 0 jest przodkiem wartości 1 w drzewie, tzn dla każdej krawędzi gdzie nastąpiła mutacja, jest to zawsze substytucja $0 \to 1$.

Definicja 3 *Dla każdej cechy j (kolumny w macierzy M) niech O_j oznacza zbiór wierszy (gatunków) dla których $M_{ij} = 1$.*

Zachodzi następujący fakt:

Twierdzenie 1 *Binarna macierz M ma kompatybilne drzewo wttw dla każdej pary cech i oraz j zbiory O_i i O_j są albo rozłączne, albo jeden zawiera się w drugim.*

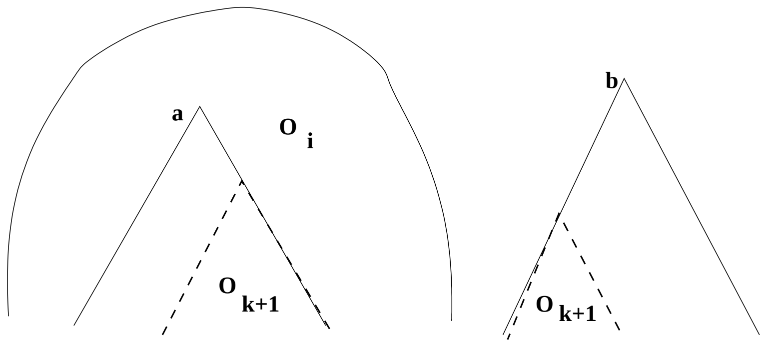
Warunek w twierdzeniu jest równoważny kompatybilności cech — inne sformułowanie twierdzenia powyżej brzmi:

> Zbiór cech binarnych jest kompatybilny wttw jest parami kompatybilny.

Załóżmy, że M ma kompatybilne drzewo, każdej spośród cech można przypisać jednoznacznie krawędź w drzewie, która odpowiada mutacji tej cechy i rozdziela drzewo na dwa poddrzewa: etykietowane 0 (powyżej) oraz etykietowane 1 (poniżej tej krawędzi). Załóżmy dalej, że istnieją 3 gatunki A, B i C, takie że $A, B \in O_i$, $C \notin O_i$, oraz $B, C \in O_j$, $A \notin O_j$. Oznacza to że A i B musza być we wspólnym poddrzewie (ze względu na cechę i) natomiast C nie należy do tego poddrzewa. Dodatkowo ze względu na cechę j B oraz C muszą być w tym samym poddrzewie: sprzeczność.

Załóżmy teraz że wszystkie pary kolumn M spełniają warunki twierdzenia. Pokażemy indukcyjnie jak zbudować kompatybilne z M drzewo. Łatwo widać jak skonstruować drzewo kompatybilne z jedną cechą: cecha ta indukuje podział na zbiorze gatunków, które trafiają do jednego z dwóch poddrzew.

Załóżmy, że zbudowaliśmy już drzewo T_k dla k cech i chcemy przekształcić je w drzewo dla $k+1$ cech. Rozważmy ojców liści z drzewa T_k i każdego z nich zastąpmy przez zbiór jego synów. Cecha $k+1$ indukuje kolejny podział na zbiorze gatunków (liści). Podział ten może rozszerzyć drzewo T_k do drzewa T_{k+1} o ile grupuje liście mające wspólnego ojca w T_k. Jeśli nie to będziemy



Rysunek 2: $O_i \cap O_{k+1} \neq \emptyset$, ale nie zachodzi $O_i \subseteq O_{k+1}$, ani $O_{k+1} \subseteq O_i$

zmuszeni etykietować dwie krawędzie cechą $k+1$ czyli nie będzie ona kompatybilna. Załóżmy, że tak się zdarzyło: cecha $k+1$ rozdziela liście mające ojców a i b w drzewie T_k. Ponieważ $a \neq b$ czyli istnieje cecha i, która je różni. Mamy $O_i \cap O_{k+1} \neq \emptyset$. Nie jest również prawdą, że $O_i \subseteq O_{k+1}$ lub $O_{k+1} \subseteq O_i$ (Rysunek 2). Czyli otrzymujemy sprzeczość.

Z dowodu Twierdzenia 1 można wywnioskować algorytm realizujący nasze zadanie w czasie $\mathcal{O}(nm^2)$

Można skonstruować bardziej efektywny algorytm działający w czasie $\mathcal{O}(mn)$ [2] — zastanów się jak przyspieszyć fazę decyzyjną.

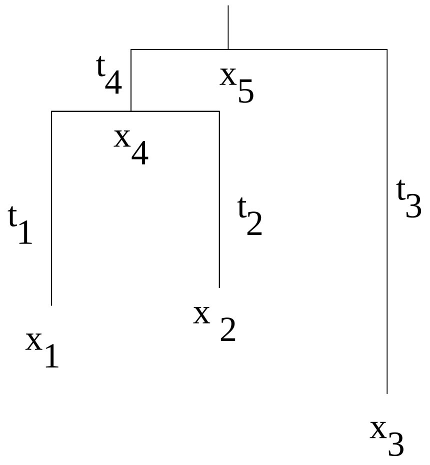
3 NP zupełność kryterium kompatybilności

Wróćmy do wspomnianego na początku problemu znalezienia drzewa, które jest kompatybilne z jak największą liczbą cech. W przypadku cech binarnych problem jest równoważny pytaniu: czy istnieje w zbiorze cech parami kompatybilny podzbiór o danej liczności.

Można pokazać, że już dla cech binarnych problem ten jest NP trudny redukując do niego problem kliki (ćwiczenia). Pozytywnym aspektem jest fakt, że istniejące dla problemu szukania kliki heurystki okazują się przydatne w problemie kompatybilności.

4 Probabilistyczne podejście do drzew filogenetycznych

Przejdziemy teraz to klasy metod, które używają modeli probabilistycznych w rekonstrukcji drzew. Podstawowym celem takiego podejścia jest klasyfikacja drzew pod względem ich wiarygodności (*ang. likelihood*), czyli prawdopodobieństwa obserwacji takich nie innych danych $\mathcal{D}$, pod warunkiem, że ich ewolucja przebiegała według drzewa T: $P(\mathcal{D}|T)$. Oznaczmy przez $P(\boldsymbol{x}|T, \boldsymbol{t})$



Rysunek 3: Przykładowe drzewo dla 3 sekwencji

warunkowe prawdopodobieństwo danego zbioru sekwencji $\boldsymbol{x} = x_1, x_2, \ldots x_n$ mając dane drzewo T o n liściach (etykietowanych sekwencjami), gdzie $\boldsymbol{t}$ oznacza długości krawędzi w drzewie. Do zdefiniowania $P(\boldsymbol{x}|T, \boldsymbol{t})$ niezbędny jest model ewolucji sekwencji po drzewie T. Załóżmy, że potrafimy zdefiniować prawdopodobieństwo $P(x|y, t)$ zdarzenia, że sekwencja y ewoluowała w sekwencję x po krawędzi długości t. Prawdopodobieństwo całego drzewa T przy założeniu, że znamy sekwencje wspólnych przodków w węzłach wewnętrznych, może być policzone jako iloczyn prawdopodobieństw ewolucyjnych wszystkich krawędzi. Dla przykładowego drzewa z Rysunku 3 to prawdopodobieństwo będzie równe:

$$P(x_1, x_2, \ldots x_5)|T, \boldsymbol{t}) = P(x_1|x_4, t_1)P(x_2|x_4, t_2)P(x_3|x_5, t_3)P(x_4|x_5, t_4)P(x_5)$$

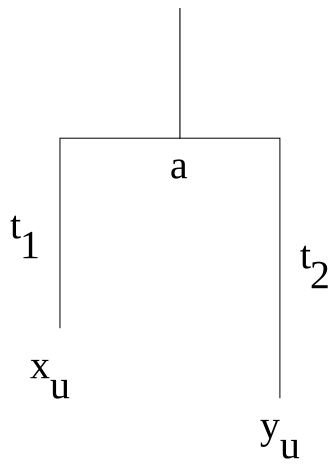
gdzie $P(x_5)$ jest prawdopodobieństwem wystąpienia sekwencji x_5 w korzeniu drzewa T. Niestety zazwyczaj nie znamy sekwencji wspólnych przodków, czyli żeby uzyskać $P(x_1, x_2, x_3|T, \boldsymbol{t})$ musimy posumować po wszystkich możliwych przodkach x_4 i x_5. Dla tak zdefiniowanego modelu probabilistycznego poszukujemy drzewa **największej wiarygodności**, czyli takiej topologii T oraz długości krawędzi $\boldsymbol{t}$, które maksymalizują $P(\boldsymbol{x}|T, \boldsymbol{t})$.

5 Probabilistyczne modele ewolucji

Do zdefiniowania $P(x|y,t)$ użyjemy uproszczonych modeli ewolucji sekwencji poznanych na wykładzie 2. Jak pamiętamy zaniedbują one insercje i delecje oraz traktują ewolucję każdego miejsca w sekwencji niezależnie. Zgodnie z tym ostatnim założeniem mamy:

$$P(x|y,t) = \prod_u P(x_u|y_u, t)$$

gdzie u indeksuje kolejne miejsca w uliniowieniu sekwencji x i y. Policzmy wiarygodność drzewa z Rysunku 4 dla dwóch sekwencji. Istnieje tylko jedno



Rysunek 4: Drzewo dla 2 sekwencji

takie drzewo (topologia), czyli musimy policzyć jak zmienia się wiarygodność wraz z długościami krawędzi t_1 i t_2. Nazwijmy wspólnego przodka x_u i y_u

przez a (w odróżnieniu od znanych nam sekwencji x oraz y jest to zmienna). Mamy:

$$P(x_u, y_u, a|T, t_1, t_2) = q_a P(x_u|a, t_1) P(y_u|a, t_2)$$

czyli jest to prawdopodobieństwo wystąpienia a w korzeniu (o którym przyjmuje się, że jest prawdopodobieństem stacjonarnym zakładanego modelu ewolucji) a następnie dwóch substytucji $a \to x_u$ oraz $a \to y_u$. Ponieważ z reguły nie znamy a musimy posumować po wszystkich możliwościach, czyli:

$$P(x_u, y_u|T, t_1, t_2) = \sum_a q_a P(x_u|a, t_1) P(y_u|a, t_2) \tag{1}$$

Zakładając, że rozważane przez nas sekwencje są długości N, otrzymujemy:

$$P(x, y|T, t_1, t_2) = \prod_{u=1}^{N} P(x_u, y_u, a|T, t_1, t_2) \tag{2}$$

Dla przykładu policzmy wiarygodność dla ewolucji dwóch sekwencji zakładając model Jukes-Cantora.

C	C	G	G	C	C	G	C	G	C	G
C	G	G	G	C	C	G	G	C	C	G

Przypomnijmy prawdopodobieństwa ewolucji residuów w t krokach w modelu JC:

$$P_{ii(t)} = \frac{1}{4}(1 + 3e^{-4pt}) \tag{3}$$

$$P_{ij(t)} = \frac{1}{4}(1 - e^{-4pt}) \qquad \text{dla} \quad i \neq j \tag{4}$$

gdzie p jest jedynym parametrem modelu i oznacza prawdopodobieństwo mutacji. Wykorzystując (1) otrzymujemy:

$$P(C, C|T, t_1, t_2) = q_C P_{ii(t_1)} P_{ii(t_2)} + q_G P_{ij(t_1)} P_{ij(t_2)} + q_A P_{ij(t_1)} P_{ij(t_2)} + q_T P_{ij(t_1)} P_{ij(t_2)} =$$

$$= \frac{1}{4}(P_{ii(t_1)} P_{ii(t_2)} + 3P_{ij(t_1)} P_{ij(t_2)})$$

Z symetrii mamy: $P(G, G|T, t_1, t_2) = P(C, C|T, t_1, t_2)$ i podobnie:

$$P(C, G|T, t_1, t_2) = P(G, C|T, t_1, t_2) = \frac{1}{4}(P_{ii(t_1)} P_{ij(t_2)} + P_{ij(t_1)} P_{ii(t_2)} + 2P_{ij(t_1)} P_{ij(t_2)})$$

Podstawiając wzory (3) oraz (4) otrzymujemy:

$$P(C,C|T,t_1,t_2) = \frac{1}{16}(1+3e^{-4p(t_1+t_2)})$$
$$P(C,G|T,t_1,t_2) = \frac{1}{16}(1-e^{-4p(t_1+t_2)})$$

Załóżmy teraz, że n_1 miejsc w sekwencji jest identycznych a na pozostałych n_2 miejscach zaszły substytucje. Ze wzoru (2) otrzymujemy:

$$P(x,y|T,t_1,t_2) = \frac{1}{16^{n_1+n_2}}(1+3e^{-4p(t_1+t_2)})^{n_1}(1-e^{-4p(t_1+t_2)})^{n_2}.$$

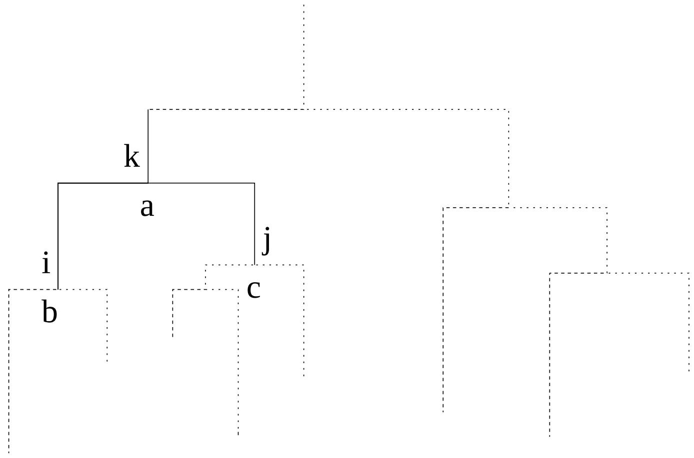
6 Wiarygodność drzewa: algorytm Felsensteina [1]

Policzmy teraz wiarygodność dla drzewa T o n liściach i długości krawędzi $\boldsymbol{t}$. Oznaczmy bezpośredniego przodka węzła i przez $\alpha(i)$. Ustalmy miejsce w sekwencji u i pomińmy dla uproszczenia w dalszych wzorkach indeks u. Prawdopodobieństwo $P(x_1, x_2, \ldots x_n|T,\boldsymbol{t})$ uzyskania tych residuów w liściach drzewa T jest uzyskane przez pomnożenie prawdopodobieństw substytucji po wszystkich krawędziach T:

$$P(x_1, x_2, \ldots x_n|T,\boldsymbol{t}) =$$
$$\sum_{a_{n+1},a_{n+2},\ldots a_{2n-1}} q_{a_{2n-1}} \prod_{i=n+1}^{2n-2} P(a_i|a_{\alpha(i)},t_i) \prod_{i=1}^{n} P(x_i|a_{\alpha(i)},t_i) \qquad (5)$$

gdzie musimy posumować po wszystkich możliwych przypisaniach residuów węzłom wewnętrznym (numerowanym od $n+1$ do $2n-1$). Rozważmy teraz algorytm (podobny do algorytmu ważonej parsymonii), który wylicza te prawdopodobieństwo dla danego drzewa, poczynając od liści. Niech $P(L_k|a)$ oznacza prawdopodobieństwo dla wszystkich liści poniżej węzła k przy założeniu, że literka w k jest równa a. Potrafimy policzyć $P(L_k|a)$ z analogicznych prawdopodobieństw zdefiniowanych dla synów k: $P(L_i|b)$ oraz $P(L_j|c)$ (porównaj Rysunek 5). Na koniec wyliczamy wiarygodność dla całych sekwencji korzystając z założenia o niezależności miejsc:

$$P(\boldsymbol{x}|T,\boldsymbol{t}) = \prod_{u=1}^{N} P(\boldsymbol{x}_u|T,\boldsymbol{t}).$$



Rysunek 5: Etykietowanie drzewa w algorytmie

Algorithm 1 Felsenstein 1981

inicjalizacja $k := 2n - 1$ (korzeń)
rekurencja dla $P(L_k|a)$
if k jest liściem **then**
  $P(L_k|a) := 1$ o ile w danym miejscu w sekwencji występuje a,
  jeśli nie: $P(L_k|a) := 0$.
else
  Policz $P(L_i|a)$, $P(L_j|a)$, dla synów i oraz j oraz dla każdego a.
  $P(L_k|a) := \sum_{b,c} P(b|a,t_i)P(L_i|b)P(c|a,t_j)P(L_j|c)$
wiarygodność miejsca u w sekwencji $:= P(\boldsymbol{x}_u|T,\boldsymbol{t}) = \sum_a P(L_{2n-1}|a)q_a$

7 Odwracalność i położenie korzenia

Jak pamiętamy koszt drzewa w kryterium parsymonii nie zależy od położenia korzenia. Okazuje się, że również wiarygodność drzewa możemy uniezależnić od korzenia przyjmując pewne założenia dotyczące modelu ewolucji. Pierwsze założenie to multiplikatywność modelu:

$$\sum_b P(a|b,t_1)P(b|c,t_2) = P(a|c,t_1+t_2),$$

która w oczywisty sposób zachodzi dla modeli markowowskich. Drugie założenie to odwracalność modelu (zachodzi dla modelu JC i modelu Kimury, jak również dla macierzy PAM):

$$\forall a, b, t \quad P(b|a,t)q_a = P(a|b,t)q_b$$

Zastanówmy się dlaczego multiplikatywność i odwracalność wystarczają żeby wiarygodność drzewa nie zależała od położenia korzenia. Załóżmy, że synowie korzenia (węzeł $2n-1$) to wierzchołki i oraz j. Z definicji $P(L_{2n-1}|a)$, które w algorytmie Felsensteina liczymy z wartości: $P(L_i|a)$ oraz $P(L_j|a)$ możemy policzyć wiarygodność całej rodziny sekwencji w miejscu u jako:

$$P(\boldsymbol{x}_u|T,\boldsymbol{t}) = \sum_a q_a P(L_{2n-1}|a) = \sum_{b,c,a} q_a P(b|a,t_i)P(c|a,t_j)P(L_i|b)P(L_j|c).$$

Teraz wykorzystując odwracalność otrzymujemy:

$$P(\boldsymbol{x}_u|T,\boldsymbol{t}) = \sum_{b,c}\left(\sum_a P(c|a,t_j)P(a|b,t_i)\right) q_b P(L_i|b)P(L_j|c).$$

Suma w nawiasie może zostać przekształcona wykorzystując multiplikatywność:

$$\sum_a P(c|a,t_j)P(a|b,t_i) = P(c|b,t_i+t_j)$$

co oznacza że prawdopodobieństwo nie zależy od symbolu a przypisanego korzeniowi, jedynie od sumarycznej długości krawędzi poniżej korzenia. Stąd wnioskujemy, że korzeń może zostać przesunięty do któregokolwiek z synów.

Literatura

[1] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleid acids.* Cambridge University Press, 1998.

[2] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing Company, 1997.

8 Zadania

1. Pokaż NP zupełność dla problemu kompatybilności (klika).
2. Policz wiarygodność dla drzewa o trzech liściach zawierającego sekwencje:

C	C	G	G	C	C	G	C	G	C	G
C	G	G	G	C	C	G	G	C	C	G
G	C	C	G	C	C	G	G	G	C	C

   w którym dwie pierwsze sekwencje są sobie bliższe.

9 Zadania ,,*domowe*''

1. Załóż, że dysponujesz algorytmem, który znajduje maksymalną klikę w grafie. Jak użyjesz tego algorytmu do znalezienia maksymalnego kompatybilnego drzewa dla zbioru cech (niekoniecznie binarnych).
2. Dla problemu obliczania wiarygodności drzewa pokazaliśmy że odwracalność modelu pozwala umieścić korzeń w dowolnym wierzchołku drzewa. Co się stanie jak umieścimy korzeń w liściu ?