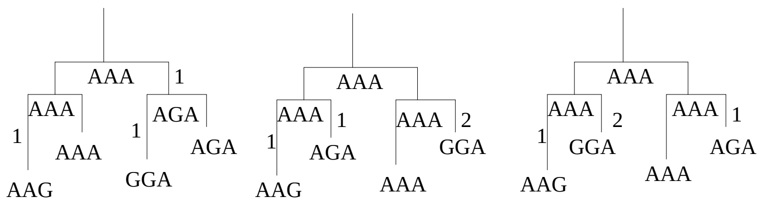
Wykład IV: Parsymonia

Anna Gambin
24.10.2002

1 Zasada parsymonii

Zasada parsymonii jest jedną z najczęściej stosowanych w algorytmach rekonstrukcji drzew filogenetycznych. Polega na znalezieniu drzewa, którego topologia "tłumaczy" pochodzenie znajdujących się w liściach sekwencji minimalną liczbą mutacji (substytucji). Jest to często stosowanana w nauce zasada zwana "brzytwą Okhama", czyli znalezieniu najprostszego poprawnego wytłumaczenia danego zjawiska. Główny pomysł różni się znacznie od rozważanych przez nas algorytmów opartych o macierz odległości: zamiast budować drzewo, przypisujemy każdej topologii koszt (czyli wyliczamy kryterium optymalności) a następnie przeszukujemy przestrzeń drzew w poszukiwaniu optymalnego.



Rysunek 1: Drzewa parsymonii dla przykładowych sekwencji

Zacznijmy od prostego przykładu i rozważmy 4 uliniowione sekwencje nukleotydów:

$$\begin{array}{ccc} A & A & G \\ A & A & A \\ G & G & A \\ A & G & A \end{array}$$

Przedstawione na rysunku trzy drzewa wyjaśniają ewolucję tych sekwencji: w wierzchołkach wewnętrznych znajdują się hipotetyczne sekwencje wspólnych przodków zbudowane w oparciu o zasadę parsymonii. Skrajne lewe drzewo ma najmniejszy koszt (trzy substytucje) podczas gdy dwa pozostałe są gorsze (ich koszt wynosi 4). W omawianych algorytmach będziemy (jak zwykle) przyjmować upraszczające założenia dotyczące ewolucji sekwencji:

- miejsca w sekwencji ewoluują niezalżnie od siebie: mutacja w jednym miejscu nie ma wpływu na inne.
- od momentu, kiedy ewolucja dwóch sekwencji/gatunków rozdziela sie w drzewie (zdarzenie specjacji) przebiega ona dalej zupełnie niezależnie.

Dzięki tym założeniom możemy każde miejsce w sekwencji traktować oddzielnie przy rozwiązywaniu następuącego problemu: dla danego drzewa (topologii) o liściach etykietowanych sekwencjami, policz minimalną liczbę mutacji, które "wyjaśniają" to drzewo. Chcemy również zrekonstruować sekwencje w wierzchołkach wewnętrznych drzewa.

Okazuje się że łatwiej zacząć od trochę ogólniejszego sformułowania czyli problemu *ważonej parsymonii*.

2 Algorymn Sankoffa: ważona parsymonia

Algorithm 1 Sankoff & Cedergren 1983

inicjalizacja $k := 2n - 1$ (korzeń)
rekurencja dla $S_k(a)$
if k jest liściem **then**
  $S_k(a) := 0$ o ile w danym miejscu w sekwencji występuje a,
  jeśli nie: $S_k(a) := \infty$.
else
  Policz $S_i(a)$, $S_j(a)$, dla synów i oraz j dla każdej wartości a.
  $S_k(a) := min_b(S_i(b) + S(a, b)) + min_b(S_j(b) + S(a, b))$
Minimalny koszt drzewa $:= min_a S_{2n-1}(a)$

W problemie ważonej parsymonii chcemy policzyć sumaryczny koszt mutacji (dla danego miejsca w sekwencji i danej topologii drzewa). Niech $S(a, B)$ oznacza koszt substytucji a na b. Ustalmy miejsce w sekwencji (w liściach znajdują się teraz pojedyncze nukleotydy) i oznaczmy przez $S_k(a)$ minimalny koszt przypisania a w węźle k.

Jeżeli chcemy dodatkowo zrekonstruować parsymonicznych wspólnych przodków wystarczy w każdym węźle pamiętać dowiązania do tych etykiet synów, które dawały minimalną wartość w kroku 8 algorytmnu. Czyli dodajemy na koniec procedury rekurencyjnej dwa kroki:

$$l_k(a) := \mathrm{argmin}_b(S_i(b) + S(a,b))$$
$$r_k(a) := \mathrm{argmin}_b(S_j(b) + S(a,b))$$

Wystarczy teraz ustalić literkę (nukleotyd) w korzeniu i zejść po wskaźnikach do liści wybierając dowolnie w sytuacji kiedy mamy wiele minimalnych dowiązań.

Koszt algorytmu wynosi $\mathcal{O}(n \cdot m \cdot |\Sigma|)$ (uzasadnij dlaczego), gdzie n jest liczbą rozważanych sekwencji, m ich długością, a Σ rozważanym alfabetem (dla sekwencji nukleotydowych $\Sigma = 4$).

3 Algorytm Fitcha

W prostszym przypadku kiedy $S(a,b) = 1$ dla $a \neq b$ oraz $S(a,a) = 0$, chcemy policzyć minimalną liczbę substytucji "wyjaśniających" ewolucję sekwencji w liściach. W tym celu w każdym wierzchołku drzewa k pamiętamy zbiór literek dających minimalny koszt (oznaczamy go R_k) oraz aktualny koszt drzewa C.

Algorithm 2 Fitch 1971

inicjalizacja $C := 0$ oraz $k := 2n - 1$ (korzeń)
rekurencja dla policzenia R_k
if k jest liściem **then**
  $R_k := a$ (jeśli w danym miejscu w sekwencji występuje a),
else
  Policz R_i, R_j, dla synów i oraz j.
  if $R_i \cap R_j \neq \emptyset$ **then**
    $R_k := R_i \cap R_j$
  else
    $R_k := R_i \cup R_j$
    $C := C + 1$
Minimalny koszt drzewa $:= C$

Podobnie jak poprzednio możemy zrekonstruować sekwencje w wierzchołkach wewnętrznych cofając się od korzenia i wybierając dla syna i tę samą literkę co w ojcu, o ile jest to możliwe, lub losową literkę ze zbioru R_i. Okazuje się, że uzyskane w ten sposób etykietowania nie wyczerpują wszystkich poprawnych możliwości (zastanów się dlaczego ?), które możnaby natomiast uzyskać algorytmem ważonej parsymonii.

Obydwa rozważane przez nas algorytmy zostały zdefiniowane dla drzew ukorzenionych, co ułatwia rekurencyjne podejście do problemu. Okazuje się, że w rzeczywistości koszt drzewa (w kryterium tradycyjnej nieważonej parsymonii) jest niezależny od położenia korzenia. Jeżeli rozpatrzymy dwie krawędzie prowadząe od korzenia do synów to w optymalnym drzewie nie mogą obydwie jednocześnie reprezentować substytucji (czyli etykieta korzenia musi być taka sama jak u jednego z synów). Oznacza to że możemy usunąć korzeń i liczyć koszt jako sumę kosztów dla wszystkich pozostałych krawędzi.

4 Poszukiwanie najlepszego drzewa

Umiemy efektywnie policzyć koszt dla danego drzewa. Dużo trudniejszym problemem jest znalezienie optymalnego parsymonicznie drzewa. Można pokazać że ten problem jest NP zupełny [3], jednak bardziej interesują nas efektywne strategie poszkiwań w przestrzeni drzew.

4.1 Strategia *Branch and Bound*

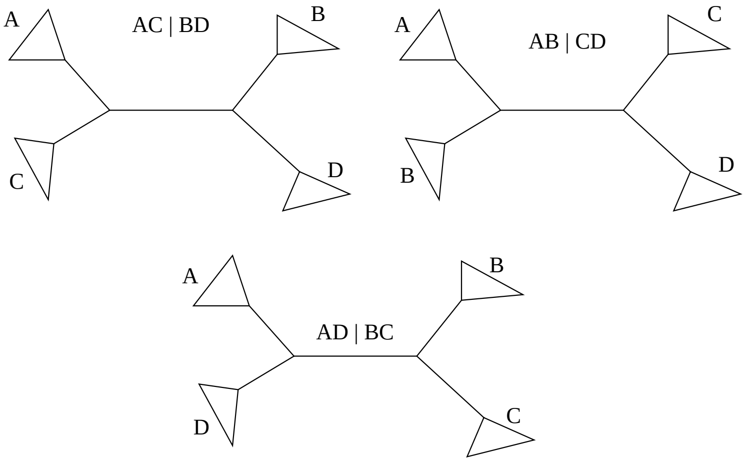
Jest to ogólna strategia stosowana przy rozwiązywaniu problemów optymalizacyjnych, dla których przestrzeń poszukiwań może być przedstawiona w postaci drzewa zawierającego w liściach rozwiązania dopuszczalne. Strategia gwarantuje znalezienie rozwiązania optymalnego, jednak pesymistyczna złożoność algorytmu jest wykładnicza.

Dla poprawności algorytmu potrzebna jest następująca własność *monotoniczności* drzewa: waga każdego węzła jest co najmniej taka jak waga jego przodków. Dzięki tej własności możemy obciąc (czyli wykluczyć z przeszukiwań) całe poddrzewo zaczepione w wierzchołku, którego waga jest większa od najmniejszej napotkanej (zakładamy tutaj, że szukamy najlżejszego liścia).

Dla drzew filogenetycznych nasza przestrzeń poszukiwań wygląda następująco:

k- ty poziom drzewa przeszukiwań składa się z drzew filogenetycznych o k liściach, każdy węzeł z tego poziomu ma $2k-1$ synów bo na tyle sposobów możemy dodać nowy węzeł (gatunek) do drzewa. Łatwo widać że tak skonstruowane drzewo jest monotoniczne, ponieważ dodanie nowego wierzchołka do drzewa nie może zmniejszyć jego kosztu mierzonego minimalną liczbą mutacji.

4.2 Strategia *Nearest Neighbor Interchanges*



Rysunek 2: Sąsiednie drzewa w metryce NNI.

Pomysł polega na zdefiniowaniu relacji sąsiedztwa w przestrzeni drzew i przeszukiwaniu tej przestrzeni przy pomocy lokalnych kroków (przechodzimy od danego drzewa do jego sąsiadów). Mając zdefiniowane sąsiedztwo możemy używać znanych heurystyk takich jak: podejście zachłanne lub metoda symulowanego wyżarzania (*ang. simmulated annealing*), które zazwyczaj pozwalają znaleźć lokalne optimum (które może się szczęśliwym zbiegiem okoliczności okazać również globalnym optimum). Dla drzew nieukorzenionych (a jak pamiętamy tylko takie wystarcza rozważać) relacja NNI zdefiniowana jest następująco: rozważamy krawędź (i, j) i cztery poddrzewa o korzeniach w synach i oraz j, oznaczmy je A, B, C, D. Jeśli A i B mają wspólnego przodka bliżej niż C i D, piszemy $AB|CD$ pozostałe sąsiednie drzewa uzyskamy zamieniając sąsiednie poddrzewa: $AC|BD$ oraz $AD|BC$. Dla drzew o 5 liściach graf relacji NNI jest znany jako graf Petersona.

Inne rozważane relacje sąsiedztwa posługują się operacją wycięcia pewnego poddrzewa i zawieszenia go w innym (dowolnym) miejscu [1].

Literatura

[1] B. DasGupta, H. Xin, M. Li, J. Tromp, L. Wang, and L. Zhang. Computing distances between evolutionary trees. In D-Z. Du and P.M. Pardalos, editors, *Handbook of Combinatorial Optimization*, pages 35–76. Kluwer Academic Publishers, 1998.

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

[3] D. Gusfield. The steiner tree problem in phylogeny. Technical Report 332, Yale University Computer Science Dep., 1984.

[4] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology.* Cambridge University Press, 1997.

5 Zadania

1. Dlaczego algorytm Fitch'a, nie znajduje wszystkich optymalnych etykietowań ? Popraw go tak ażeby znajdował.

2. **Parsymonia i drzewa Steinera** [4]

   Definicja 1 *Dla grafu* $G = (V, E)$ *i podzbioru jego wierzchołków* $X \subseteq V$ **drzewo Steinera** *jest spójnym poddrzewem* G *zawierającym wszystkie wierzchołki z* X *(może ono zawierać również wierzchołki spoza* X*). W grafach ważonych problem drzewa Steinera dla* X *polega na wyznaczeniu drzewa o minimalnej wadze.*

   Definicja 2 **Hiperkostka** *wymiaru* d *to nieskierowany graf o* 2^d *wierzchołkach, etykietowanych liczbami* $0, \ldots, 2^d - 1$*. Dwa wierzchołki hiperkostki są połączone krawędzia, jeśli reprezentacje binarne ich etykiet różnią się tylko jednym bitem.*

   Sformułuj problem wyznaczenia maksymalenego drzewa parsymonicznego w języku drzew Steinera na hiperkostce. Uogólnij alfabet binarny na dowolny skończony.

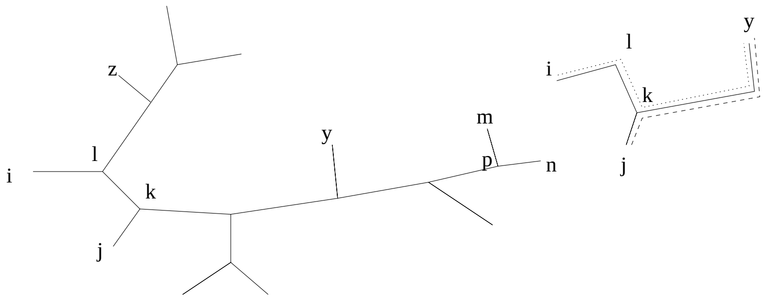
6 Zadania „*domowe*"

1. Uzupełnij lukę w dowodzie twierdzenia Neighbor-Joining: jeśli poddrzewo zawieszone w wierzchołku k posiada tylko jeden liść m (czyli nie możemy znaleźć pary sąsiadów w L_k) pokaż, że obecność innych (niż k) węzłów na ścieżce z i do j pociąga za sobą $D^u_{ij} > D^u_{jm}$, czyli sprzeczność z założeniem że D^u_{ij} jest minimalne.

2. Pokaż że minimalny koszt drzewa w kryterium ważonej parsymonii jest niezależny od położenia korzenia o ile funkcja definiująca koszt substytucji jest metryką.

7 Dowód twierdzenia neighbour-joining

Twierdzenie 1 *[2] Dla addytywnego drzewa minimalne* $D^u_{ij} = D_{ij} - u_i - u_j$ *oznacza, że liście* i *oraz* j *są sąsiednie.*



Rysunek 3: Ilustracja dowodu twierdzenia, po lewej: $d_{iy} + d_{jy} = d_{ij} + 2d_{ky}$

Załóżmy przeciwnie: D^u_{ij} jest minimalne ale i oraz j nie sąsiadują w oryginalnym drzewie, czyli na ścieżce pomiędzy nimi znajdziemy co najmniej dwa węzły k i l. Niech L_k (odpowiednio L_l) będzie zbiorem liści poddrzewa o wierzchołku k (l). Niech m i n będą parą sąsiednich liści w L_k połączonych ojcem p (a co jeśli takowych nie ma ?). Niech d_{uv} oznacza sume długości krawędzi na ścieżce pomiędzy u i v. Jeżeli u i v są liśćmi z własności addytywności $d_{uv} = D_{uv}$. Dla każdego y w L_k mamy: $d_{iy} + d_{jy} = d_{ij} + 2d_{ky}$ i podobnie $d_{my} + d_{ny} = d_{mn} + 2d_{py}$, stąd:

$$d_{iy} + d_{jy} - d_{my} - d_{ny} = d_{ij} + 2d_{ky} - 2d_{py} - d_{mn} \quad (1)$$

Podobnie dla $z \in L_l$ mamy:

$$d_{iz} + d_{jz} - d_{mz} - d_{nz} = d_{ij} - 2d_{pk} - 2d_{lk} - d_{mn} \quad (2)$$

Z definicji D^u_{ij}:

$$D^u_{ij} - D^u_{mn} = d_{ij} - d_{mn} - \frac{1}{N-2}\left(\sum_{\text{liście } u} d_{iu} + d_{ju} - d_{mu} - d_{nu}\right)$$

Łatwo sprawdzić z (1) oraz (2) że współczynniki przy d_{ij} oraz d_{mn} posumowane po wszystkich liściach wynoszą dokładnie $(N-2)$, czyli otrzymujemy:

$$D^u_{ij} - D^u_{mn} = \frac{1}{N-2}\left(\sum_{y \in L_k}(2d_{py} - 2d_{ky}) + \sum_{z \in L_l}(2d_{pk} + 2d_{lk})\right) + C$$

gdzie C jest sumą pozostałych składników z poddrzew zaczepionych na ścieżce pomiędzy i oraz j w wierzchołkach innych niż k i l. Wykorzystując fakt, że $d_{py} - d_{ky} > -d_{pk}$ mamy:

$$D_{ij}^u - D_{mn}^u > 2d_{pk}\frac{|L_l| - |L_k|}{N-2}$$

Musi być $D_{mn}^u > D_{ij}^u$ (bo założyliśmy, że D_{ij}^u jest minimalne), czyli $|L_l| < |L_k|$. Pamietamy że rola k i l była zupełnie symetryczna, czyli możemy je zamienic i otrzymamy sprzeczność: $|L_l| > |L_k|$. Stąd wnioskujemy, że założenie jest nieprawdziwe: liście i oraz j są sąsiednie.