Wykład III: metody rekonstrukcji drzew z macierzy odległości

Anna Gambin
17.10.2002

1 Charakterystyka sensownej metody rekonstrukcji drzewa

Na tym i kilku najbliższych wykładach będziemy rozważać różne metody pozwalające stworzyć drzewo modelujące ewolucję umieszczonych w liściach sekwencji molekularnych. Ustalmy co rozumiemy przez sformułowanie **rekonstrukcja drzewa** oraz jakie kryteria powinna taka metoda spełniać ażeby uznać ją za użyteczną dla biologa i sensowną dla matematyka.

Takie podejście pozwoli nam na dokonanie porównania omówionych metod. Niestety nie wypracowano jeszcze jednolitych kryteriów pozwalających oceniać różne podejścia do budowy drzew. Oprzemy się tutaj na propozycjach zawartych w [1].

Zacznijmy od zdefiniowania co oznacza określenie rekonstrukcja drzewa. Na każdą metodę składają się trzy części:

1. **kryterium optymalności**, które mierzy dopasowanie posiadanych przez nas danych (np. ewolucyjnych odległości) do danego drzewa.

2. **strategia poszukiwań** optymalnego drzewa w przestrzeni wszystkich drzew.

3. założenie dotyczące **mechanizmu ewolucji** dla danych, których używamy w rekonstrukcji.

Jako przykład rozważmy poszukiwanie drzewa "minimalnej ewolucji"(1), algorytmem poszukiwań Neighbor-Joining (2), dla odległości ewolucyjnych, gdzie przy korygowaniu wpływu wielokrotnych substytucji wykorzystano dwuparametrowy model Kimury (3).

Niestety często podejmowane są próby porównywania metod różniących się we wszystkich trzech "wymiarach", co jest zazwyczaj z góry skazane na niepowodzenie.

Sprecyzujmy teraz jakich własnosci oczekujemy od sensownej metody. W [1] wyróżniono następujące cechy:

- **efektywność** czyli zlożoność obliczeniowa algorytmów używanych w danej metodzie. Punkt (1) czyli ewaluacja kryterium optymalności dla danego drzewa nie stanowi tutaj zazwyczaj problemu: dla większości rozważanych kryteriów istnieją wielomianowe algorytmy. Niestety nie istnieją efektywne algorytmy przeszukujące całą (wykładniczą) przestrzeń drzew. Stosuje się tutaj heurystyczne strategie przeszukiwań lub próbkowanie przestrzeni.

- **zgodność**. Kolejne trzy cechy wiążą się z następującą obserwacją: biolodzy zgadzają się, że mając jako dane dość krótkie sekwencje molekularne możemy zrekonstruować drzewo zawierające pewne błedy wynikające właśnie z niedostatecznych danych. Jednak wraz ze wzrostem rozważanych danych nasza metoda powinna zbiegać do optymalnego drzewa. Taką metodę określimy jako **zgodną**. Okazuje się że większość sensownych kryteriów optymalności (1) prowadzi do zgodnej metody o ile założenia dotyczące mechanizmu ewolucji sekwencji (3) są poprawne. Niestety te same kryteria optymalności zaowocują metoda, która nie zbiega do poprawnego drzewa jeśli przyjmiemy nieprawdziwe założenia o ewolucji sekwencji.

- **zbieżność**. Kolejne nasuwające się pytanie dotyczy rozmiaru danych potrzebnych aby metoda dawała drzewo optymalne: jak długich sekwencji nukleotydów powinniśmy użyć, aby zapewnić zbieżność metody. Za cechę pozytywną uznamy szybką zbieżność, czyli poprawne drzewo otrzymane z relatywnie krótkich sekwencji.

- **stabilność** metody oznacza wpływ zaburzeń danych na otrzymane wyniki. Metode określimy jako **stabilną** jeśli pozostaje zbieżna również dla lekko zaburzonych danych. Zaburzone dane oznaczają w tym przypadku, że nie znamy dokładnie wszystkich mechanizmów ewolucyjnych i przyjmujemy upraszczające założenia (np. że kolejne miejsca w sekwencji ewoluują niezależnie od siebie i w identyczny sposób).

2 Metoda najmniejszych kwadratów

Będziemy zakładać, że dla każdej pary gatunków/sekwencji znamy ich odległość ewolucyjną. Wielkości te zgromadzone są w macierzy odlgłości D. Naszym celem jest skonstruowanie drzewa ważonego, które zawiera w liściach rozważane sekwencje i jest zgodne z macierzą D (odległości pomiędzy liśćmi w drzewie, liczone jako sumy wag po ścieżce je łączącej są równe odległościom z macierzy D). Niestety dla większości rozważanych macierzy odległości nie istnieje zgodne z nią drzewo — wtedy naszym celem jest zbudowanie drzewa, które indukuje odległości jak najbliższe tym z macierzy D.

Jedną z prostszych miar bliskości jest następujące kryterium: jeśli przez d_{ij} oznaczymy odległości indukowane przez drzewo a przez D_{ij} odległości z macierzy, naszym celem jest zbudowanie drzewa, które minimalizuje:

$$\Sigma_{i=1}^{n}\Sigma_{j\neq i}w_{ij}(D_{ij}-d_{ij})^2 \quad (1)$$

gdzie zazwyczaj wagi $w_{ij} = 1$ (kryterium Edwards & Cavalli-Sforza) lub $w_{ij} = \frac{1}{D_{ij}^2}$ (Fitch & Margoliash).

Udowodniono, że w przypadku dowolnej (rzeczywistej dodatniej) macierzy D znalezienie ważonego drzewa, które minimalizuje (1) jest NP trudne. Omówimy dzisiaj dwie wielomianowe heurystyki, które jednak rozwiązują ten problem dokładnie przy dodatkowych ograniczeniach jakie musi spełniać macierz odległości D. Dwie wspomniane klasy macierzy to macierze ultrametryczne (działa dla nich prosty i efektywny algorytm UPGMA) oraz addytywne (algorytm Neigbor-Joining).

3 Odległość ultrametryczna i hipoteza zegara molekularnego

Hipoteza zegara molekularnego zaproponowana we wczesnych latach 60-tych przez Zuckerkandla i Paulinga postuluje, że dla każdej sekwencji (DNA lub białka) częstotliwość zmian ewolucyjnych jest w przybliżeniu niezmienna w czasie [3]. Oznacza to, że liczba mutacji jakie zaszły w pewnym przedziale czasowym jest proporcjonalna do długości tego przedziału. W rzeczywistości częstotliwości mutacji są różne dla różnych białek (np. hemoglobina mutuje szybciej niż cytochrom c), oraz różne dla poszczególnych części białka.

Hipoteza zegara znacznie upraszcza wyznaczenie odległości ewolucyjnej pomiędzy gatunkami/sekwencjami. Załóżmy że potrafimy oszacować liczbę mutacji (oznaczmy ją przez k) jakie zaszły od chwili rozejścia się sekwencji A i B. Oznacza to, że w każdej z rozważanych sekwencji dokonało się około $k/2$ mutacji i ta liczba jest proporcjonalna do czasu jaki upłynął od rozejścia się ewolucji tych sekwencji. Czyli zredukowaliśmy problem oszacowania czasu jaki upłynął od zdarzenia specjacji (rozejścia się gatunków) do problemu szacowania liczby mutacji. Drzewa ewolucyjne, dla których zakłada się istnienie zegara molekularnego, mają następującą własność: odległości od korzenia do każdego z liści są równe. Takie drzewa nazywamy ultrametrycznymi. Zdefiniujemy teraz formalnie pojęcia macierzy (odległości) ultrametrycznej i drzewa ultrametrycznego.

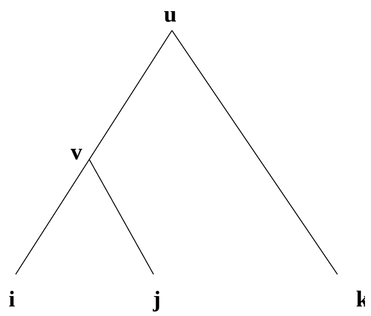
Definicja 1 *Dla symetrycznej rzeczywistej macierzy D wymiaru n x n drzewo ultrametryczne dla D to ukorzenione drzewo T spełniające:*

1. *T zawiera n liści po jednym dla każdego wiersza macierzy D.*
2. *Każdy węzeł wewnętrzny T ma co najmniej dwóch synów i jest etykietowany przez jedną z pozycji macierzy D.*
3. *Dla każdej ścieżki prowadzącej od korzenia do liścia liczby będące etykietami wierzchołków wewnętrznych tworzą ciąg ściśle malejący.*
4. *Dla każdej pary liści i oraz j etykietą ich najbliższego wspólnego przodka w drzewie T jest D_{ij}.*

Definicja 2 *Symetryczna rzeczywista macierz D jest macierzą ultrametryczną wttw dla każdych i, j, k, jeśli rozważymy wartości D_{ij}, D_{ik} oraz D_{jk} to albo wszystkie trzy są równe, albo dwie spośród nich są równe i większe od trzeciej.*

Twierdzenie 1 *Symetryczna macierz D ma drzewo ultrametryczne wttw D jest macierzą ultrametryczną.*

Dowód implikacji $\Longrightarrow$ jest naszkicowany na rysunku 1. Ażeby dokończyć dowód twierdzenia należy pokazać, że dla macierzy ultrametrycznej można



Rysunek 1: Załóżmy że macierz D ma drzewo ultrametryczne T; i, j, k wybieramy dowolnie i rozważamy minimalne poddrzewo drzewa T je zawierające. Wiemy, że etykieta węzła u musi być większa niż etykieta węzła v. Etykietą węzła u jest $D_{ik} = D_{jk}$ a etykietą v jest D_{ij}, czyli te trzy wielkości spełniają warunek odległości ultrametrycznej.

skonstruować (co więcej jednoznacznie i efektywnie) ultrametryczne drzewo. Konstrukcję tego algorytmu pozostawiamy jako ćwiczenie. Okazuję się, że buduje on drzewo efektywnie i w sposób jednoznaczny. Jako wniosek możemy sformułować następujące twierdzenie:

Twierdzenie 2 *Dla macierzy ultrametrycznej D, jej ultrametryczne drzewo jest wyznaczone jednoznacznie i może być zrekonstruowane w czasie $\mathcal{O}(n^2)$.*

Okazuje się, że nawet jeśli wątpimy w hipotezę zegara molekularnego oraz uzyskana przez nas macierz odległości nie jest ultrametryczna, możemy wykorzystać algorytmy znajdujące drzewo ultrametryczne dla efektywnej aproksymacji bardziej ,,realistycznych” drzew.

4 Odległości ultrametryczne: algorytm UPGMA

Opisany tutaj algorytm mimo dość wyrafinowanej nazwy (UPGMA jest skrótem od ang. Unweighted Pair Group Method using Arithmetic averages) jest bardzo prosty i intuicyjny. Główny pomysł polega na łączeniu wspólnym przodkiem najbliższych sobie sekwencji. Powstaje w ten sposób nowy wierzchołek w drzewie, który uczestniczy w dalszym "zlepianiu". W każdym kroku algorytmu zlepiamy ze sobą dwie najbliższe grupy sekwencji (klastry) aż pozostanie tylko jedna (klaster odpowiadający korzeniowi drzewa).

Zdefiniujmy odległość pomiędzy dwiema grupami (klastrami) sekwencji C_i oraz C_j jako:

$$D_{ij} = \frac{1}{|C_i| \cdot |C_j|} \Sigma_{p \in C_i} \Sigma_{q \in C_j} D_{pq} \tag{2}$$

Na algorytm UPGMA składa się faza inicjalizacji oraz n iteracji (w każdej z nich dwie grupy są zlepiane w jedną).

Algorithm 1 UPGMA

---

1 w wynikowym drzewie przypisz liściom gatunki (wiersze macierzy D); każdy z liści staje się jednoelementowym klastrem;

repeat

2 znajdź C_i oraz C_j o najmniejszej odległości D_{ij};

3 połącz je w jeden klaster C_k zawierający $|C_i| + |C_j|$ sekwencji;

4 dodaj w drzewie wierzchołek odpowiadający nowemu klastrowi krawędziom łączącym go z synami C_i oraz C_j nadaj wagi $\frac{D_{ij}}{2}$;

5 policz odległość pomiędzy nowym klastrem a pozostałymi (z pominięciem klastrów C_i oraz C_j) jako:

$$D_{kl} = (\frac{|C_i|}{|C_i| + |C_j|})D_{il} + (\frac{|C_j|}{|C_i| + |C_j|})D_{jl} \tag{3}$$

6 Usuń z macierzy D kolumny i wiersze odpowiadające klastrom C_i oraz C_j oraz dodaj kolumnę i wiersz dla nowego klastra C_k.

until pozostanie tylko jeden klaster

Ażeby uzasadnić poprawność algorytmu UPGMA należy udowodnić następujący lemat:

Lemat 1 *Jeżeli odlegóści pomiędzy klastrami C_i oraz C_j zdefinowane są wzorem (2) to odleglość od nowo stworzonego klastra $C_k = C_i \cup C_j$ do pozostałych wyraża się wzorem (3).*

Dodatkowo można pokazać, że algorytmn UPGMA zastosowany do macierzy ultrametrycznej zbuduje drzewo zgodne z tą macierzą (suma odległości na ścieżce pomiędzy każdą parą i, j jest równa wartości D_{ij}). Warunek ultrametryczności z Definicji 2 gwarantuje, że najbliższe sobie liście są sąsiednimi liśćmi w oryginalnym drzewie.

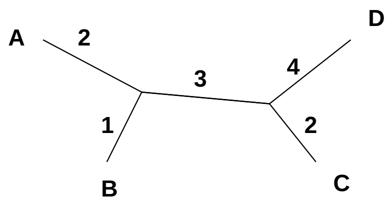
5 Odległości addytywne: algorytm Neighbor-Joining

Dane ultrametryczne to przypadek idealny i niestety rzadko spotykany w praktyce. Słabszym założeniem dotyczącym macierzy odległości jest **addytwność**.

Definicja 3 *Dla symetrycznej dodatniej macierzy D wymiaru n x n o elementach diagonalnych równych* 0*, drzewo addytywne to ważone drzewo T spełniające:*

1. *T zawiera co najmniej n węzłów etykietowanych wierszami macierzy D.*
2. *Dla każdych dwóch etykietowanych węzłów i oraz j waga scieżki pomiędzy nimi wynosi D_{ij}.*

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



Rysunek 2: Symetryczna macierz D i jej addytywne drzewo T (uwaga: T zawiera nieetykietowane wierzchołki czyli nie jest *zwarte* wg. Def. 4)

Problem addytywnego drzewa polega na znalezieniu dla macierzy D drzewa zdefiniowanego powyżej lub stwierdzeniu, że takowe nie istnieje.

Można udowodnić, że jeżeli macierz D jest ultrametryczna oraz $D_{ii} = 0 \;\; \forall i$ to D jest również addytywna (czyli ma addytywne drzewo).

Definicja 4 *Dla symetrycznej dodatniej macierzy D wymiaru n x n o elementach diagonalnych równych* 0, **zwarte** *drzewo addytywne to ważone drzewo T spełniające:*

1. *T zawiera* **dokładnie** *n węzłów etykietowanych wierszami macierzy D.*
2. *Dla każdych dwóch etykietowanych węzłów i oraz j waga scieżki pomiędzy nimi wynosi D_{ij}.*

Niech $G(D)$ bedzie grafem pełnym o n etykietowanych wierzchołkach, w którym waga krawędzi pomiędzy wierzchołkami wynosi D_{ij}. Można udowodnić następujące twierdzenie (dowód pozostawiamy jako ćwiczenie).

Twierdzenie 3 *Jeżeli T jest zwartym addytywnym drzewem dla macierzy D to T musi być jedynym minimalnym drzewem rozpinającym w grafie $G(D)$.*

Korzystając z tego twierdzenia można znaleźć zwarte drzewo addytywne dla macierzy D w czasie $\mathcal{O}(n^2)$ (zastanów się jak ?).

Omówimy teraz efektywny algorytm, który konstruuje drzewo dla macierzy, o ile macierz jest addytywna algorytm znajduje drzewo zgodne z tą macierzą. Głowny pomysł polega na połączeniu wspólnym ojcem pary sąsiednich (w oryginalnym drzewie) liści. Przy założeniu ultrametryczności wystarczyło wybrać parę najbliższych liści, warunek addytywności jest słabszy: teraz może się zdarzyć, że najbliższe liście wcale nie są sąsiednie (spróbuj narysować taki przykład). Czyli nie potrafimy wybrać pary sąsiednich liści jedynie na podstawie macierzy odległości. Okazuje się że zadziała pomysł Saitou & Nei: ażeby zrekompensować wpływ długich krawędzi prowadzących do liścia, odejmujemy średnią odległość do wszystkich innych liści, czyli łaczymy dwa klastry nie tylko bliskie sobie ale również odległe od reszty. Zdefiniujmy odległości węzła i od reszty jako:

$$u_i = \frac{1}{n-2}\Sigma_{k \neq i} D_{ik}$$

Można udowodnić (por. [2]), że para liści i, j, dla których $D_{ij} - u_i - u_j$ jest minimalne jest sąsiednia.

Algorytm wykorzystuje dwie własności addytywnego drzewa:

Algorithm 2 Neighbor-Joining

w wynikowym drzewie przypisz liściom gatunki (wiersze macierzy D); każdy z liści staje się jednoelementowym klastrem;

repeat

  znajdź i oraz j o najmniejszej wartości

$$D_{ij} - u_i - u_j$$

  połącz je w jeden klaster C_k;

  dodaj w drzewie wierzchołek odpowiadający nowemu klastrowi krawędziom łączącym go z synami C_i oraz C_j nadaj wagi

$$\begin{aligned} d_{ik} =& \frac{1}{2} D_{ij} + \frac{1}{2}(u_i - u_j) \\ d_{jk} =& \frac{1}{2} D_{ij} + \frac{1}{2}(u_j - u_i) \end{aligned}$$

  policz odległość pomiędzy nowym klastrem C_k a pozostałymi (z pominięciem klastrów C_i oraz C_j) jako:

$$D_{kl} = \frac{1}{2}(D_{ik} + D_{jk} - D_{ij}) \tag{4}$$

  Usuń z macierzy D kolumny i wiersze odpowiadające klastrom C_i oraz C_j oraz dodaj kolumnę i wiersz dla nowego klastra C_k.

until pozostanie tylko jeden klaster

- dla dowolnych trzech liści i, j, l połączonych wierzchołkiem wewnętrznym k, mamy $D_{il} = D_{ik} + D_{kl}$, $D_{jl} = D_{jk} + D_{kl}$ oraz $D_{ij} = D_{ik} + D_{kj}$ czyli:

$$D_{kl} = \frac{1}{2}(D_{ik} + D_{jk} - D_{ij})$$

  (krok [5] algorytmu).

- **warunek czterech punktów** dla dowolnych czterech liści i, j, k, l, dwie spośród odległości $D_{ij} + D_{kl}$, $D_{ik} + D_{jl}$ oraz $D_{il} + D_{jk}$ muszą być równe i większe od trzeciej (zrób rysunek).

Literatura

[1] D.J. Balding, M. Bishop, and C. (Ed.) Cannings. *Handbook of Statistical Genetics*. John Wiley & Sons, 2000.

[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] W-H. Li. *Molecular Evolution*. Sinauer Associates, 1997.

6 Zadania

1. Niech $G(D)$ oznacza graf kodowany przez macierz D (czyli graf pełny o wagach krawędzi $w_{ij} = D_{ij}$) niech M_D oznacza minimalne drzewo rozpinające w tym grafie. Udowodnij że D jest macierzą ultrametryczną wttw dla każdej pary wierzchołków i oraz j maksymalna waga krawędzi na scieżce łączącej i i j w drzewie M_D wynosi D_{ij}.

2. Załóżmy że z definicji drzewa ultrametrycznego usunęliśmy zalożenie o *ściśle* malejącej numeracji wierzchołków na ścieżce od korzenia do liścia. Teraz dla jednej macierzy może istnieć więcej niż jedno drzewo ultrametryczne: scharakteryzuj klasę tych drzew.

7 Zadania *,,domowe"*

1. Podaj $\mathcal{O}(n^2)$ algorytm, który sprawdza czy macierz n x n jest macierzą ultrametryczną.

2. Niech $G(D)$ oznacza graf kodowany przez macierz D (czyli graf pełny o wagach krawędzi $w_{ij} = D_{ij}$). Udowodnij, że jeżeli macierz D jest ultrametryczna to $G(D)$ posiada minimalne drzewo rozpinające będące scieżką o n wierzchołkach.