Wykład VII: Porównanie metod rekonstrukcji drzew.

Anna Gambin
14.11.2002

1 Zgodność i zbieżność metody

Zaczęliśmy omawianie metod rekonstrukcji drzew filogenetycznych od wymienienia własności jakie powinna spełniać taka metoda. Oczywistą własnością była **efektywność** obliczeniowa. Nie będziemy teraz przypominać złożoności czasowej poszczególnych metod. Jest jasne, że metody oparte o macierz odległości wygrywają z metodą parsymonii, a ta z kolei jest szybsza od metody największej wiarygodności (heurystyki przeszukujące przestrzeń drzew w metodzie parsymonii są w miarę szybkie).

Inną cechą wymienioną przez nas była **stabilność** metody, czyli wpływ zaburzeń danych na otrzymane wyniki. Tę cechę również pominiemy w dalszych rozważaniach: wyniki analityczne są tutaj bardzo trudne i niepełne, w większości przypadków przeprowadza się symulacje.

Dużo więcej można powiedzieć na temat **zgodności** oraz szybkości **zbieżności** danej metody. Statystycy nazywają daną metode estymacji (estymator) zgodną jeśli zbiega do prawdziwej wartości wraz ze wzrostem danych użytych do estymacji. Rekonstrukcja drzewa filogenetycznego może być widziana jako estymacja prawdziwego drzewa. Oczywiście zgodność metody jest cechą wysoce pożądaną. Interesującą własnością jest też szybkość zbieżności metody estymacji dla wzrastających danych.

1.1 Model CF (Cavender-Felsenstein/Cavender-Farris)

Ażeby badać zgodność i zbieżność poszczególnych metod musimy mieć matematyczny model ewolucji sekwencji "po drzewie". Najprostszym jest markowowski model ewolucji zwany modelem CF (por. [3]). Zakładamy w nim, że każde miejsce w sekwencji może przyjmować tylko 2 stany. Ustalamy korzeń drzewa i jednorodny rozkład prawdopodobieństwa na jego stanach (czyli stan korzenia może z prawdopodobieństwem $\frac{1}{2}$ być równy 0 lub 1). Dla każdej krawędzi drzewa e mamy ustalone prawdopodobieństwo mutacji po tej krawędzi $p(e)$, gdzie $0 < p(e) < \frac{1}{2}$. Niech p oznacza funkcję przypisującą prawdopodobieństwa mutacji krawędziom drzewa T ($p : E(T) \to (0, \frac{1}{2})$). Czyli każda krawędź drzewa odpowiada łańcuchowi Markowa o macierzy przejścia:

$$M(e) = \begin{pmatrix} 1-p(e) & p(e) \\ p(e) & 1-p(e) \end{pmatrix}$$

Zakładamy (oczywiście bardzo upraszczając), że każde miejsce w sekwencji ewoluuje niezależnie i identycznie zgodnie z powyższą rodziną łańcuchów.

1.2 Miara błędu topologii drzewa

Drzewa możemy reprezentować na wiele sposobów, w pewnych zastosowaniach jedne definicje są wygodniejsze od innych. Zastanówmy się jak możemy mierzyć dokładność metod rekonstrukcji drzewa. W literaturze przeważa miara oparta na **dwupodziałach** (*ang. bipartition)* drzewa. Dla drzewa o liściach etykietowanych zbiorem S każda krawędź indukuje podział S na dwa podzbiory (liście znajdujące się w dwóch spójnych składowych drzewa jakie powstaną po usunięciu tej krawędzi). Podział $S = A \cup B$ nazwiemy dwupodziałem i bedziemy oznaczać $A|B$. Jest oczywiste, że zbiór wszystkich dwupodziałow drzewa $\beta(T)$ reprezentuje je w sposób jednoznaczny i umożliwia rekonstrukcję drzewa w czasie wielomianowym.

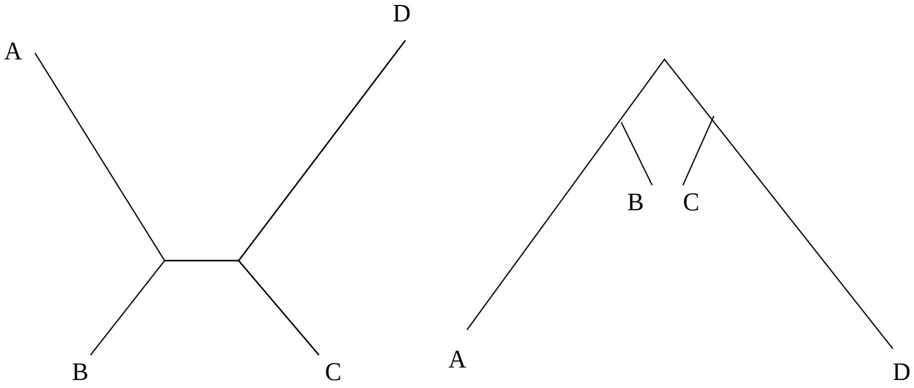
Załóżmy że badamy dokładność metody rekonstrukcji drzewa filogenetycznego, która przybliżyła prawdziwe drzewo T drzewem T^*. Jeśli $\beta(T) = \beta(T^*)$ nasza metoda bezbłednie znalazła poprawną topologię. Kiedy $\beta(T) \neq \beta(T^*)$ możemy wyróżnić dwa rodzaje błędów:

- dwupodziały należące do $\beta(T^*) \setminus \beta(T)$ nazywamy **fałszywymi pozytywami** (FP);
- natomiast dwupodziały należące do $\beta(T) \setminus \beta(T^*)$ to **fałszywe negatywy** (FN).

Jako miarę dokładności metody rekonstrukcji drzewa przyjmuje się najczęściej FN lub sumę FN + FP. Zauważmy, że dla drzew binarnych metoda dająca FP = 0 może zwracać zupełnie bezwartościowe drzewo (gwiazdę).

2 Metoda parsymonii nie jest zgodna

Rozważmy prosty przykład, dzięki któremu zauważono, że metoda parsymonii może nie być zgodna kiedy pozwalamy na różną częstotliwość zmian ewolucyjnych dla różnych krawędzi w drzewie [1]. Załóżmy, że występują



Rysunek 1: Niezgodność w metodzie MP: równoległa mutacja na dwóch długich krawędziach

w drzewie dwie bardzo długie gałezie o liściach A oraz D, natomiast dwie pozostałe są dużo krótsze. Długość krawędzi odpowiada w tym przykładzie prawdopodobieństwu mutacji "po tej krawędzi", czyli im krawędź dłuższa tym większe prawdopodobieństwo mutacji. Sekwencje na końcach krawędzi B i C mają dużą szansę pozostać bardzo podobne do ich wspólnego przodka, czyli również podobne do siebie nawzajem. Z drugiej strony podobieństwo sekwencji na końcach długich krawędzi do swoich sąsiadów w drzewie jest dość trudne do oszacowania. Okazuje się, że zjawisko równoległych mutacji sprawia, że z dużym prawdopodobieństwem stają się one bardziej podobne do siebie niż do B lub C. Kryterium parsymonii zaproponuje w takim przypadku zupełnie błedną klasyfikacją B i C do jednego klastra. Drzewa o takiej i podobnej strukturze należą do tak zwanego obszaru Felsensteina, na którym metoda parsymonii nie jest zgodna. Dla naszego przykładu czterolistnego drzewa wraz ze wzrostem długości rozważanych sekwencji rośnie prawdopodobieństwo podania przez metodę parsymonii błędnego drzewa ((AD)(BC)) zamiast ((AB)(CD)).

Możnaby podejrzewać, że analizowany przez nas przykład drzewa jest dosyć patologiczny (duże różnice w długościach krawędzi); niestety pokazano, że istnieją drzewa o pięciu liściach spełniające założenia zegara molekularnego, dla których metoda parsymonii zwraca błędne wyniki.

3 Metody rekonstrukcji drzew z macierzy odległości

Przypomnijmy, że addytywna macierz D jest równoważna drzewu, w którym długość najkrótszej scieżki między i oraz j wynosi D_{ij}. Umówmy się, że każda symetryczna dodatnia macierz o zerowych elementach diagonalnych jest przez nas uważana za macierz odległości. Metoda konstruująca drzewo na podstawie tej macierzy odwzorowuje ją w macierz addytywną (czyli kodowane przez nią drzewo).

Zdefiniujmy dwie własności metody, o których udowodnimy że są wystarczające aby metoda była zgodna: kombinatoryczną zgodność i ciągłość.

Powiemy, że metoda M jest **kombinatorycznie zgodna** jeśli dla macierzy addytywnych D jest identycznością, czyli $M(D) = D$.

Większość metod jest kombinatorycznie zgodna z wyjątkiem tych, które starają się zrekonstruować drzewo ultrametryczne. Metoda poszukująca drzewa ultrametrycznego dla macierzy addytywnej, może zmodyfikować tę macierz, czyli również zmienić topologię drzewa.

Definicja ciągłości metody uwzględnia metrykę, wględem której metoda jest ciągła. Załóżmy, że pracujemy z metryką L_∞ zdefiniowaną następująco:

$$L_\infty(D, D') = \max_{ij} |D_{ij} - D'_{ij}|.$$

Powiemy, że metoda macierzowa M jest **ciągła** dla macierzy D jeżeli:

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad (L_\infty(D, D') < \delta \implies L_\infty(M(D), M(D')) < \epsilon).$$

Definicja 1 *Powiemy, że metoda macierzowa jest* **rozsądna** *jeżeli jest jednocześnie kombinatorycznie zgodna oraz ciągła dla macierzy addytywnych odpowiadających drzewom binarnym.*

3.1 Nieskończone sekwencje

Celem metod macierzowych jest skonstruowanie na podstawie zbioru sekwencji addytywnej macierzy kodującej ważone drzewo, w którego liściach znajdują sią te sekwencje. Okazuje się, że wyliczenie addytywnych odległości pomiędzy sekwencjami jest niemożliwe, kiedy dysponujemy danymi o ograniczonej długości. Możemy natomiast tak zdefiniować odległości pomiędzy sekwencjami, że wraz ze wzrostem długości coraz lepiej aproksymują one odległości addytywne. Ta definicja odległości opiera się na pojęciu *transformacji korygującej odległość*. Odległości poddane takiej transformacji spełniają następującą własność: *dla rosnących długości sekwencji macierz odległości zbiega do macierzy addytywnej.*

Zastosowanie transformacji korygującej zapewnia, że metody macierzowe są zgodne. Rozważmy przykład takiej transformacji dla modelu CF. Potrzebny nam będzie prosty do udowodnienia fakt:

Fakt 1 *W modelu CF prawdopodobieństwo mutacji p między dwoma wierzchołkami połączonymi najkrótszą ścieżką $p_1, p_2, \ldots p_k$ wyraża się następująco:*

$$p = \frac{1}{2}\left(1 - \prod_{i=1}^{k}(1 - 2p_i)\right). \tag{1}$$

Mając dane drzewo T w modelu CF o sekwencjach długości k w liściach, przez $H(i,j)$ oznaczymy odległości Hamminga pomiędzy sekwencjami i oraz j. Niech $h^{ij} = H(i,j)/k$ oznacza średnie zróżnicowanie sekwencji. Oznaczmy przez E^{ij} prawdopodobieństwo mutacji danego miejsca pomiędzy sekwencjami i oraz j (czyli wartość oczekiwaną h^{ij}).

Macierz D tworzymy następująco: skorygowana odległość

$$D_{ij} = -\frac{1}{2}\log(1 - 2E^{ij}).$$

Jeśli przypiszemy teraz każdej krawędzi e drzewa T dodatnią wagę:

$$w(e) = -\frac{1}{2}\log(1 - p(e)),$$

na podstawie (1) widzimy, że D_{ij} jest dokładnie równe sumie wag krawędzi na ścieżce $i \rightsquigarrow j = \{p_1, p_2, \ldots p_k\}$:

$$D_{ij} = -\frac{1}{2}\log\left(1 - 2\left(\frac{1}{2}\left(1 - \prod_{j=1}^{k}(1-2p_j)\right)\right)\right) =$$

$$= -\frac{1}{2}\log\left(\prod_{j=1}^{k}(1-p_j)\right) = \sum_{e\in(i\rightsquigarrow j)} w(e).$$

Pokazaliśmy, że kombinatorycznie zgodna metoda zastosowana do macierzy odległości policzonej na podstawie dowolnie długich sekwencji skorygowanych odpowiednią dla danego modelu ewolucji transformacją z prawdopodobieństwem równym 1 zrekonstruuje poprawną topologię.

Niestety w rzeczywistości dysponujemy sekwencjami o skończonej i niezbyt wielkiej długości; musimy teraz zastanowić się kiedy metoda osiąga poprawną topologię dla skończonych sekwencji. Do tego niezbędna okazuje się własność ciągłości metody, czyli będziemy zajmować się metodami rozsądnymi według Definicji 1. Przytoczone poniżej twierdzenia pochodzą z pracy [2].

3.2 Otoczenia macierzy addytywnych

Chcemy udowodnić, że każda rozsądna metoda macierzowa jest zgodną metodą rekonstruującą drzewa binarne. W tym celu zastanowimy się kiedy dwie różne macierze addytywne definiują tę samą topologię drzewa.

Przypomnijmy następujące twierdzenie, zwane warunkiem czterech punktów Bunemana:

Twierdzenie 1 *Macierz D jest addytywna wttw dla każdej czwórki i, j, k, l, maksimum z trzech sum $D_{ij} + D_{kl}$, $D_{ik} + D_{jl}$ oraz $D_{il} + D_{jk}$ jest osiągane przez co najmniej dwie spośród nich.*

Jako wniosek z twierdzenia otrzymujemy wielomianowy (chociaż nie optymalny) algorytm konstruujący drzewo na podstawie macierzy addytywnej. Wystarczy rozważyć każdą czwórkę liści — zależność:

$$D_{ij} + D_{kl} < D_{ik} + D_{jl} = D_{ij} + D_{jk}$$

oznacza, że topologia poddrzewa indukowanego przez te liście jest $ij|kl$ (czyli ścieżka $i \rightsquigarrow j$ nie przecina ścieżki $k \rightsquigarrow l$). Teraz wystarczy poskładać te czwórkowe "puzzle" w jedno drzewo, zaczynając od znalezienia sąsiednich liści, potem poddrzew, etc.

Następne twierdzenie podaje warunki przy których dwie macierze addytywne definiują to samo drzewo (topologię):

Twierdzenie 2 *Dwie macierze addytywne D i D' definiują tę samą topologię wttw dla każdej czwórki i, j, k, l, $D_{ij}+D_{kl}$ jest minimalną sumą spośród trójki sum par odległości wttw $D'_{ij}+D'_{kl}$ jest minimalna z odpowiednich sum w macierzy D'.*

Zaskakującym wnioskiem z tego twierdzenia jest fakt, że dookoła każdej macierzy addytywnej definiującej drzewo binarne istnieje pewne otoczenie, w którym wszystkie macierze addytywne definiują tę samą topologię.

Twierdzenie 3 *Niech D będzie macierzą addytywną definiującą ważone drzewo binarne T, oraz niech x będzie najmniejszą wagą krawędzi w T. Niech D' będzie inną macierzą addytywną definiującą drzewo T'. Jeśli $L_\infty(D, D') < \frac{x}{2}$ to topologie T i T' są identyczne.*

Potrafimy teraz udowodnić główne twierdzenie:

Twierdzenie 4 *Każda rozsądna metoda konstruująca drzewo na podstawie macierzy odległości jest zgodną metodą dla drzew binarnych.*

Ażeby udowodnić powyższe twierdzenie załóżmy, że T jest drzewem binarnym, które chcemy zrekonstruować (na podstawie macierzy odległości D), a D' definiującą je macierzą addytywną. Niech x będzie wagą najlżejszej krawędzi w T, a M rozważaną metoda, która jest zarówno kombinatorycznie zgodna jak i ciągła dla macierzy addytywnych definiujących drzewa binarne. Z ciągłości M wiemy, że istnieje δ spełniająca dla macierzy D':

$$L_\infty(D, D') \leq \delta \quad \Longrightarrow \quad L_\infty(M(D), M(D')) < \frac{x}{2}.$$

Metoda M przekształca macierze w macierze addytywne czyli mamy zagwarantowane, że $M(D)$ ma taką samą topologię co $M(D')$. Z kolei wiemy, że

M jest kombinatorycznie zgodna czyli $M(D') = D'$. Reasumując $M(D)$ jest macierzą addytywną, która definiuje topologię szukanego drzewa T.

Udało nam się pokazać że większość metod, które konstruują macierze addytywne jest zgodna (wyjątek stanowią metody konstruujące macierze ultrametryczne). O wiele trudniej zbadać szybkość zbieżności tych metod, tzn jak długie powinny być sekwencje, ażeby nasza metoda znajdowała macierz addytywną z odpowiednio małego otoczenia macierzy, której poszukujemy.

Dowód ostatniego twierdzenia sugeruje metodę porównywania szybkości zbieżności różnych metod: dla każdej metody M i drzewa binarnego T oraz jego macierzy addytywnej D', istnieje pewne dodatnie otoczenie δ, takie że dla każdej macierzy odległości $D_0 \in N(D', \delta) = \{D : L_\infty(D, D') < \delta\}$, $M(D_0)$ konstruuje macierz addytywną definiującą topologię T. Możemy stąd wnioskować, że im większe jest maksymalne δ o tej własności, tym łatwiej metoda osiąga poprawną topologię.

4 Szybkość zbieżności metod macierzowych

Ograniczymy się do rozważenia metody Neighbor Joining. Przytoczymy twierdzenie z pracy [2] pozwalające szacować szybkość zbieżności tej metody:

Twierdzenie 5 *Niech T będzie drzewem w modelu CF, w którym prawdopodobieństwa mutacji krawędzi należą do przedziału $[f, g]$. Wtedy dla każdego $\epsilon > 0$ istnieje stała c, taka że dla n sekwencji długości k w liściach drzewa T, gdzie*

$$k > \frac{c \cdot \log n}{f^2(1-2g)^{2\,diam(T)}},$$

z prawdopodobieństwem $1-\epsilon$ algorytm Neighbor-Joining zastosowany do skorygowanych odległości zrekonstruuje poprawną topologię.

Średnica drzewa T, którą oznaczamy przez $\mathrm{diam}(T)$ jest zdefiniowana jako liczba krawędzi w najdłuższej ścieżce w drzewie T. Jest oczywste, że średnica małych drzew jest mała, a większych większa. Obliczono jaką jest oczekiwana średnica losowego drzewa: wynosi ona $\mathcal{O}(\sqrt{n})$ dla rozkładu jednostajnego, oraz $\mathcal{O}(\log n)$ dla rozkładu Yule-Hardinga. Czyli długość sekwencji, które gwarantują na podstawie Twierdzenia 5 poprawne wyniki może rosnąć szybciej niż wielomianowo wraz z rosnącą liczbą gatunków (liści). Okazuje się, że te dość pesymistyczne oszacowania nie znajdują potwierdzenia w praktyce. Wykonano wiele symulacji, które pokazują dużo szybszą, niż udowodniona analitycznie, zbieżność.

5 Podsumowanie i problemy otwarte

Podsumujmy naszą dyskusję następującymi wnioskami i nadal satysfakcjonująco nierozwiązanymi problemami.

- Znalezienie optymalnego rozwiązania w metodzie parsymonii jest NP trudne. Poza tym metoda nie jest zgodna czyli nawet jeśli potrafimy znajdować rozwiązania optymalne, to istnieją drzewa, dla których metoda parsymonii zrekonstruuje niepoprawną topologię z prawdopodobieństwem 1 nawet mając do dyspozycji dowolnie długie sekwencje.
- Metoda Neighbor-Joining jest wielomianowa, ale analityczne oszacowania szybkości zbieżności pokazują, że nie gwarantuje ona poprawnej rekonstrukcji drzewa dla sekwencji o realistycznej długości.
- Metoda Maximum Likelihood (największej wiarygodności) jest dla wielu modeli zgodną metodą estymacji, natomiast jej zastosowania są bardzo ograniczone ze względu na ogromną złożoność obliczeniową.
- Pokazaliśmy przykład oszacowania szybkości zbieżności dla metody NJ. Ciekawym zagadnieniem jest polepszenie tych oszacowań, które w porównaniu do wyników eksperymentalnych wydają się bardzo zgrubne.
- Ważnym problemem otwartym jest oszacowanie błędu topologii jaki popełnia dana metoda rekonstrukcji drzewa działając na niedostatecznie długich sekwencjach.
- Wszystkie oszacowania szybkości zbieżności spotykane w literaturze dotyczą metod macierzowych (opartych o macierze odległości); nie są znane narzędzia pozwalające szacować szybkość zbieżności dla innych metod rekonstrukcji drzew (np. metody parsymonii w jej obszarze zbieżności).

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] P. Erdoös, M. Steel, L. Szekely, and T. Warnow. Inferring big trees from short sequences. In *ICALP*, 1997.

[3] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation. 2000.

6 Zadania

1. Zaproponuj inną niż FP + FN miarę błędu topologii drzewa.
2. Udowodnij Fakt 1 (przez indukcję po długości ścieżki).
3. Zaprojektuj algorytm rekonstrukcji drzewa na podstawie topologii jego czwórek.
4. Udowodnij Twierdzenia 2 oraz 3.