Drzewa filogenetyczne

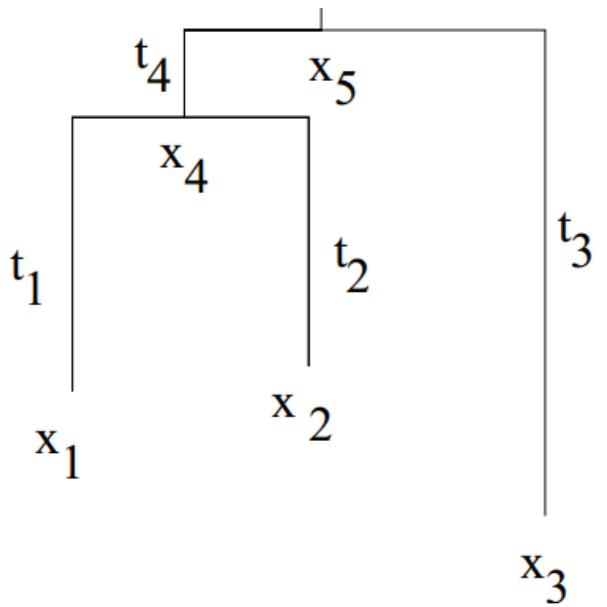
WBO 2015/2016

Paweł Górecki

Wersja 1.05

Maksymalizacja wiarygodności (ML)

- Drzewo ukorzenione binarne T o węzłach $\{1, 2, \ldots, 2n-1\}$ (węzły $1, 2, \ldots, n$ to liście, $2n-1$ oznacza korzeń) wraz z długościami krawędzi $t = \{t_1, t_2, \ldots, t_{2n-2}\}$, gdzie t_i oznacza długość krawędzi łączącej węzeł i z jego ojcem.
- Sekwencje $X = \{x_1, x_2, \ldots x_{2n-1}\}$ są przypisane do węzłów drzewa T, gdzie x_i to sekwencja skojarzona z węzłem i.



Obrazek: J.Tiuryn

Maksymalizacja wiarygodności (ML)

Zadanie: oblicz

$$P(x|T,t)$$

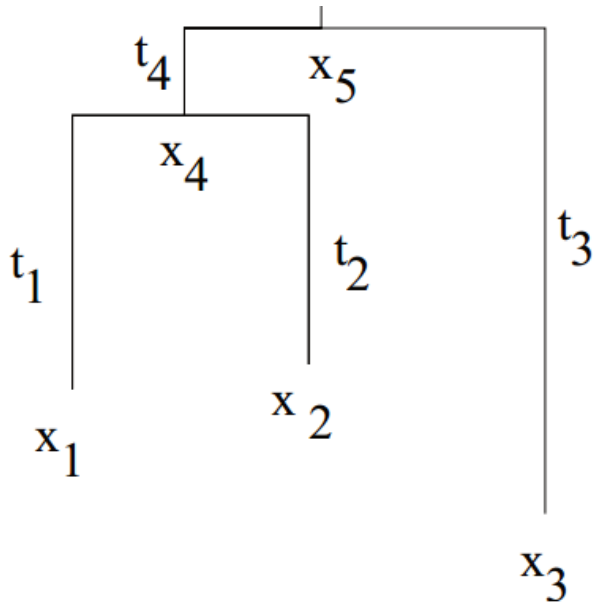
tj. p-stwo warunkowe zbioru X przy danych T i t.

To p-stwo obliczamy wg wzoru:

$$P(x|T,t) := P(x_{2n-1}) \prod_{i<2n-1} P(x_i|x_{p(i)},t_i),$$

gdzie $p(i)$ to ojciec węzła i, $P(a|b,t)$ oznacza p-stwo przekształcenia sekwencji b w a w czasie t, a $P(x_{2n-1})$ to p-stwo sekwencji x_{2n-1} w korzeniu drzewa.

Maksymalizacja wiarygodności (ML)



$$P(x|T,t) = P(x_1|x_4,t_1)P(x_2|x_4,t_2)P(x_3|x_5,t_3)P(x_4|x_5,t_4)P(x_5)$$

Obrazek: J.Tiuryn

Maksymalizacja wiarygodności (ML)

W praktyce nie znamy T, t ani sekwencji przypisanych do węzłów wewnętrznych. Zaczniemy od prostego przypadku gdy znamy drzewo i sekwencje na liściach.

Założenia:

- pozycje są niezależne,
- znamy model ewolucji sekwencji (np. JC69).

Wówczas (u to pozycja):

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

tutaj $P(x_u|y_u,t)$ obliczamy z modelu ewolucji sekwencji.

Maksymalizacja wiarygodności (ML)

Ostatecznie p-stwo $P(x_1, x_2, \ldots, x_n | T, t)$, gdzie T to drzewo binarne o n liściach, wynosi:

$$\sum_{a_{n+1}, a_{n+1}, \ldots, a_{2n-1}} q_{a_{2n-1}} \prod_{i=n+1}^{2n-2} P(a_i | a_{p(i)}, t_i) \prod_{i=1}^{n} P(x_i | a_{p(i)}, t_i)$$

Pytanie: jaka jest złożoność obliczania tego p-stwa? Czy można to poprawić?

Algorytm Felsensteina

Wiarygodność pozycji u to

$$P(x_u|T,t) = \sum_a P(L_{2n-1}|a)q_a,$$

gdzie q_a to p-stwo stacjonarne nukleotydu a oraz $P(L_k|a)$ obliczane jest wg:

- jeśli k to liść to $P(L_k|a) := 1$ jeśli na pozycji u w sekwencji k-tej jest a wpp $P(L_k|a) := 0$.
- wpp niech i oraz j to synowie k, dla każdego nukleotydu a:

$$P(L_k|a) := \sum_{b,c} P(b|a,t_i)P(L_i|b)P(c|a,t_i)P(L_j|c).$$

Ostatecznie

$$P(x_1, x_2, \ldots, x_n | T, t) = \prod_u P(x_u | T, t).$$

Pytanie: jaka jest złożoność obliczania tego p-stwa? Porównaj z poprzednim wynikiem.

Co z korzeniem?

Poznane do tej pory modele spełniają założenia:

- model jest multiplikatywny:

$$\sum_b P(a|b,t)P(b|c,t') = P(a|c,t+t'),$$

- oraz model jest odwracalny:

$$P(a|b,t)q_b = P(b|a,t)q_b.$$

Z tego wynika, że pozycja korzenia nie ma znaczenia - liczy się tylko sumaryczna odległość korzenia od jego dzieci. Wniosek: w obliczeniach korzeń możemy przesunąć do jednego z dzieci. Ponadto drzewa obliczane metodą ML są **nieukorzenione.**

ML - podsumowanie

W praktyce nie znamy T i długości krawędzi. Następujący problem jest NP-trudny:

- Dany zbiór sekwencji uliniowionych $x_1, \ldots, x_n$.
- Znajdź drzewo T z długościami krawędzi t, które maksymalizuje p-stwo $P(x_1, \ldots, x_n | T, t)$.

W konsekwencji rekonstrukcja drzewa z uliniowienia jest realizowana za pomocą heurystyk (programy np. dnaml, proml z pakietu Phylip, phyml i wiele innych).