Fast Tree Distance Technical Details

Xinqi Li, Dec 2021

Profile Distance

The profile distance at each position is the average dissimilarity of the characters. The uncorrected distance between two profiles is then the average of these position-wise distance, weighted by the product of the proportion of nongaps in each of the two profiles.

- The distance between the profiles at position l

  $$\Delta_l(A,B)=\sum_{\alpha}\sum_{\beta}f_{Al}(\alpha)f_{Bl}(\beta)D(\alpha,\beta)$$

  , where D is the dissimilarity matrix on characters, $f_{Al}(\alpha)$ is the frequency of character α in the profile of A at position l
- The profie distance is weighted at each position:

  $$\Delta(A,B)=\frac{\sum_{l=1}^{L}\Delta_l(A,B)w_l(A)w_l(B)}{\sum_{l=1}^{L}w_l(A)w_l(B)}$$

  , where $w_l(A)$ is the proportion of non-gaps in the profile of A at position l. For example,

  ```
  Profile(A)    ACGTACGTACGT
  Profile(B)    A-CGACGTAC-T

  Profile(AB)   A[ccg/-gt]ACGTAC[g/-]T
  ```

  $w(A)=[1,1,1,1,1,1,1,1,1,1,1,1,1]$
  $w(B)=[1,0,1,1,1,1,1,1,1,1,1,0,1]$
  $w(AB)=[1,1/2,1,1,1,1,1,1,1,1,1,1/2,1]$

FastTree don't compute the correct distance from the profiles. Instead, it corrects distance when merge profiles.

Distance between Internal Nodes

$$d_u(i,j)=\Delta(i,j)-u(i)-u(j)$$

$u(i)$ is the up-distance of node i, $\Delta(i,j)$ is the profile distance.

- Corrected distance is the Jukes-Cantor distance: $d = -\frac{3}{4} log(1 - \frac{4}{3} d_u)$. Turncates the corrected distance to maximum of 3.0 substitutions per site, and for sequences that do not overlap because of gaps, Fasttree uses this maximum distance.

Out-distance $r(i)$

- Without gaps:

$$r(i) = \frac{\sum_{j \neq i} d_u(i,j)}{n-2}$$

$$\sum_{j \neq i} d_u(i,j) = \sum_{j \neq i} (\Delta(i,j) - u(i) - u(j))$$

,where

$$\sum_{j \neq i} \Delta(i,j) = n\Delta(i,T) - \Delta(i,i)$$

T is the total profile - the average of all active nodes' profiles. $\Delta(i,i)$ is the average distance between children of i, including self-comparisons. Then we get

$$\sum_{j \neq i} d_u(i,j) = n\Delta(i,T) - \Delta(i,i) - (n-2)u(i) - \sum_j u(j)$$

, $\sum_j u(j)$ is the total up-distance(sum of up distance of all active nodes).

- With gaps:

$$\sum_{j \neq i} \Delta(i,j) = (n-1)\Delta(i, T-i)$$

$$\Delta(i, T-i) = \frac{\sum_j \sum_{l=1}^{L} \Delta_l(i,j) w_l(i) w_l(j)}{\sum_{j \neq i} \sum_{l=1}^{L} w_l(i) w_l(j)}$$

  Then we can get

$$\sum_{j \neq i} d_u(i,j) = (n-1)\Delta(i, T-i) - (n-2)u(i) - \sum_j u(j)$$

Update Profile & Nodes

No weight, set $\lambda = 1/2$.

Update Weight λ

Compute the "variances" that are used to computed the weights of the joins and use a "variance correction" $v(i)$ analogous to the up-distance.
$v(i) = 0$ for leaves and the variance values for pairs of leaves:

$$V(l_1, l_2) = d_u(l_1, l_2)$$

The variance of internal nodes is

$$V(ij, k) = \lambda V(i, k) + (1 - \lambda)V(j, k) - \lambda(1 - \lambda)V(i, j)$$

$$v(ij) = \lambda v(i) + (1 - \lambda)v(j) + \lambda(1 - \lambda)v(i, j)$$

Given these variances, we weights the join of i, j so as to minimize the variance of the distance estimates for the new node ij,

$$\lambda = \frac{1}{2} + \frac{\sum_{k \neq i,j}(V(j, k) - V(i, k))}{2(n - 2)V(i, j)}$$

$$V(i, j) = \Delta(i, j) - v(i) - v(j)$$

$$\sum_{k \neq i,j}(V(j, k) - V(i, k)) = (n - 2)(v(i) - v(j)) + \sum_{k \neq i,j}\Delta(j, k) - \sum_{k \neq i,j}\Delta(i, k)$$

, where n is the number of active nodes before the join takes place.

Update Profiles

Join node i and j, for the parent node ij, the profile is a weighted average, with weight λ

$$P_l(ij) = \lambda P_l(i) + (1 - \lambda)P_l(j)$$

Update up-distance $r(i)$

$$u(ij) = \lambda(u(i) + d_u(i, ij)) + (1 - \lambda)(u(j) + d_u(j, ij))$$

, where $u(i) = 0$ for all leaves.

$$d_u(ij, i) = \frac{d_u(i, j) + r(i) - r(j)}{2}$$