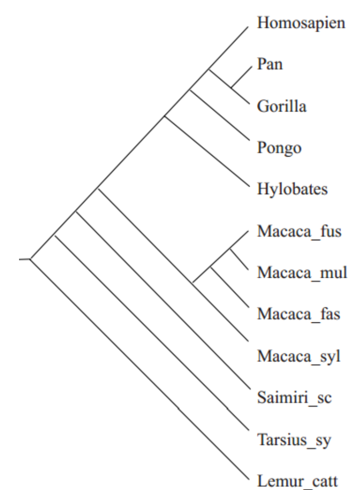
Averaging Phylogenetic Trees

Presenter: Suyi Wang

Phylogenetic Tree Space

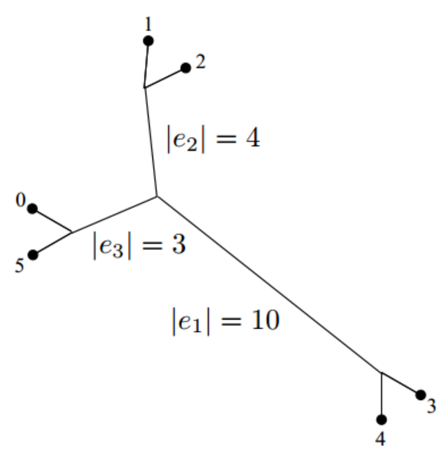
- A phylogenetic n-tree is:
  - A Tree with n leaves
  - Leaves: represent distinguishable species
  - Interior vertices: #degree >= 3
  - Edges: weighted
    - Max: 2*n – 1 edges

```
'Lemur_catta'      AAGCTTCATAGG
'Tarsius_syrichta'AAGTTTCATTGG
'Saimiri_sciureus'AAGCTTCACCGG
'Macaca_sylvanus' AAGCTTCTCCGG
'Macaca_fascicul.'AAGCTTCTCCGG
'Macaca_mulatta'  AAGCTTTTCTGG
'Macaca_fuscata'  AAGCTTTTCCGG
'Hylobates'       AAGCTTTACAGG
'Pongo'           AAGCTTCACCGG
'Gorilla'         AAGCTTCACCGG
'Pan'             AAGCTTCACCGG
'Homo_sapiens'    AAGCTTCACCGG
```



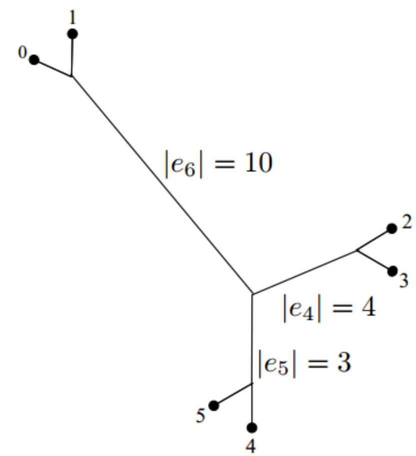
Phylogenetic Tree Space

- Splits:
  - Associates with an edge e
  - A partition (Xe, Xe*) of leaves



T

splits
$e_1 : \{0,1,2,5\} \,|\, \{3,4\}$
$e_2 : \{0,3,4,5\} \,|\, \{1,2\}$
$e_3 : \{0,5\} \,|\, \{1,2,3,4\}$

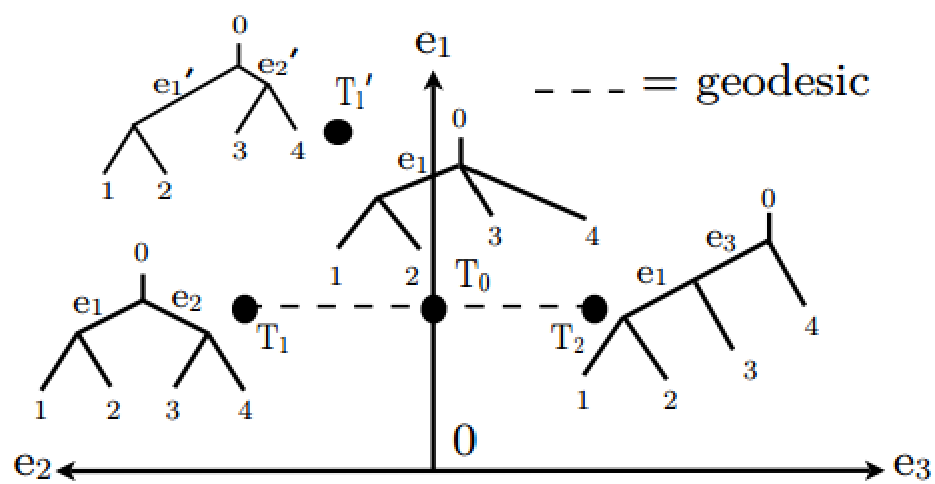


T'

splits
$e_4 : \{0,1,4,5\} \,|\, \{2,3\}$
$e_5 : \{0,1,2,3\} \,|\, \{4,5\}$
$e_6 : \{0,1\} \,|\, \{2,3,4,5\}$

Phylogenetic Tree Space

- Trees are uniquely defined by compatible set of splits [Semple 03]
  - Suggests a tree space
  - Axis – length of edges
  - 2n – 1 orthants



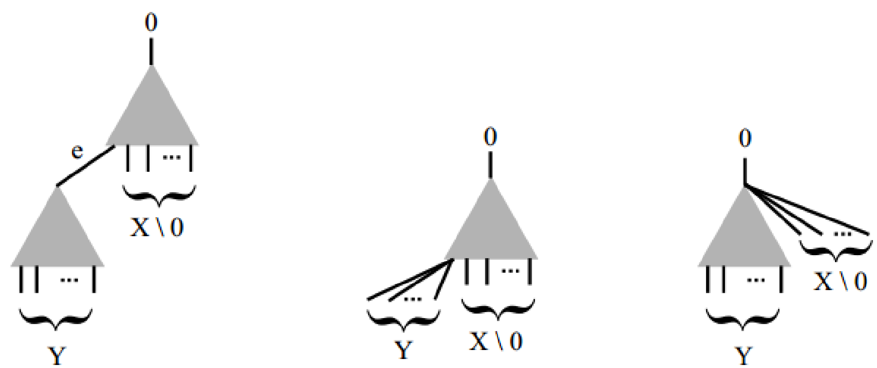
Geodesic distance

- Computing geodesic distances in tree space [Owen 07]
  - Shortest path from T1 to T2
  - Sum (sub path in each orthant)
  - Unique [Billera 01]

- Same orthant
  - Euclidean distance

Geodesic distance

- Trees share a common edge
  - [Vogtmann 07] [Owen 08]

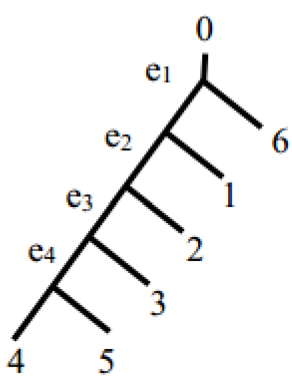


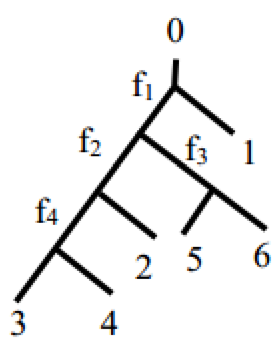
(a) Tree T_i. (b) Tree T_i^X. (c) Tree T_i^Y.

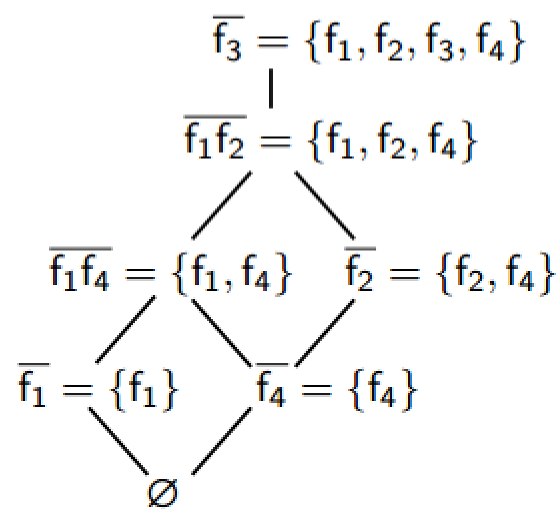
$$d(T_1, T_2) = \sqrt{d(T_1^X, T_2^X)^2 + d(T_1^Y, T_2^Y)^2 + (|e|_{T_1} - |e|_{T_2})^2}$$

Geodesic distance

- Trees share no common edge
  - A path space: $K(\Sigma_1, \Sigma_2)$
  - Try all possible orthants series.
  - A dynamic programming







Averaging in tree space

- Tree space T={T1, T2, ..., Tr}

- Variance S(X, T)
  - Sum of squares distances from X to each tree in T

- Mean
  - X that minimize the variance

Averaging in tree space

- Iterative approach [Strum 03]
  - Random sample on T for k times – {T1…Tk}
  - Current result u_i, i=[1, k]
  - u_i+1 = 1/(k+1) from u_i to Ti
  - Converge to mean
- A descent method [Millier 12]
  - To be continued…