The space of trees

David Epstein and Jon Ingram

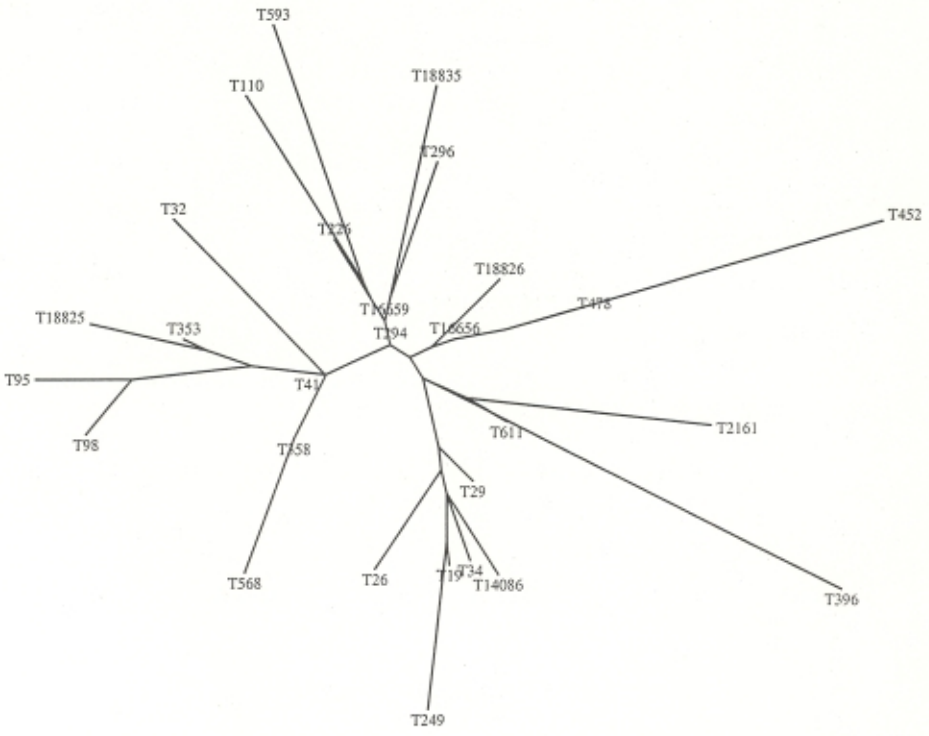
Mathematics Institute
University of Warwick
dbae@maths.warwick.ac.uk
jingram@maths.warwick.ac.uk

Working with Chris Dowson, microbiologist at Warwick. 28 strains 469 nucleotides (also called bases)
(Show only 64 bases, taken from middle of 469. Show only 2 strains)

First strain C T A T C A G C C G A C G
C T G G C T G A A G A A A T G G G C A
A G C T G C A [A] G A G C G C A T C A C
G T C G A C C A A G A A G G G

Second strain C T A T C A G C C G A C G
C T G G C T G A A G A A A T G G G C A
A G C T G C A [G] G a A C G C A T C A C
G T C G A C C A A G A A G G G

Use differences between sequences to make a "most likely" tree.



Five different sites give five different trees.

Applications

- Bill Martin (Düsseldorf, Botany).
  70 trees, each with 19 leaves.

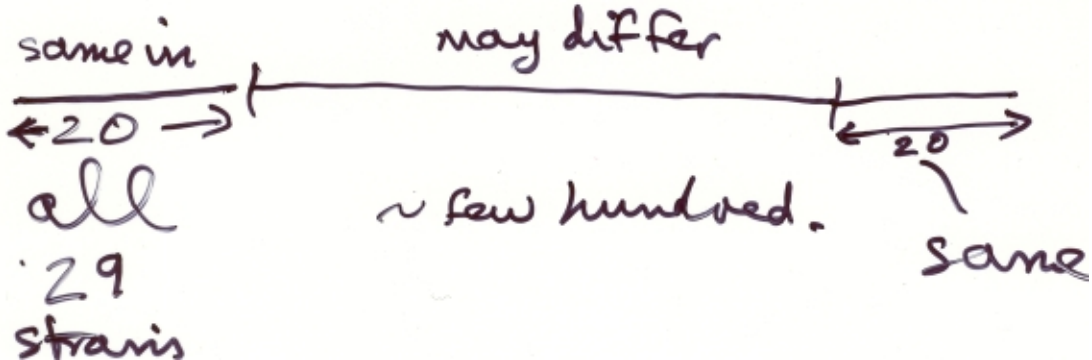
- Possible project: Automated search for horizontal transfer of DNA fragments.

- Multiple alignment: 19 sequences. *How reliable is the alignment?*

- Markov chain Monte Carlo. Measure variability of "best" trees.

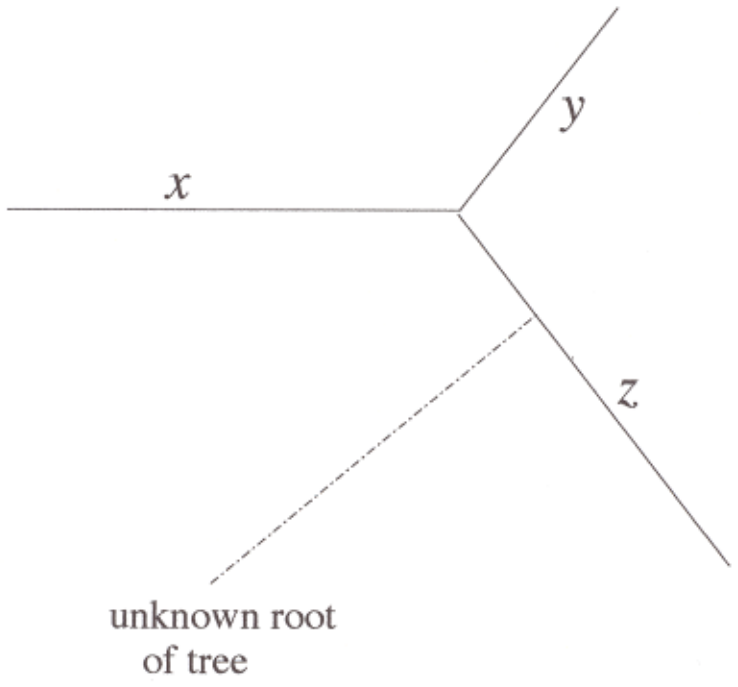
- Summary statistic from *a posteriori* distribution of possible trees.

S-1



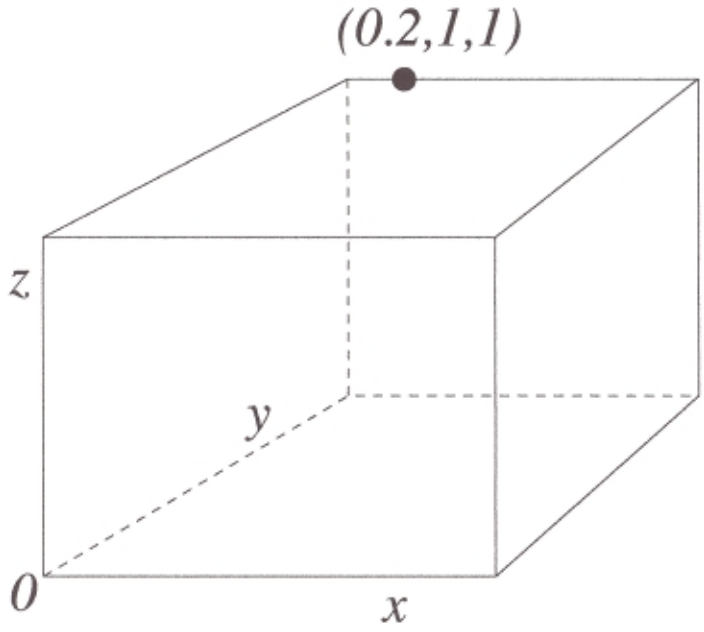
Three leaves

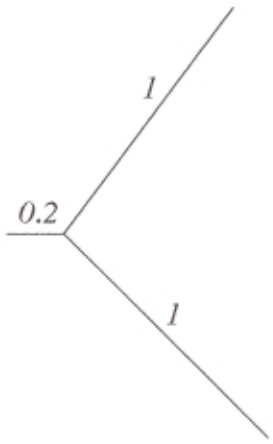
x, y and z are positive numbers representing *number of mutation events*.



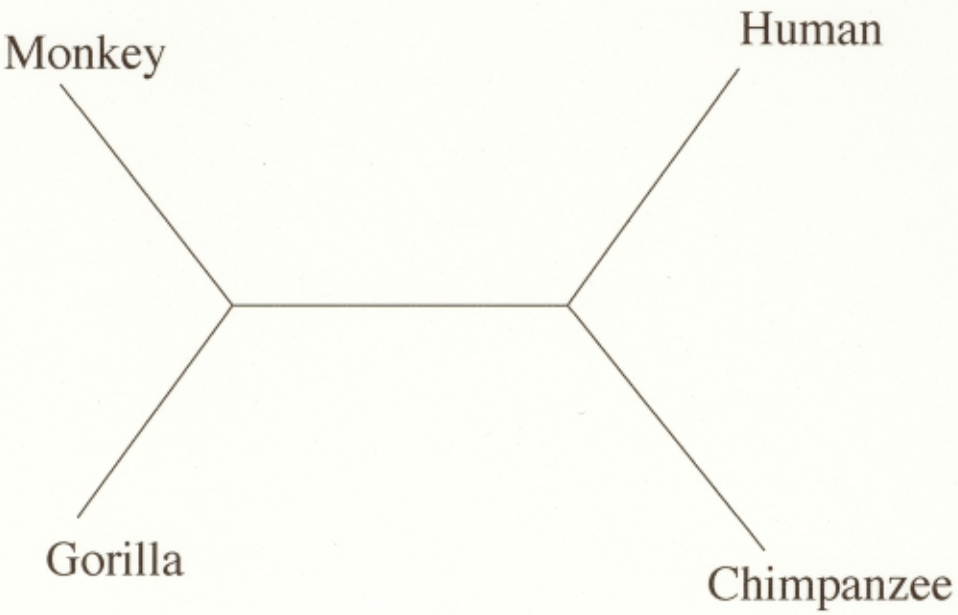
One tree is represented by one point (x, y, z) in ordinary (euclidean) space, with all three coordinates non-zero and positive.

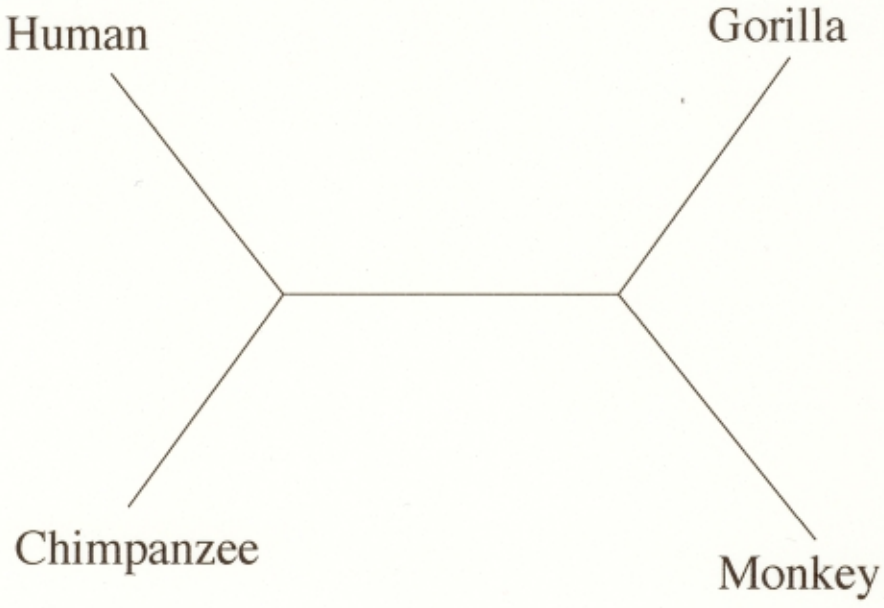
A single point in 3-dimensional space (left) represents a tree with three leaves (right)



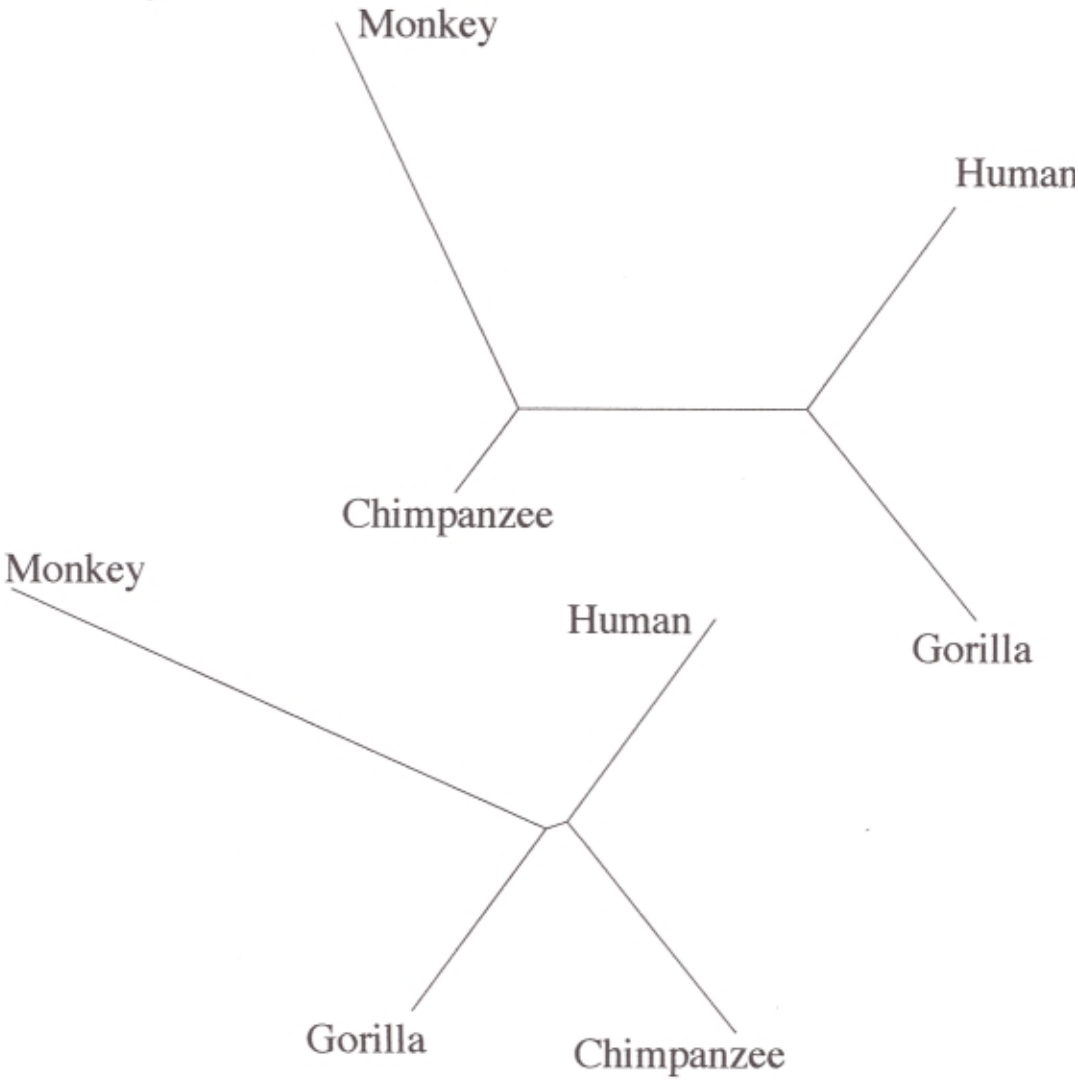


Two indistinguishable and identical trees:

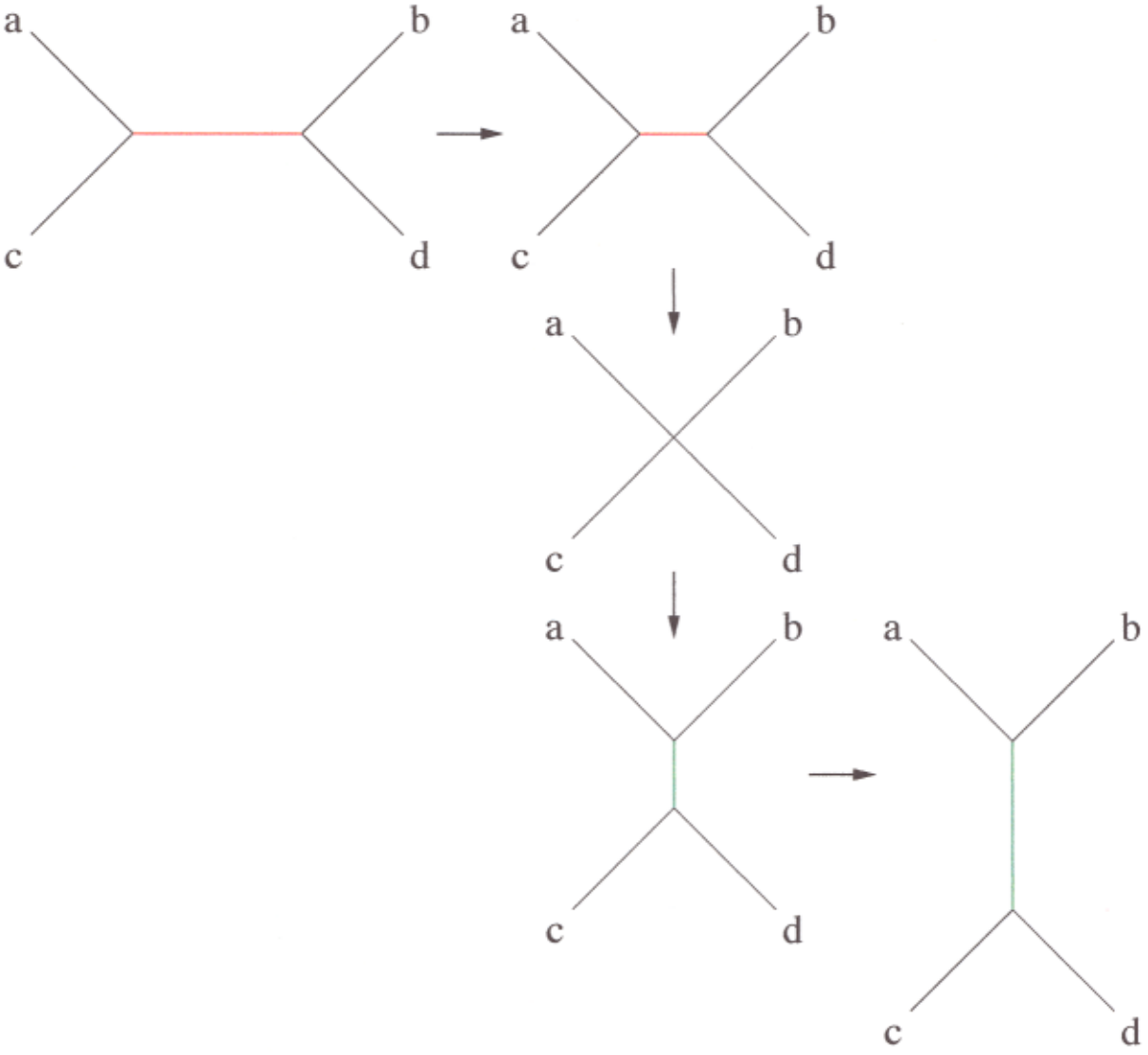




Two different trees:
How different?
Quantify!!

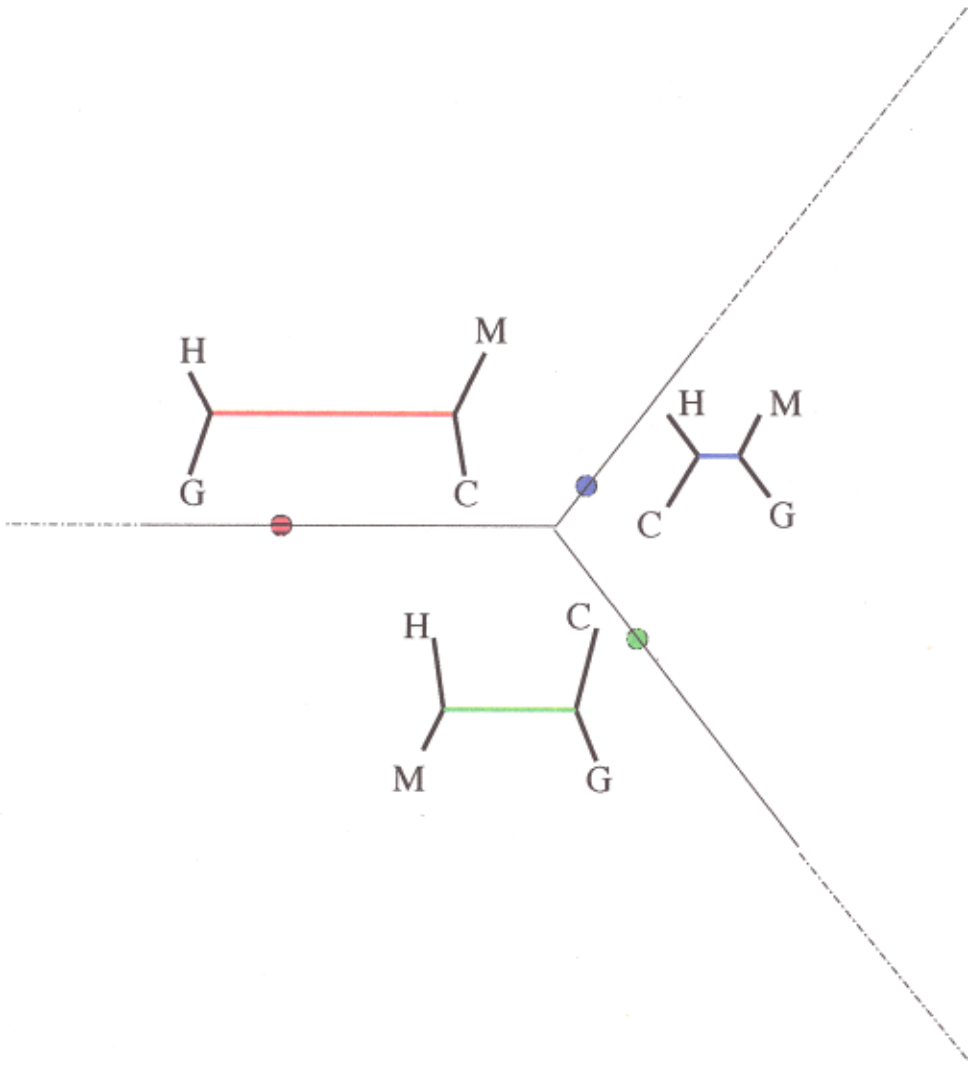


Moving from one tree to another.

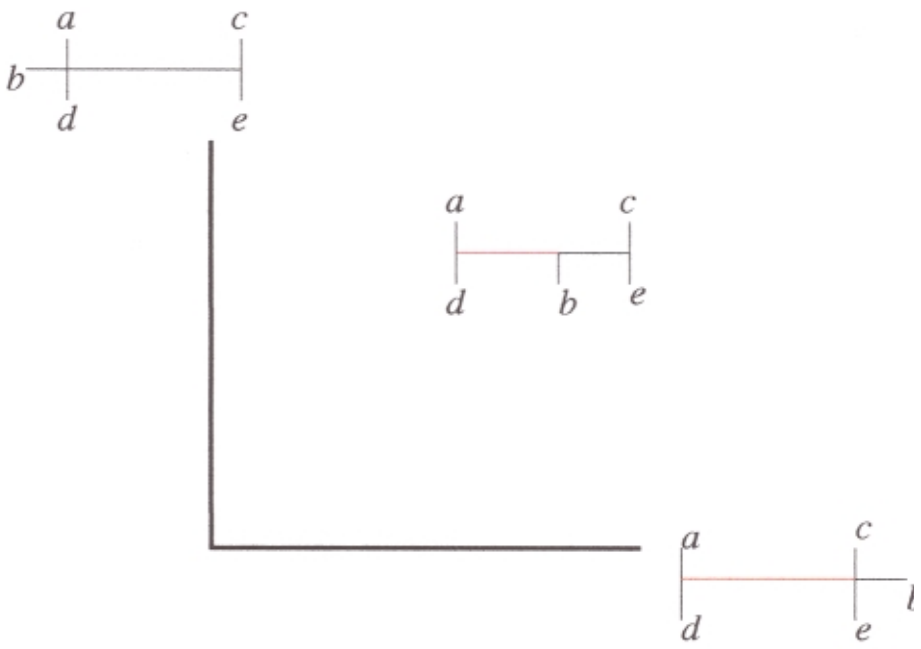


One internal edge

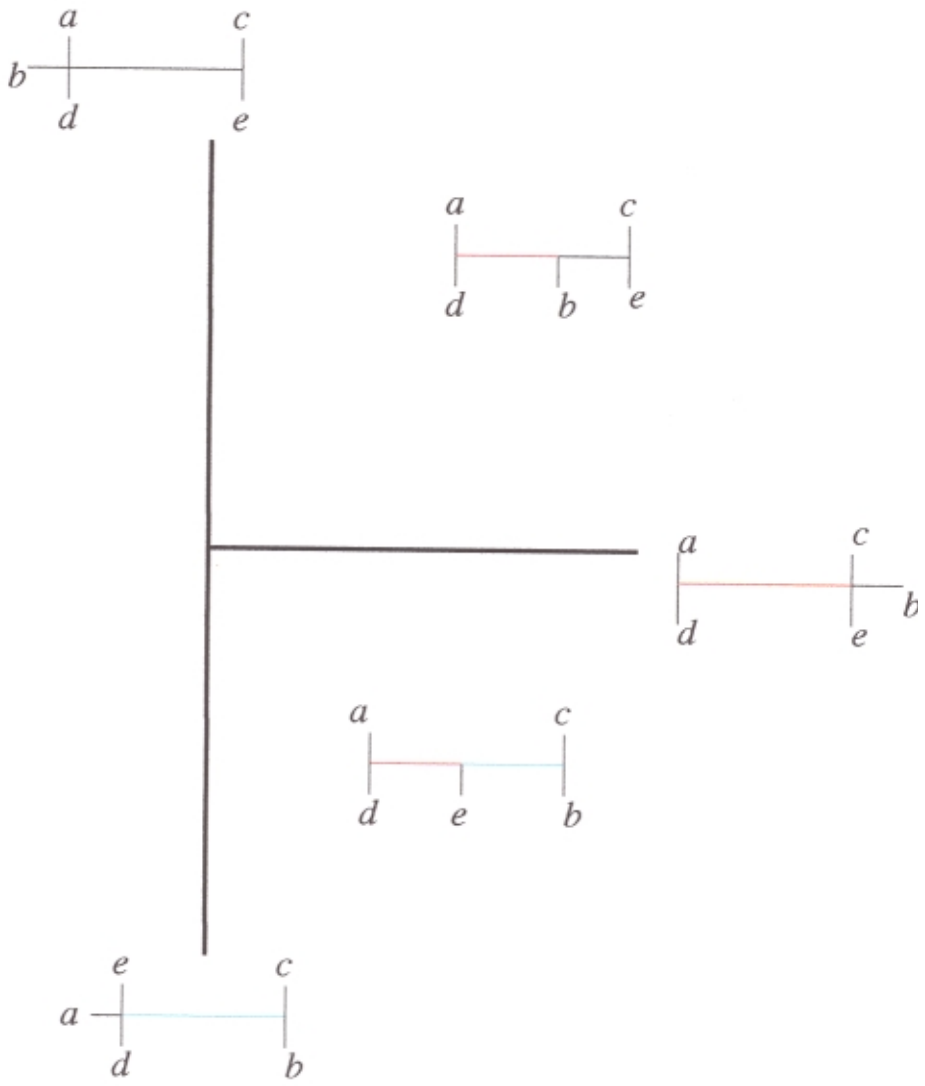
Suppress lengths of external edges.



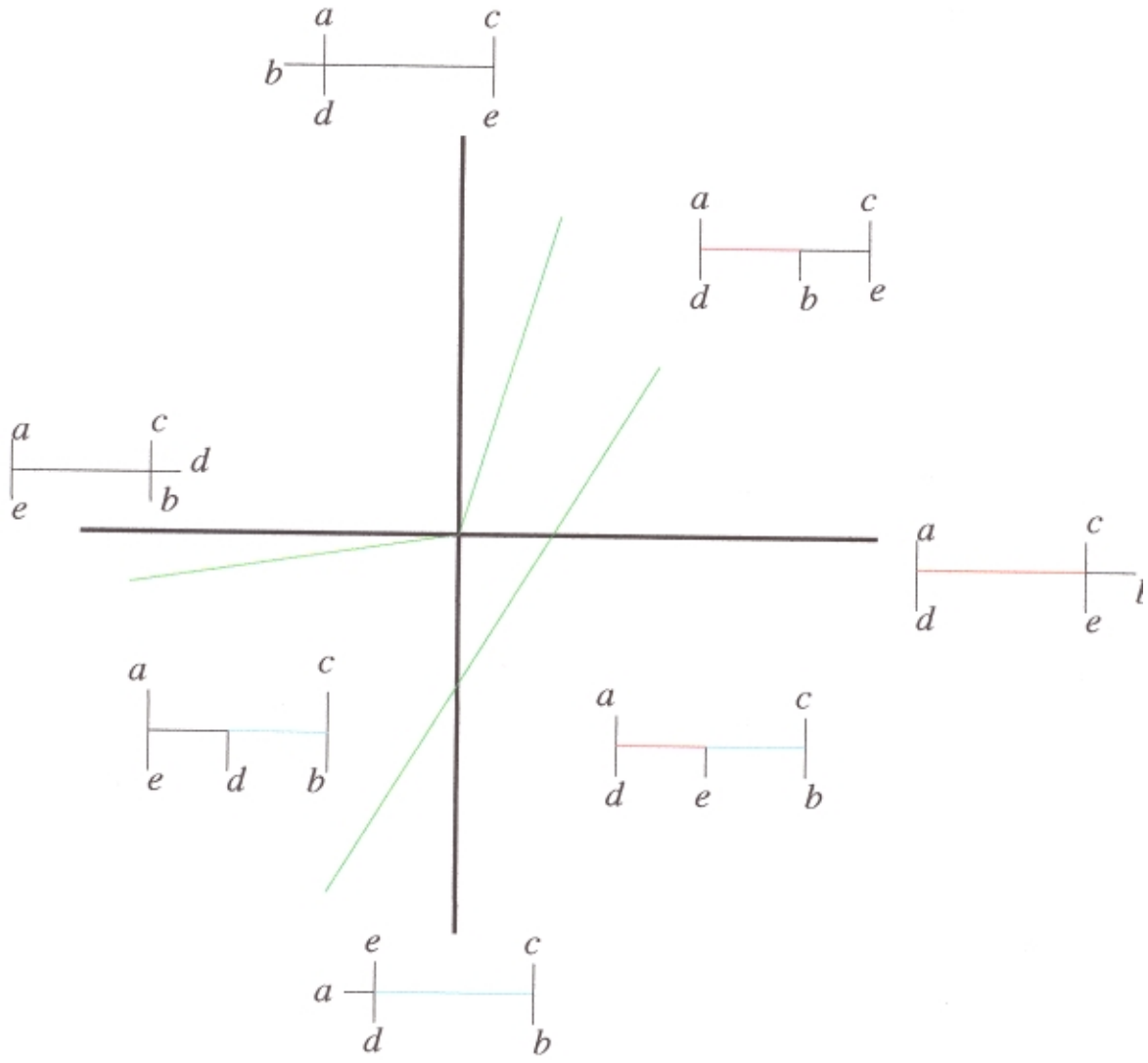
Two internal edges

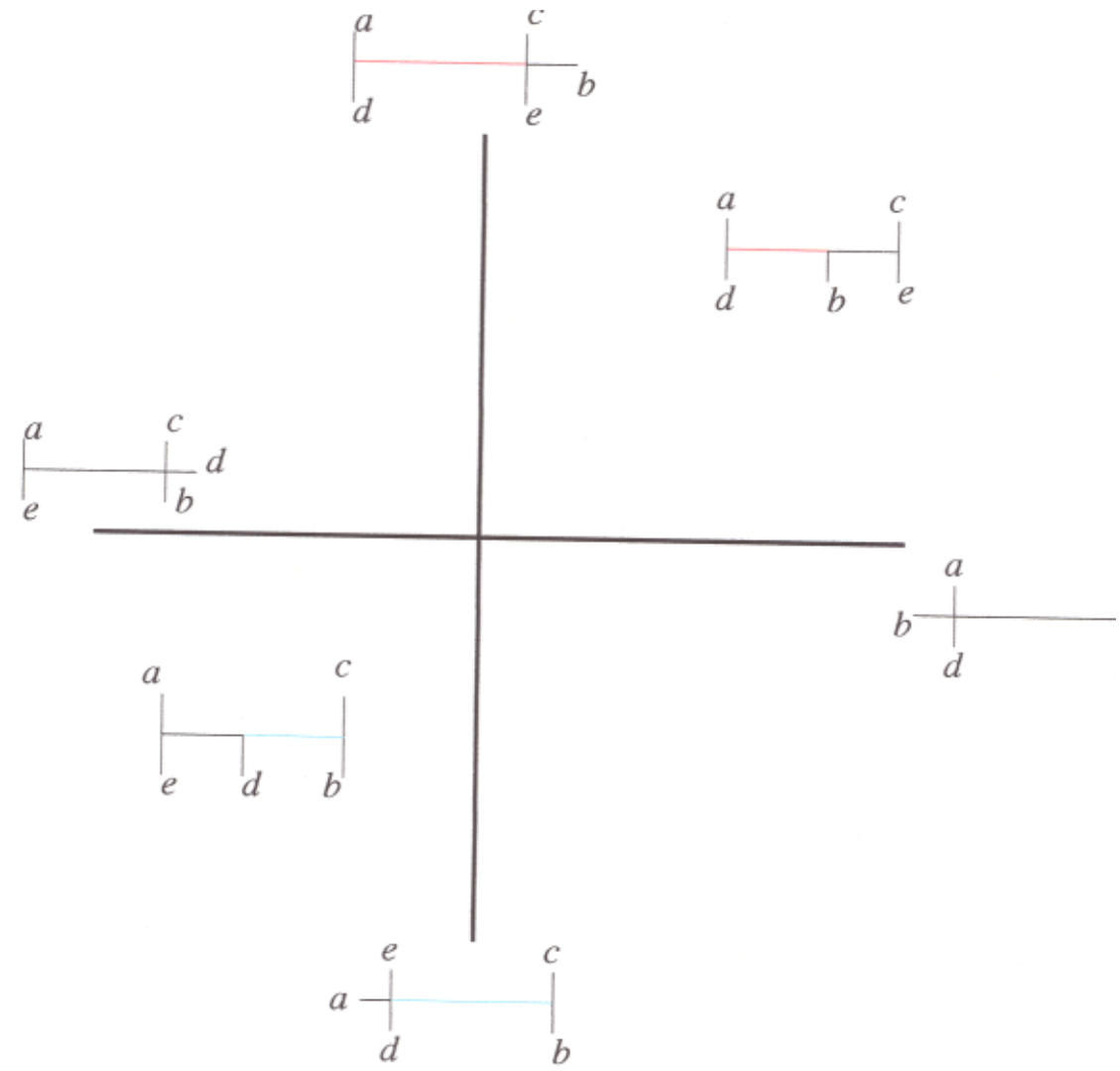


Two internal edges



Two internal edges





The wrong permutation

Geodesics

Edge common to two trees.
Compatible edges.

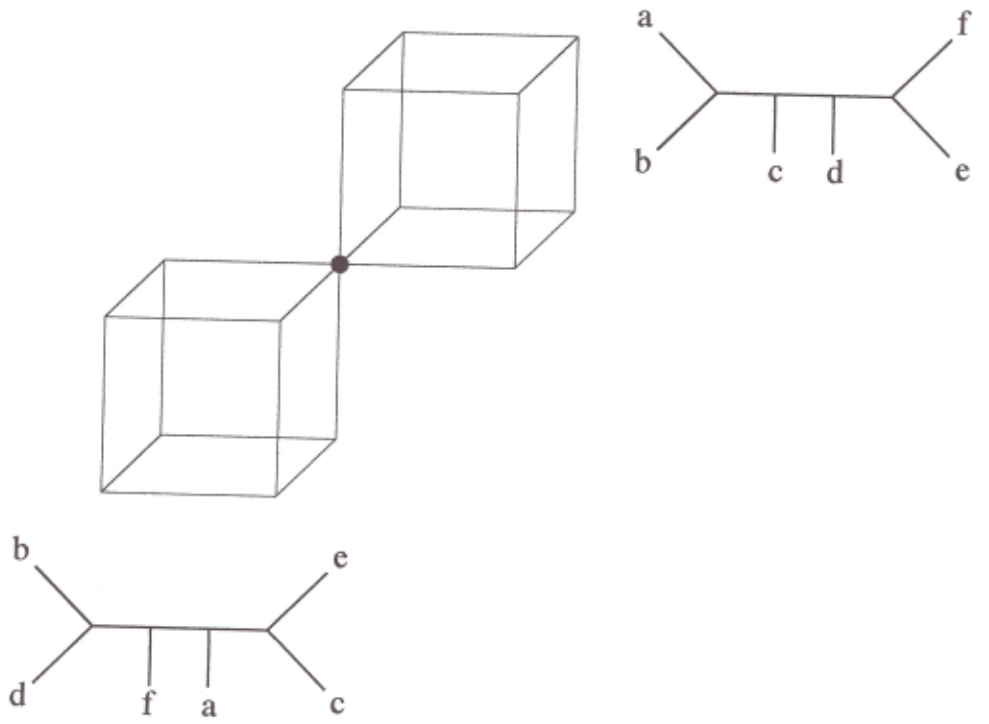
Weighted T_1 and T_2 $(n+3)$ leaves.
Assume no internal edges in common.
n internal edges $\rightarrow$ n coordinates.

Theorem. (Billera, Holmes, Vogtmann) *Label the positive axes in $\mathbb{R}^n$ by the edges of T_1 and the negative axes by the edges of T_2. Let $t_1 \in \mathbb{R}^n$ (all coordinates positive) represents T_1, t_2 (all coordinates negative) represents T_2. By choosing appropriate permutations of the n positive axes, the geodesic from T_1 to T_2 goes through the part of tree space in $\mathbb{R}^n$.*

Theorem. (Epstein, Ingram). *Given T_1 and T_2 as before, one can find in time $O(n^2)$ a permutation of the axes, such that*

- *All trees with the topology of T_1 are represented by points in the positive orthant.*
- *All trees with the topology of T_2 can be represented by points in the negative orthant.*
- *Each tree using any of the edges of T_1 and/or any of the edges of T_2 are represented by points in $\mathbb{R}^n$.*

As a corollary, any geodesic in tree space from any tree with the topology of T_1 to any tree with the topology of T_2 lies in $\mathbb{R}^n$.

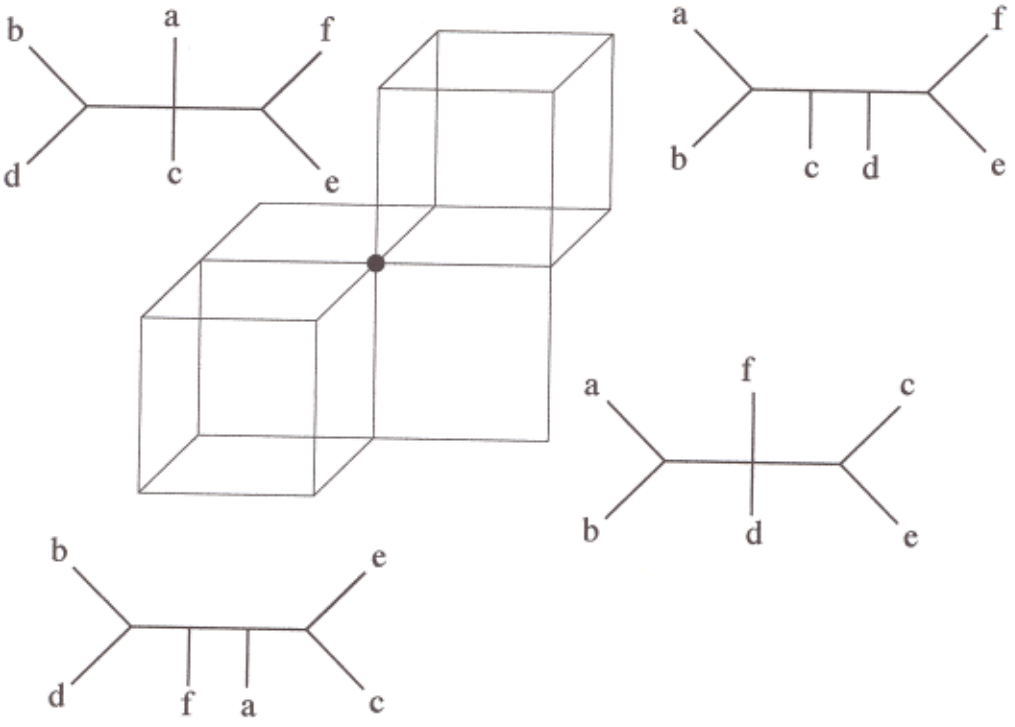


Edges in upper tree:

$$(ab|cdef), \quad (abc|def), \quad (abcd|ef).$$

Edges in lower tree:

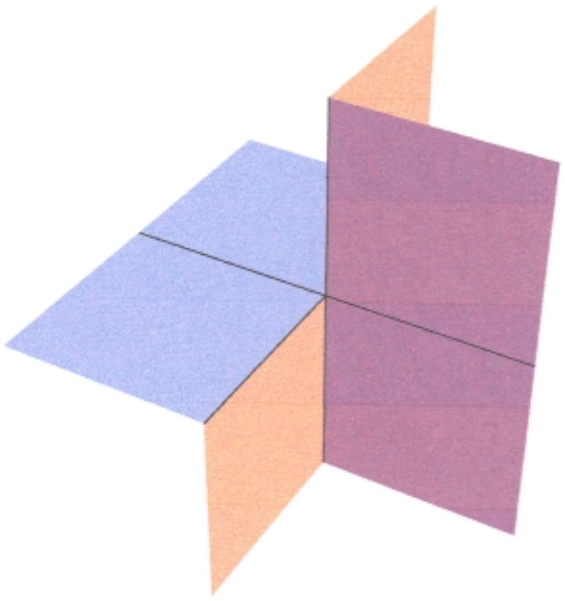
$$(bd, acef), \quad (bdf|ace), \quad (abdf|ce).$$

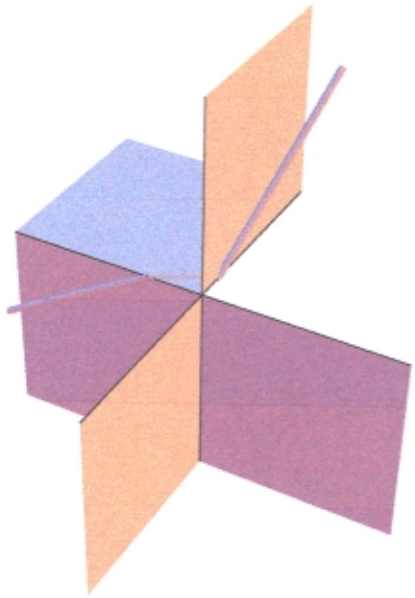

b
a
f
d
c
e

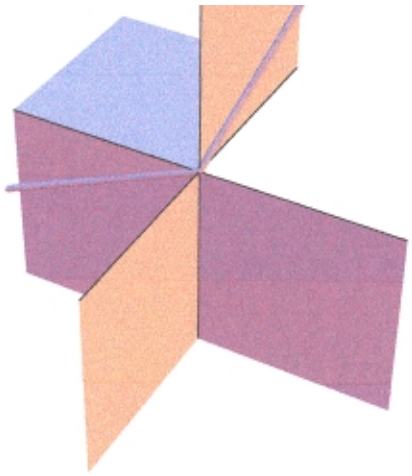
a
f
b
c
d
e

a
f
c
b
d
e

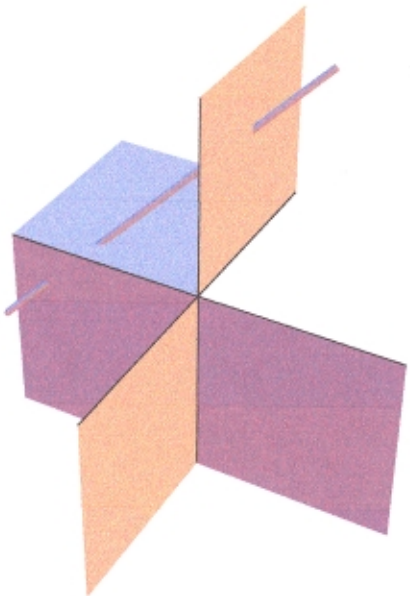
b
e
d
f
a
c







Longer



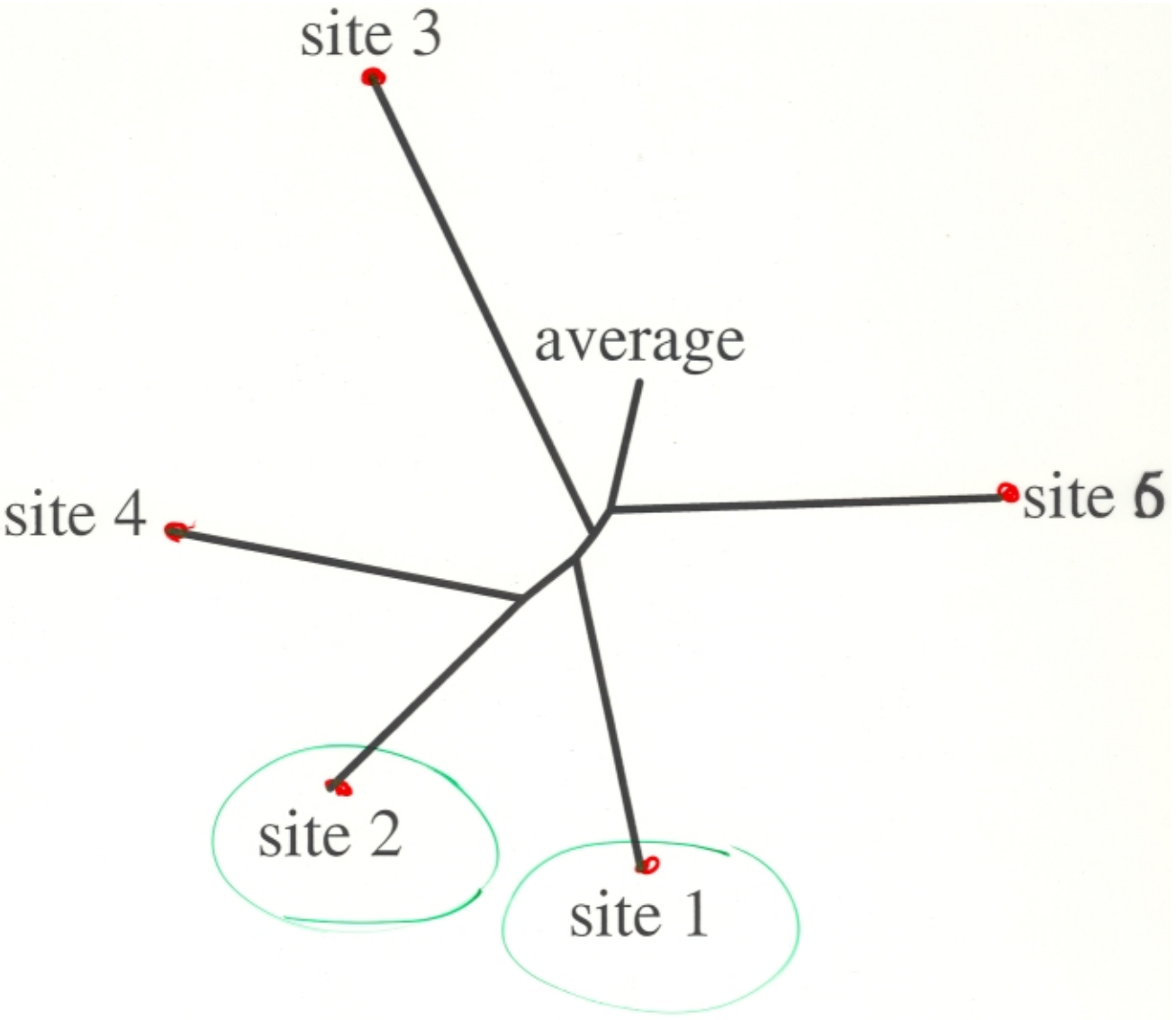
Shorter

The average of a finite set of points can be defined in many ways.

Given $t_1, \ldots, t_N$ be N points in tree space.

$$f(v) = d(v, t_1)^2 + \ldots + d(v, t_N)^2.$$

Theorem. f has a unique minimum. This can be computed in a time proportional to N times the time taken to compute the distance between two trees.


site 3
average
site 4
site 6
site 2
site 1