Computing the BHV metric
David Epstein and Jon Ingram
Mathematics Institute
University of Warwick
dbae@maths.warwick.ac.uk
jingram@maths.warwick.ac.uk
Chris Dowson, Warwick microbiologist.
28 strains 5 sites; each about 450 nucleotides.
Fix site, find ”most likely” 28-leaved tree.
5 sites → 5 trees.
Picture of 1 tree out of 5.
5 trees in cluster tree, “average” = concatenated. BHV-metric = $\ell^2$-pathmetric metric on 5 sites + concatenated site.
Computation could be done with $\ell^1$-pathmetric.
IN: $S, T$ weighted trees with same leaves. OUT BHV-geodesic.
(internal) edge $\equiv$ split of set of leaves into disjoint subsets (each with $\geq 2$ elts).
Step 1 (BHV): find common edges and cut along them. $S, T$ have no edges in common.

Moral: Need to work with large irreducible trees.
Two internal edges
Two internal edges

6-a
Two internal edges
The wrong permutation
Edges in upper tree:

\[(ab|cdef), \ (abc|def), \ (abcd|ef).\]

Edges in lower tree:

\[(bd, acef), \ (bdf|ace), \ (abdf|ce).\]
**Theorem.** (Billera, Holmes, Vogtmann) *Label the positive axes in \( \mathbb{R}^n \) by the edges of \( T \) and the negative axes by the edges of \( S \). Let \( t \in \mathbb{R}^n \) (all coordinates positive) represent \( T \), \( s \) (all coordinates negative) represent \( S \). By choosing appropriate permutations of the \( n \) positive axes, the geodesic from \( T \) to \( S \) goes through part of tree space in \( \mathbb{R}^n \).*

**Theorem.** (BHV) and, independently, (Epstein, Ingram). *One can find in time \( O(n^2) \) a permutation of the axes, such that each tree using ANY of the edges of \( T \) and/or ANY of the edges of \( S \) is represented by a point of \( \mathbb{R}^n \). As a corollary, ANY geodesic in tree space from ANY tree with the topology of \( T \) to ANY tree with the topology of \( S \) lies in \( \mathbb{R}^n \).*
Heart of the algorithm. (Unsuccessful attempt at) divide-and-conquer gives an algorithm which seems fast in practice for computing approximate distance. Approximation improves as program runs.

Conjecture: IN: weighted trees $S$ and $T$. OUT: BHV-geodesic is NP-hard.