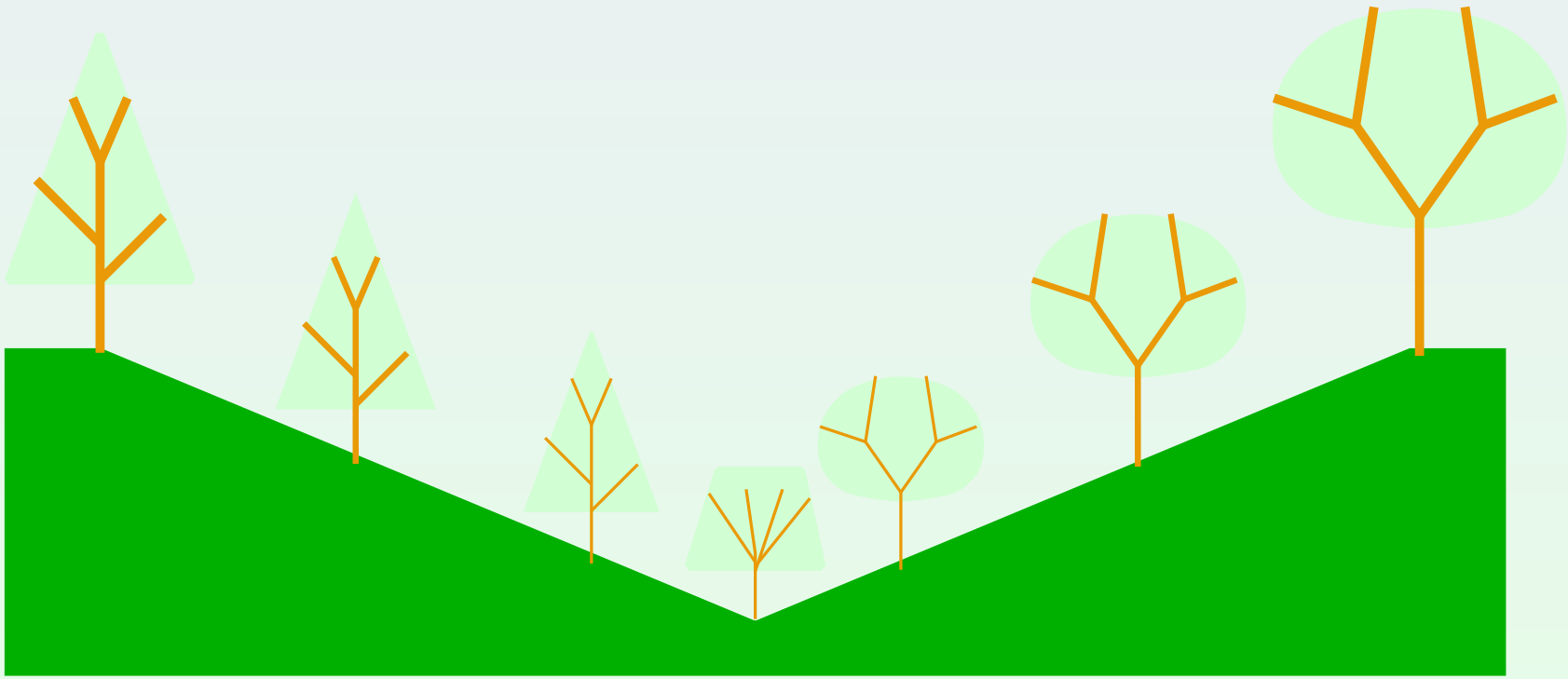
An algorithm for computing the geodesic distance between phylogenetic trees

Anne Kupczok and Steffen Kläre

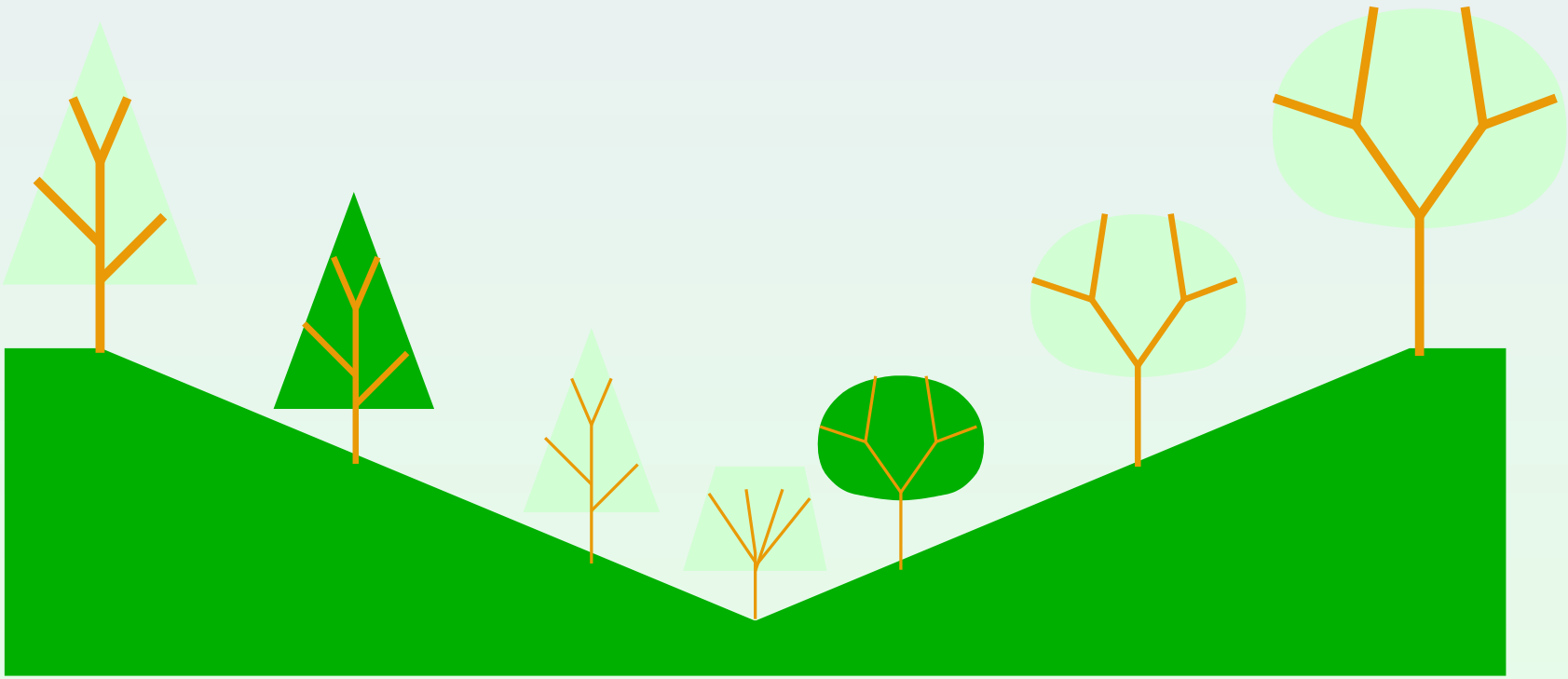
Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

December 18th, 2007

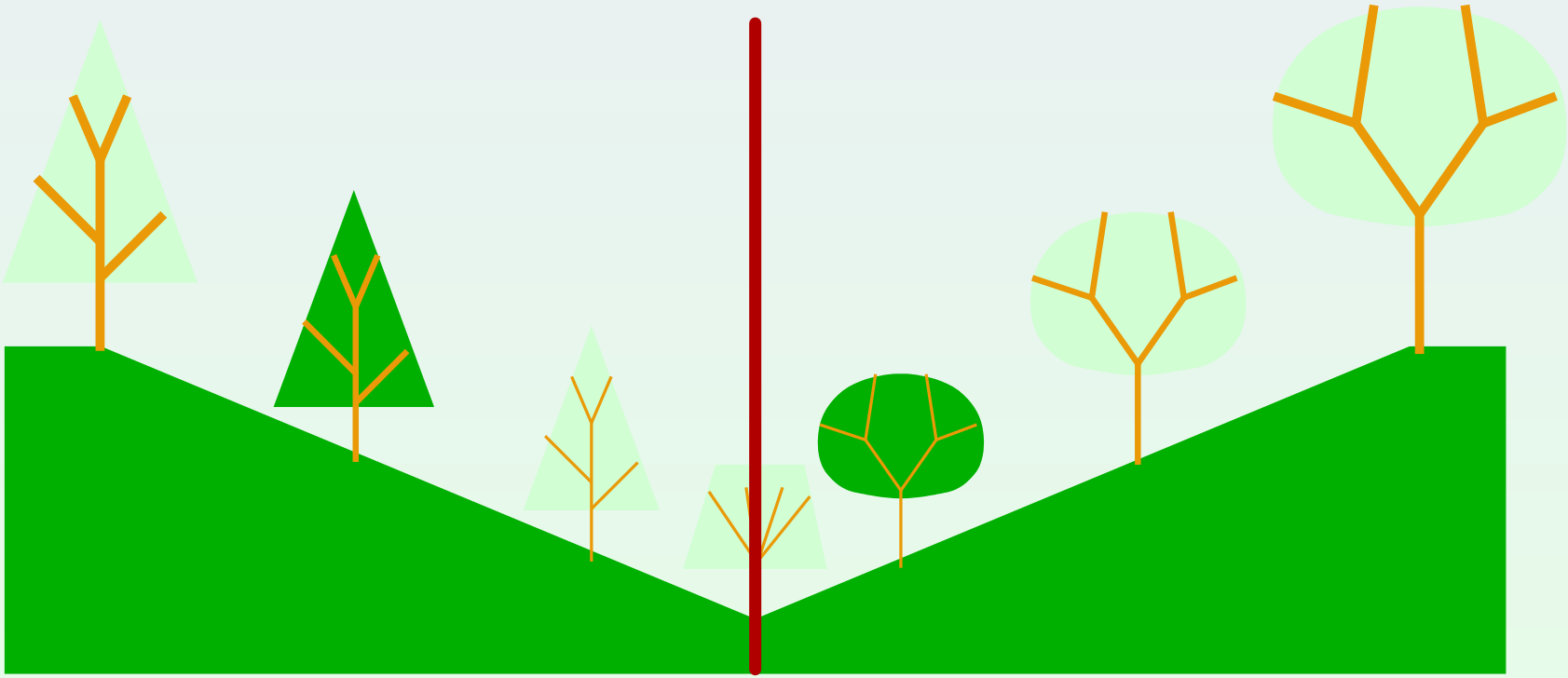
Why yet another tree distance?



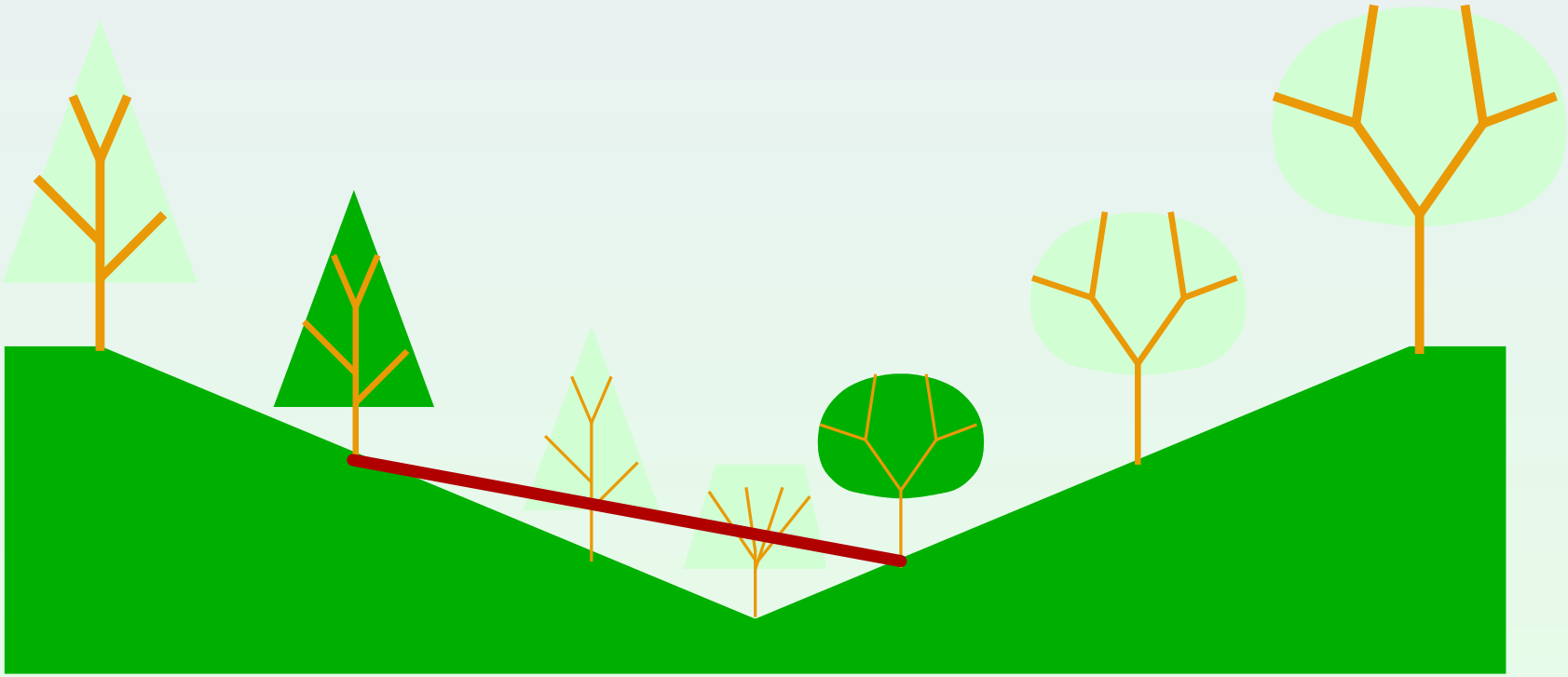
Why yet another tree distance?



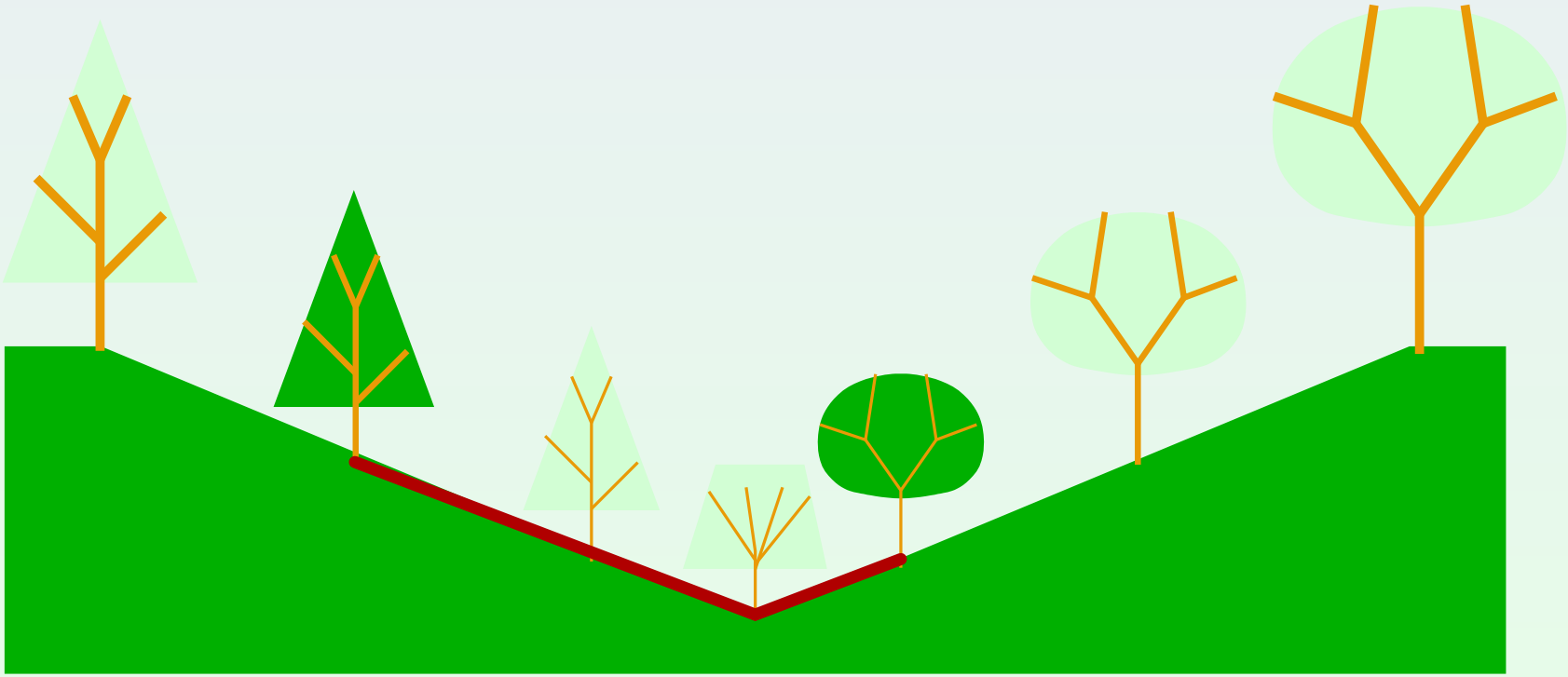
Robinson-Foulds distance



Branch-score distance

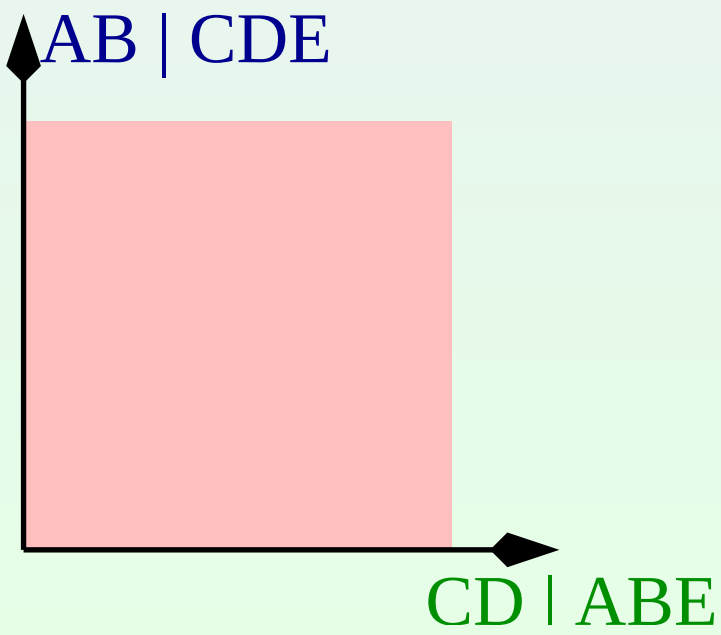


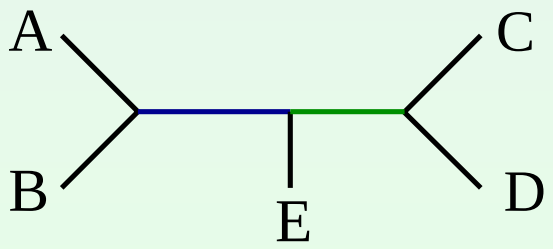
Geodesic distance



The tree space

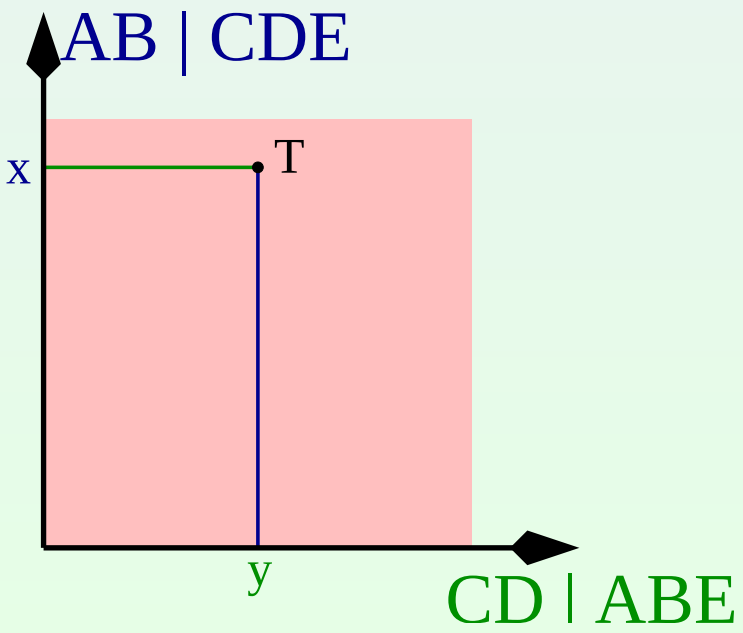
- An (unrooted, bifurcating) **topology** $\mathcal{T}$ for n taxa corresponds to an **orthant** $\mathbb{R}_+^{2n-3}$
  - The unit vectors correspond to the $2n-3$ splits

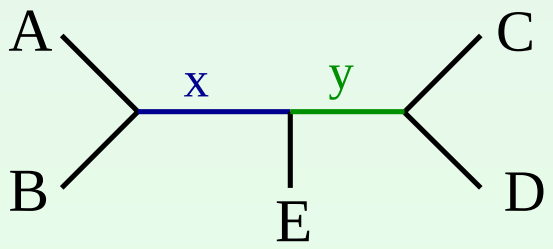




The tree space

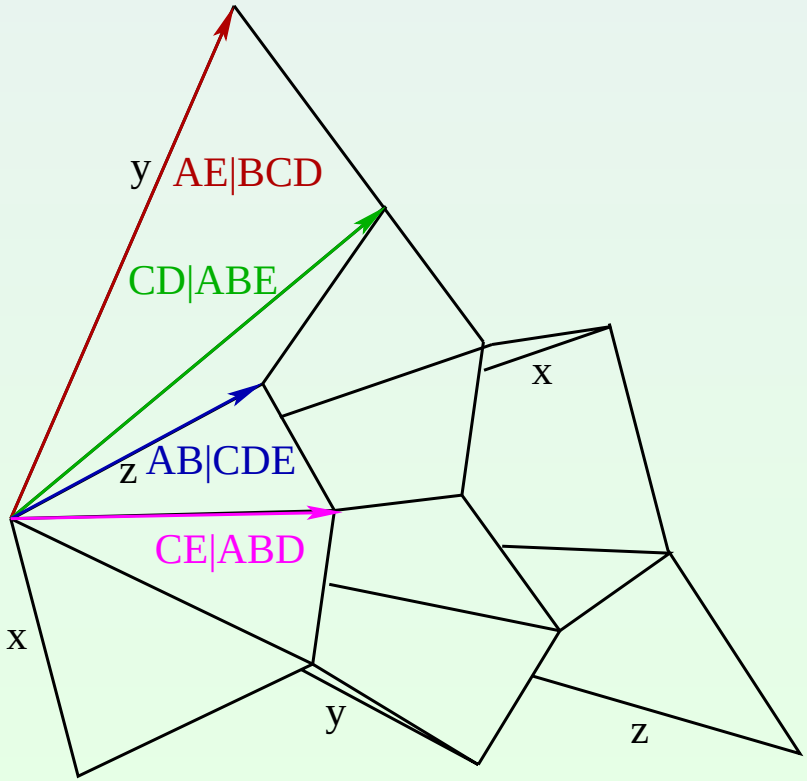
- An (unrooted, bifurcating) **topology** $\mathcal{T}$ for n taxa corresponds to an **orthant** $\mathbb{R}_+^{2n-3}$
  - The unit vectors correspond to the $2n-3$ splits
- A **tree** T with $n-3$ internal and n external branch lengths is a point in that orthant





The tree space

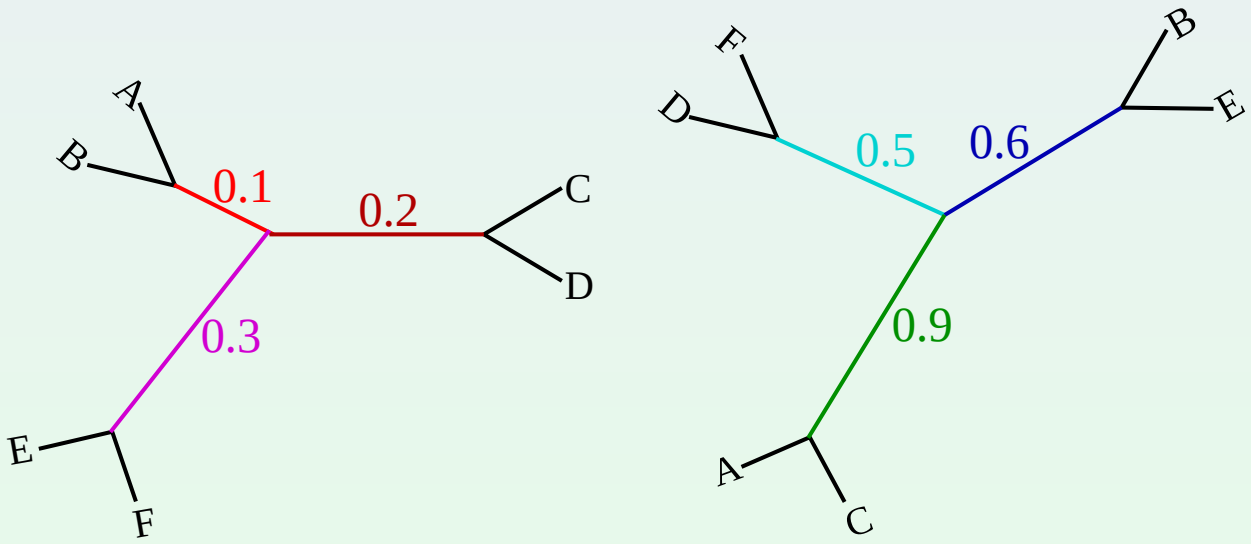
- The **tree space** for n taxa contains all possible topologies
- Its dimension is the number of splits: $2^{n-1} - 1$
- Topologies are connected by less resolved topologies
- The unique shortest path between two points is called **geodesic**



Tree space for 5 taxa (2 splits) taken from: Billera, Holmes and Vogtmann: "Geometry of the space of phylogenetic trees" *Adv. Appl. Math.*, 27, 2001

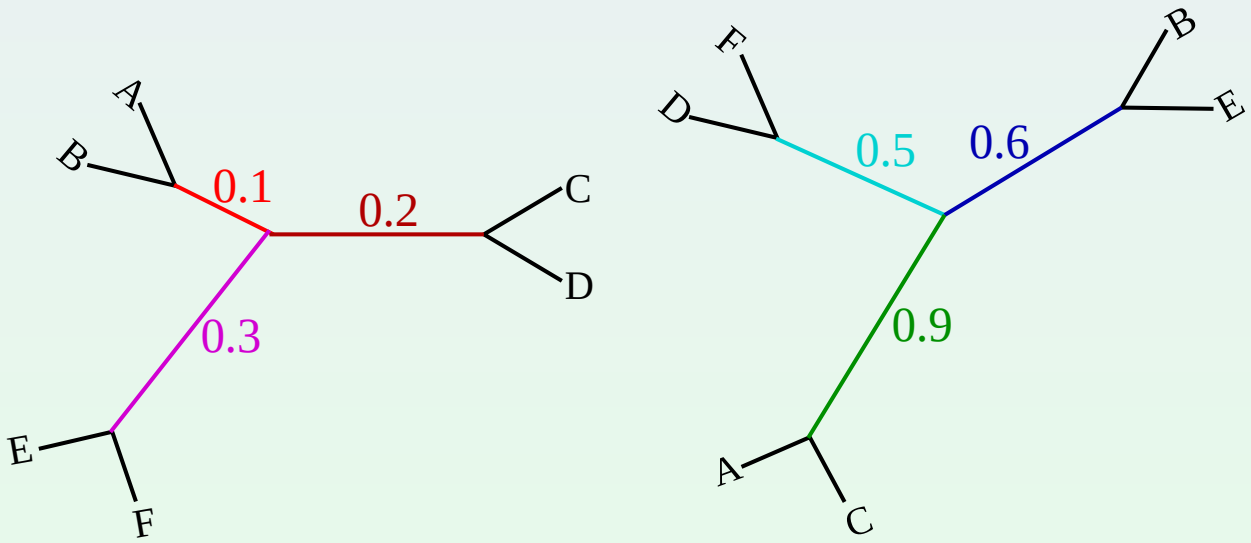
Trees and splits

- **Now**: Geodesic path connecting two weighted trees T_1 and T_2



Trees and splits

- Now: Geodesic path connecting two weighted trees T_1 and T_2
- Dimension d is the number of splits only in one tree



- Different splits:
- $\mathcal{S}_1 = (\mathtt{AB}|\mathtt{CDEF}, \mathtt{CD}|\mathtt{ABEF}, \mathtt{EF}|\mathtt{ABCD})$
- $\mathcal{S}_2 = (\mathtt{AC}|\mathtt{BDEF}, \mathtt{FD}|\mathtt{ABCE}, \mathtt{BE}|\mathtt{ACDF})$
- $d = |\mathcal{S}_1| = |\mathcal{S}_2| = 3$

The set of legal topologies

- Legal toplogies are $2d$-dimensional binary vectors
- A 1 indicates that a split is present
- All present splits must be compatible
- The toplogy is maximal (no 1 can be added)
- The two given toplogies:

$$\mathcal{T}_1 = (\overbrace{1,\ldots,1}^{d},\overbrace{0,\ldots,0}^{d})$$
$$\mathcal{T}_2 = (\underbrace{0,\ldots,0}_{d},\underbrace{1,\ldots,1}_{d})$$

The set of legal topologies

- Legal toplogies are $2d$-dimensional binary vectors
- A 1 indicates that a split is present
- All present splits must be compatible
- The toplogy is maximal (no 1 can be added)
- The two given toplogies:

$$\mathcal{T}_1 = (\overbrace{1,\ldots,1}^{d},\overbrace{0,\ldots,0}^{d})$$
$$\mathcal{T}_2 = (\underbrace{0,\ldots,0}_{d},\underbrace{1,\ldots,1}_{d})$$

Example:

$\mathcal{S} = (\mathtt{AB|CDEF}, \mathtt{CD|ABEF}, \mathtt{EF|ABCD}, \mathtt{AC|BDEF}, \mathtt{FD|ABCE}, \mathtt{BE|ACDF})$

Topologies: $(1,0,0,0,1,0)$, $(0,1,0,0,0,1)$, $(0,0,1,1,0,0)$

The set of legal topologies

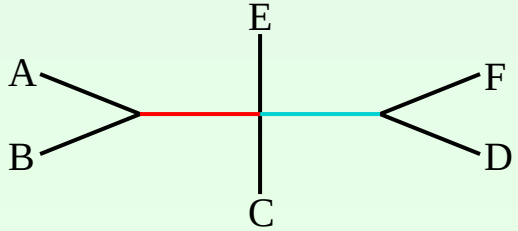
- Legal toplogies are $2d$-dimensional binary vectors
- A 1 indicates that a split is present
- All present splits must be compatible
- The toplogy is maximal (no 1 can be added)
- The two given toplogies:

$$\mathcal{T}_1 = (\overbrace{1,\ldots,1}^{d},\overbrace{0,\ldots,0}^{d})$$
$$\mathcal{T}_2 = (\underbrace{0,\ldots,0}_{d},\underbrace{1,\ldots,1}_{d})$$

Example:

$\mathcal{S} = (\mathtt{AB|CDEF}, \mathtt{CD|ABEF}, \mathtt{EF|ABCD}, \mathtt{AC|BDEF}, \mathtt{FD|ABCE}, \mathtt{BE|ACDF})$

Topologies: $(1,0,0,0,1,0)$, $(0,1,0,0,0,1)$, $(0,0,1,1,0,0)$



The directed acyclic graph of legal topologies

Two topologies are connected

$\Updownarrow$

Some of the first d splits are removed (L) and some of the last d splits are added (R)

(1,1,1,0,0,0)

(1,0,0,0,1,0) (0,1,0,0,0,1) (0,0,1,1,0,0)

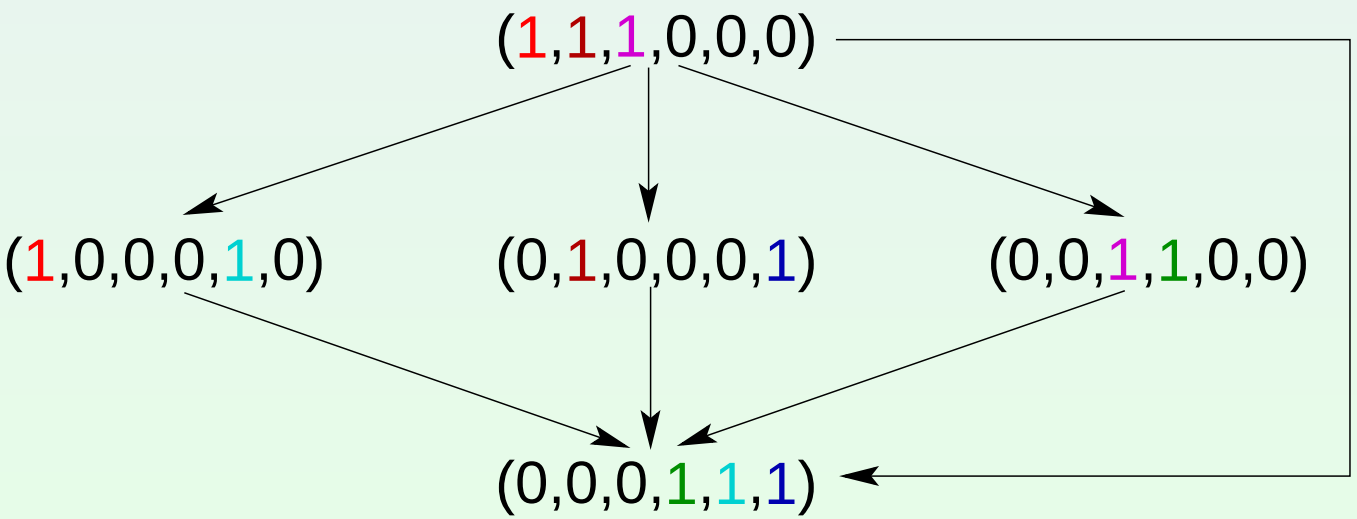
(0,0,0,1,1,1)

The directed acyclic graph of legal topologies

Two topologies are connected

$\Updownarrow$

Some of the first d splits are removed (L) and
some of the last d splits are added (R)

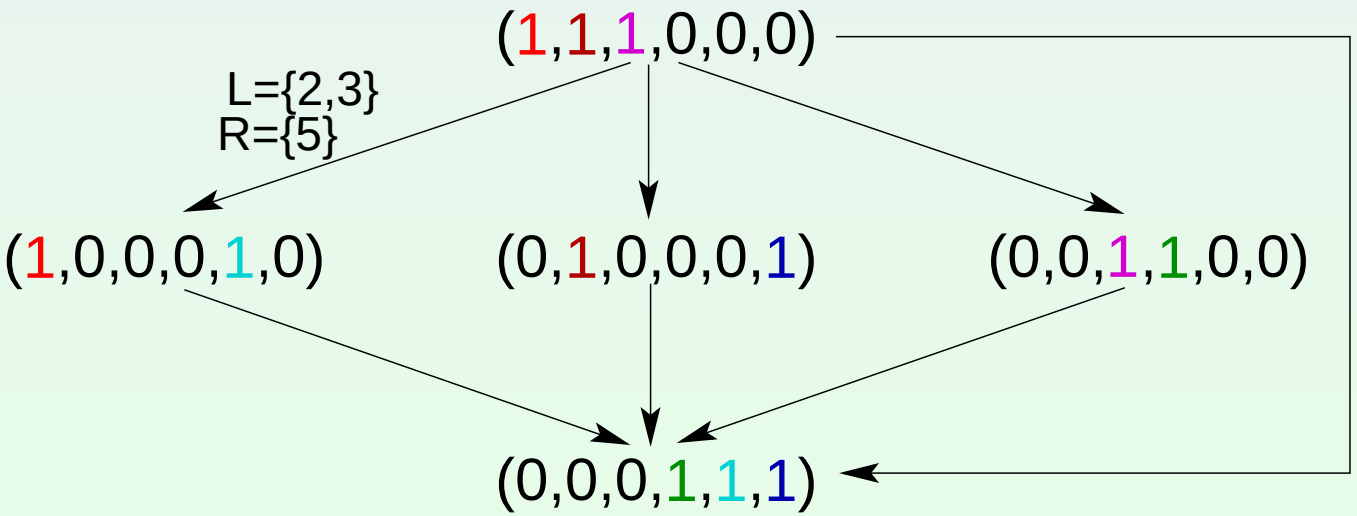


The directed acyclic graph of legal topologies

Two topologies are connected

$\Updownarrow$

Some of the first d splits are removed (L) and some of the last d splits are added (R)

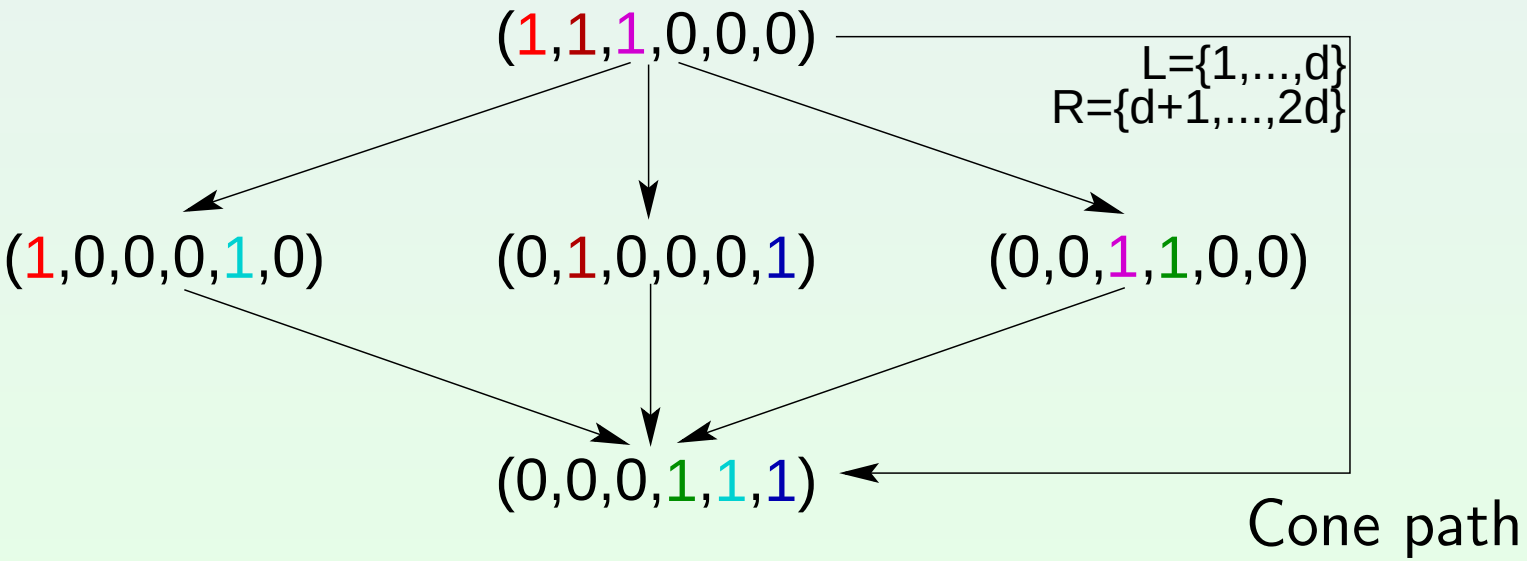


The directed acyclic graph of legal topologies

Two topologies are connected

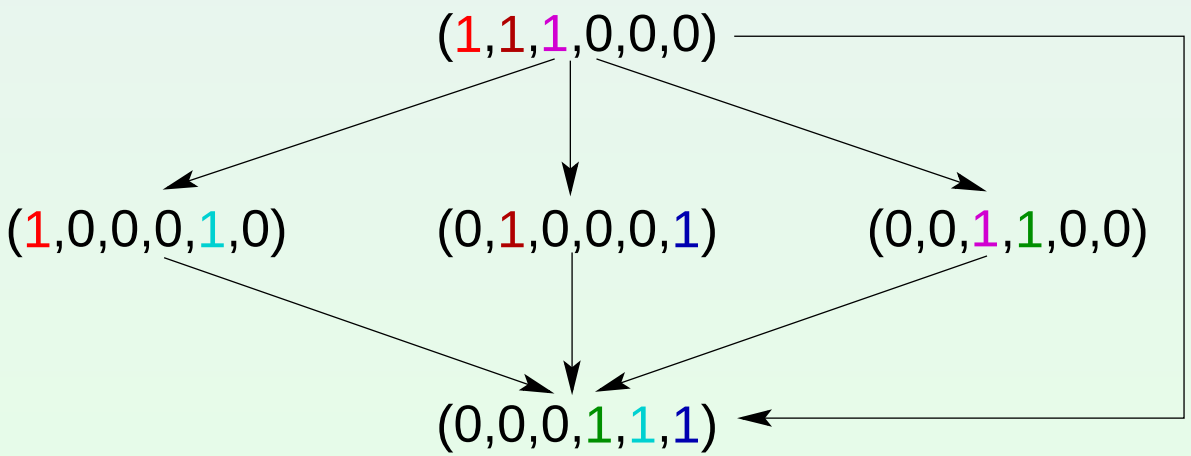
$\Updownarrow$

Some of the first d splits are removed (L) and
some of the last d splits are added (R)



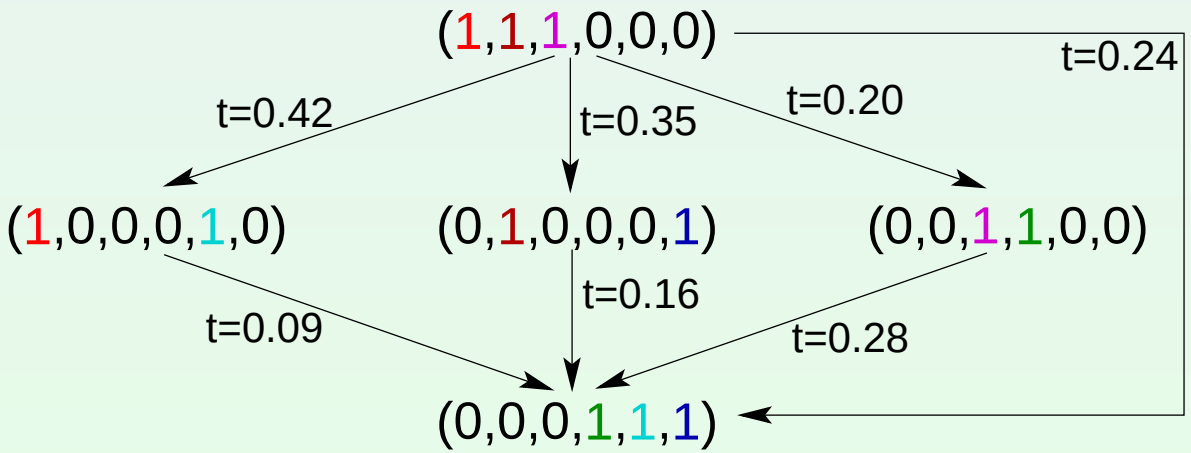
Transition times

- The path is **parametrized with constant speed** by a piecewise linear function g with $g(0) = T_1$ and $g(1) = T_2$
- For edge e in the DAG: **transition time** $t_e = \frac{\|T_1(L_e)\|}{\|T_1(L_e)\|+\|T_2(R_e)\|}$ (Karen Vogtmann, Technical report, Cornell University)



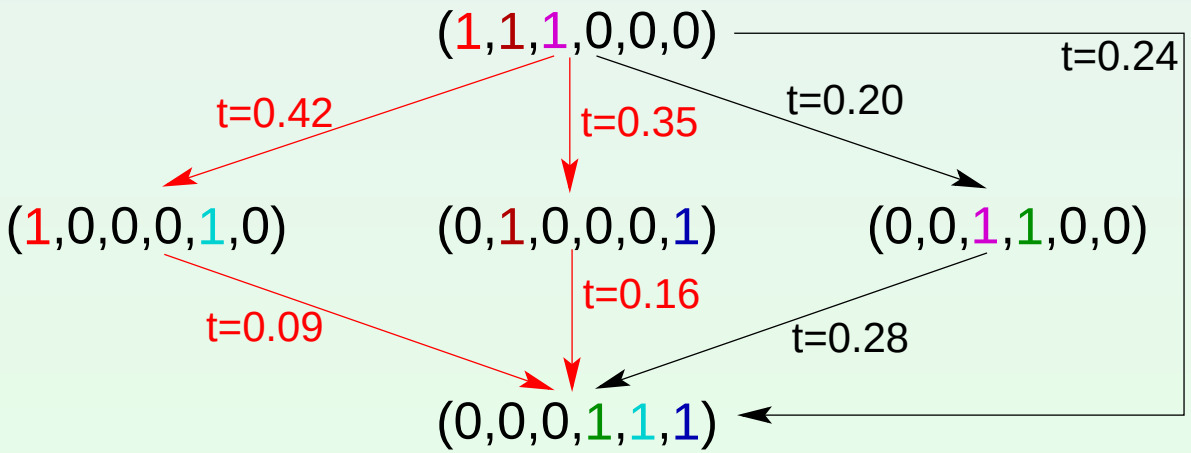
Transition times

- The path is **parametrized with constant speed** by a piecewise linear function g with $g(0) = T_1$ and $g(1) = T_2$
- For edge e in the DAG: **transition time** $t_e = \frac{\|T_1(L_e)\|}{\|T_1(L_e)\|+\|T_2(R_e)\|}$ (Karen Vogtmann, Technical report, Cornell University)



Transition times

- The path is **parametrized with constant speed** by a piecewise linear function g with $g(0) = T_1$ and $g(1) = T_2$
- For edge e in the DAG: **transition time** $t_e = \frac{\|T_1(L_e)\|}{\|T_1(L_e)\|+\|T_2(R_e)\|}$ (Karen Vogtmann, Technical report, Cornell University)



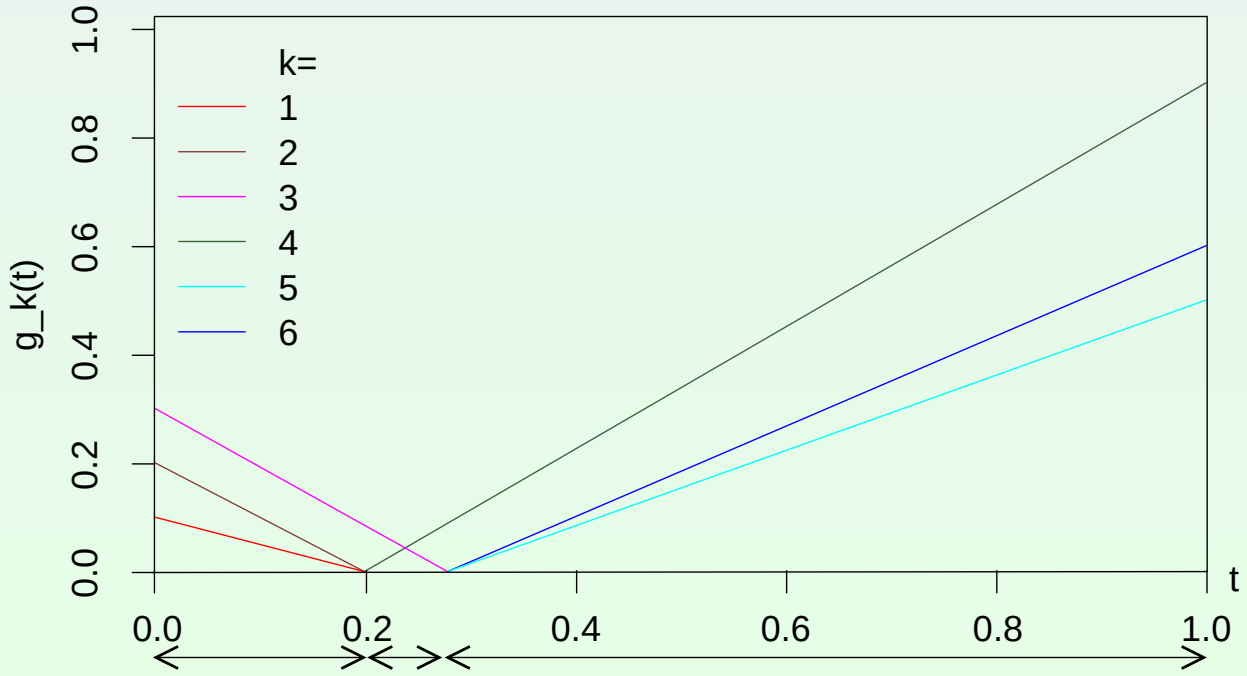
- For a sequence of topologies, the transition times must be increasing $\longrightarrow$ some sequences turn out to be **illegal**

The length of the path

- For every legal path in the DAG, the length is computed $\rightarrow$ geodesic path has shortest length
- $(1,1,1,0,0,0) \overset{t=0.24}{\longrightarrow} (0,0,0,1,1,1)$
- $(1,1,1,0,0,0) \overset{t=0.2}{\longrightarrow} (0,0,1,1,0,0) \overset{t=0.28}{\longrightarrow} (0,0,0,1,1,1)$

The length of the path

- For every legal path in the DAG, the length is computed → geodesic path has shortest length
- $(1,1,1,0,0,0) \overset{t=0.24}{\longrightarrow} (0,0,0,1,1,1) \quad ||g|| = 1.57$
- $(1,1,1,0,0,0) \overset{t=0.2}{\longrightarrow} (0,0,1,1,0,0) \overset{t=0.28}{\longrightarrow} (0,0,0,1,1,1)$
  $||g|| = 1.56$



Computational aspects

- The DAG allows a clever enumeration of topologies
- Transition times are computed when generating an edge
- The number of topologies is exponential in d
  $\longrightarrow$ The algorithm is worst-case exponential in d
- Input trees need not be bifurcating

Linear-time approximations

- **Lower bound**:
  Branch-score distance: $d = ||T_1 - T_2||$ (no path in tree space)

- **Upper bound**:
  Cone path: edge connecting $\mathcal{T}_1$ and $\mathcal{T}_2$ directly in DAG

Linear-time approximations

- **Lower bound**:
  Branch-score distance: $d = ||T_1 - T_2||$ (no path in tree space)

- **Upper bound**:
  Cone path: edge connecting $\mathcal{T}_1$ and $\mathcal{T}_2$ directly in DAG

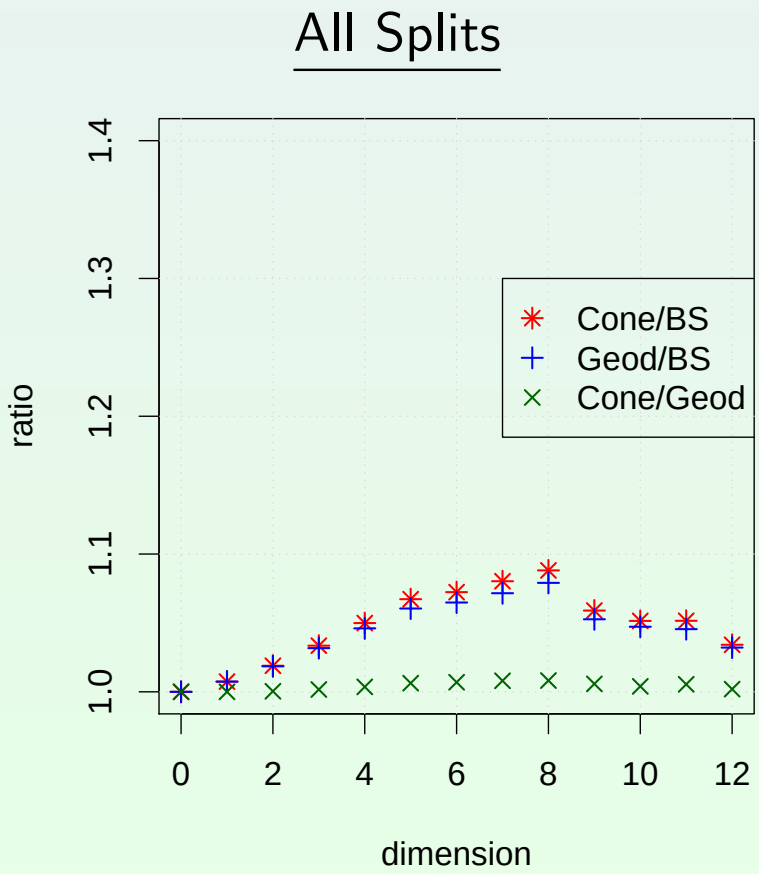
- The bounds differ at most in a factor of $\sqrt{2}$
  (Amenta et al. Approximating geodesic tree distance. Inf. Process. Lett., 103(2), 2007)

Comparison of the approximations

- Inparanoid database: orthologs from 20 Metazoa species + yeast outgroup (216 orthologs $\longrightarrow$ ML trees with phyML)
- 118 trees without internal polytomies $\longrightarrow$ 6903 pairs

Comparison of the approximations

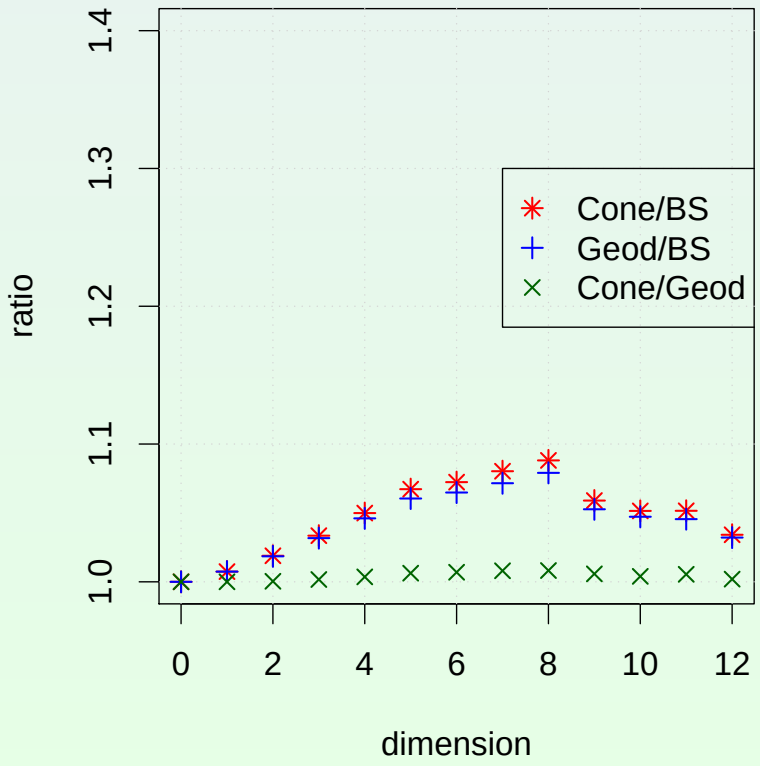
- Inparanoid database: orthologs from 20 Metazoa species + yeast outgroup (216 orthologs $\rightarrow$ ML trees with phyML)
- 118 trees without internal polytomies $\rightarrow$ 6903 pairs



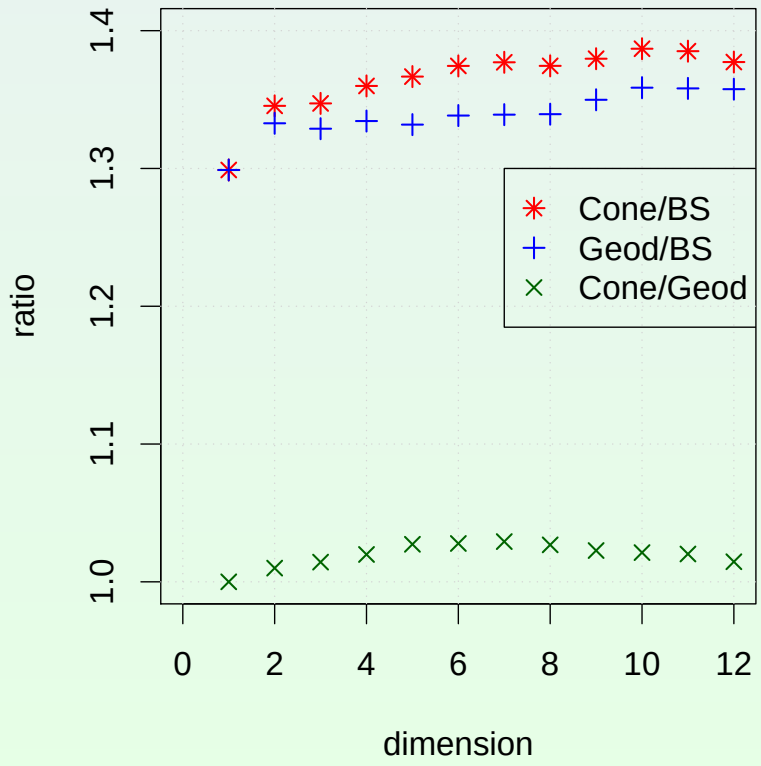
Comparison of the approximations

- Inparanoid database: orthologs from 20 Metazoa species + yeast outgroup (216 orthologs → ML trees with phyML)
- 118 trees without internal polytomies → 6903 pairs

All Splits



Different Splits



Summary

- Algorithm for the geodesic path connecting two weighted trees
- Exponential in the number of different splits
- Cone path can be computed in linear time and is a good approximation of the geodesic path

Acknowledgements

- Karen Vogtmann

- CIBIV:
  Arndt von Haeseler
  Ingo Ebersberger
  Gregory Ewing

- WWTF (funding)