

# Identifying Lateral Transfer with Neighbor-Nets

Marc Thuillard

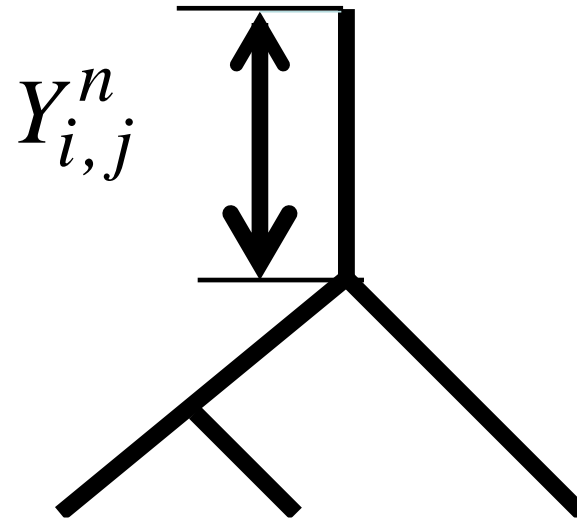
Creuze 9, CH-2072 St-Blaise, Switzerland

Vincent Moulton

School of Computing Sciences

University of East Anglia, City, Norwich

n: Reference Taxon



$$Y_{i,j}^n = 1/2 \cdot (d_{i,n} + d_{j,n} - d_{i,j})$$

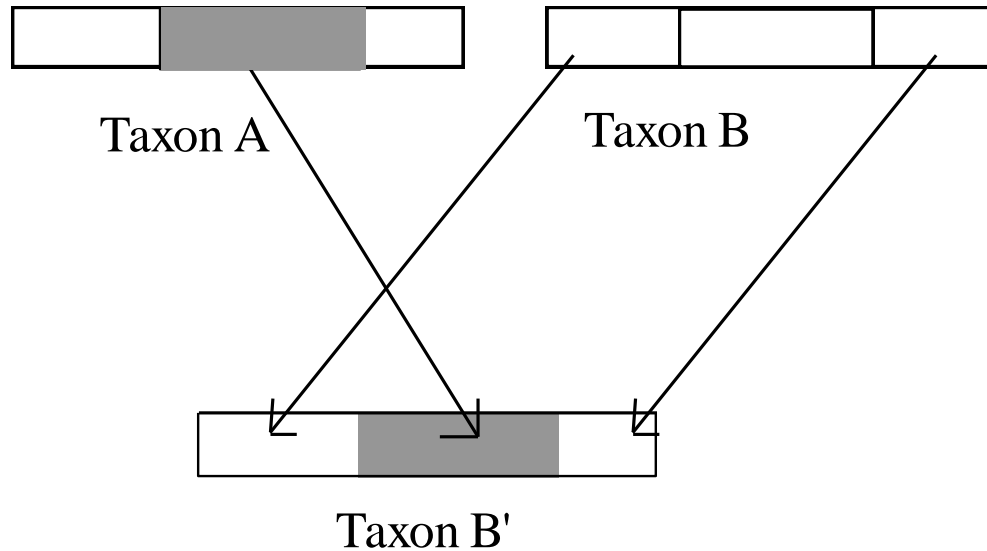
# Perfect Order

Kalmanson Inequalities:

$$Y_{i,j}^n \geq Y_{i,k}^n \quad (i \leq j \leq k)$$

$$Y_{k,j}^n \geq Y_{k,i}^n$$

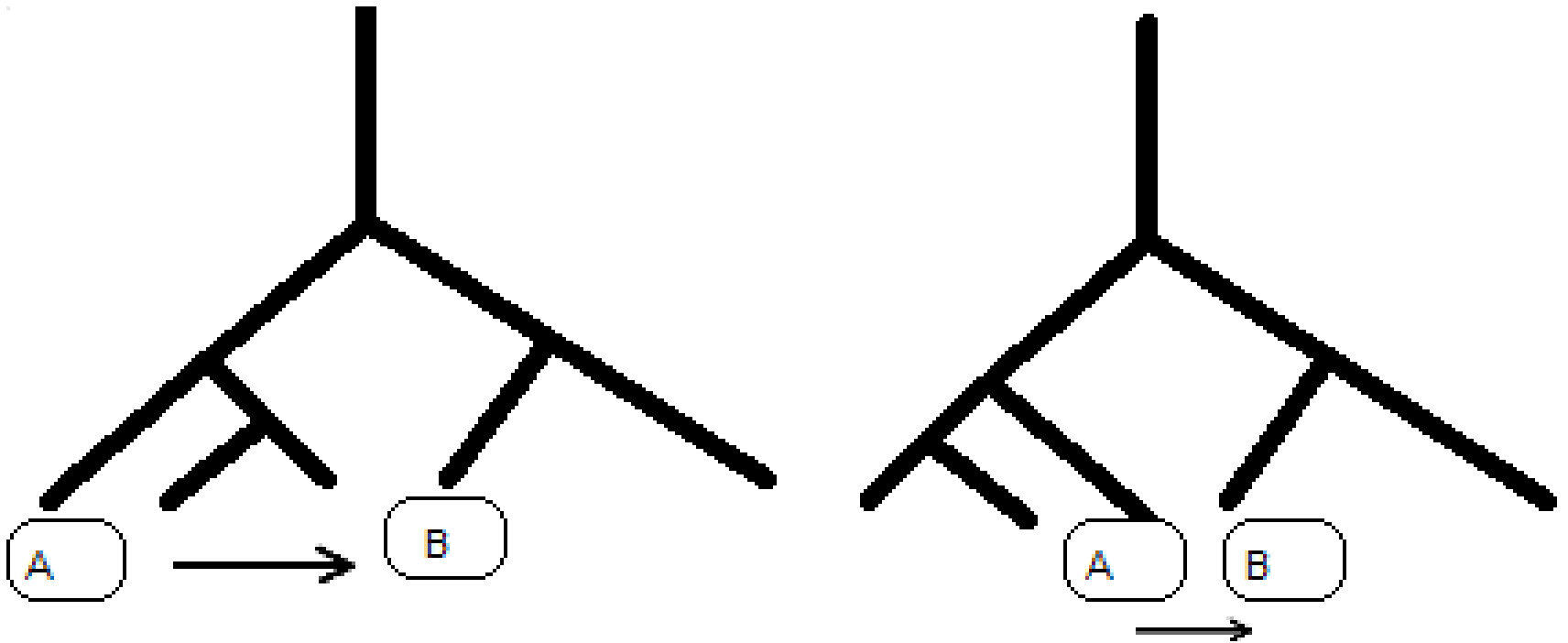
# Model of Lateral Transfer



$$\hat{d}_{B,C} = \alpha \cdot d_{A,C} + (1 - \alpha) \cdot d_{B,C}$$

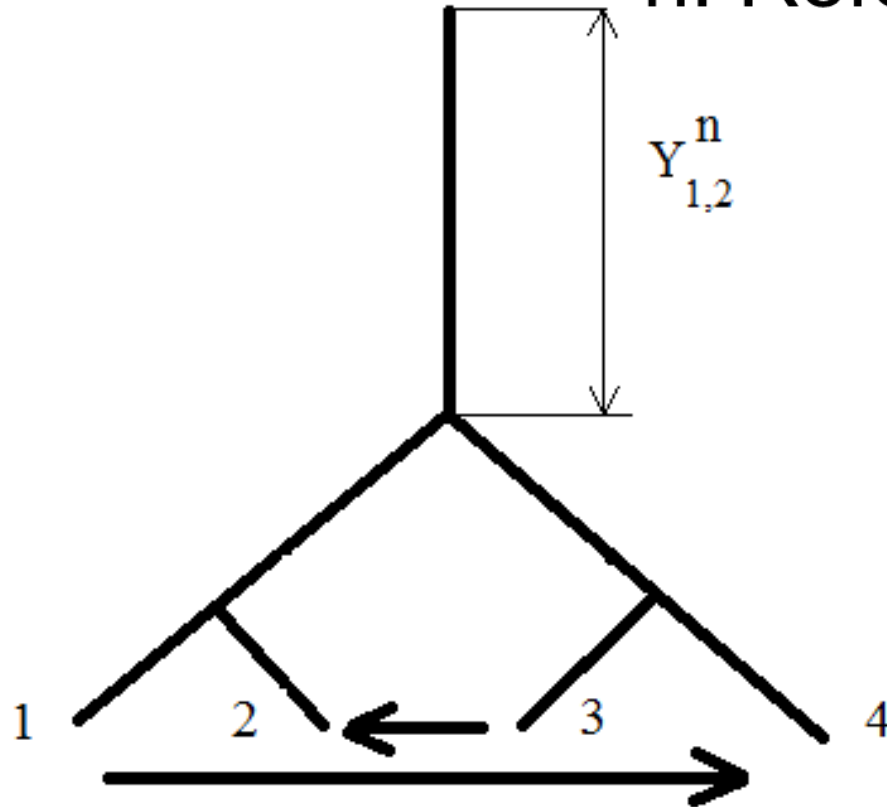
$$\hat{d}_{B,C} = -3/4 \ln(1 - 4/3 (\alpha p_{A,C} + (1 - \alpha) p_{B,C}))$$

# Lateral Gene Transfer



Transfer between consecutive nodes preserves best order

n: Reference Taxon



# Minimizing the Contradiction

Kalmanson Inequalities:

$$Y_{i,j}^n \geq Y_{i,k}^n \quad Y_{k,j}^n \geq Y_{k,i}^n \quad (i \leq j \leq k)$$

Minimizing the contradiction:

$$C = \sum_{\substack{k > j \geq i \\ i, j, k \neq n}} (\max((Y_{i,k}^n - Y_{i,j}^n), 0))^\beta + \sum_{\substack{k \geq j > i \\ i, j, k \neq n}} (\max((Y_{i,k}^n - Y_{j,k}^n), 0))^\beta$$

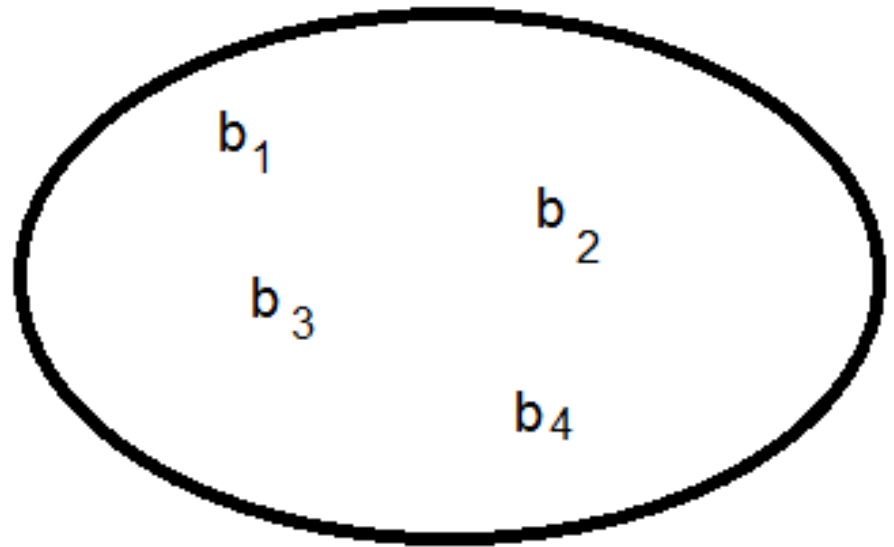
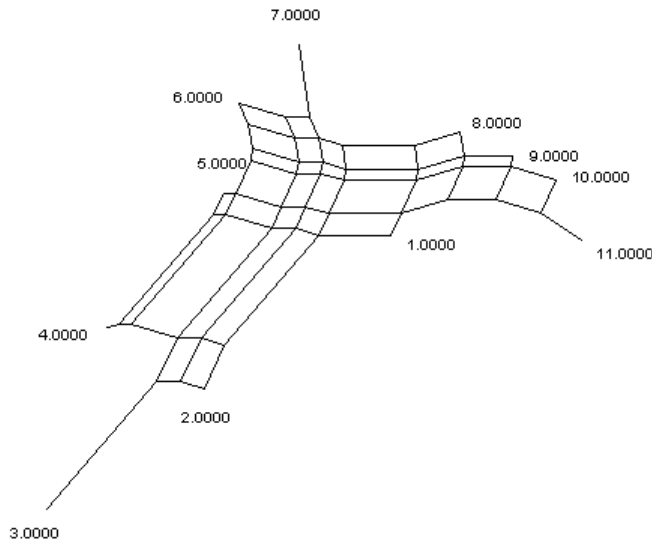
# Algorithm Part 1

1. Choose an arbitrary subset  $A$  consisting of 3 taxa
2. A new taxon  $x_i$  is added to the subset  $A$  if the order generated by Neighbor-Net satisfies  $(C_{i,j} < \varepsilon \text{ for all } 1 \leq i, j \leq n)$  for some small  $\varepsilon$  chosen by the user
3. Repeat 2 until no further taxa can be added without violating the necessary conditions.



# Outer Planar Network

$$(x_1, x_2, \dots, x_{i-1}, x_i, \dots, x_n)$$



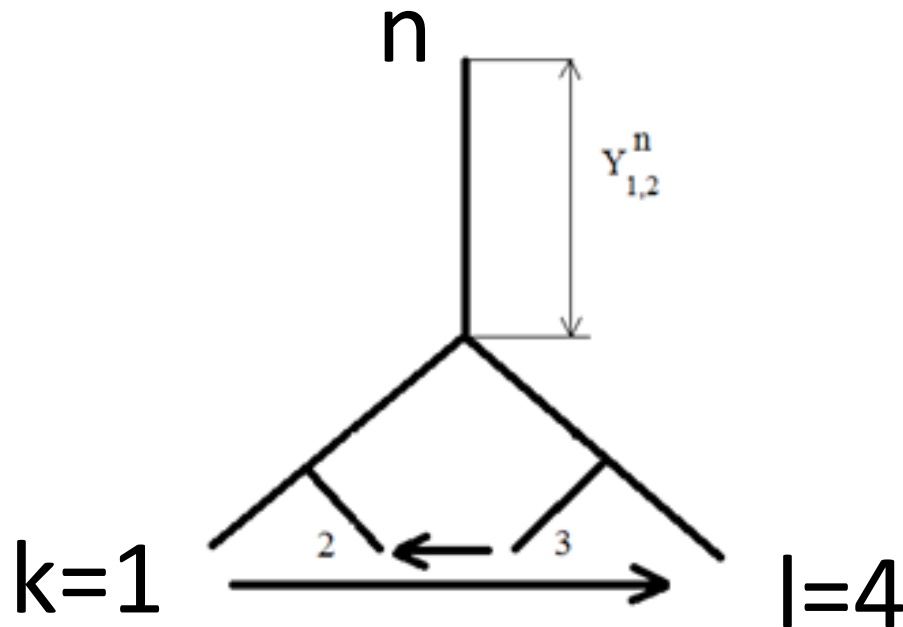
$B(i)$ : distance matrix defined on order  $(x_1, x_2, x_{i-1}, b, x_i, \dots, x_n)$

Definition: A distance matrix in Farris space is said to correspond to a weakly-perturbed tree if the matrix is of the form:

$$NK = (1-\alpha) Y^n + \alpha T_{k \rightarrow l} Y^n \quad (0 < \alpha < 1) \text{ with}$$

$$Y^n = Y_{Tree}^n + \sum_i T_{k_i \rightarrow l_i} Y_{Tree}^n, \quad |(k_i - l_i)| = 1, \quad k_i \neq k_j, l_j, k, l; \quad l_i \neq l_j, k_j, k, l; \quad k, l \neq n.$$

Example:



# Algorithm: Part 2

1. Determine  $i_{min}(x_n)$ , the largest index  $i$  with  $1 \leq i \leq n$  so that  $B_{m_1, m_2}(i) \geq B_{m_1, m_3}(i)$ ,  $B_{m_3, m_2}(i) \geq B_{m_3, m_1}(i)$  holds for  $1 \leq m_1 \leq m_2 \leq m_3 \leq i$ .
2. Determine  $i_{max}(x_n)$ , the smallest index  $i$  with  $1 \leq i \leq n$  so that  $B_{m_1, m_2}(i) \geq B_{m_1, m_3}(i)$ ,  $B_{m_3, m_2}(i) \geq B_{m_3, m_1}(i)$  holds for  $1 \leq m_1 \leq m_2 \leq m_3 \leq n$ .
3. If  $i_{min}(x_n) \leq i_{max}(x_n)$  choose a taxon  $x_m$  with  $m$  in  $(i_{min}(x_n), \dots, i_{max}(x_n) - 1)$ . Keeping the relative positions the same, reorder the set  $A$  so that  $x_m$  comes last in the order, and repeat Steps 1 and 2 with  $n=m$  to give  $i_{min}(x_m), i_{max}(x_m)$ .

$(x_1, x_2, x_{i-1}, x_i, \dots, x_{n-2}, x_{n-1}, x_n)$

Input:  $b$

$(x_1, b, x_n)$

$i=2$

*Kalmanson*

$(x_1, x_2, b, x_n):$

$i=3$

*Kalmanson ? if yes next*

$(x_1, x_2, \dots, x_j, b, x_n)$

$i=j+1$

*Kalmanson ? If no*

Output :  $i_{\min}(n) = j$

$(b, x_{n-1}, x_n)$

$i=n-2$

*Kalmanson*

$(b, x_{n-2}, x_{n-1}, x_n)$

$i=n-3$

*Kalmanson ? if yes next*

....

Output :  $i_{\max}(n)$

# Properties of «Weakly-perturbed Tree»

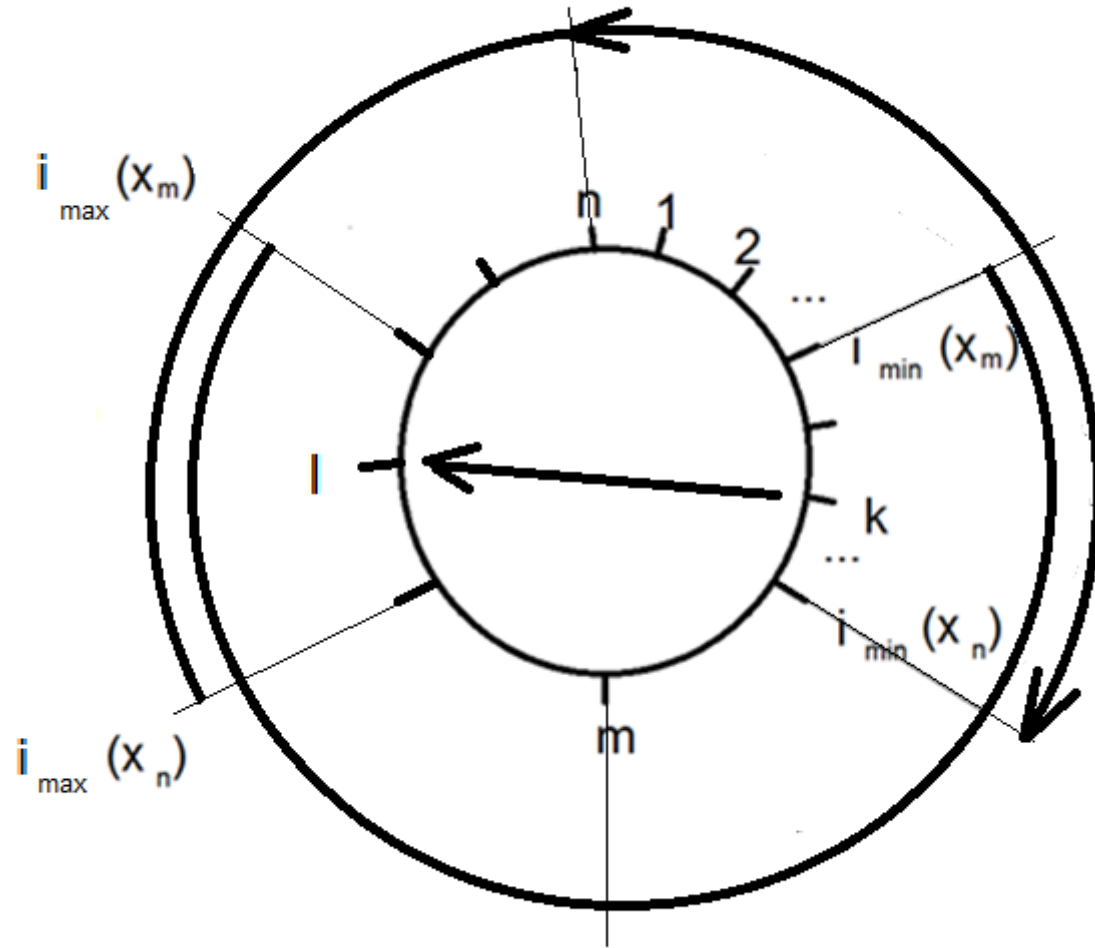
1. If  $|(k - l)| > 3$

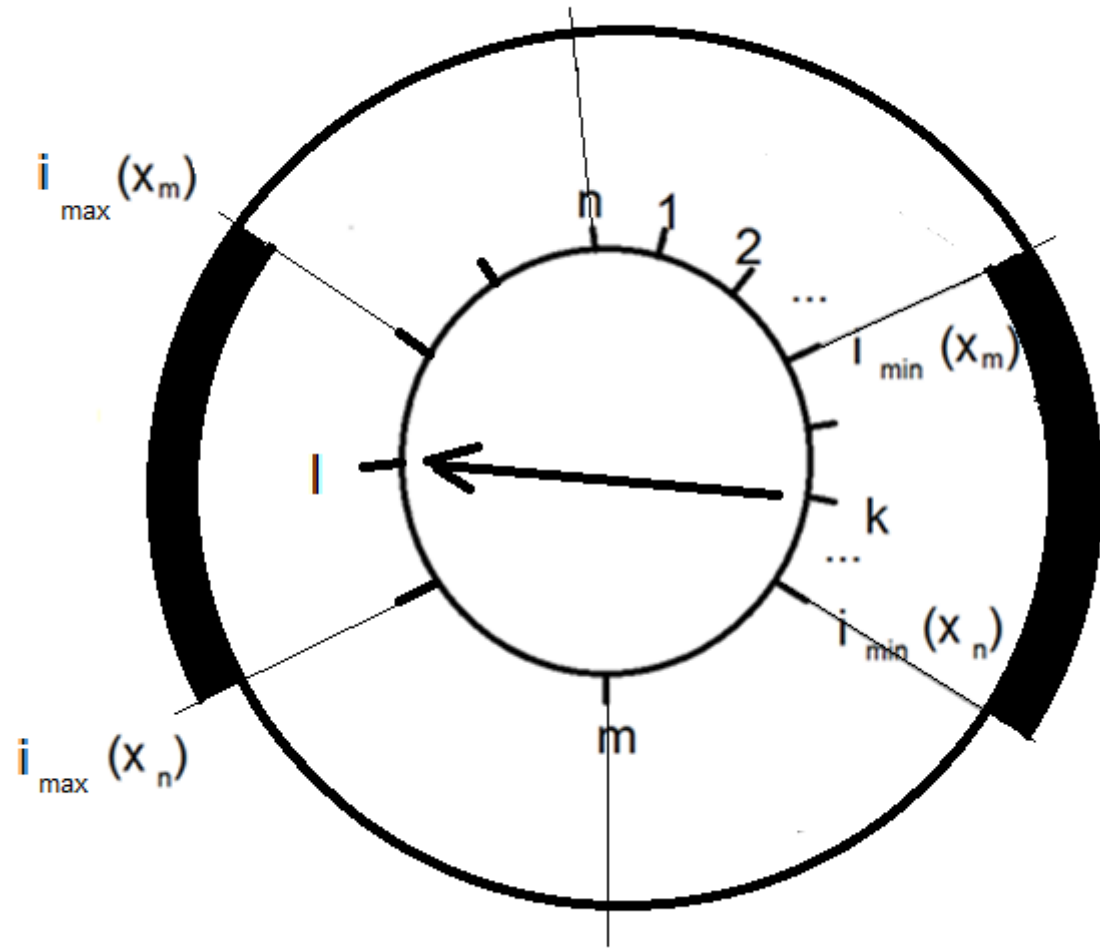
then  $b$  is either the taxon  $k$  or the taxon  $l$  of the lateral transfer

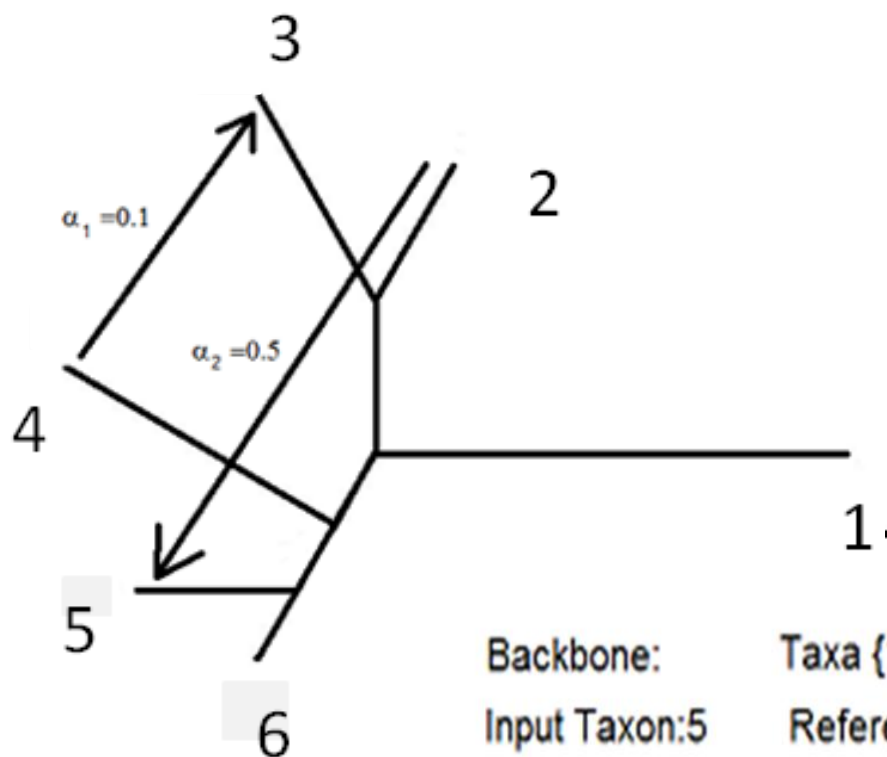
$$T_{k \rightarrow l}$$

2.  $[i_{\min}(x_n), i_{\max}(x_n)] \subseteq [k, l]$

$$k < l$$







$$\hat{Y}_{ij}^1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 7.8 & 6 & 8 & 6 \\ 0 & 7.8 & 11 & 6.5 & 6.35 & 6.1 \\ 0 & 6 & 6.5 & 11 & 6.5 & 7 \\ 0 & 8 & 6.35 & 6.5 & 9.5 & 7 \\ 0 & 6 & 6.1 & 7 & 7 & 10 \end{pmatrix}$$

Taxa: 1 2 3 4 5 6

Backbone: Taxa {1,2,3,4,6}  
 Input Taxon: 5 Reference Taxon: 1

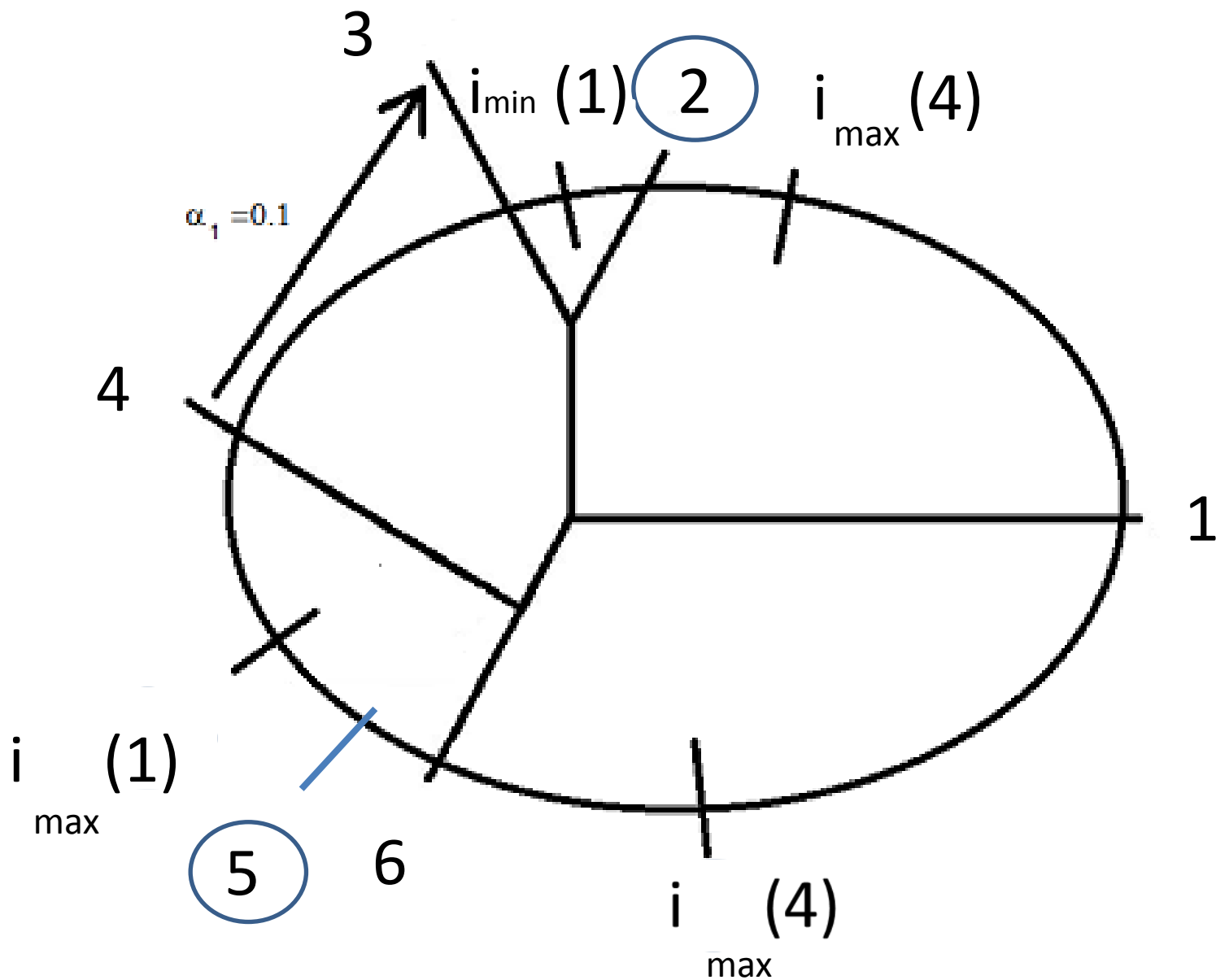
$$B(i_{\max} - 1) = \begin{pmatrix} 9.5 & 6.5 & 7 \\ 6.5 & 11 & 7 \\ 7 & 7 & 10 \end{pmatrix}$$

Taxa: 5 4 6

$$B(i_{\max}) = \begin{pmatrix} 9.5 & 7 \\ 7 & 10 \end{pmatrix}$$

Taxa: 5 6





# Example: Whole Genomes

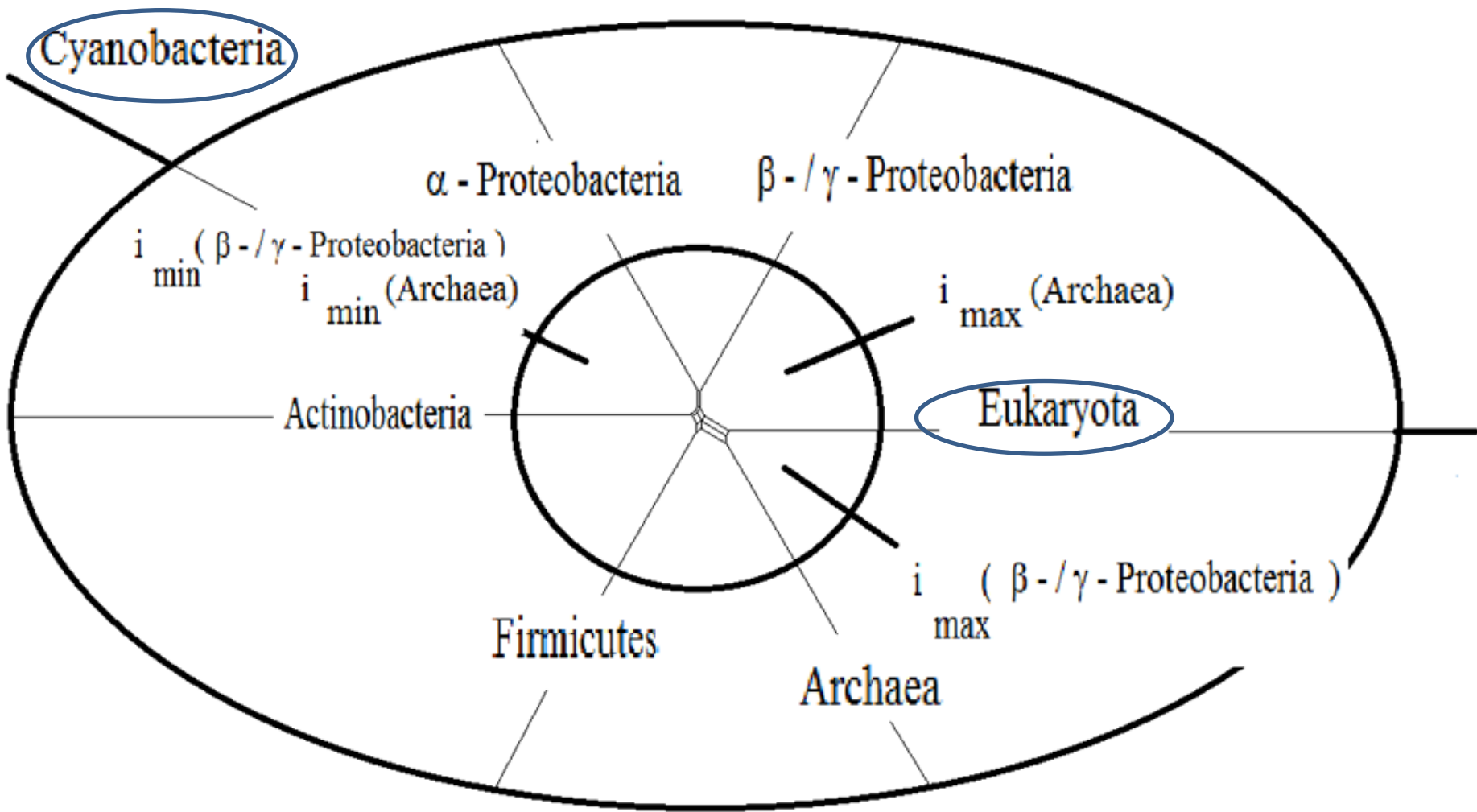
184 Taxa reduced to a 20x20 distance matrix by averaging distances ('Archaea', 'Eukaryota', 18 Phyla or Classes for 'Bacteria').

Distance computed using Genome Conservation Approach (Kunin et al. Nucleic Acids Research 33: 616-621, 2005.)

Backbone: 6 taxa

Input: b= Cyanobacteria

Output: Transfer between 'Cyanobacteria' and 'Eukaryota'  
'Cyanobacteria' adjacent to 'Actinobacteria' and ' $\alpha$ -  
Proteobacteria' in a circular order



## References:

Thuillard, M., Moulton V., Journal of Bioinformatics and Computational Biology, in Press

Bryant D, Moulton V, Molecular Biology and Evolution 21: 255-265, 2004.

Thuillard M, Evolutionary Bioinformatics 4: 237-247, 2008.