

# Computing the Transposition Distance between Phylogenetic Trees

Ricardo Alberich    Francesc Rosselló    **Gabriel Valiente**

Department of Mathematics and Computer Science  
University of the Balearic Islands

Algorithms, Bioinformatics, Complexity and Formal Methods Research Group  
Technical University of Catalonia

London Stringology Day and London Algorithmic Workshop  
King's College London, UK, 7–8 February 2007

## Abstract

Unlike most metrics for phylogenetic tree comparison that are based on elementary edit operations, the transposition distance between phylogenetic trees can be computed in polynomial time. As a matter of fact, a linear time algorithm is known for computing the transposition distance between two fully resolved phylogenetic trees. In this talk, we recall a bijection between fully resolved phylogenetic trees and matching permutations, present an embedding of general phylogenetic trees with  $n$  leaves labeled  $1, \dots, n$  into the symmetric permutation group  $\mathcal{S}_{2n-2}$ , and present a linear time algorithm for computing the transposition distance between two general phylogenetic trees.

# University of the Balearic Islands

## *Research Institute of Health Science*

- Ricardo Alberich



- Francesc Rosselló



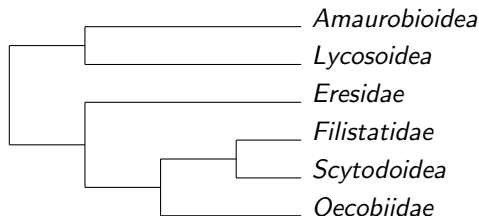
# Contents

- Background
  - Phylogenetic Trees
  - Comparison of Phylogenetic Trees
  - Resolving the Tree of Life
- Matching Representation of Phylogenetic Trees
  - Full Resolution of a Phylogenetic Tree
  - Perfect Matchings and Phylogenetic Trees
- Transposition Distance between Phylogenetic Trees
  - Transposition in the Matching Representation
  - Matching Distance
  - Transposition Distance
- Algebraic Distance between Phylogenetic Trees
  - Algebraic Distance
  - Computing the Algebraic Distance
- Conclusions

# Background

## Phylogenetic Trees

- A **phylogenetic tree** is denoted by  $T = (V, E)$ , where  $V$  is the set of nodes and  $E \subset V \times V$  is the set of edges
- A phylogenetic tree is said to be **fully resolved** if all nodes have either zero or two children
- Sample fully resolved phylogenetic tree



- The **taxon** associated with a leaf node  $v \in V$  is denoted  $\ell(v)$

# Background

## *Comparison of Phylogenetic Trees*



# Background

## *Comparison of Phylogenetic Trees*



# Background

## *Comparison of Phylogenetic Trees*



*Quercus lobata*



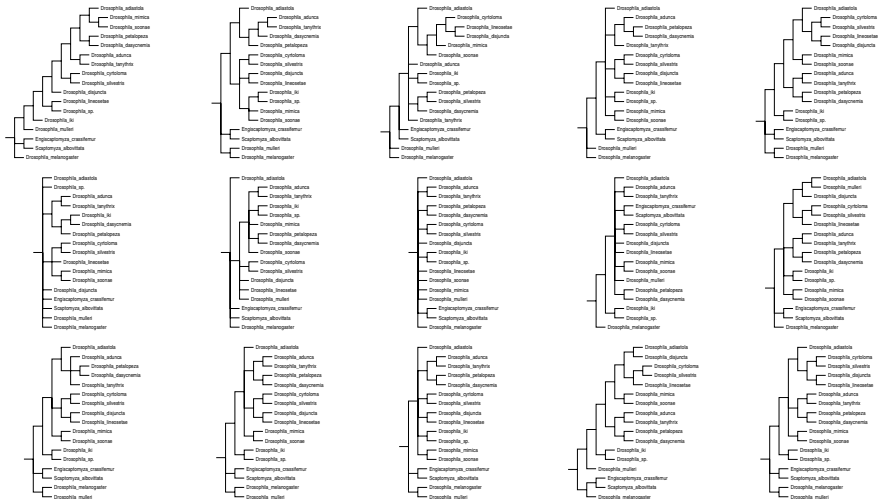
*Quercus alba*



*Prunus x yedoensis*

# Background

## Comparison of Phylogenetic Trees



# Background

## *Comparison of Phylogenetic Trees*

- TreeBASE is the main repository of published phylogenetic analyses
- The comparison of phylogenetic trees is essential to performing phylogenetic queries on databases of phylogenetic trees such as TreeBASE
  - Find trees that share a specified subtree (TreeSearch)
  - Compute tree dissimilarity scores (TreeRank)
- Computation of a phylogenetic query by TreeRank takes  $O(n^2)$  time

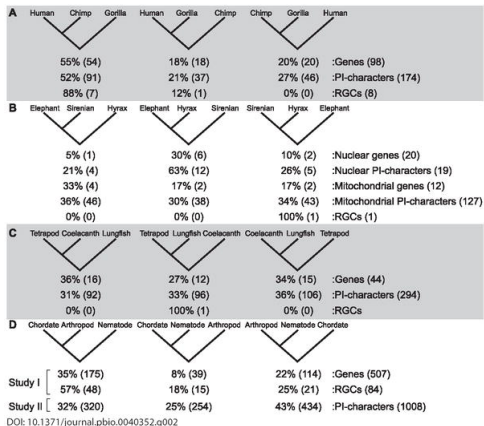
# Background

## *Comparison of Phylogenetic Trees*

- **Nearest neighbor interchange distance** Smallest number of nearest neighbor interchange operations needed to transform one tree into the other
- **Subtree transfer distance** Smallest number of subtree transfer operations needed to transform one tree into the other
- **Partition distance** Number of split differences between the trees
- **Quartet distance** Number of quartet topology differences between the trees
- **Nodal distance** Difference of the distances between each pair of taxa in one tree and in the other
- **Transposition distance** Smallest number of transposition operations needed to transform one tree into the other

# Background

## Resolving the Tree of Life



- Rokas, Carroll *PLOS Biol.* 4 (2006) 1899

# Matching Representation of Phylogenetic Trees

## *Full Resolution of a Phylogenetic Tree*

- Phylogenetic analyses often produce phylogenies with polytomies
  - TreeBASE contains 2,592 phylogenies over 36,593 taxa: 1,725 with polytomies and only 867 fully resolved
- A phylogenetic tree with polytomies can be turned into a fully resolved phylogenetic tree in a canonical way, such that two isomorphic phylogenetic trees have exactly the same full resolution
  - Natural correspondence between general trees and those binary trees that have a root but no right subtree
  - Phylogenetic trees have unique leaf labels

# Matching Representation of Phylogenetic Trees

## *Perfect Matchings and Phylogenetic Trees*

- A **partition of  $\{1, \dots, 2n\}$  into 2-subsets** is a set of  $n$  pairwise disjoint unordered pairs
- $u \prec v$  denotes that node  $u$  is a predecessor of node  $v$  in sorted bottom-up order
- The **matching representation**  $M(T)$  of a fully resolved phylogenetic tree  $T = (V, E)$  with  $n$  leaves labeled  $1, \dots, n$ , is the partition of  $\{1, \dots, 2n - 2\}$  into 2-subsets defined as follows. Let the internal nodes of  $T$  be labeled as  $\ell(v) = \max\{\ell(u) \mid u \prec v\} + 1$ . Then,  $M(T) = \{\{\ell(v), \ell(w)\} \mid (u, v), (u, w) \in E \text{ for some } u \in V\}$
- Diaconis, Holmes *PNAS* 95 (1998) 14600

# Matching Representation of Phylogenetic Trees

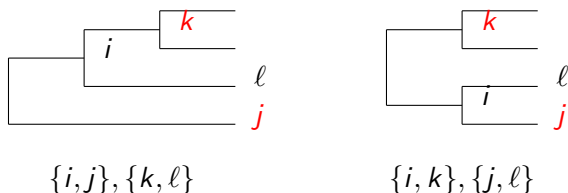
## *Perfect Matchings and Phylogenetic Trees*

- The matching representation can be obtained during a **bottom-up traversal** of the phylogenetic tree
- The internal nodes of  $T$  are labeled according to the following scheme: the parent in  $T$  of the labeled nodes  $v, w \in V$  with smallest  $\ell(v)$  or  $\ell(w)$  is assigned label  $n + 1$ , the one with next-smallest child label is assigned label  $n + 2$ , and so on
- $M(T) = \{ \{ \ell(v), \ell(w) \} \mid (u, v), (u, w) \in E \text{ for some } u \in V \}$

# Transposition Distance between Phylogenetic Trees

## *Transposition in the Matching Representation*

- Generalization of transposition in permutations to partitions into 2-subsets
- Let  $M$  be a partition of  $\{1, \dots, 2n\}$  into 2-subsets, and let  $\{i, j\}, \{k, \ell\} \in M$ . The **transposition of  $M$  at  $j$  and  $k$**  is the replacement of  $\{i, j\}$  by  $\{i, k\}$  and  $\{k, \ell\}$  by  $\{j, \ell\}$  in  $M$
- Sample transposition at  $j$  and  $k$



# Transposition Distance between Phylogenetic Trees

## *Transposition in the Matching Representation*

- Transpositions are sufficient to transform any two partitions of  $\{1, \dots, 2n\}$  into 2-subsets to each other
- Given two partitions  $M_1$  and  $M_2$  of  $\{1, \dots, 2n\}$  into 2-subsets, there exists a set of transpositions that transform  $M_1$  into  $M_2$

# Transposition Distance between Phylogenetic Trees

## *Matching Distance*

- Transpositions are sufficient to transform the matching representations of any two fully resolved phylogenetic trees over the same taxa to each other
- The **matching distance**  $MD(T_1, T_2)$  between two fully resolved phylogenetic trees  $T_1$  and  $T_2$  over the same taxa, is the minimum number of transpositions needed to transform  $M(T_1)$  into  $M(T_2)$

# Transposition Distance between Phylogenetic Trees

## *Matching Distance*

- Let  $M_1$  and  $M_2$  be partitions of  $\{1, \dots, 2n\}$  into 2-subsets, and let  $G = (V, E)$  be the undirected graph with vertex set  $V = \{1, \dots, 2n\}$  and edge set  $E = M_1 \cup M_2$ . Let also  $C$  be the set of connected components of  $G$ . Then,  $MD(M_1, M_2) = |E|/2 - |C|$
- Each connected component  $A$  of  $G$  consists of either a single 2-subset or an alternating cycle of 2-subsets coming in turn from  $M_1$  and  $M_2$

$$MD(M_1, M_2) = \sum_{A \in C} (|A|/2 - 1) = \left( \sum_{A \in C} |A| \right) / 2 - |C| = |E|/2 - |C|$$

# Transposition Distance between Phylogenetic Trees

## *Transposition Distance*

- The **transposition distance**  $TD(T_1, T_2)$  between two fully resolved phylogenetic trees  $T_1$  and  $T_2$ , is the matching distance between the topological restriction  $T_1|L$  of  $T_1$  and  $T_2|L$  of  $T_2$  to their common taxa  $L = L_1 \cap L_2$
- For phylogenetic trees with overlapping taxa, the transposition distance can be extended by also taking non-common taxa and contracted edges (in the topological restriction of the trees to their common taxa) into account
- The transposition distance can be **normalized** to a value between zero and one by dividing by the total size of the trees, which is an upper bound on the transposition distance

# Algebraic Distance between Phylogenetic Trees

## *Algebraic Distance*

- Defined for phylogenetic trees over the same taxa
  - $T$  has leaves  $\mathcal{L}(T)$ , leaf  $x$  has label  $\ell_T(x)$
- Matching representation
  - $M(T) = \{\ell_T(\text{children}(x)) \mid x \in V - \mathcal{L}(T)\}$
- Permutation representation
  - $\pi(T) = \prod_{x \in V - \mathcal{L}(T)} \kappa(\ell_T(\text{children}(x))) \in \mathcal{S}_{2n-2}$
- Transposition distance
  - $d(T_1, T_2) = |\pi(T_2)^{-1}\pi(T_1)|$

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

- Defined for phylogenetic trees over the same taxa
  - Topological restriction
- Matching representation
  - Bottom-up traversal
- Permutation representation
  - Bottom-up traversal
- Transposition distance
  - Number of non-isolated vertices minus number of alternating cycles

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

### Definition

The **directed graph** associated to a permutation  $\pi \in \mathcal{S}_m$  is the graph  $G_\pi = (\{1, \dots, m\}, Q_\pi)$  with

$$Q_\pi = \{(i, j) \mid i \neq j \text{ and } \pi(i) = j\}$$

The directed graph  $G_{\pi^{-1}}$  associated to the inverse  $\pi^{-1}$  of a permutation  $\pi$  is obtained by reversing all arrows in  $G_\pi$

$$Q_{\pi^{-1}} = Q_\pi^{-1} \quad G_{\pi^{-1}} = G_\pi^{-1}$$

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

### Definition

Given two permutations  $\pi_1, \pi_2 \in \mathcal{S}_m$ ,  $G_{\pi_1} + G_{\pi_2}^{-1}$  is the 2-colored-arcs multigraph with set of nodes  $\{1, \dots, m\}$ , set of red arcs  $Q_{\pi_1}$  and set of blue arcs  $Q_{\pi_2}^{-1}$

### Definition

A node of  $G_{\pi_1} + G_{\pi_2}^{-1}$  is **unbalanced** when it is isolated in one, and only one, of the graphs  $G_{\pi_1}, G_{\pi_2}^{-1}$  (which means that it is fixed by one, and only one, of the permutations  $\pi_1, \pi_2$ )

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Theorem

For every unbalanced node  $i$  of  $G_{\pi_1} + G_{\pi_2}^{-1}$

- (1) If  $i$  is isolated in  $G_{\pi_2}$  and  $(i_0, i), (i, i_1) \in Q_{\pi_1}$  with  $i_0 \neq i_1$ , then replacing the red arcs  $(i_0, i)$  and  $(i, i_1)$  by a single red arc  $(i_0, i_1)$  increases  $d(\pi_1, \pi_2)$  by 1
- (2) If  $i$  is isolated in  $G_{\pi_2}$  and  $(i, i_1), (i_1, i) \in Q_{\pi_1}$ , removing the red arcs  $(i, i_1)$  and  $(i_1, i)$  increases  $d(\pi_1, \pi_2)$  by 1
- (3) Similar properties hold if  $i$  is isolated in  $G_{\pi_1}$  but not in  $G_{\pi_2}$  and we modify the set of blue arcs

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

### Theorem

*If  $G_{\pi_1} + G_{\pi_2}^{-1}$  has no unbalanced node, then*

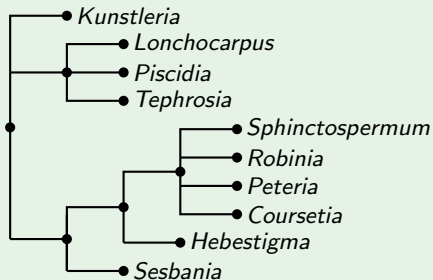
$$d(\pi_1, \pi_2) = N(G_{\pi_1}, G_{\pi_2}^{-1}) - A(G_{\pi_1}, G_{\pi_2}^{-1})$$

*where  $N(G_{\pi_1}, G_{\pi_2}^{-1})$  is the number of non-isolated nodes of  $G_{\pi_1} + G_{\pi_2}^{-1}$  and  $A(G_{\pi_1}, G_{\pi_2}^{-1})$  is the number of alternating cycles in  $G_{\pi_1} + G_{\pi_2}^{-1}$*

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

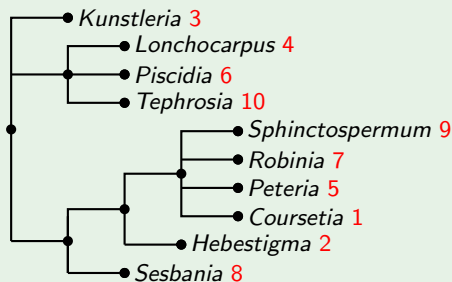
Example (TreeBASE study S11x6x95c08c52c19)



# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

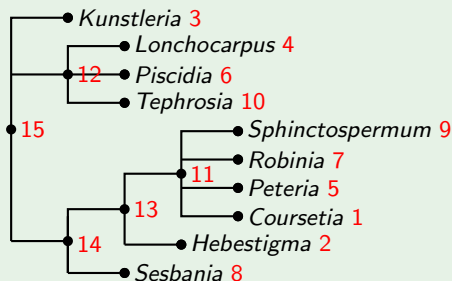
Example (TreeBASE study S11x6x95c08c52c19)



# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example (TreeBASE study S11x6x95c08c52c19)

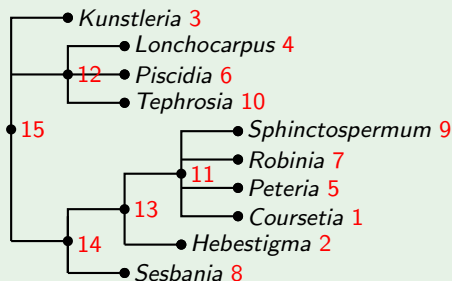




# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example (TreeBASE study S11x6x95c08c52c19)

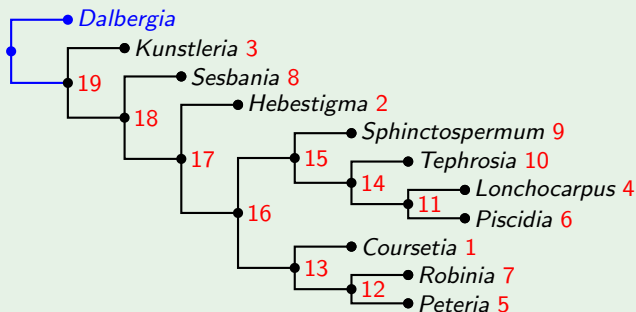


$$\begin{aligned} \pi(T_1) &= (1, 5, 7, 9)(4, 6, 10)(2, 11)(8, 13)(3, 12, 14) \\ &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 5 & 11 & 12 & 6 & 7 & 10 & 9 & 13 & 1 & 4 & 2 & 14 & 8 & 3 \end{pmatrix} \end{aligned}$$

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

Example (TreeBASE study S11x5x95c16c31c52)

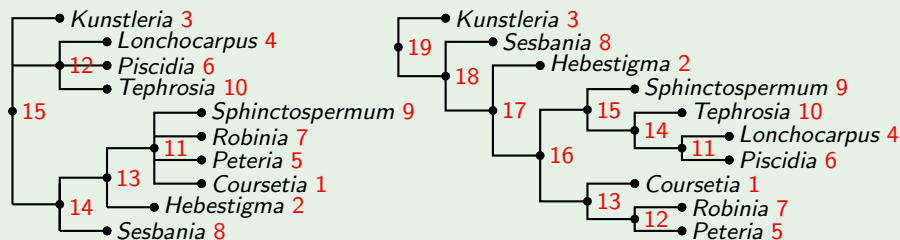


$$\begin{aligned} \pi(T_2) &= (4, 6)(5, 7)(1, 12)(10, 11)(9, 14)(13, 15)(2, 16)(8, 17)(3, 18) \\ &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\ 12 & 16 & 18 & 6 & 7 & 4 & 5 & 17 & 14 & 11 & 10 & 1 & 15 & 9 & 13 & 2 & 8 & 3 \end{pmatrix} \end{aligned}$$

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example



$$\pi(T_1) = (1, 5, 7, 9)(4, 6, 10)(2, 11)(8, 13)(3, 12, 14)$$

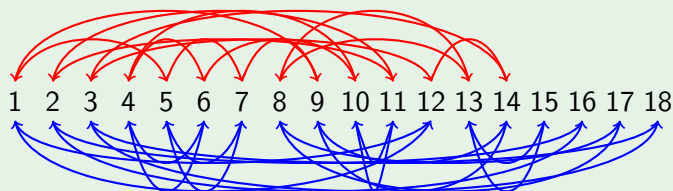
$$\pi(T_2) = (4, 6)(5, 7)(1, 12)(10, 11)(9, 14)(13, 15)(2, 16)(8, 17)(3, 18)$$

$$d(T_1, T_2) = 9 + 8 - 5 = 12$$

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

### Example

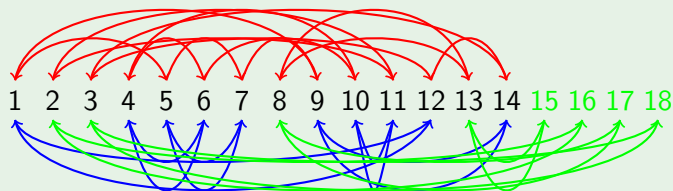


- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

### Example

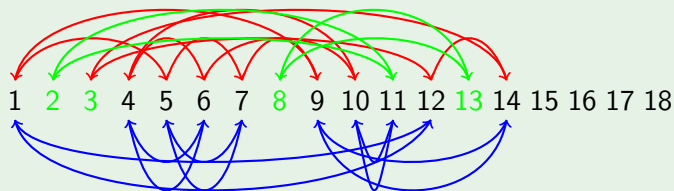


- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

# Algebraic Distance between Phylogenetic Trees

## *Computing the Algebraic Distance*

### Example

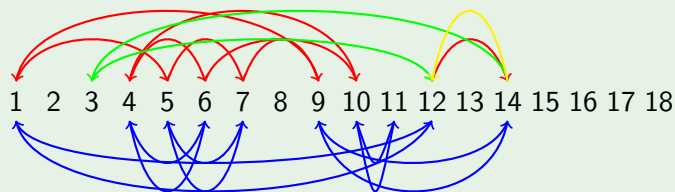


- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example

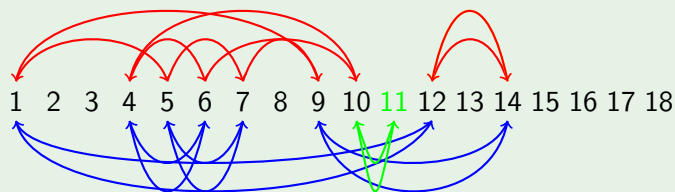


- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example

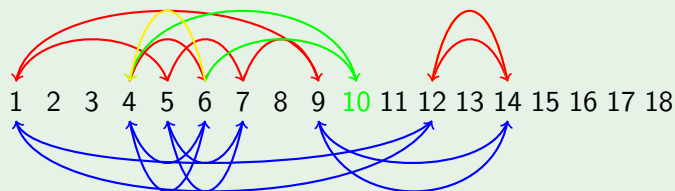


- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example

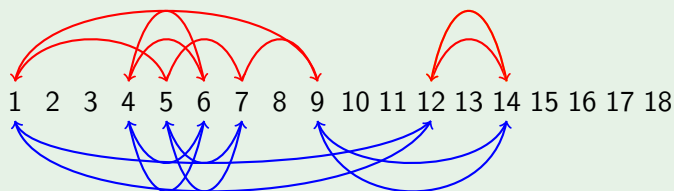


- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

# Algebraic Distance between Phylogenetic Trees

## Computing the Algebraic Distance

### Example



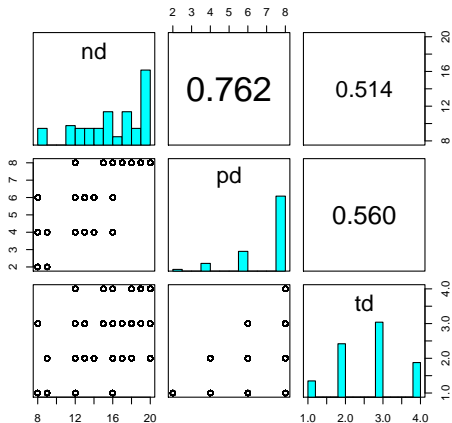
- alternating cycle 4—6—4
- alternating cycle 4—6—4
- alternating cycle 5—7—5
- alternating cycle 1—5—7—9—14—12—1
- alternating cycle 1—12—14—9—1

## Conclusions

- The transposition distance between fully resolved phylogenetic trees is based on a well-known bijection between perfect matchings and phylogenetic trees
- The matching representation, matching distance, and transposition distance can all be computed in  $O(n)$  time
- The algebraic distance between phylogenetic trees, which generalizes the transposition distance between fully resolved phylogenetic trees, can also be computed in  $O(n)$  time
- Further work is needed to establish additional properties of the transposition and algebraic distances

# Conclusions

- Correlation for 105 unrooted phylogenies with 6 taxa



## Conclusions

- G. Valiente. A fast algorithmic technique for comparing large phylogenetic trees. In *Proc. 12th Int. Symp. String Processing and Information Retrieval*, volume 3772 of *Lecture Notes in Computer Science*, pages 371–376. Springer-Verlag, 2005
- 10th Meeting on Computer Algebra and Applications (EACA 2006)

