Fast Approximation Algorithms for Tree Distance Problems

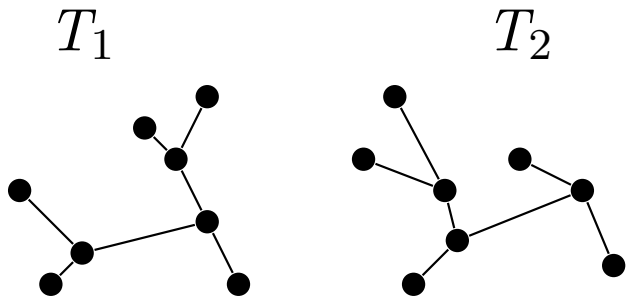
Mikael Gast

Institut für Informatik
Universität Bonn

March 1, 2010

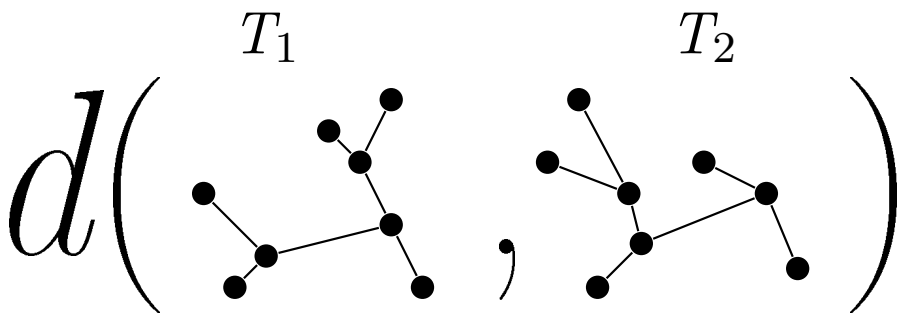
universitätbonn

Given two Trees...



universitätbonn

Given two Trees...

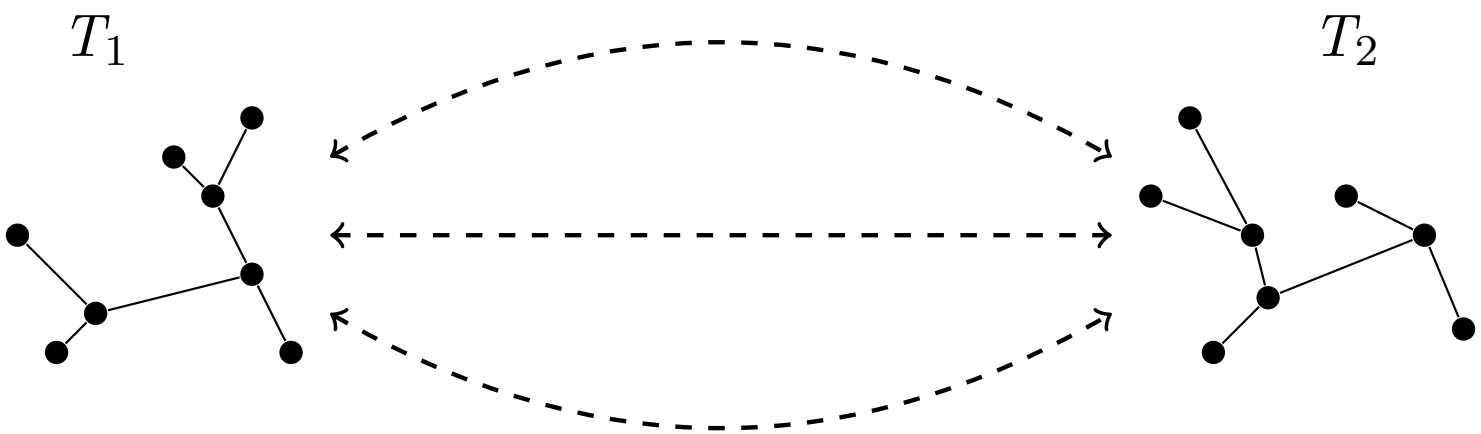


How to measure a Distance d between them?

universitätbonn

Tree Distances

- Different classification of distance-measures
  - → Edit distances, top-down distances, alignment distances, common-subtree distances
- Numerous areas of application
  - → Graph transformation, information processing, image processing, pattern recognition, signal processing, chemistry, systematic and molecular biology



universitätbonn

Tree Distances

In systematic and molecular biology (phylogeny):

- Transition and/or distance measuring between *phylogenetic trees*
- Via *edit distance* or *isolated-subtree distance*
  - → Edit operations: inserting, deleting, relabeling of nodes, edges or subtrees
  - → Common-subtrees: largest common subtrees, maximum agreement forest
- Often of high computational complexity

universitätbonn

Table of Contents

universitätbonn

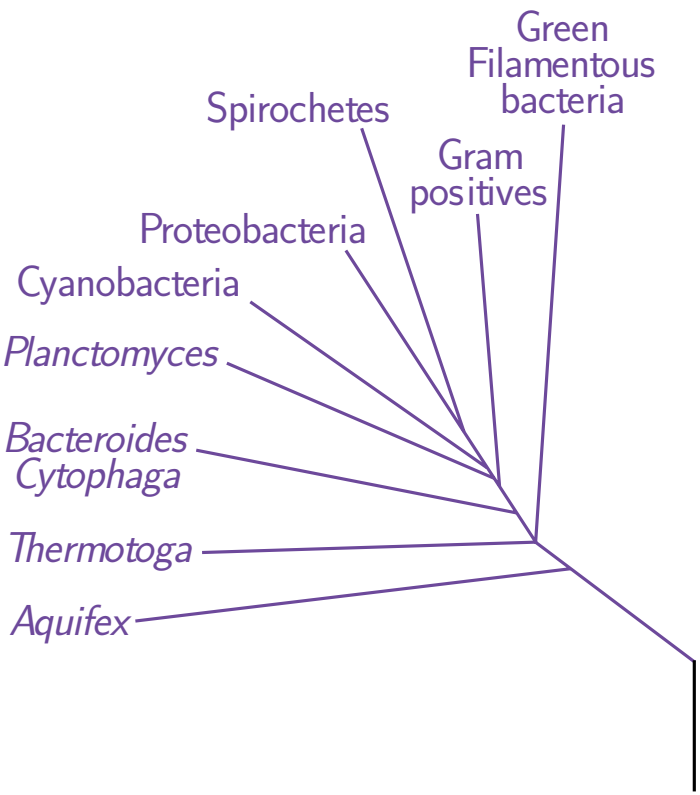
Phylogenetic Trees

- Evolutionary tree – model of representation
- Set of lifeforms or species (so called *Taxa*) at leaf-level
- Internal nodes describe (hypothetic) ancestral history

universitätbonn

Phylogenetic Trees

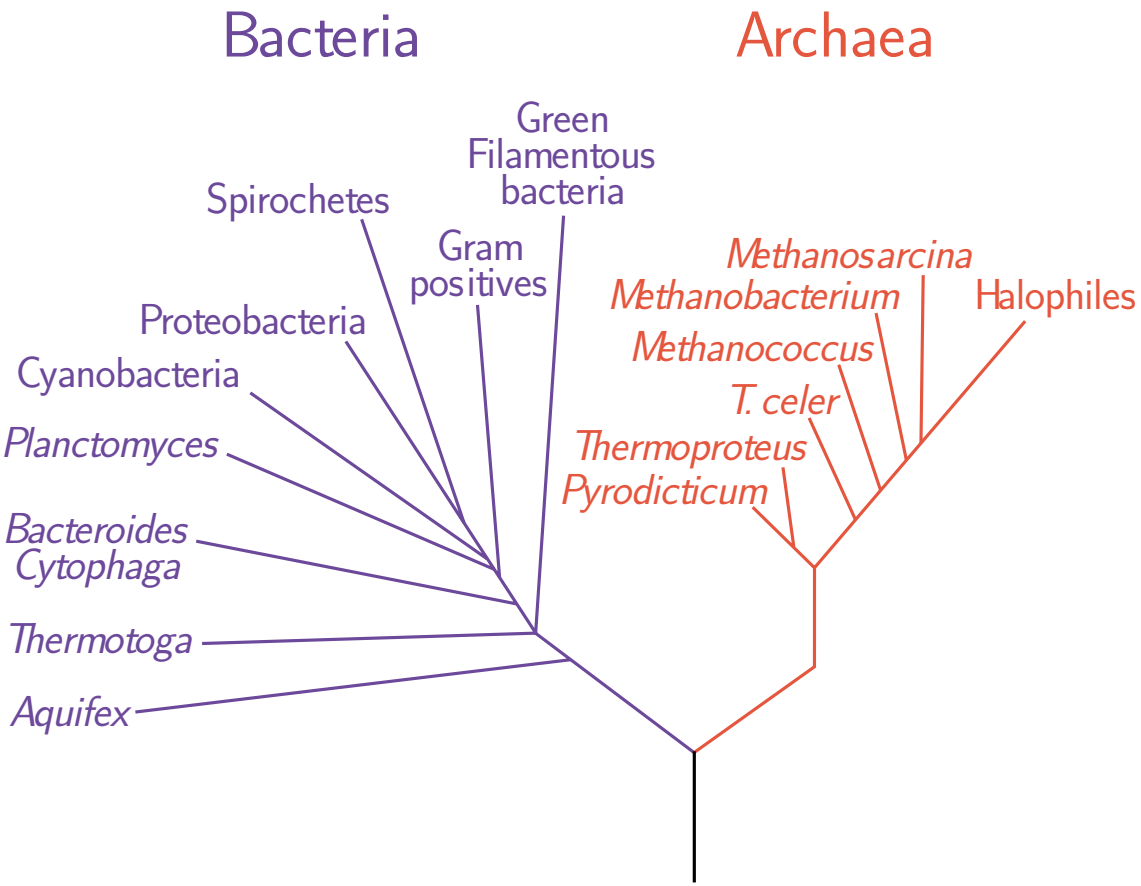
- Evolutionary tree – model of representation
- Set of lifeforms or species (so called *Taxa*) at leaf-level
- Internal nodes describe (hypothetic) ancestral history



universitätbonn

Phylogenetic Trees

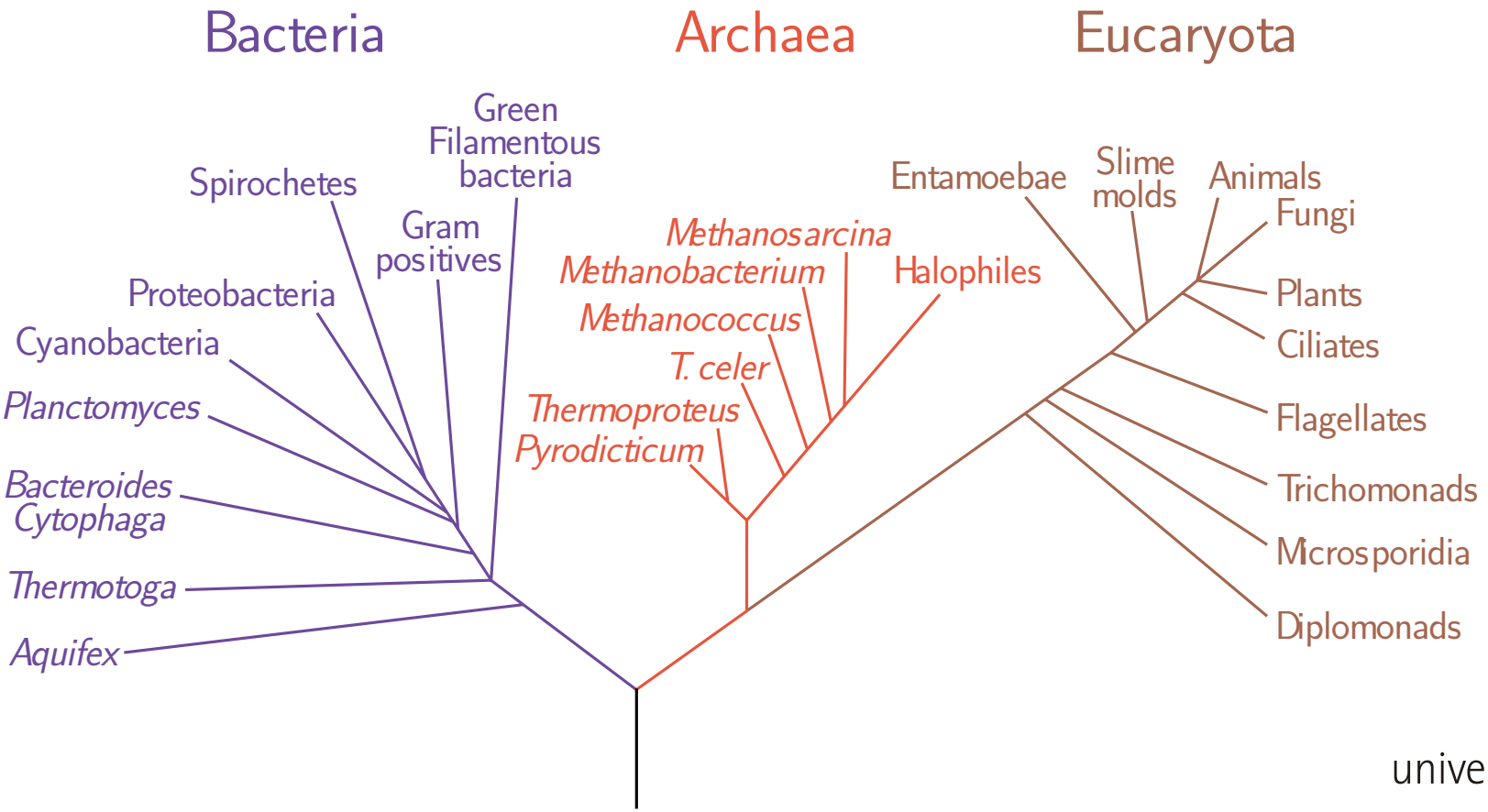
- Evolutionary tree – model of representation
- Set of lifeforms or species (so called *Taxa*) at leaf-level
- Internal nodes describe (hypothetic) ancestral history



universitätbonn

Phylogenetic Trees

- Evolutionary tree – model of representation
- Set of lifeforms or species (so called *Taxa*) at leaf-level
- Internal nodes describe (hypothetic) ancestral history



universitätbonn

Reconstructing Phylogenetic Trees

Problem

Input: Set S of Taxa with pairwise distances and "model of evolution".
Output: (3-regular) tree T with set S at leaf-level and topology reflecting the "model of evolution", a so called *phylogeny* for S.

- Internal nodes are *hypothetical taxonomical units* (HTU's)
- Edges represent ancestral connections
- Edge-weights or path-lengths describe evolutionary distances

universitätbonn

Reconstruction Methods

Example reconstruction methods:

- Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [Sokal and Michener, 1958]
- Maximum-Parsimony [Fitch, 1971]
- Maximum-Likelihood [Felsenstein, 1981]
- Neighbor-Joining [Saitou, 1987]
- String Insertions & Deletions (indel) [Braverman et al., 2009]

  ⋮

universitätbonn

Comparing/Matching Phylogenetic Trees

Problem

Input: Two phylogenies T_1, T_2 over the same set of Taxa S.
Output: (Minimum) distance $d(T_1, T_2)$ between T_1, T_2 with respect to *distance metric* d.

Motivation:

- Measuring – What is the distance between two reconstructed trees over the same set of Taxa?
- Evaluation – Which is the *"best"* tree under the currently viewed model of evolution?
- Transformation – How do we represent underlying *mutations* and *reticulation-events*?

universitätbonn

Comparison Metrics

Subtree-Transfer distance metrics:

- Tree Bisection and Recombination (TBR)
- Subtree Prune and Regraft (SPR)
- Nearest Neighbor Interchange (NNI)

universitätbonn

Comparison Metrics

Subtree-Transfer distance metrics:

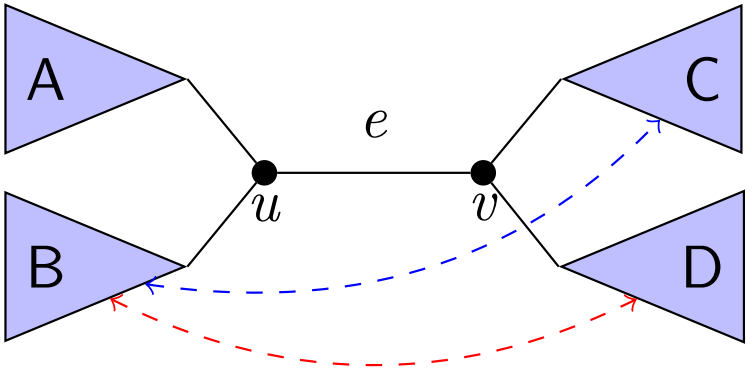
- Tree Bisection and Recombination (TBR)
- Subtree Prune and Regraft (SPR)
- Nearest Neighbor Interchange (NNI)

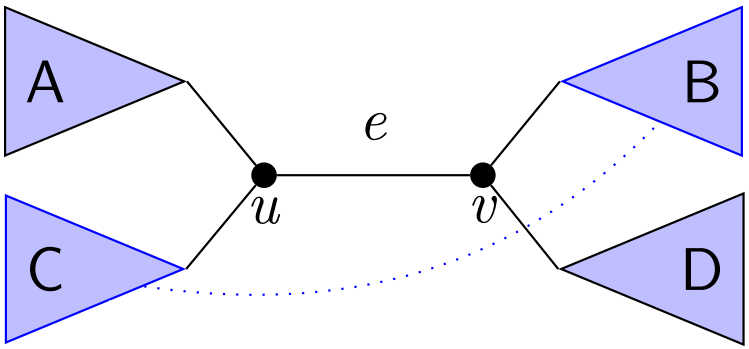
universitätbonn

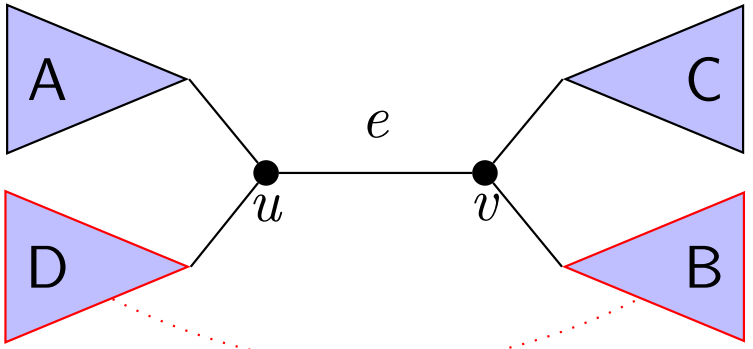
The NNI-distance

The NNI-distance measure:

- Introduced by D.F. Robinson 1971
- Crossover (interchange) operation on subtrees







universitätbonn

The NNI-distance

Definition (NNI-distance)

The NNI-distance $d_{\text{NNI}}(T_1, T_2)$ of T_1, T_2 is the *minimum length* of a sequence of NNI-operations that transforms T_1 into T_2.

(*Minimum cost* in case of weighted phylogenies and $d_{\text{NNI}}(T_1, T_2) = \infty$ in case no such sequence exists).

Theorem (DasGupta et al., 2000)

Let T_1, T_2 be phylogenies for S and k an integer. It is NP-complete to decide if $d_{\text{NNI}}(T_1, T_2) \leq k$.

universitätbonn

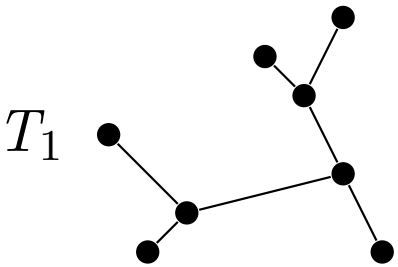
DasGupta's Approximation Algorithm

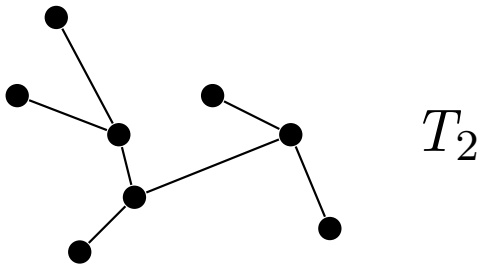
Theorem (DasGupta et al., 2000)

Let T_1, T_2 be phylogenies for S. Then $d_{\mathsf{NNI}}(T_1, T_2)$ and the corresponding sequence of NNI-operations can be approximated within $O(n^2)$ time and approximation ratio $4(1 + \log n)$.

- Time consumption governed by pre-computational steps and sorting procedures
- Approximation ratio reflects upper bound on a number of NNI-operations needed for sorting and transformation

universitätbonn





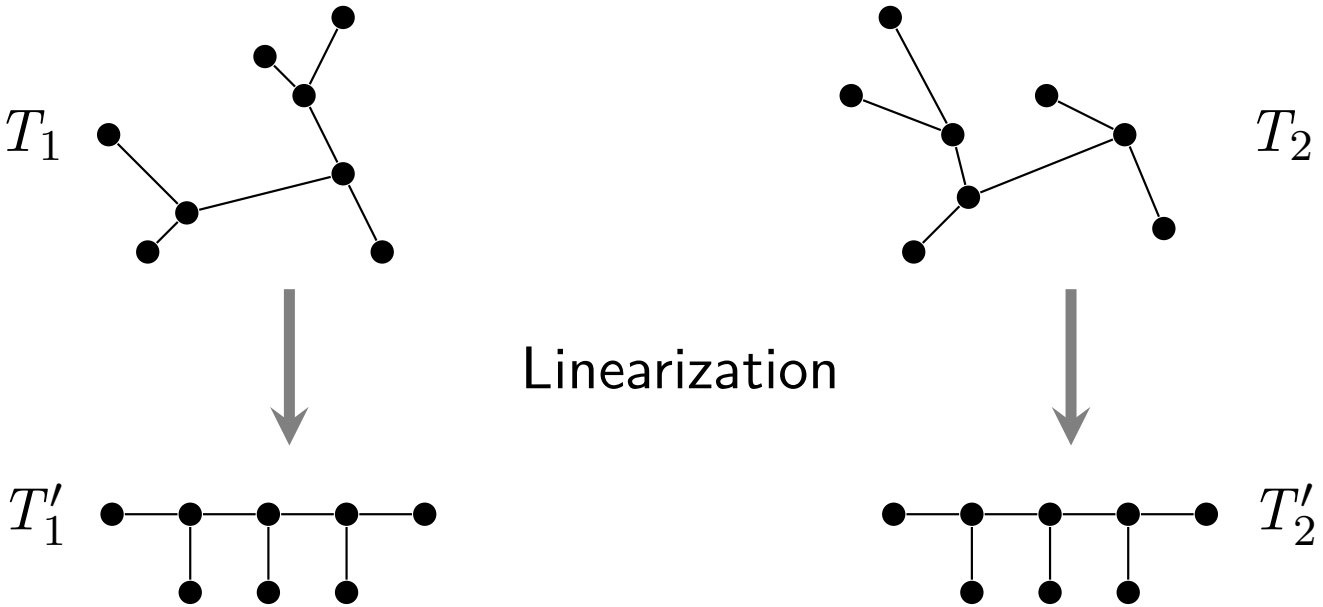
1

2

3

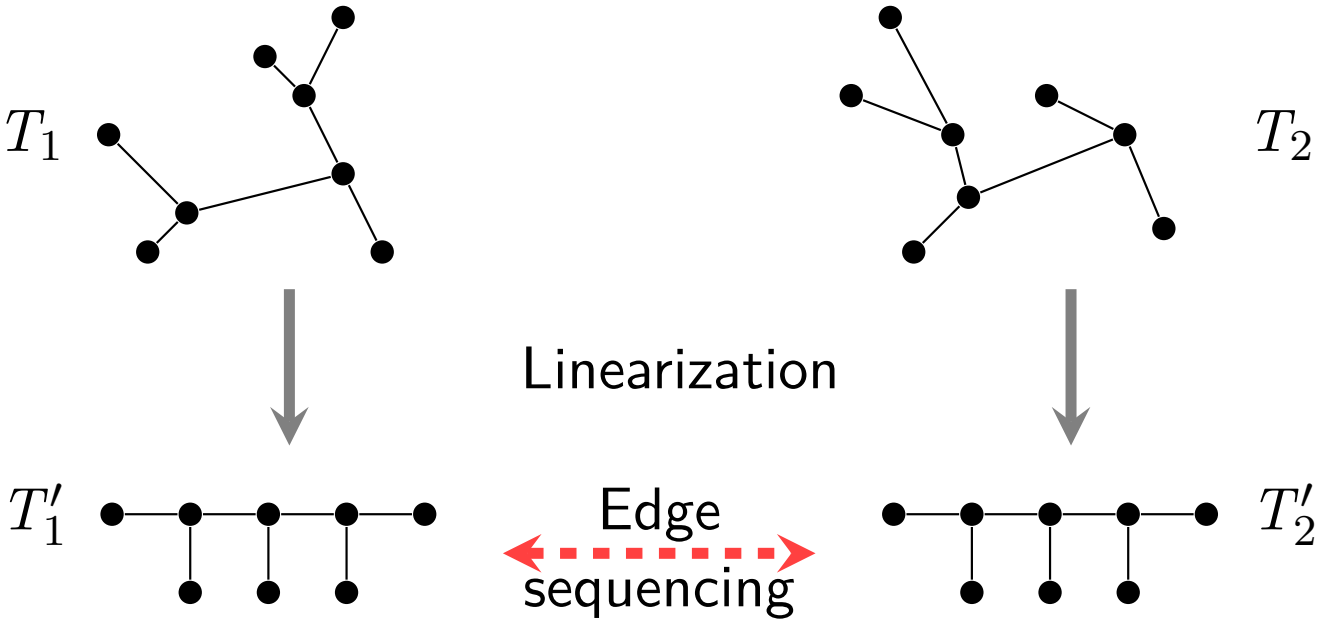
4

universitätbonn



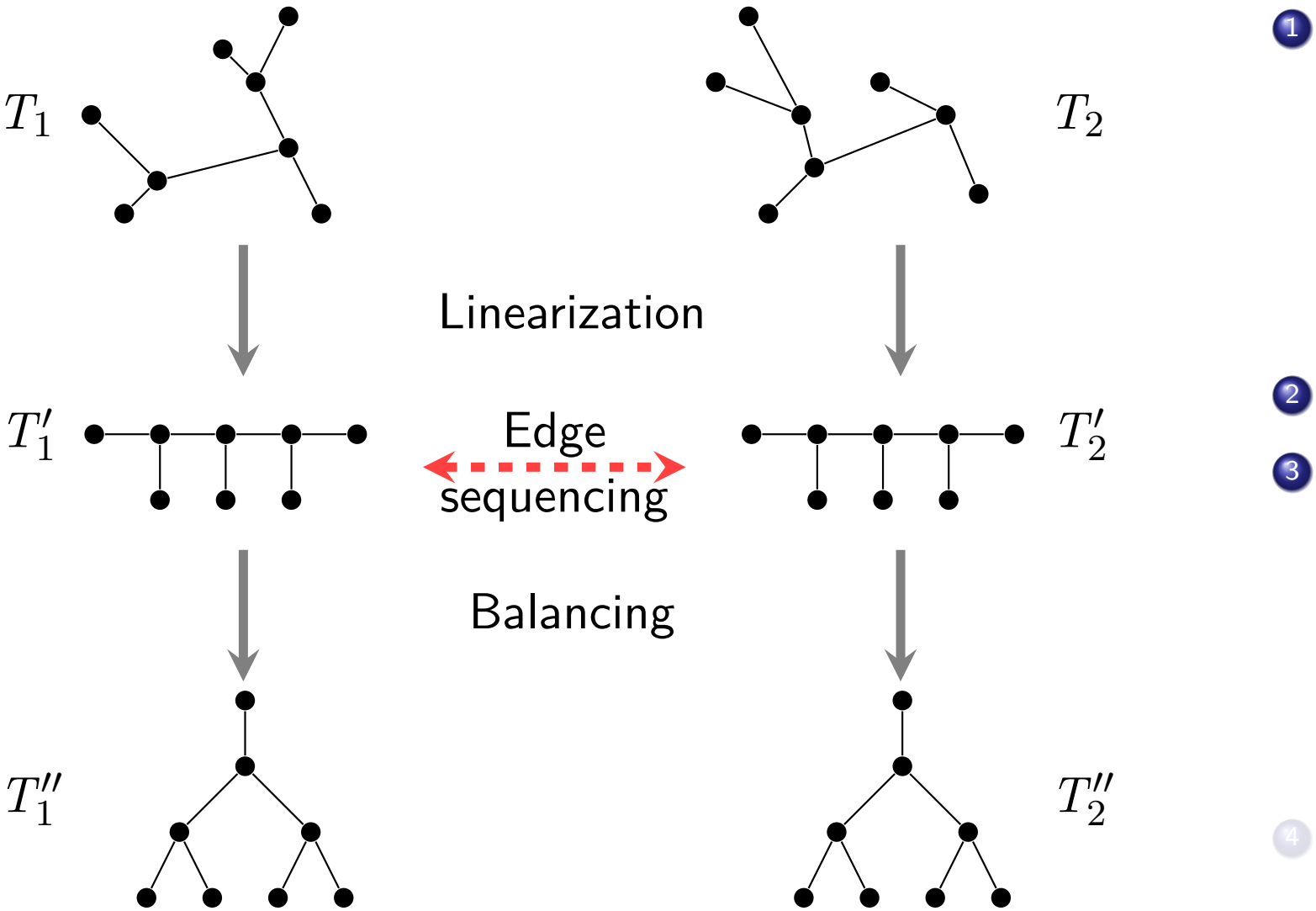
1. 2. 3. 4.

universitätbonn



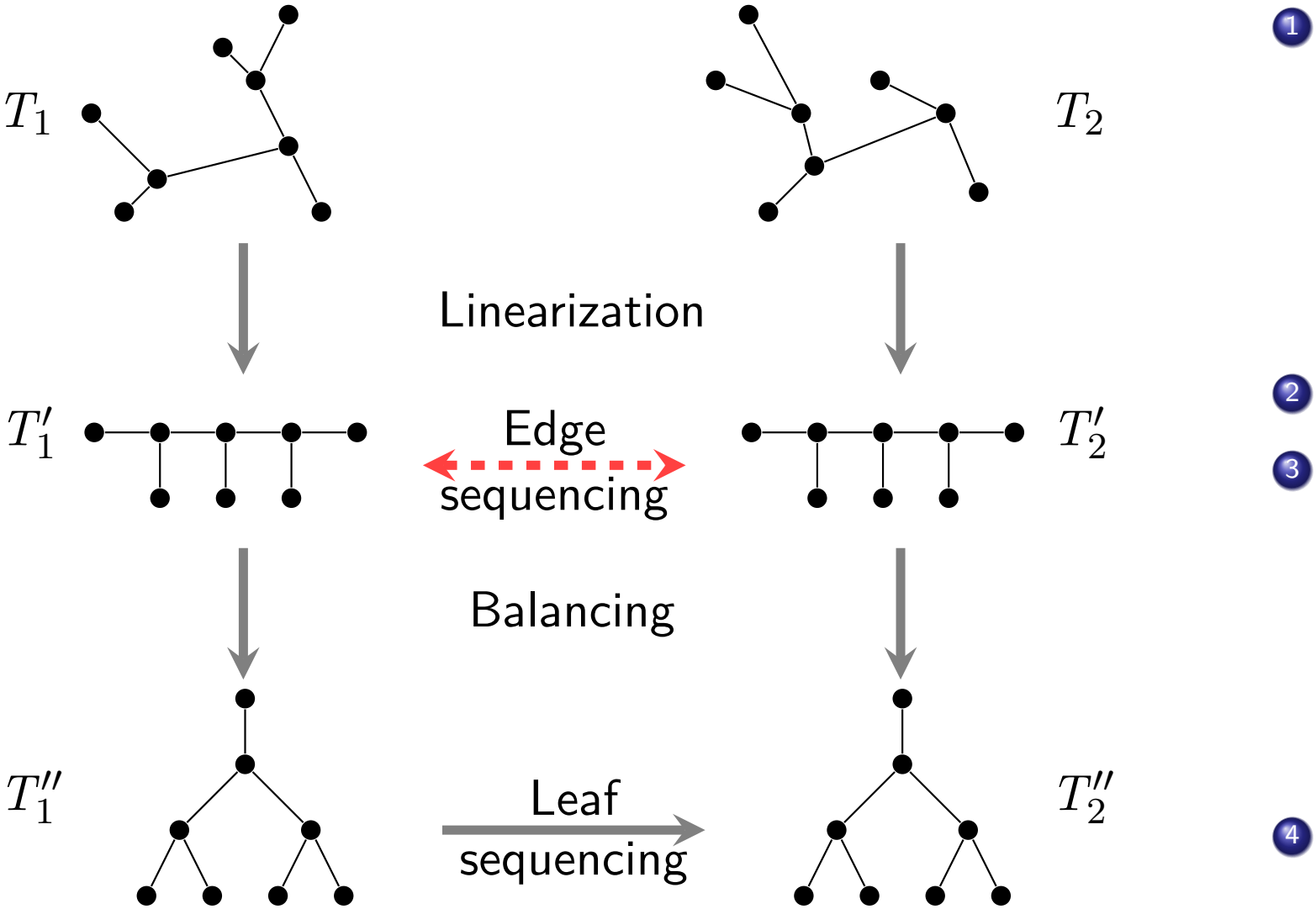
1. 2. 3. 4.

universitätbonn

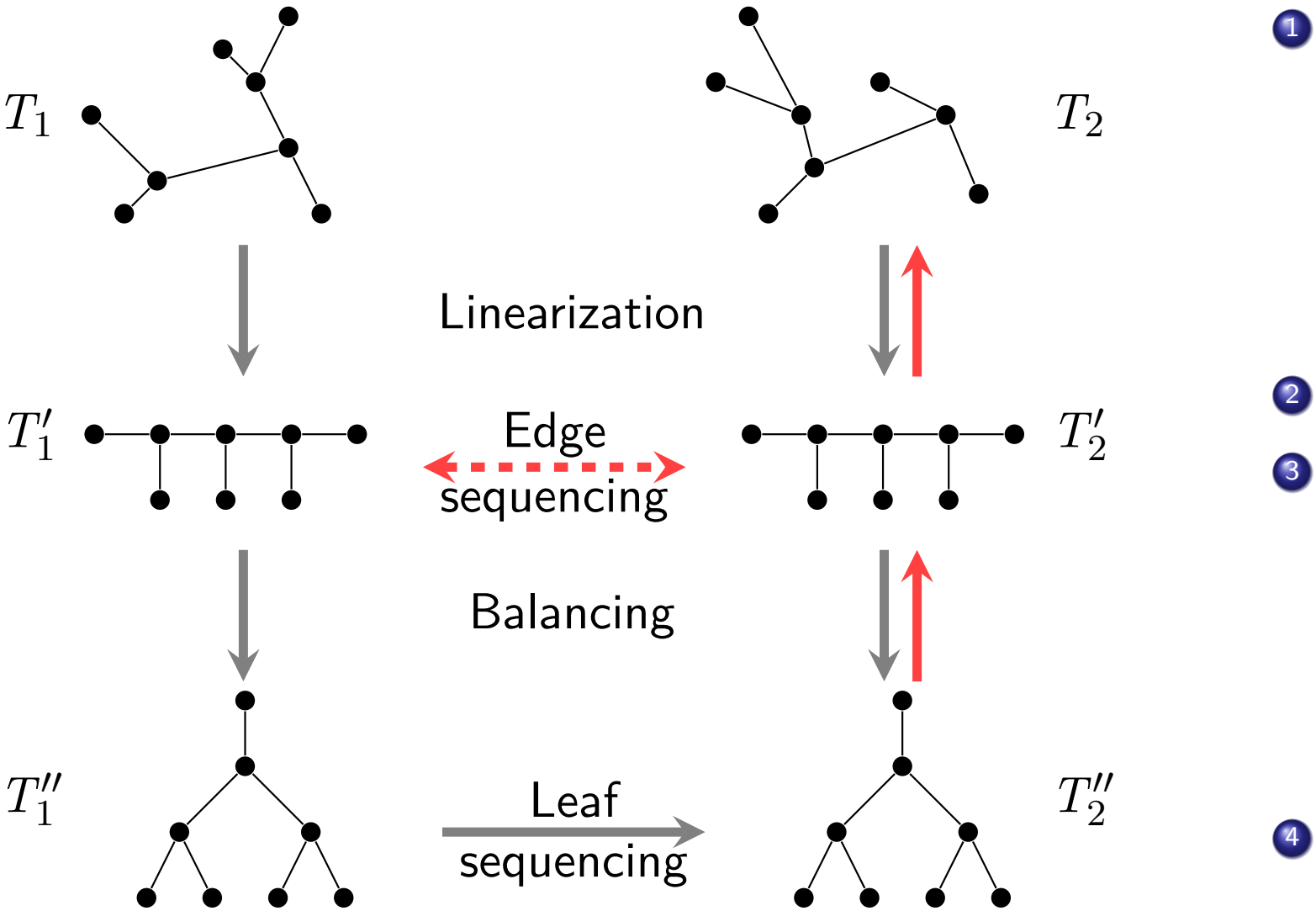


1
2
3
4

universitätbonn



universitätbonn



universitätbonn

Our Result/Parallel Algorithm

Theorem (G. and Hauptmann, 2010)

Let T_1, T_2 be phylogenies for S. Then $d_{\mathsf{NNI}}(T_1, T_2)$ and the corresponding sequence of NNI-operations can be approximated on a CRCW-PRAM with $O(n)$ processors within $O(\log n)$ time and approximation ratio $4(1 + \log n)$.

Method:

- Efficient parallelization of the four main steps of DasGupta's algorithm
- Efficient parallel computation of good edge-pairs (resp. non-shared edges) for problem decomposition

universitätbonn

Parallel Algorithm

Step 1: Linearizing trees

Problem

Input: Phylogeny T for S.
Output: Sequence of NNI-operations that transforms T into a linear tree T', s.t. every internal node is adjacent to at least one leaf.

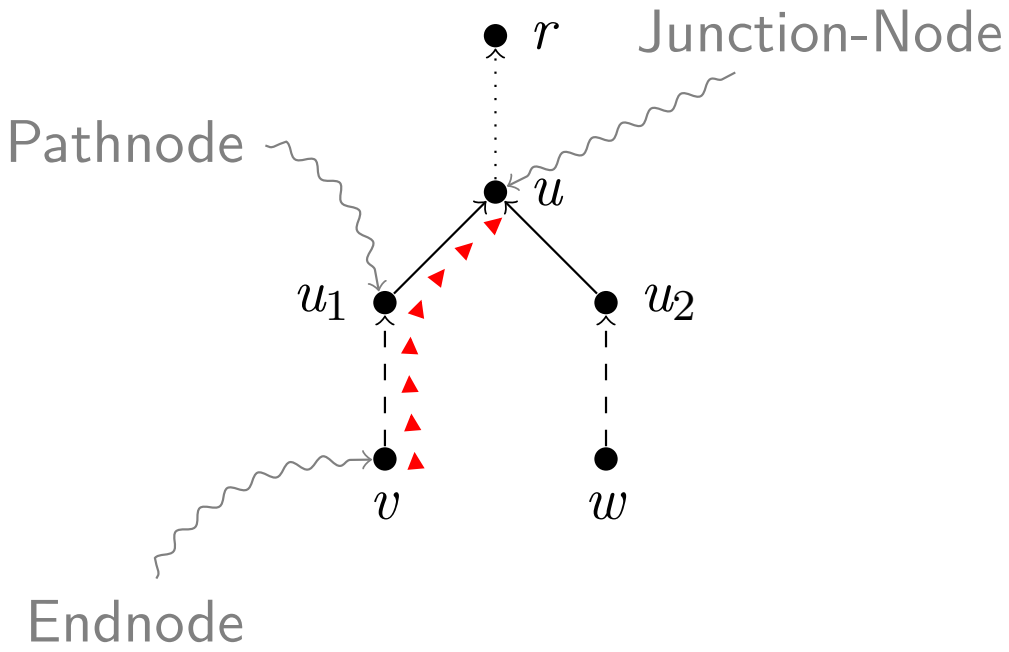
Classification of internal nodes:

- *Pathnodes*, adjacent to one leaf
- *Endnodes*, adjacent to two leaves
- *Junction-Nodes*, adjacent to no leaf but only internal nodes

universitätbonn

Step 1: Linearizing trees

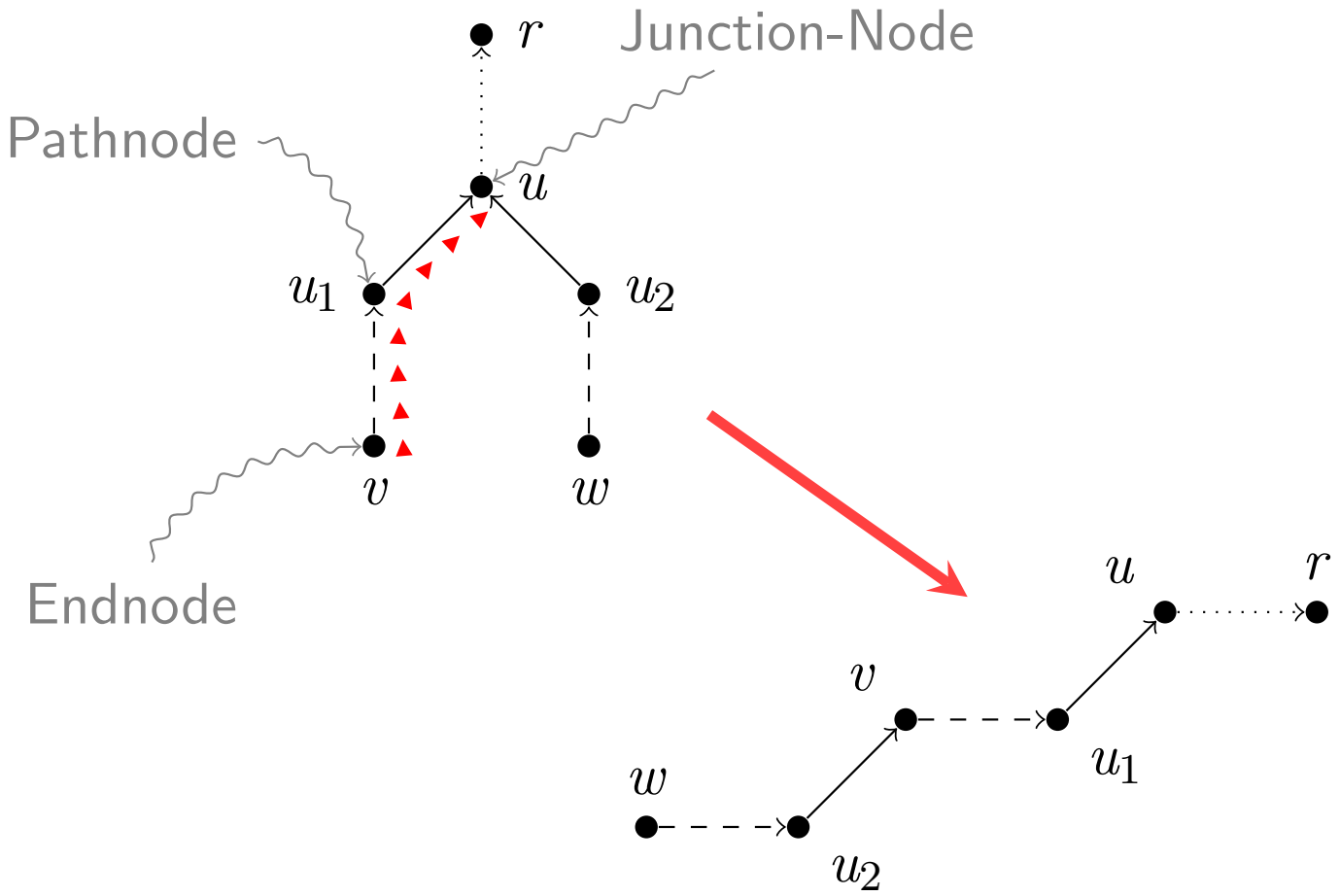
Inserting linearized subtrees and thereby resolving Junction-Nodes:



universitätbonn

Step 1: Linearizing trees

Inserting linearized subtrees and thereby resolving Junction-Nodes:



universitätbonn

Step 1: Linearizing trees

Lemma (Linearization)

The linearization of a tree can be computed in $O(\log n)$ time on a CRCW-PRAM with n processors.

- Parallel processing of all Endnodes
- Number of Endnodes halves with every finalization of a linearization step

$\Rightarrow$ Number of steps bounded by $\lceil \log n \rceil$

universitätbonn

Summary & Further Research

Main results:

- Efficient parallelization of sorting and transformation steps of DasGupta's algorithm
- Exponential running time improvement $O(n^2) \rightsquigarrow O(\log n)$ via parallel computation
- Parallel extraction of the sequence of NNI-operations used for transformation
- Efficient parallel computation of *good edge-pairs* for problem decomposition

universitätbonn

Further Research

Possible future research directions:

- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

Further Research

Possible future research directions:

- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

Further Research

Possible future research directions:

- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

Further Research

Possible future research directions:

- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

Further Research

Possible future research directions:

- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

Further Research

Possible future research directions:

- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

Further Research

Possible future research directions:

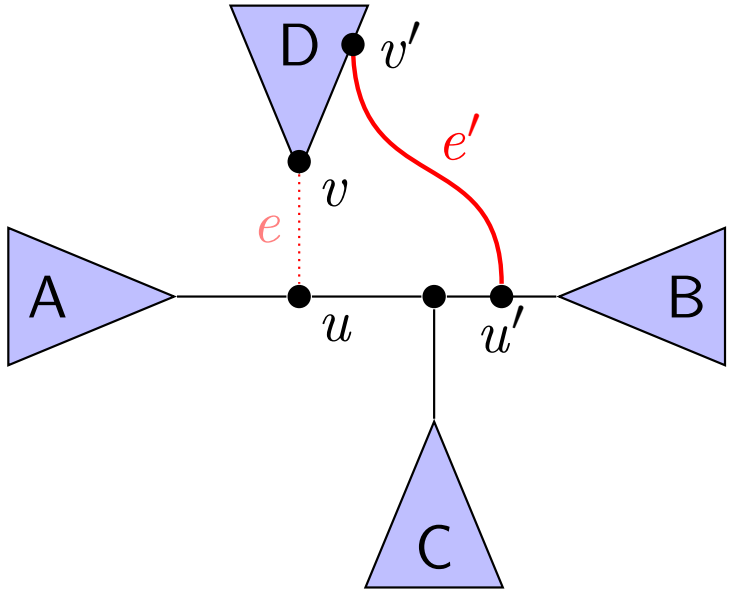
- New methods, paradigms and improvements of the above algorithm and approximation ratio
- Implementation of above algorithm and tackle questions regarding efficiency, scalability
- Parallel computation of SPR and TBR distance problems
- Adaption of good edge-pair heuristics for general splitting of problem instances
- Algorithms (and complexity) in phylogenetic networks
- Finding/defining new metrics for constructing and comparing phylogenetic trees or networks

universitätbonn

The TBR-distance

Tree Bisection and Reconnection:

- TBR-distance problem is NP-hard [Allen and Steel, 2000; Hein, 1996]
- Fixed parameter tractable (FPT) when parameterized with d_{TBR}

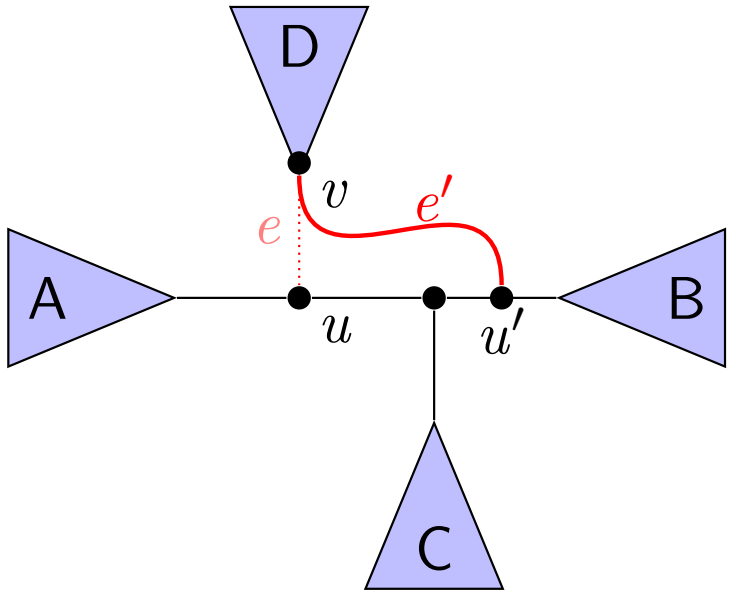


universitätbonn

The SPR-distance

Subtree Prune and Regraft:

- Rooted rSPR-distance problem is NP-hard (correspondence to size of *maximum agreement forest*) [Bordewich and Semple, 2004]
- FPT when parameterized with d_{rSPR}
- Decision problem is NP-complete (reduction from *Exact Cover by 3-Sets*)
- Approximation algorithm with ratio 3 and running time $O(n^5)$ [Bordewich, McCartin and Semple, 2007]



universitätbonn

Phylogenetic networks

Phylogenetic networks:

- Modelling reticulation events (hybridization, horizontal gene transfer, recombination, gene duplication/loss)
- Constructing and comparing phylogenetic networks
- Discuss restrictions to get computationally tractable problems
- Parallel and parameterized algorithms

universitätbonn