

Recovering True Rearrangement Events on Phylogenetic Trees

Hao ZHAO and Guillaume BOURQUE

Genome Institute of Singapore





Outline

- Motivation
- Concepts
- Method
- Simulation and application
- Conclusion and discussion



Motivation

- gene order data provides a whole genome view on phylogeny
- performance criteria
 - Ability to infer correct tree topologies
 - Number of events recovered
 - Quality of ancestral reconstructions
- **A new criteria**
 - Accuracy of predicted rearrangement events
 - Multiple sorting scenarios for even 2 genomes



Basic concepts

- Representation of a chromosome
 - Signed permutation (e.g., 1 2 3 or 1 -3 -2)
 - Each integer corresponds to a gene
 - The sign corresponds to a gene's strand
- A genome can be
 - Uni-chromosomal
 - Multi-chromosomal



Rearrangement events

■ Reversal: 1 2 3 4 \implies 1 -3 -2 4

■ Translocation:

1 2 3 4 1 2 7 8
5 6 7 8 5 6 3 4

\implies

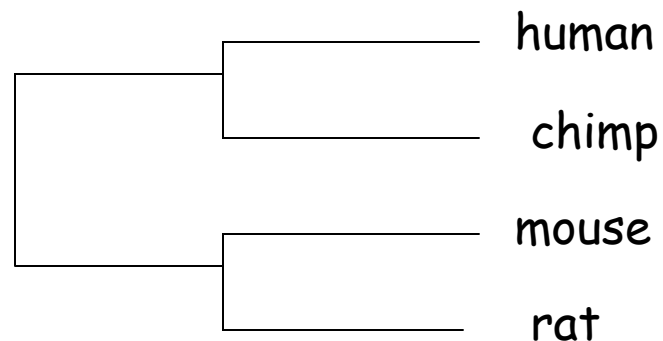
■ Transposition: 1 2 3 4 5 6 7 \implies 1 2 5 6 3 4 7

■ Fission/fusion

1 2 3 4 \longleftarrow 1 2
 \longrightarrow 3 4

Phylogenetic trees

- evolutionary relationship of species
 - Unrooted, binary
 - Leaves are contemporary genomes
 - Internal nodes are ancestors





Inferring rearrangement events

- Given a set of genomes G_1, G_2, \dots, G_m and the phylogenetic tree T , infer the ancestral events on T .
 - The genomes have equal gene content and each gene is unique in each genome
 - reversals, transpositions modeled
- NP-hard to infer most-parsimonious scenario for even 3 genomes (Caprara, '99)



related works

- GRAPPA (Moret, et al, 2001)
 - Only reversals considered
- MGR (Bourque and Pevzner, 2002)
 - Reversal, translocation, fission/fusion allowed
- BADGER (Larget, 2005)
 - MCMC-based
 - Computationally prohibitive

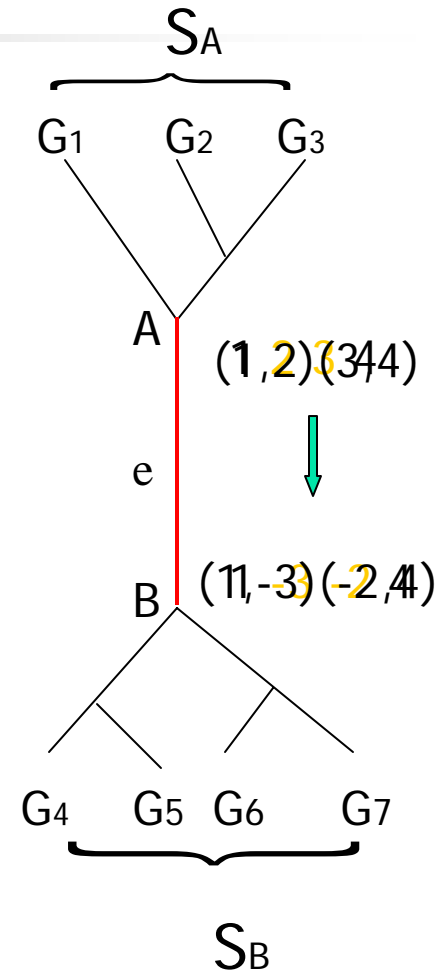


Our contribution

- Efficient Method to Recover Ancestral Events (EMRAE)
 - Only output highly reliable events
 - Better our understanding of the evolutionary mechanisms
 - Model transpositions

Our idea

- *Adjacency*: two adjacent genes
 - e.g. 1 2 3 (1,2) (2,3)
- A reversal affects 4 adjs
 - (1,2) (3,4) in A
 - (1,-3) (-2,4) in B
- *Conserved adjs* for edge (A,B)
 - (1,2) (3,4) in S_A , not in S_B
 - (1,-3) (-2,4) in S_B , not in S_A
 - All_vs_All condition
- 3 adjs to infer transpositions



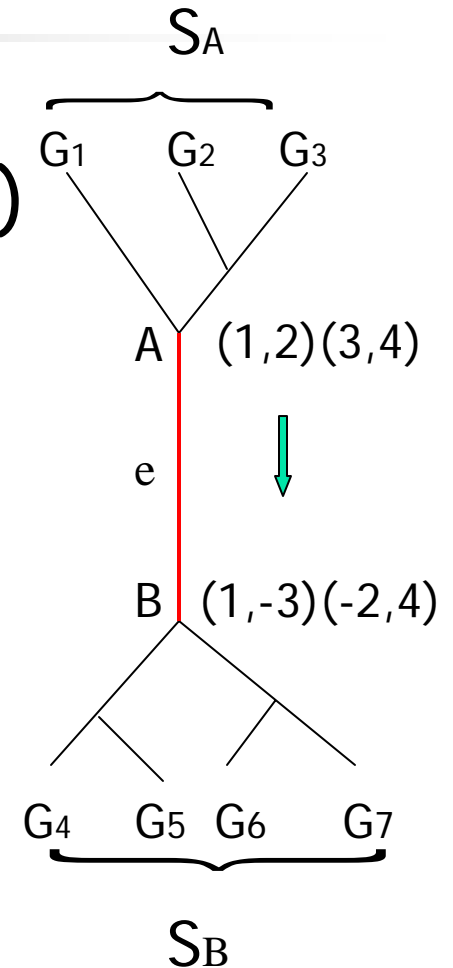
A simple algorithm: All_vs_All

- Given $(1,2)$ $(3,4)$ for S_A , $(1,-3)$ $(-2,4)$ for S_B

- trivial to recover the reversal

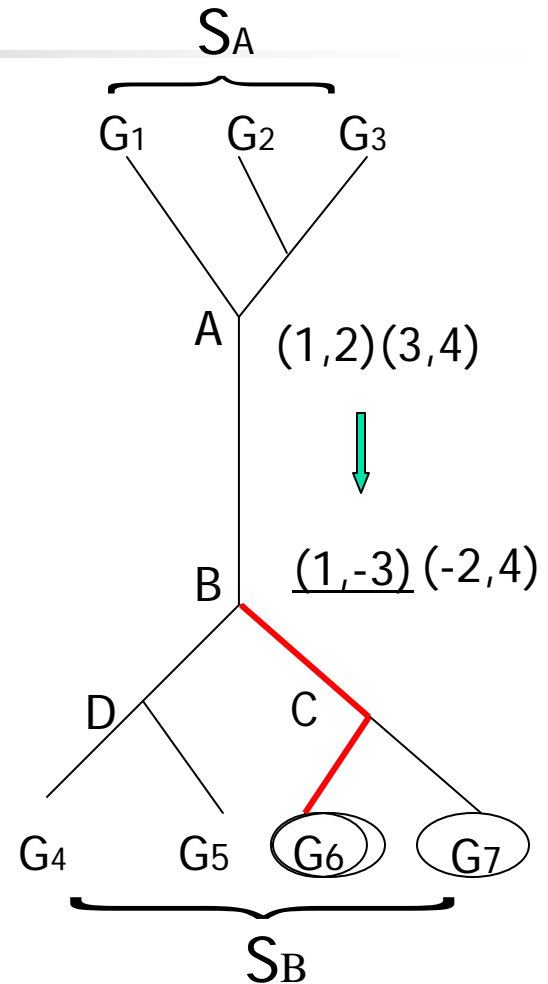
- All_vs_all

- Step 1. extract conserved adjacencies
- Step 2. track back ancestral events



More complicated scenarios

- Break point reuse
 - Case 1. $(1, -3)$ missing in $G6$
 - Case 2. $(1, -3)$ slides away
- Improvement
 - Relax conserved adjacencies
 - "Refinement" to recover sliding adjacencies

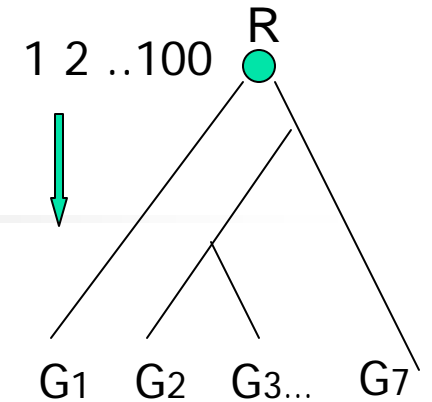




Algorithm (EMRAE)

- Step 1. For each $e = (A, B)$, compute conserved adjacencies for e
- Step 2. refine each edge e
- Step 3. Look for reversals and transpositions

Simulation (random breakage)



1. Generate a rooted tree
2. Put an identity permutation on root R
3. Assign each edge a number $0 < k < 2r$
 - r : evolutionary rate (average nb of events on an edge)
4. From R along each edge, perform k random rearrangements
5. Collect the genomes and tree as input



Performance Criteria

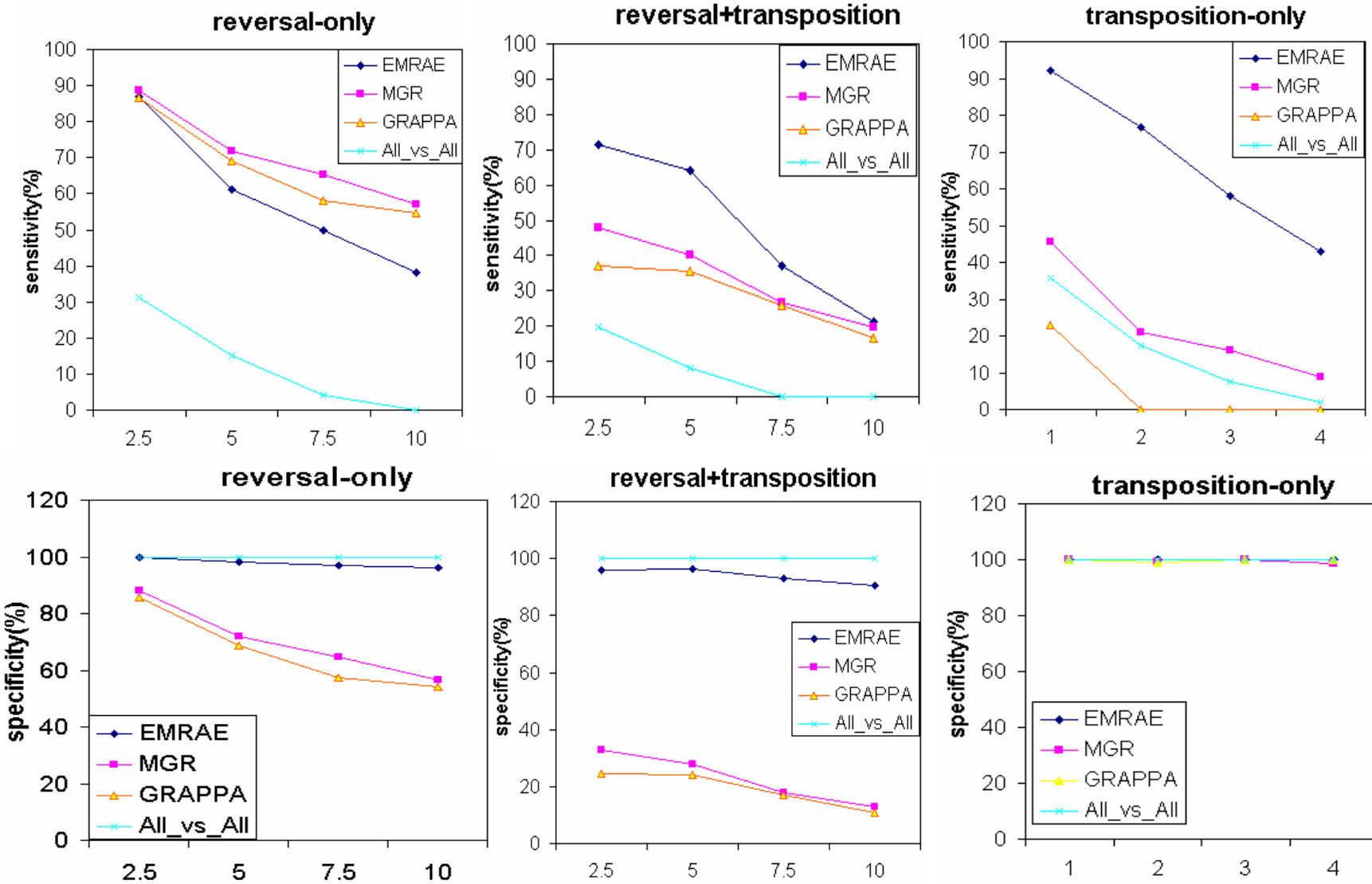
■ $\textit{sensitivity} = \frac{\text{Nb of predicted true events}}{\text{Nb of all true events}} \times 100$

- Sensitivity measures the proportion of predicted real events

■ $\textit{specificity} = \frac{\text{Nb of predicted true events}}{\text{Nb of all predicted events}} \times 100$

- Specificity measures the **quality** of predicted events

Performance of EMRAE

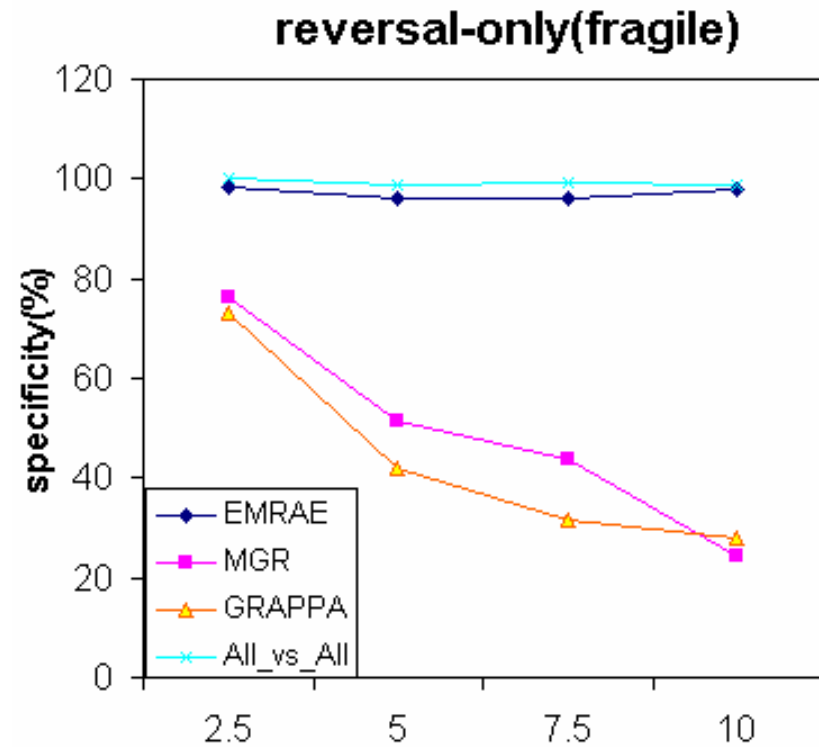
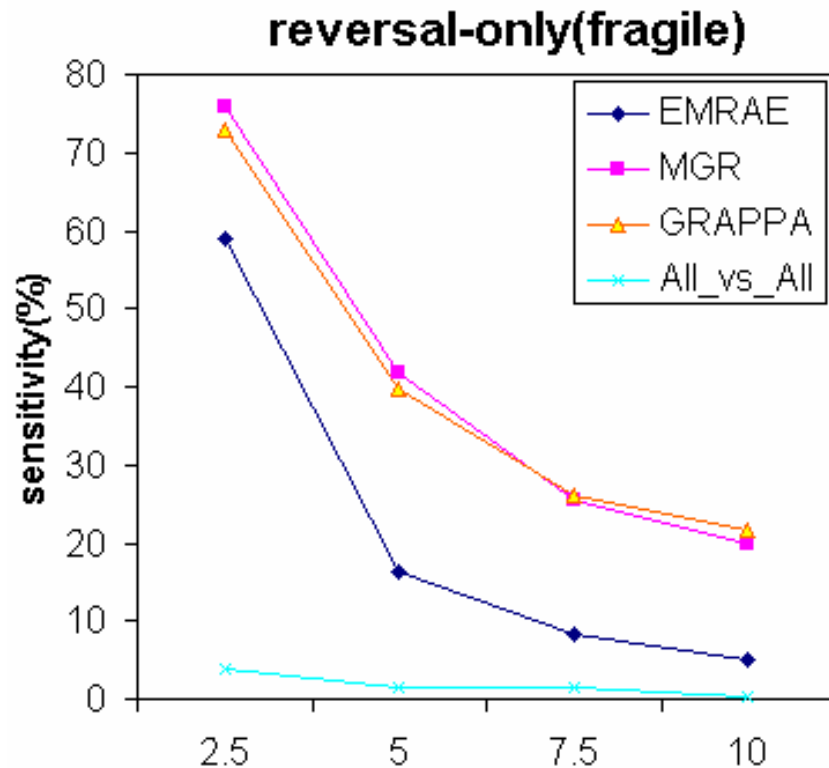




EMRAE's performance

- Comparison between EMRAE, MGR, and GRAPPA
 - EMRAE has comparable sensitivity
 - EMRAE has much higher specificity
 - EMRAE's predictions are highly reliable

Performance of EMRAE on the fragile breakage model



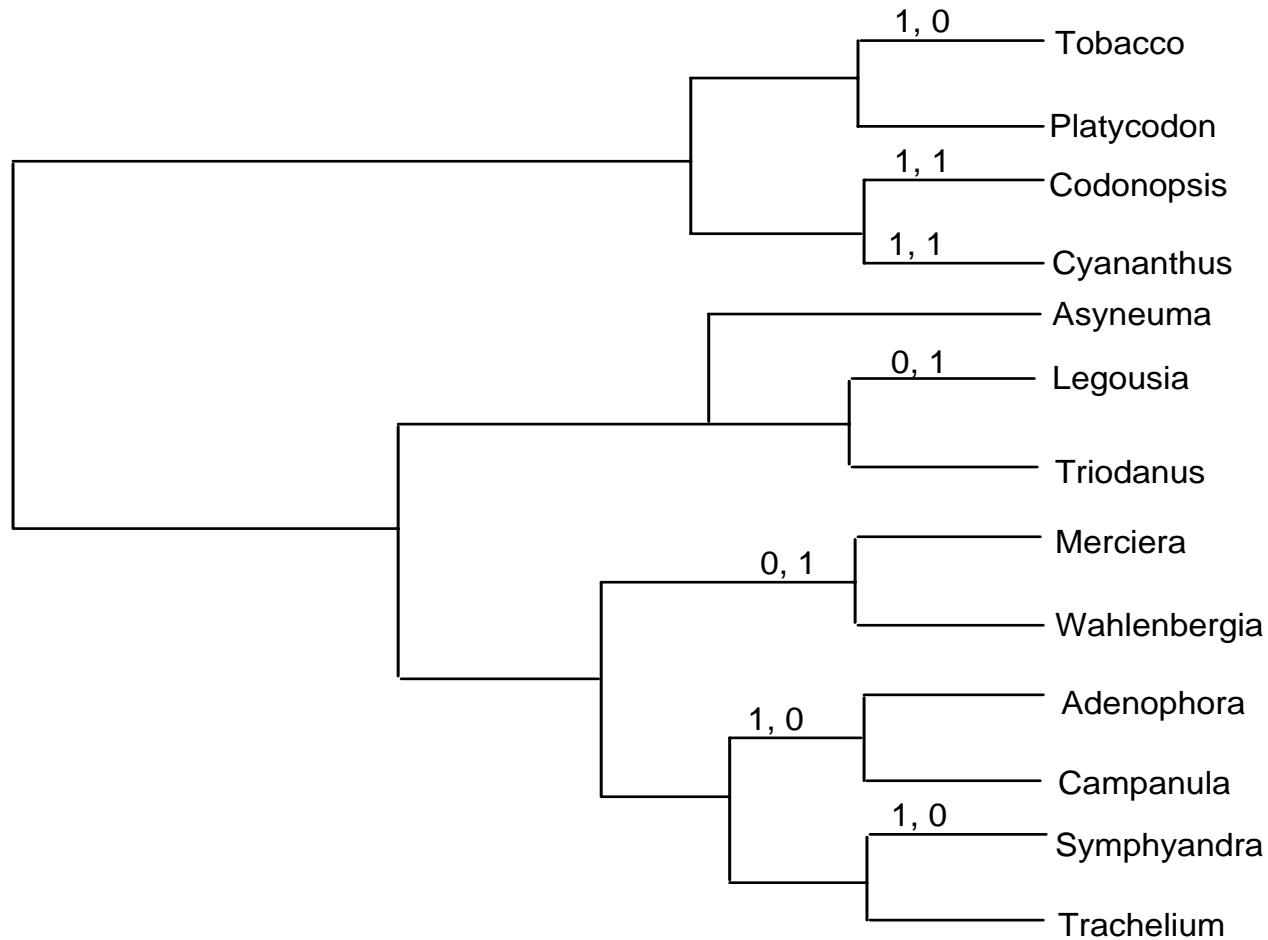
EMRAE's specificity remains very high.



Real data (I)

- *Campanulaceae* data set (Cosner, et al, 2000)
 - 13 genomes, 107 genes
 - MGR: 62 reversals, 1 transposition.
 - EMRAE: 5 reversals, 4 transpositions.
 - 3 reversals and 1 transposition shared
 - EMRAE has 2 distinct reversals and 3 transpositions
 - EMRAE's predictions are not only a subset of MGR

EMRAE's predictions

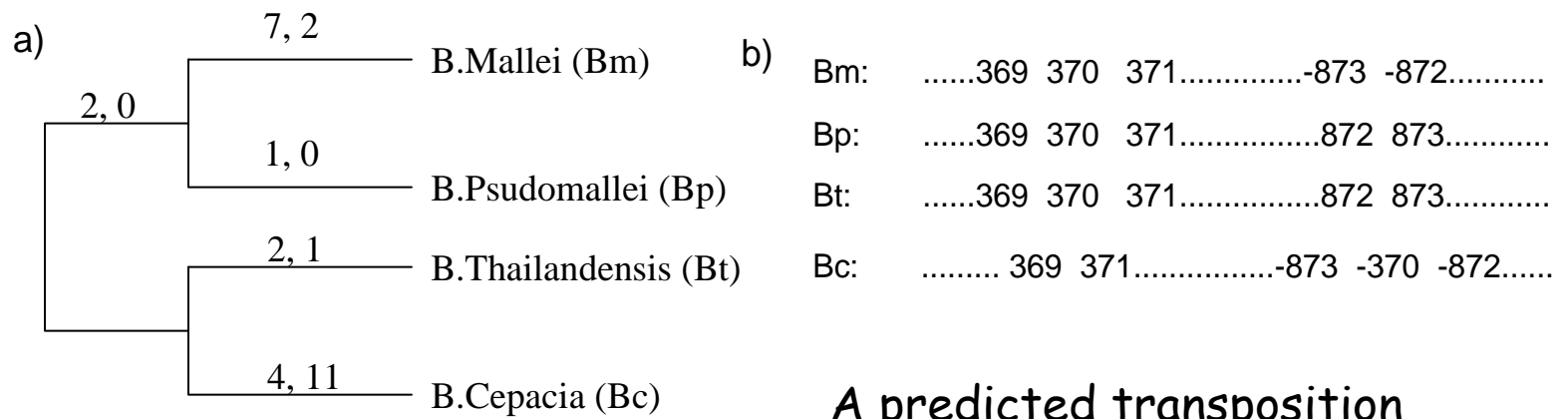


EMRAE inferred 9 events: 5 reversals and 4 transpositions.

The two nbs on an edge is the nb of reversals and transpositions predicted by EMRAE.

Data set (II)

- 4 bacterial genomes Bm, Bp, Bt, Bc (Lin et al. 2007)
 - each has 2 chroms, 2435 genes
 - MGR: 237 reversals, 1 transposition
 - EMRAE: 16 reversals, 18 transpositions
 - 14 reversals and 1 transposition shared





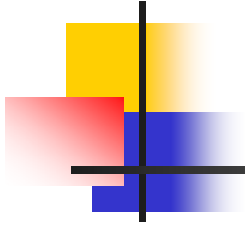
Conclusion

- An efficient method with comparable sensitivity and high specificity
- EMRAE models transpositions which appear to be common in two real data sets
- EMRAE predicted less events than MGR but
 - Based on the simulations we expect the events predicted for real data sets to be reliable
 - True events are not hidden in a complete but ambiguous scenario



Discussion

- Extension to predict other events
 - Translocations, fissions/fusions..
- Application to larger genomes
 - In-depth analysis of reliable events
 - Insights into features causing rearrangements



Thank you! 😊



Appendix

- The next pages are for the Q&A.
- Page 27: reversals mimic a transposition
- Page 28: relaxation the definition of conserved adjacencies
- Page 29: slideing adjacencies
- Page 30: how to measure MGR and GRAPPA



3 reversals mimic a transposition

- Transposition

1 2 3 4 5 6 \longrightarrow 1 4 5 2 3 6

- 3 reversals

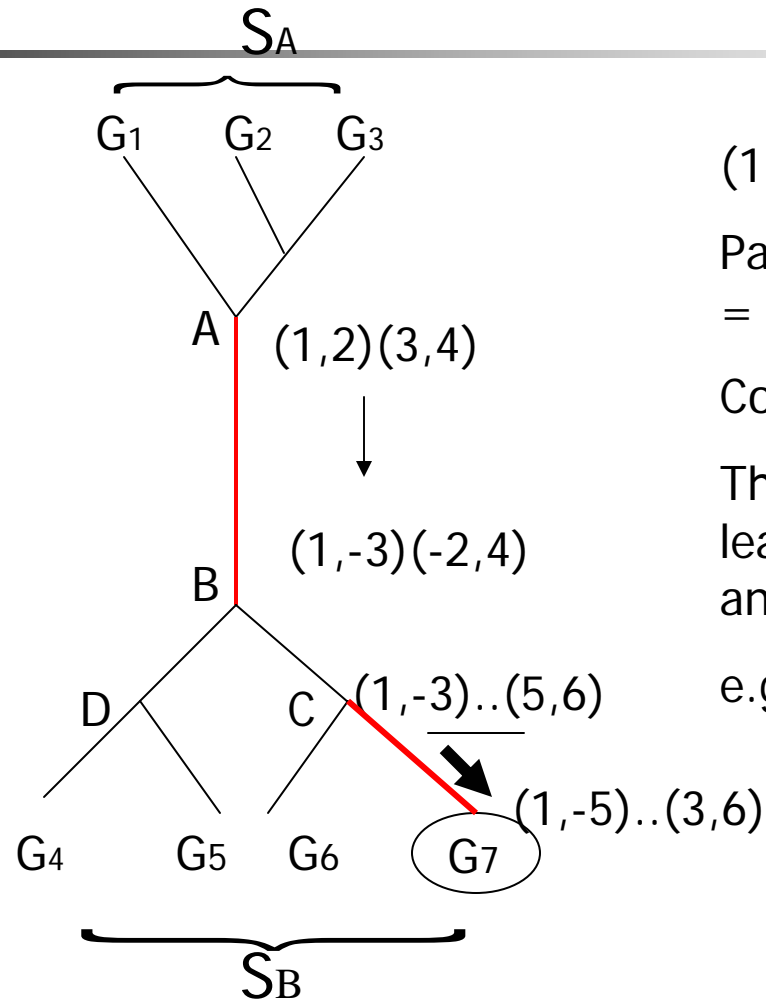
1. 1 2 3 4 5 6 \longrightarrow 1 -5 -4 -3 -2 6

2. 1 -5 -4 -3 -2 6 \longrightarrow 1 4 5 -3 -2 6

3. 1 4 5 -3 -2 6 \longrightarrow 1 4 5 2 3 6

- There are 6 ways to mimic a transposition with 3 reversals.

Quartet condition (relaxation)



(1,-3) not in G7

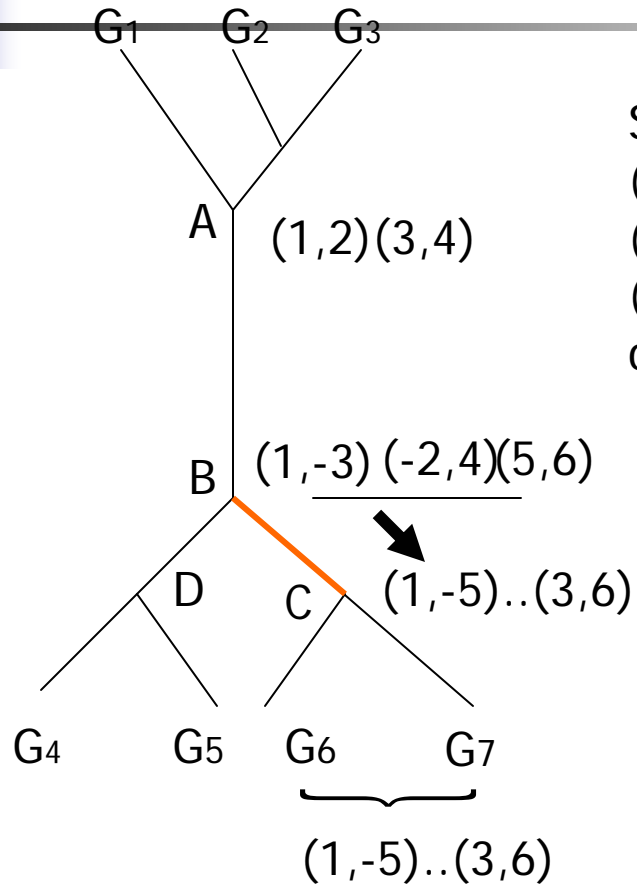
Partition SB into $S_{Bl} = \{G4, G5\}$, $S_{Br} = \{G6, G7\}$.

Conserved adjacency a:

There is at least G_i in S_{Bl} and at least G_j in S_{Br} , such that a in G_i and a in G_j .

e.g. (1, -3) in G4 and G6

Sliding adjacency



$S_{Bl} = \{G_4, G_5\}$, $S_{Br} = \{G_6, G_7\}$.

$(1,-3)$ disrupted on edge (B,C)

$(1,-3)$ not in G_6 and G_7

$(1,-3)$ slides to be a conserved adjacency of edge (B,D)



MGR and GRAPPA for transpositions

- MGR and GRAPPA aim to infer ancestral permutations
- Using GRIMM (Tesler 2003) infer events for each edge
- A transposition mimicked by 3 reversals