Finding Most Compatible Phylogenetic Trees over Multi-State Characters

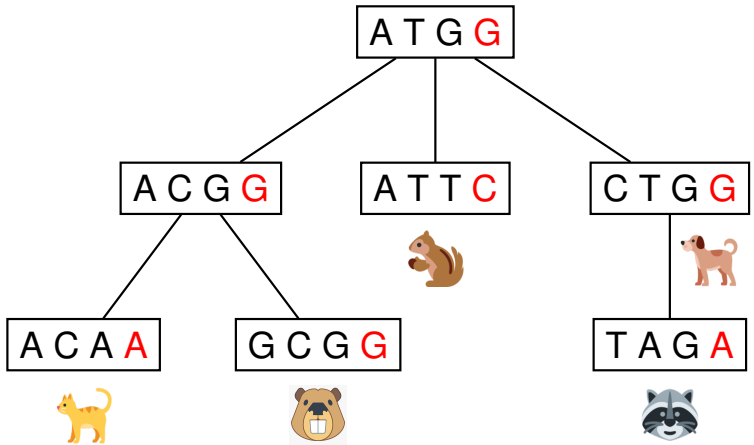
Tuukka Korhonen Matti Järvisalo

HIIT, Department of Computer Science, University of Helsinki, Finland

AAAI 2020
New York
February 10, 2020

Outline

- Problem:
  - Given data about taxa (species), construct an evolutionary tree
- Solution:
  - Graph-theoretical approach based on triangulations + MaxSAT
- Experimental evaluation:
  - Outperforms previous approaches

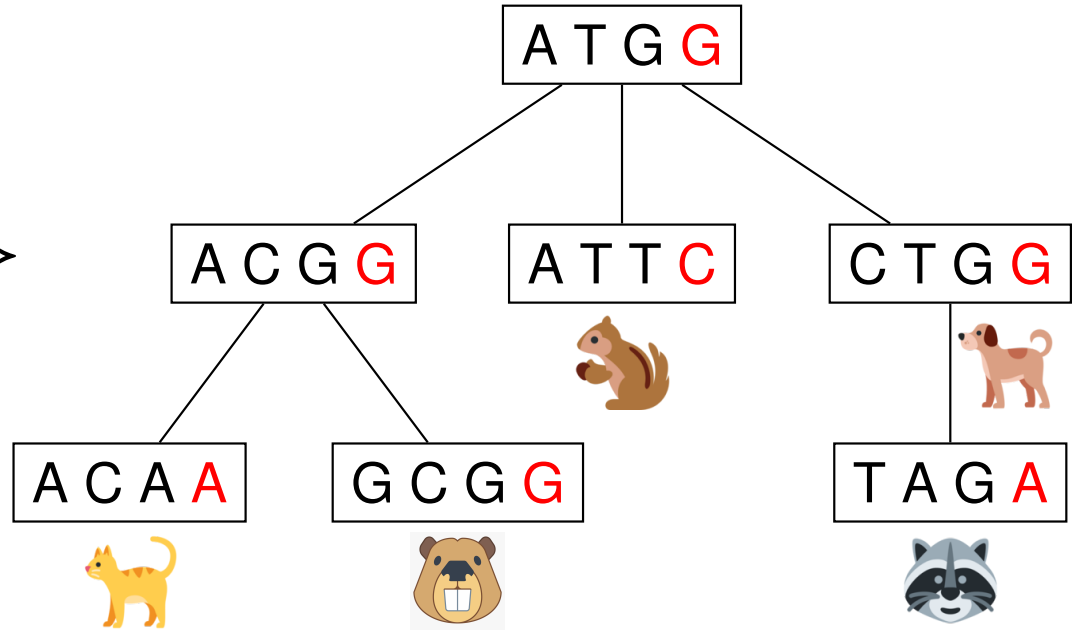


Maximum compatibility problem

- Input:
  - $n \times m$ matrix corresponding to n taxa with m characters
  - *Multi-state* $\Leftrightarrow$ characters take on $k > 2$ values
- Output:
  - An evolutionary tree that is *compatible* with as many characters as possible
- NP-complete to test if a tree compatible with *all* characters exists

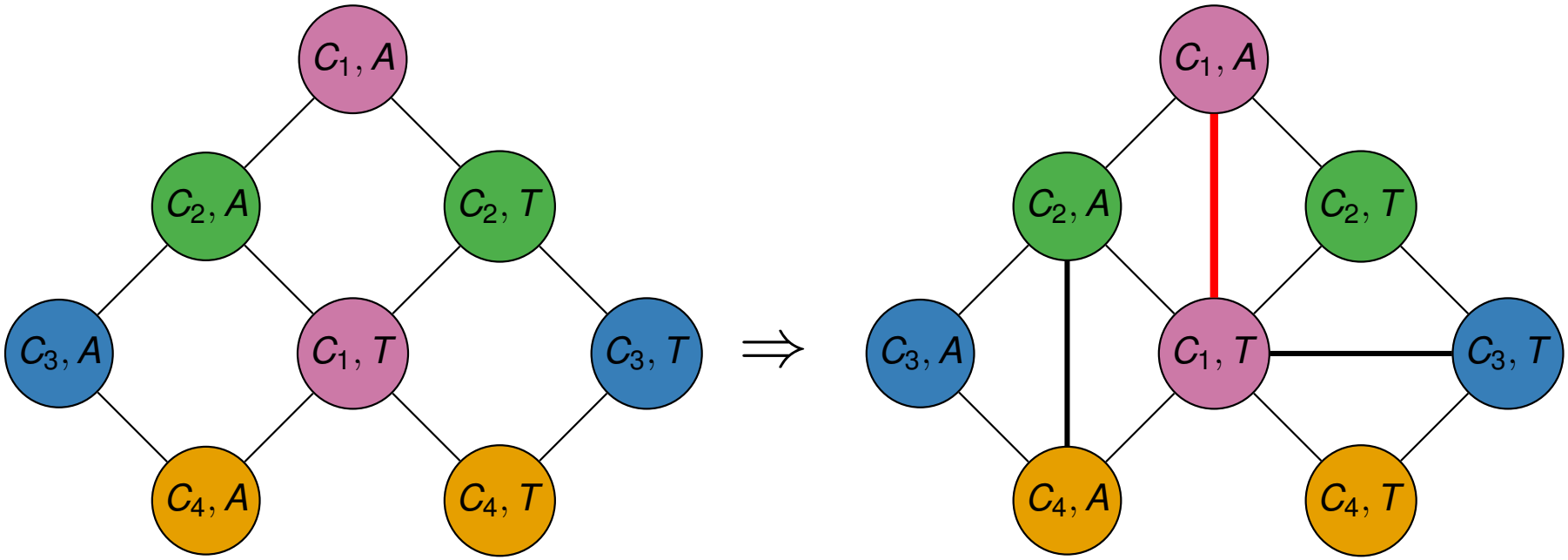
	C_1	C_2	C_3	C_4
Cat	A	C	A	A
Dog	C	T	G	G
Beaver	G	C	G	G
Raccoon	T	A	G	A
Squirrel	A	T	T	C

$\Rightarrow$



Equivalent Formulation

- Input:
  - Graph G with $m \cdot k$ vertices, colored with m colors.
- Output:
  - *Triangulation* of G that breaks the least amount of colors
  - Color is broken if edge is added between two vertices of the color



Our approach - Background

- Implementations of the Bouchitté–Todinca approach recently observed to be efficient in computing triangulations
  - Treewidth [Tamaki, 2019]
  - Total table size, notions of hypertreewidth [Korhonen et al., 2019]
  - Enumeration of minimal triangulations [Ravid et al., 2019]

- *Binary* maximum compatibility and compatibility of *all* characters formulated with BT [Gysel, 2014]

Minimal triangulations

- Minimal triangulations of graphs can be computed via potential maximal cliques [Bouchitté and Todinca, 2001]
- Potential maximal clique = vertex set that is maximal clique in some minimal triangulation
- Meta-algorithm for finding an optimal minimal triangulation:
  1. Enumerate Π, the potential maximal cliques
  2. Dynamic programming over Π in time $O(|\Pi| poly(n))$

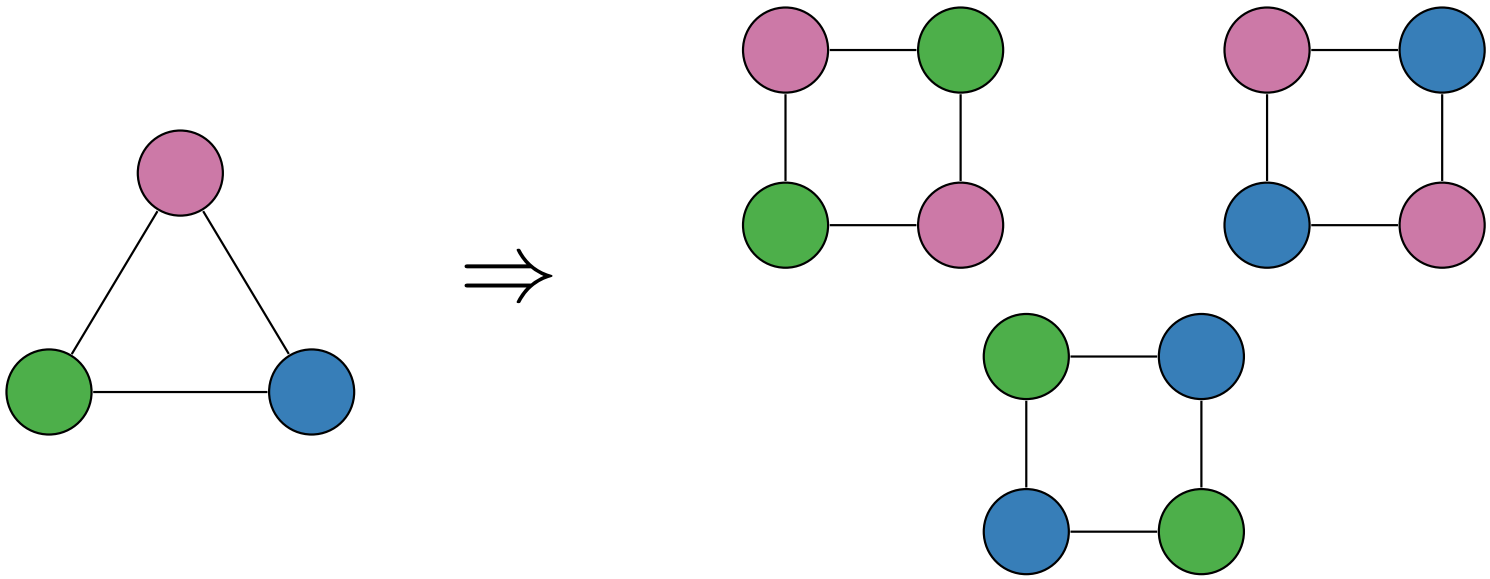
Contributions

Contributions

- We consider maximum compatibility of **multi-state** characters
  - Motivated e.g. by DNA ($k = 4$) and amino acid ($k = 20$)
- BT dynamic programming cannot be applied in multi-state case without superpolynomial overhead unless P=NP
- New hybrid BT + MaxSAT approach
  1. Enumerate potential maximal cliques
  2. Encode dynamic programming as MaxSAT with soft decision variables for broken colors
  3. Solve with MaxSAT solver

Why MaxSAT

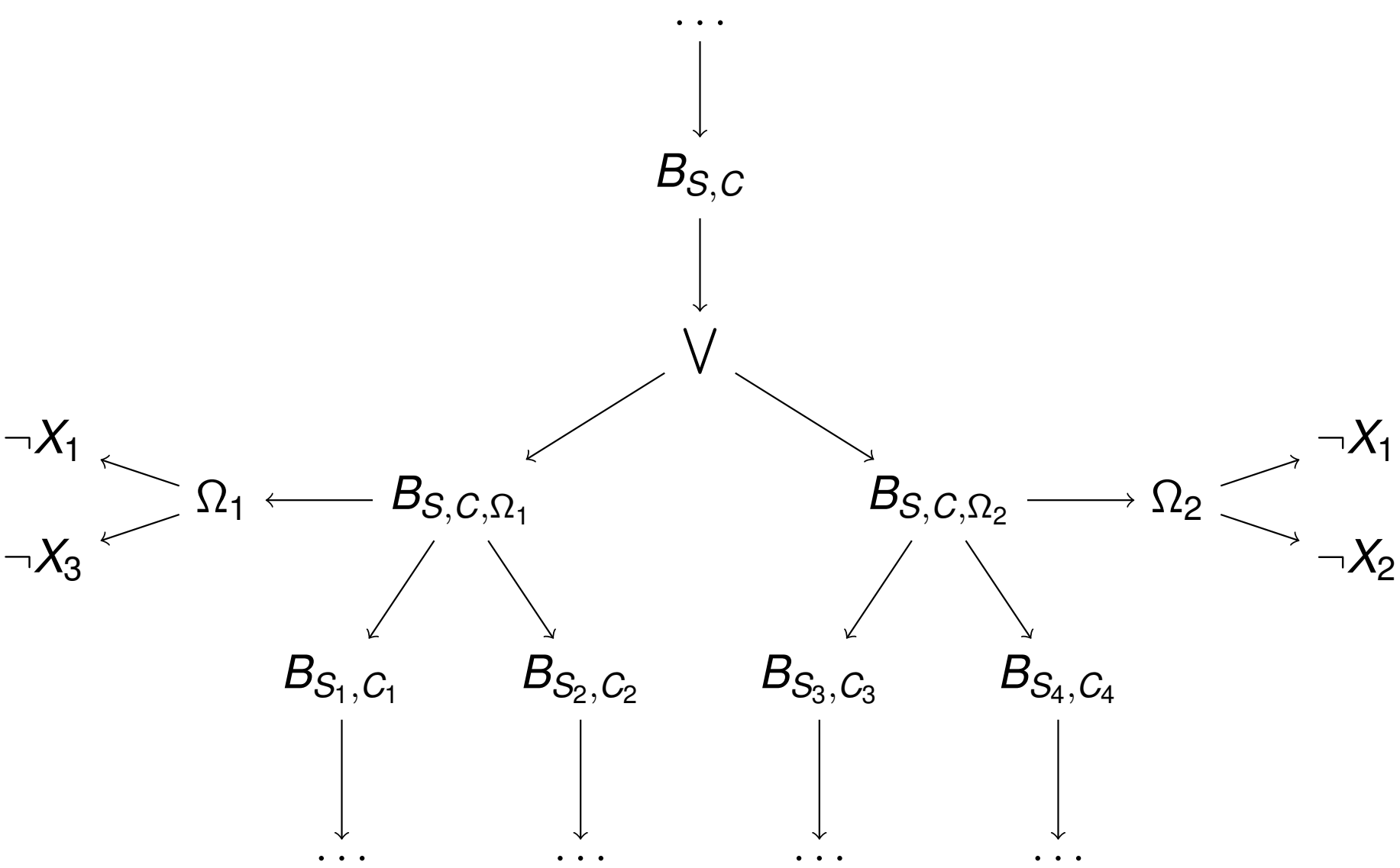
- Reduction from vertex cover to multi-state maximum compatibility with $|\Pi| = O(m)$



MaxSAT encoding

- Variables $X_1, \ldots, X_m$ represent colors
- Variables $\Omega_1, \ldots, \Omega_p$ represent PMCs
- Soft clauses $(X_1), \ldots, (X_m)$ to minimize broken colors
- Filling a PMC into a clique breaks colors inside it:
  - $\Omega_i \rightarrow \neg X_j$ if two vertices with color j inside Ω_i
- Hard clauses to force the PMCs to form a minimal triangulation

MaxSAT encoding



Encoding – The takeaways

- $O(|\Pi|mk)$ variables and clauses
- *Horn-MaxSAT*
  - SAT calls in hitting set algorithm will have linear time complexity

Experimental evaluation

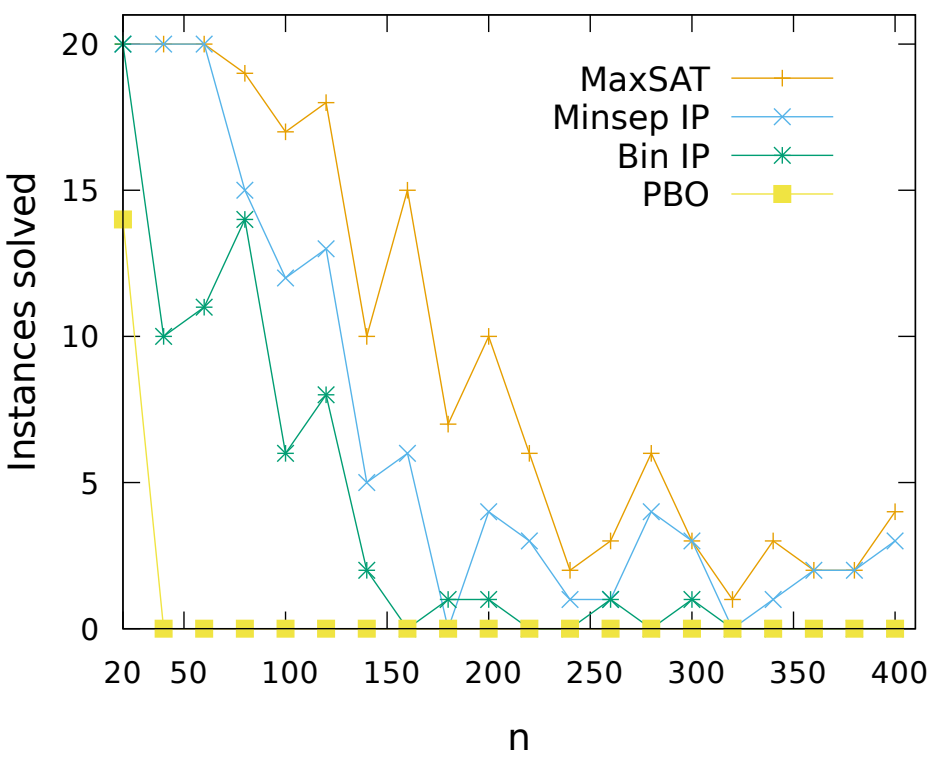
Experiments – Other approaches

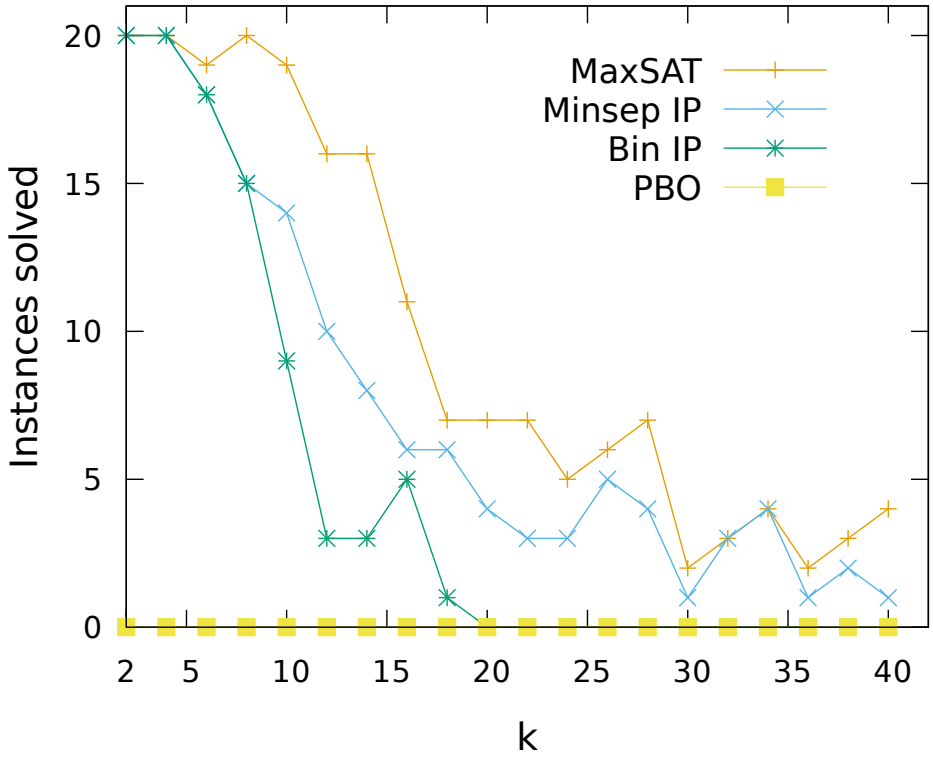
- ASP [Brooks et al., 2007]
  - Directly encodes the structure of the tree
- PBO [Miranda et al., 2014]
  - Improvement to the ASP encoding
- Bin IP [Stevens and Gusfield, 2010]
  - Reduces the multi-state case to binary case with additional constraints
  - Operates on the character matrix
- Minsep IP [Gysel and Gusfield, 2011]
  - Graph-theoretic approach with minimal separators instead of potential maximal cliques

Experiments – Setup

- 1840 generated instances
  - From well known model [Hudson, 2002]
- Typically the optimal solution consists of 60-90% of characters
- Same preprocessing techniques applied for all approaches
- MaxHS as MaxSAT solver, CPLEX as IP solver, MiniSat+ as PBO solver

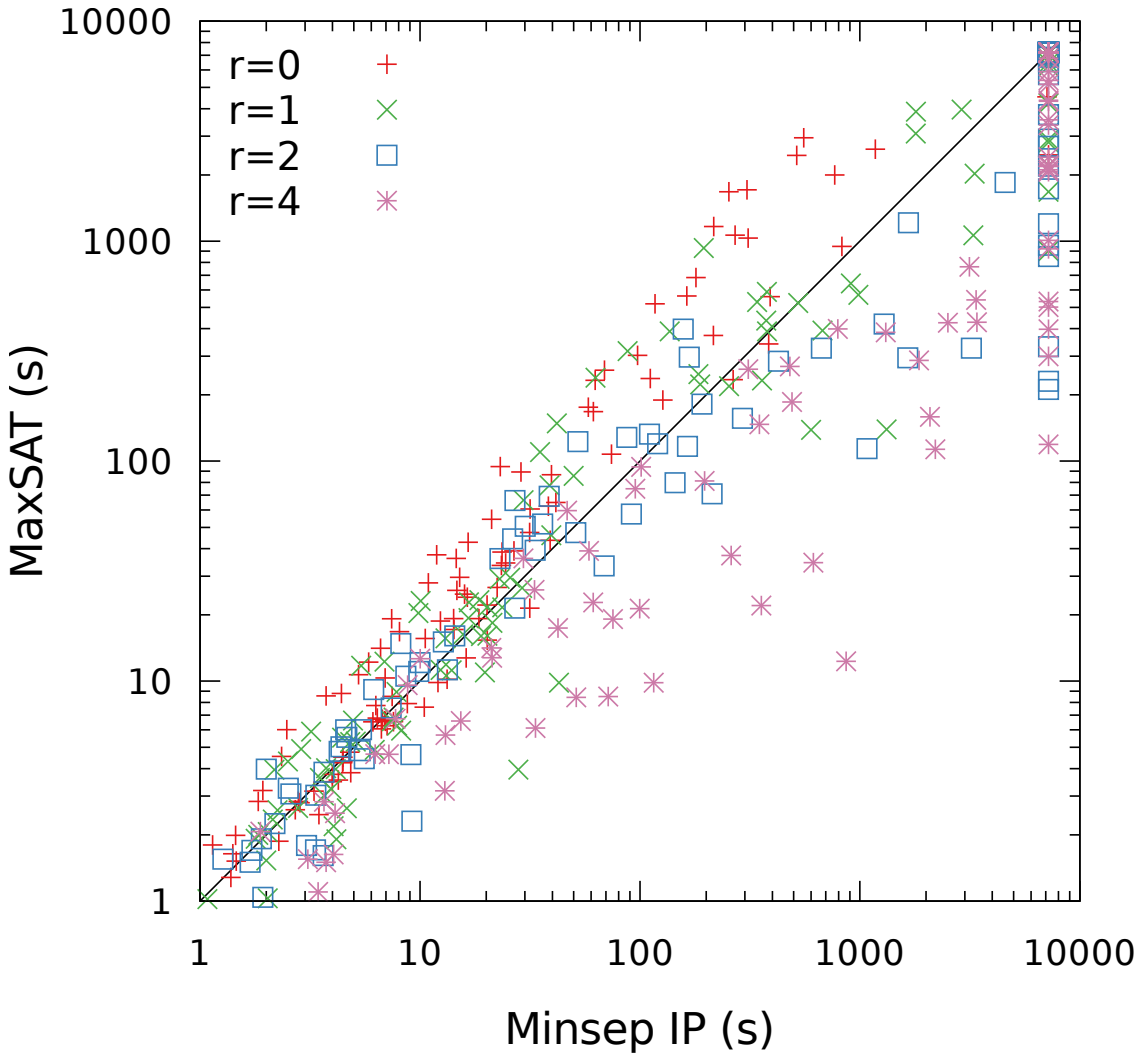
Experiments





- $n = m$: number of taxa and characters
- k: number of character states

Experiments



- r: how far the data is from perfect phylogeny

Summary

- Problem:
  - Finding phylogenetic trees with most compatible characters
- Solution:
  - Dynamic programming over potential maximal cliques encoded in MaxSAT
- Performance:
  - Outperforms earlier approaches that are based on declarative encodings
- Enumeration of PMCs is the runtime bottleneck
  - Timeouts during PMC enumeration: 828
  - Timeouts during MaxSAT solving: 5

The end

Thank you for your attention!

Bibliography

Vincent Bouchitté and Ioan Todinca. Treewidth and minimum fill-in: Grouping the minimal separators. *SIAM Journal on Computing*, 31(1):212–232, 2001.

Daniel R Brooks, Esra Erdem, Selim T Erdoğan, James W Minett, and Don Ringe. Inferring phylogenetic trees using answer set programming. *Journal of Automated Reasoning*, 39(4):471, 2007.

Rob Gysel. Minimal triangulation algorithms for perfect phylogeny problems. In *Proc. LATA*, volume 8370 of *LNCS*, pages 421–432. Springer, 2014.

Rob Gysel and Daniel Gusfield. Extensions and improvements to the chordal graph approach to the multistate perfect phylogeny problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):912–917, 2011.

Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

Tuukka Korhonen, Jeremias Berg, and Matti Järvisalo. Solving graph problems via potential maximal cliques: An experimental evaluation of the Bouchitté–Todinca algorithm. *Journal of Experimental Algorithmics*, 24(1):1–9, 2019.

Miguel Miranda, Inês Lynce, and Vasco Manquinho. Inferring phylogenetic trees using pseudo-boolean optimization. *AI Communications*, 27(3):229–243, 2014.

Noam Ravid, Dori Medini, and Benny Kimelfeld. Ranked enumeration of minimal triangulations. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 74–88, 2019. doi: 10.1145/3294052.3319678. URL `https://doi.org/10.1145/3294052.3319678`.

Kristian Stevens and Dan Gusfield. Reducing multi-state to binary perfect phylogeny with applications to missing, removable, inserted, and deleted data. In *Proc. WABI*, volume 6293 of *LNCS*, pages 274–287. Springer, 2010.

Hisao Tamaki. Positive-instance driven dynamic programming for treewidth. *Journal of Combinatorial Optimization*, 37(4): 1283–1311, 2019.