Finding Most Compatible Phylogenetic Trees over Multi-State Characters

Tuukka Korhonen and Matti Järvisalo

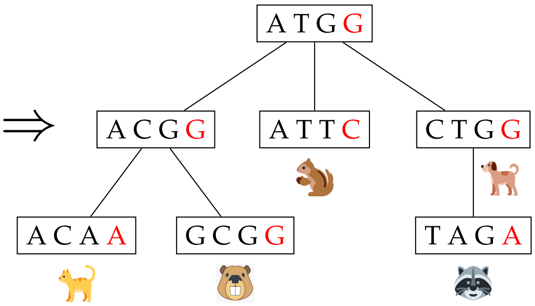
UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

Problem Definition

- Input: Data about species: $n \times m$ matrix corresponding to n taxa with m characters
- Output: Evolutionary tree that is compatible with as many characters as possible (**maximum compatibility problem**)
- Testing if a tree compatible with all characters (**perfect phylogeny**) exists is NP-complete

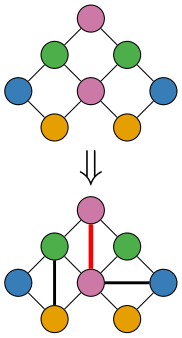
	C_1	C_2	C_3	C_4
Cat	A	C	A	A
Dog	C	T	G	G
Beaver	G	C	G	G
Raccoon	T	A	G	A
Squirrel	A	T	T	C

⇒



Graph-theoretic formulation

- Input matrix corresponds to a colored graph where colors correspond to the characters
- Find a **triangulation** of the graph that breaks the least number of colors
- A color is broken if an edge is added between two vertices of the color
- **Bouchitté-Todinca algorithm** characterizes minimal triangulations and enables finding optimal triangulations

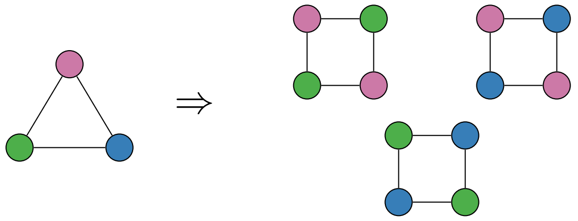


Contributions

- We show that Bouchitté-Todinca algorithm cannot be applied in multi-state maximum compatibility without superpolynomial overhead unless P = NP.
- We propose new hybrid approach, using potential maximal cliques of BT algorithm, but replacing dynamic programming with MaxSAT encoding.
- We experimentally compare to three prior approaches and outperform them.

BT algorithm

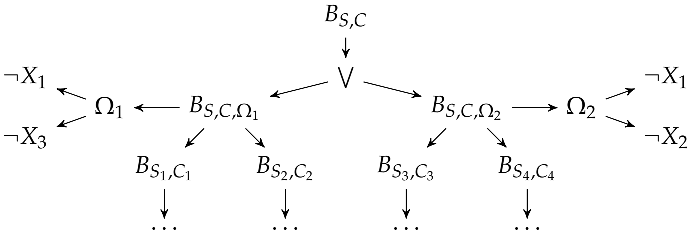
- Overview of BT algorithm:
  1. Enumerate *potential maximal cliques* $\Pi(G)$
  2. Find optimal triangulation by dynamic programming over *blocks* using PMCs in time $O(|\Pi(G)|poly(n))$
- Works for deciding if **all** characters are compatible and for maximum compatibility of **binary** characters
- Not directly applicable to maximum compatibility of **multi-state** characters. Reduction from vertex cover:



BT + MaxSAT Hybrid

- Encode phase 2 with decision variables $X_1, \ldots, X_m$ about which colors to break

  ⇒ *Horn-MaxSAT* encoding with size $O(|\Pi(G)|mk)$
- Full algorithm:
  1. Translate character-state matrix into a colored graph
  2. Enumerate $\Pi(G)$, the potential maximal cliques of the graph
  3. Encode BT dynamic programming via $\Pi(G)$
  4. Solve with MaxSAT solver to maximize $\sum X_i$



Results

Comparison to previous approaches:

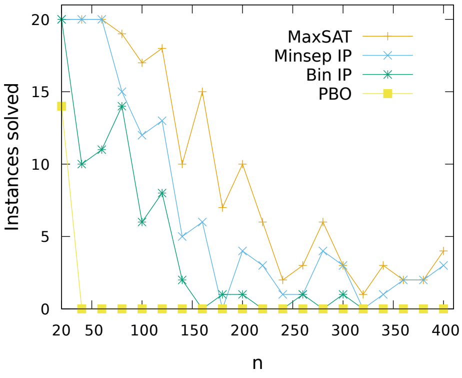
- PBO [Miranda et al., 2014]
- Bin IP [Stevens and Gusfield, 2010]
- Minsep IP [Gysel and Gusfield, 2011]

The bottleneck is PMC enumeration:
Average time: (solved instances)

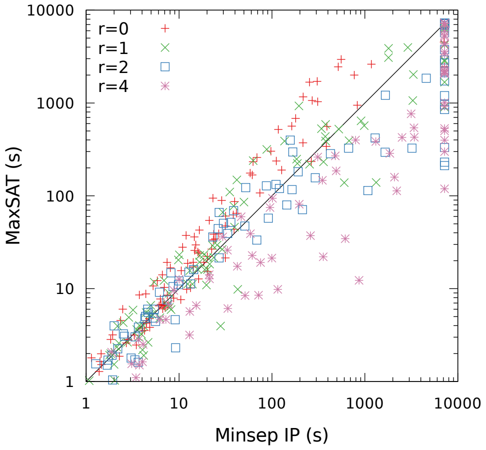
- PMC enumeration: 669s
- MaxSAT solving: 51s

Timeouts during:

- PMC enumeration: 828
- MaxSAT solving: 5



$n = m$: number of taxa and characters



r: how far the data is from perfect phylogeny