Phylogenetic simulations over constraint-based grammar formalisms

Andrew Lamont
Indiana University
alamont@indiana.edu

Jonathan North Washington
Indiana University
jonwashi@indiana.edu

Acknowledgements
Noor Abo Mokh, Daniel Dakota, Sandra Kübler, Lwin Moe, Larry Moss, Dragomir Radev
Audiences at Compmorphon 2015 and the CLingDing discussion group at IU
The three anonymous reviewers

Quantifying genetic distance

- Linguistic phylogeny attempts to estimate the evolutionary history of languages
- Traditional methods use cognate sets as comparanda
- Recent work on the efficacy of grammatical parameters
  - Dunn et al. (2005)
  - Longobardi & Guardiano (2009) successfully used syntactic parameters in a computational phylogeny of 23 Indo-European and 6 non-Indo-European languages
  - Eden (2013) used phonological stress parameters on the same languages and further demonstrated that weighted phonotactic constraints can be utilized as well

Optimality Theory

- In Optimality Theory (OT), input-output relations are determined by interaction of conflicting violable constraints.
- Example constraints:
  - NoCoda: codas are prohibited
  - Max: deletion is prohibited ("*maximise output*")
- Possible grammars using these constraints:
  - Grammar 1: NoCoda ≫ Max
    /pat/ ⟶ [pa] (*i.e., delete to avoid codas*)
  - Grammar 2: Max ≫ NoCoda
    /pat/ ⟶ [pat] (*i.e., retain codas, no deletion*)
- Constraint rankings are directed acyclic graphs (DAGs)
- OT diverges from Eden's (2013) constraint implementation:
  - Assumes a universal constraint set Con
  - Allows constraints to be unranked

Constraint pair pseudo-parameters

- For each pair of constraints, C_1, C_2, we define a dominance relation R following Antilla and Cho (1998):

$$R(C_1, C_2) = \begin{cases} 1 & \text{if } C_1 \gg C_2 \\ 0 & \text{otherwise} \end{cases}$$

- R(NoCoda, Max) conveys whether a language deletes codas ($R = 1$) or retains them ($R = 0$)
- Some R values are less meaningful
  - e.g., R(NoCoda, Max-Voice) cannot be set directly, as these two constraints do not directly conflict. (Max-Voice prohibits the deletion of voice features.)
- Pseudo-parameters have a numeric advantage over traditional parameters:
  - n parameters yields only n points of comparison
  - A set of n constraints yields $\sim n^2$ pseudo-parameters

Constraint ranking | Domination matrix | Pseudo-parameters

$C_1 \gg C_2 \gg C_3 \longrightarrow$

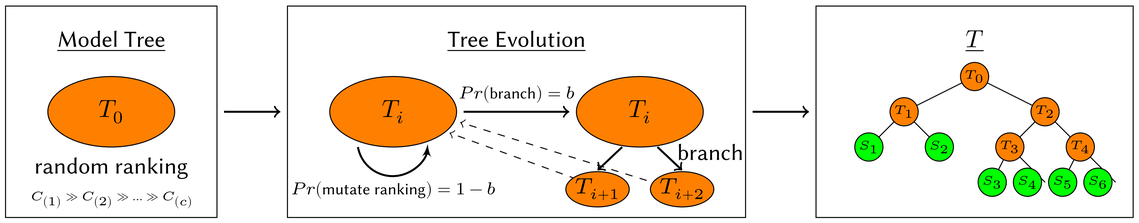
	C_1	C_2	C_3
C_1	0	1	1
C_2	0	0	1
C_3	0	0	0

$\longrightarrow$ $R(C_1,C_1)=0$, $R(C_1,C_2)=1$, $R(C_1,C_3)=1$, $R(C_2,C_1)=0$, $R(C_2,C_2)=0$, $R(C_2,C_3)=1$, $R(C_3,C_1)=0$, $R(C_3,C_2)=0$, $R(C_3,C_3)=0$

Methodology

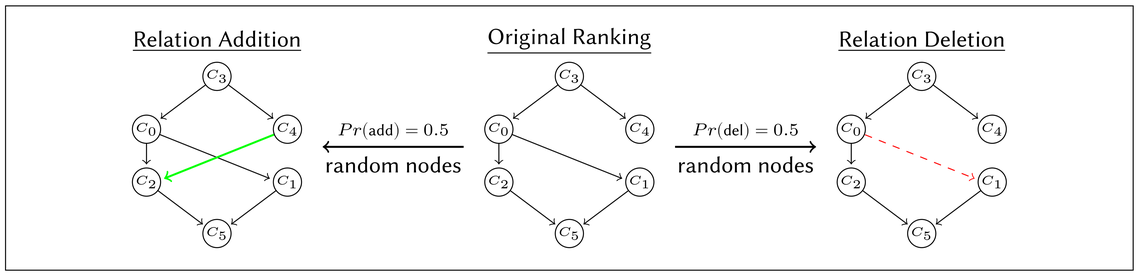
Gold standard tree

- Simulations follow the procedure described by Nichols and Warnow (2008)
- Parameters vary with respect to the number of constraints c, the size of the set of leaf nodes S, and the branching probability b
- T evolves until it reaches a minimum number of leaf nodes



Ranking mutation

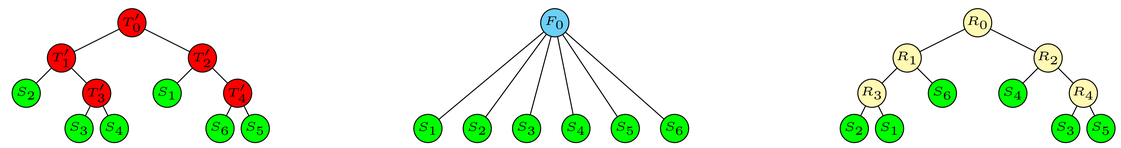
- Constraint rankings are represented with directed acyclic graphs (DAGs)
- Mutation either adds an edge in the DAG or removes an edge
- The mutated source and target nodes are chosen randomly



Test trees

- Constraint rankings of languages in the set of leaves S are decomposed into pseudo-parameter vectors
- The test tree T' is produced by hierarchically clustering S, using Euclidean distance over the pseudo-parameter vectors
- T' is compared against two baseline trees

Test tree from clustering (T') | Flat tree baseline (BF) | Random tree baseline (BR)

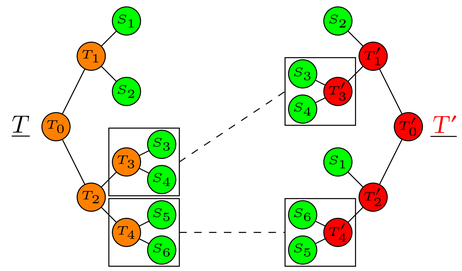


Evaluation

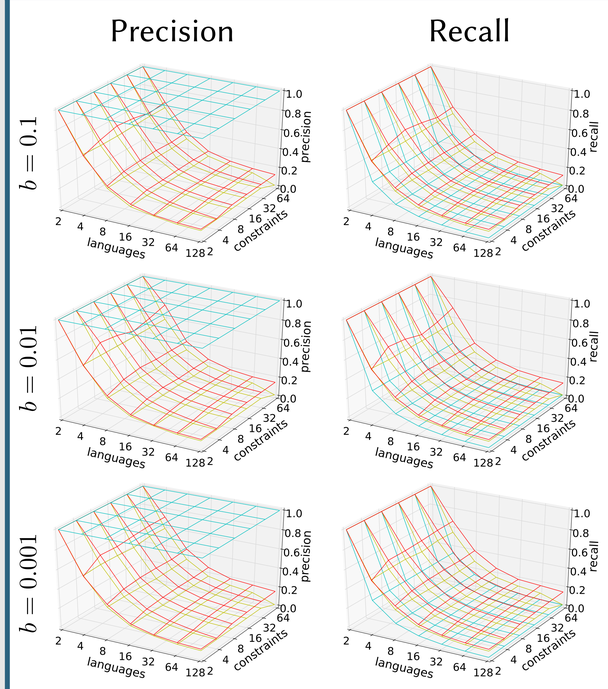
- Unlabeled precision and recall was calculated between T and the three test trees T', BF, BR
- Two internal nodes are considered matching if they dominate the same set of leaves
- We consider an experiment successful when T' is more accurate than the baseline trees

$$\text{Precision} = \frac{\text{\# matching nodes}}{\text{\# nodes in test tree}} = \frac{|\{T'_3, T'_4\}|}{|\{T'_0, T'_1, T'_2, T'_3, T'_4\}|} = \frac{2}{5}$$

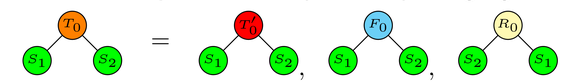
$$\text{Recall} = \frac{\text{\# matching nodes}}{\text{\# nodes in gold tree}} = \frac{|\{T'_3, T'_4\}|}{|\{T_0, T_1, T_2, T_3, T_4\}|} = \frac{2}{5}$$



Results



- Our method (T') outperforms flat (BF) and random (BR) baseline trees in most conditions
- All 3 trees have perfect accuracy with only 2 languages:



- BF always has perfect precision because it only contains the root node
- The accuracy of T' increases as the number of constraints increases, because the number of comparisons increases
- T' is more accurate with lower branching probabilities, as changes accumulate and propagate through subtrees
- The accuracies of all three trees decrease as the number of languages increase. This follows from the number of available hypotheses: the number of internal nodes is on the order of the number of leaves

Future Directions

- Extend this approach to real language data, e.g., language families with extensive phonological work
- Stress data is particularly propitious
  - Large attested typology (e.g., the StressTyp2 database)
  - There are about 14 core constraints (Kager, 1999)
  - Compare constraint-based phylogeny directly with Eden's (2013) parameter approach
- Certain linguistic changes are more likely to occur than others (e.g. from phonetic pressures). Enriching our system with these biases should improve its performance over real language data