Phylogenetic simulations over constraint-based grammar formalisms

Andrew Lamont and **Jonathan North Washington**
Indiana University
{alamont,jonwashi}@indiana.edu

Abstract

Computational phylogenetics has been shown to be effective over grammatical characteristics. Recent work suggests that constraint-based formalisms are compatible with such an approach (Eden, 2013). In this paper, we report on simulations to determine how useful constraint-based formalisms are in phylogenetic research and under what conditions.

1 Introduction

Popular computational methods for phylogenetic research (estimating the evolutionary histories of languages) primarily involve comparisons over cognate sets (Nichols and Warnow, 2008). Recent works (Dunn et al., 2005; Longobardi and Guardiano, 2009) indicate that comparing sets of grammatical parameters can be effective as well. However, generating a large number of meaningful parameters remains a formal obstacle. In this paper we argue that constraint-based grammar formalisms may be exploited for parameter generation, and explore to what extent such research is feasible.

Because the use of constraint-based grammars in phylogenetics is relatively novel, we do not know *a posteriori* how many constraints and how many languages must be considered for a computational approach to be successful. If a minimum threshold is established that is methodologically prohibitive (e.g. if such systems were only accurate given a set of 1,000 languages), we can abandon this approach as infeasible. By initially experimenting with simulated data, we establish a footing for future empirical studies.

In this paper, we report on simulations which consistently outperform two baseline models. Significantly, these results obtained with a modest number of constraints $c \geq 4$ and languages $l \geq 4$.

1.1 Grammatical parameters in phylogenetics

Longobardi and Guardiano (2009) argue that grammatical features, such as whether a language expresses a pre- or postpositional genitive, if chosen carefully, present certain advantages over lexically-based comparisons in phylogenetic work. Grammatical parameters comprise a universal set of discrete options applicable to any set of languages, especially within frameworks such as Principle and Parameters (Chomsky, 1981). Using grammatical features for phylogenetic work can be a way to avoid any difficulties associated with the collection and identification of cognate sets.

However, unlike cognate sets, there is no *a priori* assumption that correspondences between parameter settings are meaningful genetically. Instead, meaningful correspondence derives from the low probability that two languages match in a number of parameter settings by chance. Successful work therefore depends on the construction of a large set of grammatical parameters; larger sets are predicted to produce more accurate results.

1.2 Constraint-based grammar formalisms

In constraint-based theories of grammar like Optimality Theory (OT) (Prince and Smolensky, 2004), input-output relations are determined by the interaction of conflicting violable constraints.

To take a common example from phonology, a

Proceedings of NAACL-HLT 2016, pages 102–108,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

language may require that all syllables are open, deleting consonants that would otherwise surface in coda position. In an OT analysis of such a language, the constraint NoCoda, which prohibits codas, dominates the constraint Max, which prohibits deletion (written NoCoda $\gg$ Max). This encodes that satisfying NoCoda is more important than satisfying Max, though there may be additional interacting constraints complicating the analysis.

In an OT framework, the set of constraints, Con, is assumed to be universal—its members typically being grounded in typological and psycholinguistic data. Differences between grammars are encoded as different language-specific rankings of the constraint set.

OT is most often used in phonology, but has been applied widely in various linguistic sub-disciplines, including syntax, sociolinguistics, and semantics (McCarthy, 2008). Constraint-based frameworks can therefore encode diverse grammatical phenomena with minimal representation, a constraint ranking being simply a directed acyclic graph over Con.

With the exception of Eden (2013), constraint-based phylogenetic research has not yet, to our knowledge, been attempted. It remains an open question whether such a representation is useful in phylogenetics and if so, under what conditions.

1.3 Parameterizing Con

Following Longobardi and Guardiano's (2009) parametric approach, we adapt constraint rankings into binary pseudo-parameters, decomposing language-specific rankings into vectors of pairwise dominance relations (Antilla and Cho, 1998):

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

That is, for every pair of constraints C_1, C_2, R returns a binary value corresponding to whether C_1 dominates C_2 directly or transitively.[1] An example of a constraint ranking and its corresponding R values is shown is shown in Figure 1.

$$C_1 \gg C_2 \gg C_3$$

$$\begin{array}{cccc} & C_1 & C_2 & C_3 \\ C_1 & 0 & 1 & 1 \\ C_2 & 0 & 0 & 1 \\ C_3 & 0 & 0 & 0 \end{array}$$

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

Figure 1: A constraint ranking, its representation as a matrix, and as a set of binary pseudo-parameters.

We consider these to be *pseudo*-parameters because certain constraint pairs may only interact under very specific circumstances or not at all. The ranking of NoCoda and Max, for example, is meaningful under a large number of circumstances: $R(\text{NoCoda}, \text{Max})$ corresponds to whether a grammar deletes consonants that would otherwise surface in coda position. The ranking of NoCoda and Max-Voice (which prohibits deleting a voice feature), on the other hand, is less meaningful because these constraints are not expected to conflict directly (deleting a voice feature does not create an open syllable, and therefore cannot avoid a violation of NoCoda). Nevertheless, $R(\text{NoCoda}, \text{Max-Voice})$ may be determined via transitivity. R values therefore range from representations of a language's grammatical characteristics to higher-level artifacts of the theory as applied to its grammar. Weighting R values accordingly may be a fruitful topic for future research.

Pseudo-parameters pose certain advantages. For a set of n constraints, the size of the corresponding set of pseudo-parameters is on the order of n^2. This dramatically increases the number of comparisons one is able to make between languages with a modest number of empirically motivated constraints, as compared to a parameter set *tout court*. Because constructing a set of constraints or parameters is taxing, an approach that maximizes the impact of each additional constraint is advantageous. With pseudo-parameters, constraint $n + 1$ contributes n points of comparison, whereas parameter $n + 1$ contributes only 1 point of comparison.

A theory-internal advantage of this approach is that it faithfully represents even complex constraint rankings. Some models of OT allow for constraints to be unranked. The pseudo-parameter representation handles unranked constraints without issue, thus allowing wide theoretical coverage.

[1] When $R(C_1, C_2)$ returns 0, it ambiguously encodes either a non-relation between C_1 and C_2 or the dominance relation $C_2 \gg C1$. However, because R is not a symmetric relation, this ambiguity is resolved when one considers $R(C_2, C_1)$. Additionally, two constraints C_1, C_2 are unranked relative to one another if $R(C_1, C_2) = R(C_2, C_1) = 0$—i.e., there is no dominance relation, direct or transitive, between C_1 and C_2.

2 Related Work

Computational phylogenetic systems have taken a diverse set of inputs. Subgroup classification using cognate comparisons has been used by Ringe et al. (2002) for Indo-European (IE) and Bowern and Atkinson (2012) for Pama-Nyungan, among others. Both syntactic and phonological grammatical-level information have also been used effectively for computational phylogenetics.

Longobardi and Guardiano (2009) used sixty-three binary syntactic parameters for a phylogeny of twenty-two IE languages and six non-IE languages. Their generated trees largely agreed with the historical relations determined by traditional comparative linguistic methods. In a second experiment using fifteen languages for which lexical data were available, they found large overlap between trees generated using syntactic parameters and lexical data.

Eden (2013) replicated this study using thirteen typologically grounded parameters related to phonological stress over nineteen of the languages used by Longobardi and Guardiano (2009) as well as an additional five, demonstrating that the grammatical parameters need not be limited to the domain of syntax. A second experiment using phonotactic constraints over six languages yielded more variable results than the first experiment. The constraints used were generated by a phonotactic constraint learner (Hayes and Wilson, 2008), which differs from classic OT in several key regards: in this model, constraints are language-specific; constraints are weighted probabilistically, not ranked; and constraints only reference surface forms, not input-output relations. To utilize a single constraint set, the one hundred thirteen highest-weighted constraints that were persistently generated by the phonotactic learner across the six languages were chosen and reweighted in each language. Each language therefore had a grammar consisting of the same set of constraints. The rankings of these constraints were compared using Spearman's correlation coefficient.

Eden's (2013) study broke ground in using constraint-based grammars; however, there were certain limitations. The phonotactic learner requires a representative input corpus of at least 3,000-6,000 words, impeding the incorporation of under-resourced languages. Further, the generated constraint set is problematically language-specific. Only one constraint generated for English, for example, was active in the other five languages.

Our approach diverges from Eden's (2013) theory-internally and in scope. We assume an *a priori* universal constraint set, and our pseudo-parameter approach allows for constraints to be unranked relative to one another. We could in principle measure inter-language distance with rank correlations over topologically sorted constraint rankings, but unranked constraints are predicted to lead to highly variable results. Because our experiments in this study are over simulated languages, we are not limited by available linguistic descriptions.

3 Method

To investigate whether constraint-based formalisms are useful in phylogenetic work and under what assumptions, we conducted a large number of simulations following the procedure described by Nichols and Warnow (2008):

1. Produce a model tree T;
2. T evolves from the root, producing a set of leaves S;
3. S is used as input to the phylogeny estimation procedure, producing T';
4. T' is compared to T to calculate error.

Simulations varied with respect to the number of constraints, the size of S, and the rate of evolution.[2]

3.1 Model Tree

In these simulations, $\textsc{Con}$ is defined as a set of c constraints $C_1, C_2, \ldots, C_c$. The model tree T (gold standard) is initialized with a root-node language consisting of a randomly generated full ranking of $\textsc{Con}$ such that every constraint is ranked relative to every other constraint: $C_{(1)} \gg C_{(2)} \gg \ldots \gg C_{(c)}$. For c constraints, there are $c!$ possible full rankings. From this root-node, T evolves into a larger tree.

3.2 Tree Evolution

In our simulations, language change is modeled by constraint reranking (Cho, 1998), although this oversimplifies the complex processes observed in actual

[2] Our code and full numerical results are available at `https://github.com/lmaoaml/recon`.

data.[3] T evolves accordingly. At each evolutionary step, a leaf language either randomly changes or splits into two daughter languages inheriting the same constraint ranking according to the branching probability b.[4] The lower b is, the more changes on average a language will undergo before branching. A change entails either the addition or removal of a domination relation between two random constraints. Evolution continues until T contains a predetermined minimum number of leaves.

3.3 Phylogeny Estimation

The constraint rankings of the languages in the set of leaves S are decomposed into pseudo-parameter vectors. Inter-language distance is calculated by taking the Euclidean distance between vectors. We use Euclidean distance because it has been reported to perform well among fifteen vector similarity measures of typological distance (Rama and Prasanth, 2012), and our initial experiments found no major differences between measures. The inter-language distances serve as input to the phylogeny estimation procedure.

Because tree evolution in our model proceeds according to a lexical clock (i.e., changes accumulate over time)—or more precisely a grammatical clock—we use the Unweighted Pair Group Method with Arithmetic mean (UPGMA), a hierarchical clustering method that utilizes the average distance between clusters, as a phylogeny estimation procedure (Nichols and Warnow, 2008). For speed, we use the implementation in fastcluster (Müllner, 2013) with average linkage. The result of phylogeny estimation is a binary tree T', which is compared to T to measure accuracy.

3.4 Evaluation

Because we have access to T, the gold standard tree, we diverge from the partially qualitative evaluations of Longobardi and Guardiano (2009) and Eden (2013) and adopt a purely quantative evaluation metric based on precision and recall (van Rijsbergen, 1979). As in standard precision and recall, we measure the proportion of correct items relative to T' and T respectively. We define correct items to be matching subtrees rooted by internal nodes as shown in Figure 2. Two subtrees are counted as matching if they dominate the same set of leaves.

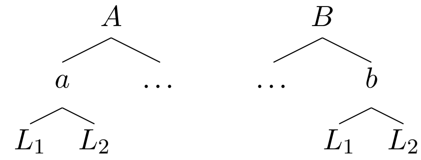


Figure 2: Subtree a in A matches subtree b in B.

T' is then compared against two null hypothesis baseline trees, BF and BR.

BF is a flat tree composed of a single internal node dominating the entire set of languages S as in Figure 3. BF encodes the empirical null hypothesis that S contains no subgroups.

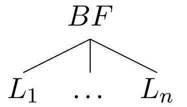


Figure 3: A random baseline tree BF with n leaves

BR is a randomly constructed binary tree encoding the null hypothesis that the phylogeny estimation procedure does not outperform chance groupings.

Precision and recall are calculated between T and the three test trees. We consider an experiment successful when T' is more accurate than BF and BR.

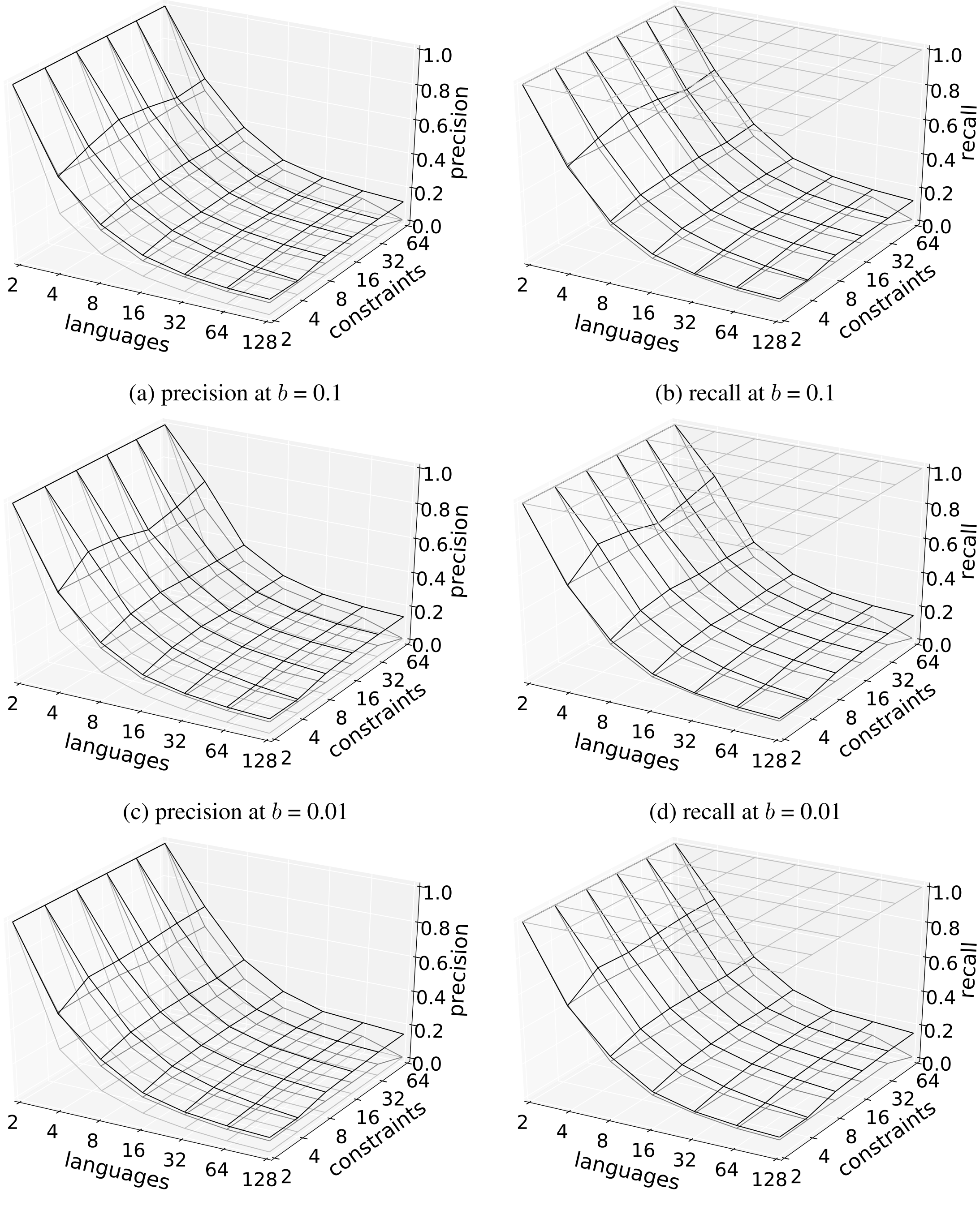
4 Results

Simulations were run across a wide range of settings. The number of constraints ranged exponentially from 2 to 64. The number of languages likewise ranged exponentially from 2 to 128. For each setting pair, we report precision and recall for BF, BR, and T' averaged over 1,000 independent iterations. Simulations were run with branching probability b set to 0.1, 0.01, and 0.001, as shown in Figure 4. Low branching probabilities yield more differences even between closely related languages (in the authors' opinion, this more accurately reflects actual language data).

Overall, the simulations were successful, albeit modestly. T' had a higher recall than BF and a higher precision and recall than BR in all cases except simulations with 2 constraints and 4 languages. The margin between T' and BR is promising - it indicates that this method can yield positive results.

[3] See Holt (2015) for an overview of approaches to language change in OT.

[4] T is limited to binary branching for simplicity, but this is not a necessary assumption for the methodology.



(a) precision at $b = 0.1$

(b) recall at $b = 0.1$

(c) precision at $b = 0.01$

(d) recall at $b = 0.01$

(e) precision at $b = 0.001$

(f) recall at $b = 0.001$

Figure 4: Precision and recall of BF (—), BR (—), and T' (—) with $b = 0.1, b = 0.01$, and $b = 0.001$.

5 Discussion

With 2 languages, as modeled in Figure 5, all hypotheses have perfect precision and recall. For T' and BR, because the order of the leaves does not matter, there is only one way to group 2 languages. Similarly, because there is not additional internal structure, BF has perfect recall. Trivially, BF always has perfect precision because the entire tree is the only subtree it identifies.

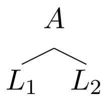


Figure 5: A tree A with two leaf nodes

As is expected, neither BF nor BR are affected by the number of constraints or b. However, as the number of languages increases, the probability that BR correctly identifies substructure decreases.

The accuracy of T' does interact with the numbers of constraints and languages as well as with b.

There is an overall trend that T' is more accurate with larger numbers of constraints, in accordance with the trend that phylogenetic algorithms' performances correlate with the amount of available data. This is especially clear when $b = 0.001$. Extending this method to real language data is expected to produce more accurate results with a larger number of constraints; however, this effect plateaus. Even with as few as eight constraints, our method scores around .100 higher precision and recall than BR. Ranking a set of eight constraints is within the scope of typical OT analyses.

The accuracy of T' negatively correlates with b, indicating that more grammatical distance is useful. This makes sense, as innovative traits passed down through subtrees aid in grouping.

Both precision and recall of T' decrease as the number of languages increases. We expect recall to decrease as the number of subtrees in T increases, which is the case with BF. Likewise, with more possible subtrees, the clustering algorithm makes more mistakes, leading to lower precision. These mistakes may additionally follow from the clustering method. With a large number of languages, the diversity within clusters may be especially large, leading to similar average distances between clusters, which can result in unpredictable performance of the linkage function. However, the effect of number of languages is not more pronounced with smaller values of b. With smaller b, there are more changes to the languages and we might expect more diversity. If this is an effect of the algorithm, we expect more error high in the tree than at the leaf level. It would be worthwhile to experiment with different linkage functions at different levels in the tree.

Our method assumes that all constraint rerankings are equally likely, which is not the case in real languages; e.g., phonological evolution is frequently shaped by phonetic biases. Given that our method was successful, we anticipate that incorporating known diachronic biases will radically improve performance on natural language data.

6 Conclusion and Future Work

Our method yielded positive results for the simulations reported on in this paper. This suggests that constraint-based formalisms may be used successfully in computational phylogenetics, though this remains to be verified with natural language data. These experiments serve to establish a baseline for the use of constraint-based grammars in phylogenetic research. We believe that the results show promise for the addition of constraint-based research to the phylogenetic toolkit, though additional work is required to fully understand its usefulness.

In the future, we plan to examine the effect of different clustering algorithms, and extend this approach to actual language data. One propitious domain is the phonology of stress, because a large number of languages have already been analysed using a set of 14 core constraints (Kager, 1999). Furthermore, it presents an opportunity to compare directly a constraint-based approach with a parametric approach, such as Eden's (2013) phylogenetic results based on stress parameters.

Acknowledgements

For their helpful discussion and feedback, we thank Noor Abo Mokh, Daniel Dakota, Sandra Kübler, Lwin Moe, Larry Moss, Dragomir Radev, and audiences at the Workshop on Computational Phonology and Morphology 2015 and the CLingDing discussion group at Indiana University, as well as three anonymous reviewers.

References

Arto Antilla and Young-mee Yu Cho. 1998. Variation and change in optimality theory. *Lingua*, 104(1-2):31–56.

Claire Bowern and Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of pama-nyungan. *Language*, 88(4):817–845.

Young-mee Yu Cho. 1998. Language change as reranking of constraints. In Richard M. Hogg and Linda van Bergen, editors, *Historical Linguistics 1995: Volume 2: Germanic linguistics*, pages 45–62. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris Publications, Holland.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072.

Elizabeth Eden. 2013. Measuring language distance through phonology: parameters or constraints? *UCL Working Papers in Linguistics*, 25:222–250.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

D. Eric Holt. 2015. Historical sound change in optimality theory: Achievements and challenges. In Patrick Honeybone and Joseph Salmons, editors, *The Oxford Handbook of Historical Phonology*, chapter 31, pages 545–562. Oxford University Press, Oxford.

René Kager. 1999. *Optimality Theory*. Cambridge University Press, Cambridge.

Giuseppe Longobardi and Christina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706.

John J. McCarthy. 2008. *Doing Optimality Theory: Applying Theory to Data*. Blackwell Publishing, Malden, MA.

Daniel Müllner. 2013. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–18.

Johanna Nichols and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing.

Taraka Rama and Kolachina Prasanth. 2012. How good are typological distances for determining genealogical relationships among languages? *Proceedings of COLING 2012; Posters*, pages 975–984.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indoeuropean and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.