The application of chordal graphs to inferring phylogenetic trees of languages

Jessica Enright and **Grzegorz Kondrak**
Department of Computing Science
University of Alberta
Edmonton, AB, T6G 2E8, Canada
{enright,kondrak}@cs.ualberta.ca

Abstract

Phylogenetic methods are used to build evolutionary trees of languages given character data that may include lexical, phonological, and morphological information. Such data rarely admits a perfect phylogeny. We explore the use of the more permissive conservative Dollo phylogeny as an alternative or complementary approach. We propose a heuristic search algorithm based on the notion of chordal graphs. We test this approach by generating phylogenetic trees from three datasets, and comparing them to those produced by other researchers.

1 Introduction

Reconstructing the histories of language families is one of the principal tasks of historical linguistics. A linguistic phylogenetic tree conveys the evolution of a language family. The family tree can be constructed on the basis of characteristics that are common to sets of languages. which include lexical, phonological, and morphological affinities. Of particular importance are cognates — words that originate from the same ancestral word, and are distinct from words that are "borrowed", i.e. transferred between languages at some point in history. For example, English *brother* is cognate with German *bruder* and Russian *brat*, all of which come from a Proto-Indo-European word reconstructed as *bhrāter*, while *cousin* is a borrowing from French.

The task of inferring phylogenetic trees of languages is complicated by the pervasiveness of borrowings, which frequently occur when languages are in close contact. For example, English is a Germanic language but the majority of its vocabulary is borrowed from other branches of Indo-European. Several approaches have been proposed to incorporate borrowings into the tree building process (Minett and Wang, 2003). In particular, Nakhleh et al. (2005a) introduce the concept of a phylogenetic *network*, which is obtained by augmenting a putative tree with edges representing contact between languages. They present a method to calculate the minimum number of borrowings required to admit that tree. However, the method does not actually construct a tree from the data, and it may be computationally intractable when the number of borrowings is large.

In this paper, we propose to apply a conservative Dollo phylogeny (CDP) as a model of linguistic phylogenetics. The approach was originally developed by Przytycka et al. (2006) in computational biology. Since it is NP-Hard to compute the minimum number of deletions required for a dataset to conform to a CDP (Lewis and Yannakakis, 1980), we propose a heuristic search algorithm based on the notion of chordal graphs. Our algorithm produces an output tree that minimizes the number of borrowings directly from the data. In addition, it has the potential of being significantly faster to compute than the more commonly known perfect phylogeny. Our approach produces plausible phylogenetic trees on three different datasets.

This paper is structured as follows. In Section 2, we outline the required background, including several graph theoretic notions and alternative phylogenies. In Section 3, we describe our heuristic search algorithm to compute the minimum set of data entries that are inconsistent with a CDP. We also describe a number of preprocessing steps that we take in order to make our problem more computationally feasible. In Section 4, we describe the experiments on three datasets, and compare the resulting phylogenetic trees to those produced by other researchers. In Section 5, we describe an extension to our heuristic search. We conclude with future work and a summary in Sections 6 and 7.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 545–552,
Chiang Mai, Thailand, November 8 – 13, 2011.

2 Background

In this section, we outline the notions of perfect and Dollo phylogenies, and several graph-theoretic notions, including intersection graphs and chordal graphs.

2.1 Perfect phylogeny

A *character* represents a property of languages. In this paper, we consider only binary characters, which have two possible states: 1 and 0. For example, a presence or absence of a particular cognate can be considered a character. The information encoded by a set of characters is used for constructing phylogenetic trees. We say that a character *back evolved* if after evolving from 0 state to 1 state, it subsequently is lost and switches back on the tree from 1 state to 0 state. We say that a character has *parallel evolution* if it evolves twice on the tree from state 0 to state 1 independently. We say that a character is *borrowed* if it has been transferred from one branch to another by contact between linguistic groups.

Given a set of binary characters, we say that a rooted tree with languages as the leaf nodes is a *perfect phylogeny* if for each character there exists a binary labeling such that the root node is labeled with a zero, and all nodes sharing the same label are connected. This implies that each character evolves exactly once, and that there is no back-mutation or borrowing. For example, the tree in Figure 1 is not a perfect phylogeny because characters one and two back-evolve. It is possible to recognize whether a set of characters admits a perfect phylogeny in polynomial time (Felsenstein, 2004).

The character data representing actual languages rarely admit a perfect phylogeny because back mutation, parallel evolution, and borrowing often occur in the course of linguistic evolution. Instead, we are usually interested in establishing for a given character data how far away it is from admitting a perfect phylogeny. *Maximum parsimony* attempts to minimize the overall number of evolutionary events required on a tree to explain the character data, where an evolutionary event is a switch of a character from one state to another (Felsenstein, 2004). Because maximum parsimony is NP-Hard (Day et al., 1986), many approximate approaches have been proposed for this task. Nakhleh et al. (2005b) provide an excellent survey of linguistic phylogenetic methods.

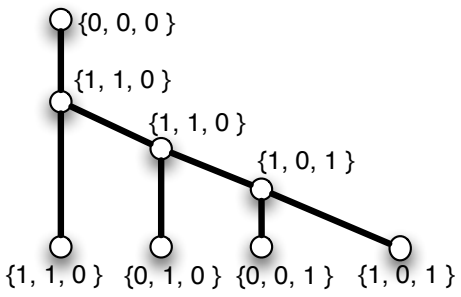


Figure 1: An example of a tree with three characters that is a conservative Dollo phylogeny (CDP), but not a perfect phylogeny.

Nakhleh et al. (2005a) propose perfect *phylogeny networks* as a way of simplifying the phylogeny problem. A perfect phylogeny network is a graph (not necessarily a tree) such that every character exhibits a perfect phylogeny on at least one of the subtrees of that graph. This approach is particularly powerful in modeling borrowing; however, it requires the underlying genetic tree to be defined beforehand. The edges added to the tree represent contact between languages. Unfortunately, even given a phylogenetic tree and character data, determining the minimum number of edges one must add to produce a perfect phylogeny network is NP-Hard (Day et al., 1986). Nakhleh et al. (2005a) mention that applying the perfect phylogeny network approach to their Indo-European language dataset is tractable only because very few edges need to be added to their tree to produce a perfect phylogeny network.

2.2 Dollo phylogenies

Given a set of binary characters, we say that a rooted tree with languages as the leaf nodes is a *Dollo phylogeny* if for each character there exists a binary labeling such that the root node is labeled with a zero, and all nodes sharing the label 1 are connected (Farris, 1977). In essence, each character evolves exactly once, but, in contrast to a perfect phylogeny, an arbitrary number of back-mutations are allowed. Unfortunately, every set of character data admits a Dollo phylogeny. Clearly, the notion of Dollo phylogeny is too permissive to be useful in linguistic phylogenetics.

Przytycka et al. (2006) propose the notion of a *conservative Dollo phylogeny* (CDP), which is a Dollo phylogeny satisfying an additional con-

dition: any two characters that occur together in their 1 states at an internal node must also occur together in their 1 states at some leaf node. For example, the tree in Figure 1 is a CDP.

In the context of language evolution, the CDP condition implies that for any two characters in some ancestral language, there exist corresponding evidence in the form of a known language possessing both of those characters. This is a very strong requirement for which numerous linguistic counter-examples can be found, but it is much less strong, and therefore more more likely to be satisfied, than the requirement for a perfect phylogeny. We expect the CDP condition to guide our heuristic search algorithm towards more realistic phylogenetic reconstructions, especially in cases where a number of diverse languages share a relatively small set of reliable characters.

Few non-trivial datasets representing language families admit a CDP. This may be attributed either to borrowing or to the violation of the CDP condition. Since we have no way distinguish between the two explanations, in the remainder of this paper we will simply refer to such events as borrowings. In most cases, our objective is to establish the minimum number of those instances.

2.3 Chordal graphs

In this section we define the notions of chordal and intersection graphs that underlie our heuristic search algorithm.

Graph $G = (V, E)$, where E is the set of edges, and V is the set of vertices, is an *intersection graph* of a family of sets $\mathcal{R}$ if there is a one-to-one and onto function (*bijection*) $\mathcal{F}$ between V and $\mathcal{R}$ such that

$$\forall s, t \in R : (F(s), F(t)) \in E \text{ iff } s \cap t \neq \emptyset \quad (1)$$

Informally, each vertex in the intersection graph represents a set, and two vertices are connected by an edge if and only if the two corresponding sets intersect. Figure 2 shows an example of an intersection graph. Given sets, we can compute their intersection graph in linear time.

A *chord* of a cycle is an edge between two non-consecutive vertices of a cycle. A *chordless cycle* is a cycle with no chords. A *chordal graph* is a graph with no chordless cycles of length greater than three. Figure 3 shows an example of a chordal graph. Rose et al. (1976) provide a linear-time recognition algorithm for chordal graphs.

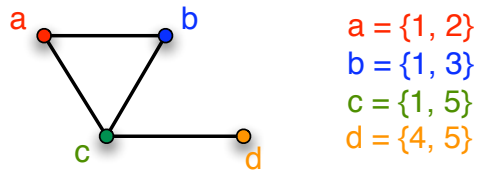


Figure 2: An example of a family of sets on the right and an intersection graph of those sets on the left.

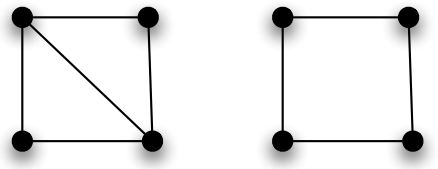


Figure 3: An example chordal graph on the left, and a graph that is not chordal on the right.

We can consider a character in phylogeny data as a set composed of the individuals or languages that possess that character. For example, the set for a cognate is the set of languages that contain that cognate. Przytycka et al. (2006) prove that a set of characters admits a CDP if and only if their intersection graph is chordal. Therefore, it is possible to determine whether a set of characters admits a CDP in linear time. We employ this result in order to infer phylogenetic trees of languages.

3 Heuristic search

Our search algorithm takes the intersection graph of the character data as input, and finds the minimum number of vertices (characters, in this case) that must be removed to produce a chordal graph. We take advantage of the fact that a character dataset admits a CDP if and only if the intersection graph of the character data is chordal. Our heuristic breadth-first search is guaranteed to find the minimum number of the inconsistent vertices.

One key observation allows this search to execute in reasonable time:

Observation 1. *Let $G = (V, E)$ be a graph. Let C be a chordless cycle of G. Let V' be a set of vertices such that removing V' from G results in a chordal graph. Then V' must include at least one vertex on C.*

Algorithm 1 Our main heuristic search algorithm.
search(Graph $G = (V, E)$, List *currentSolution*)

```
1: Vertex v ← isChordal(G);
2: if v = null then
3:    Print currentSolution
4:    return
5: end if
6: Vector candidates ← getCandidates(G, v)
7: for all Vertex u in candidates do
8:    Add u to currentSolution
9:    Graph G' ← G[V\{u}]
10:   search(G', currentSolution)
11:   Remove u from currentSolution
12: end for
```

Algorithm 2 Candidate generator.
getCandidates(Graph $G = (V, E)$, Vertex v)

```
1: Cycle C ← the vertices of the shortest chordless cycle of length ≥ 4 containing v
2: if |C| > 4 then
3:    Cycle C_4 ← a chordless cycle of length 4 if one exists in G
4:    if C_4 is not null then
5:       return C_4
6:    end if
7: end if
8: return C
```

Proof. Assume that V' contains none of the vertices in C. Then all vertices of C are in $V \backslash V'$. Therefore C is present in the graph obtained by removing V' from G, which is chordal by definition of V', a contradiction. □

By applying the above observation inductively, at each stage of our search, we need only consider as successor states the states produced by removing a vertex in a chordless cycle of our graph.

The pseudo-code of our heuristic search is shown in Algorithm 1. Subroutine *isChordal* takes a graph as a parameter, and returns a vertex that belongs to a chordless cycle, or $null$ if G is chordal. The subroutine implements the algorithm based on lexicographic breadth-first search proposed by Rose et al. (1976). The subroutine *getCandidates* for selecting the candidate nodes to be considered in the search is formalized in Algorithm 2. It gets all vertices on a chordless cycle, and identifies them as candidates for removal. In order to guarantee optimality, we need to recursively consider removing each of these vertices. Observation 1 makes this search computationally feasible in experimental data. In the worst case, the algorithm has exponential running time in the size of the input data, which is what we expect in the case of an exact algorithm applied to an NP-complete problem.

3.1 Language grouping

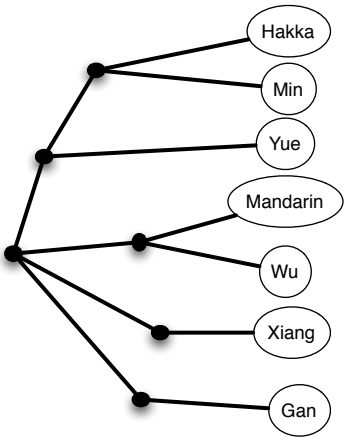
In order to make our experiments computationally tractable, we follow Nakhleh et al. (2005a) in combining sets of languages into single units. For example, we consider the Germanic languages as a single group because we are confident that their most recent common ancestor is not an ancestor of any other language. The operation of language grouping is performed as a preprocessing step to the construction of the intersection graph of the characters.

Beyond achieving the goal of decreasing computation time, we expect that the application of language grouping will actually make our data closer to admitting a CDP in a way consistent with true evolutionary history. Consider two characters s and t that intersect in the context of language grouping L, but not without. Then s and t are not present in any of the same languages, but there are two languages $l_i, l_j \in L$ such that l_i has character s but not t, and language l_j has character t but not s. If s and t are only present within the language grouping, they are not informative when language family grouping is used. However, if both s and t are present at an internal node ancestral to language grouping L, then this will make the data closer to admitting a CDP by decreasing the number of borrowings that we need to posit.

3.2 Tree extraction

Once our search has found a minimum set of vertices to remove from our graph in order to make it chordal, we extract a phylogenetic tree from the resulting graph. Gavril (1974) shows that chordal graphs are the intersection graphs of subtrees of a tree, and gives a polynomial-time algorithm to build a corresponding family of subtrees. Following Przytycka et al. (2006), we use the union of those subtrees as a phylogenetic tree for our languages.

Figure 4: The tree given by our algorithm for Chinese dialect cognate data.



4 Experiments

In order to assess the suitability of the CDP approach to linguistic phylogeny we performed experiments on three datasets.

4.1 Chinese dialects

The data consist of 15 cognates across seven Chinese dialects compiled by Minett and Wang (2003). The set can be characterized as relatively small and clean.

The tree produced by our algorithm (Figure 4, which is non-binary, is completely consistent with one of the five binary trees (Type III) of Minett and Wang (2003). Also, our tree shares the grouping of Hakka and Min with all five of their proposed trees.

Minett and Wang (2003) show that in order to explain their data, at least seven borrowings must have occurred. Our algorithm gives the same number.

The experiment confirms that our method is sound, and produces results that are open to further elucidation.

4.2 CPHL subset

The dataset consists of 22 phonological characters and 13 morphological characters for 24 Indo-European languages from the Computational Phylogenetics in Historical Linguistics (CPHL) project[1]. We decided to exclude the lexical characters which are the most likely to be borrowed.

For example, one phonological character identifies languages that underwent the loss of initial y when it was followed by e. Three languages (Hittite, Luvian, and Lycian) which exhibit that sound change are encoded as character 1, while the remaining Indo-European languages are encoded as character 2.

Figure 5 shows the tree produced by our method on the CPHL subset. No characters needed to be removed from the intersection graph of the characters to yield a chordal graph, which can be interpreted that our CDP assumption is reasonable.

There are several differences between the tree in Figure 5 and the tree presented on the website of the CPHL project. First, Albanian is grouped with the Armenian and Greek languages rather than with the Germanic languages. Second, there is a node of high degree linking four language subfamilies. This result implies that the character set is not sufficiently large to establish a more detailed relationship between those subfamilies.

Nakhleh et al. (2005a) use the CPHL dataset to build perfect phylogenetic networks. However, their approach requires a phylogenetic tree as input, and minimizes borrowing events over only that tree. In contrast, our approach minimizes borrowing events over all possible phylogenetic trees without enumerating them.

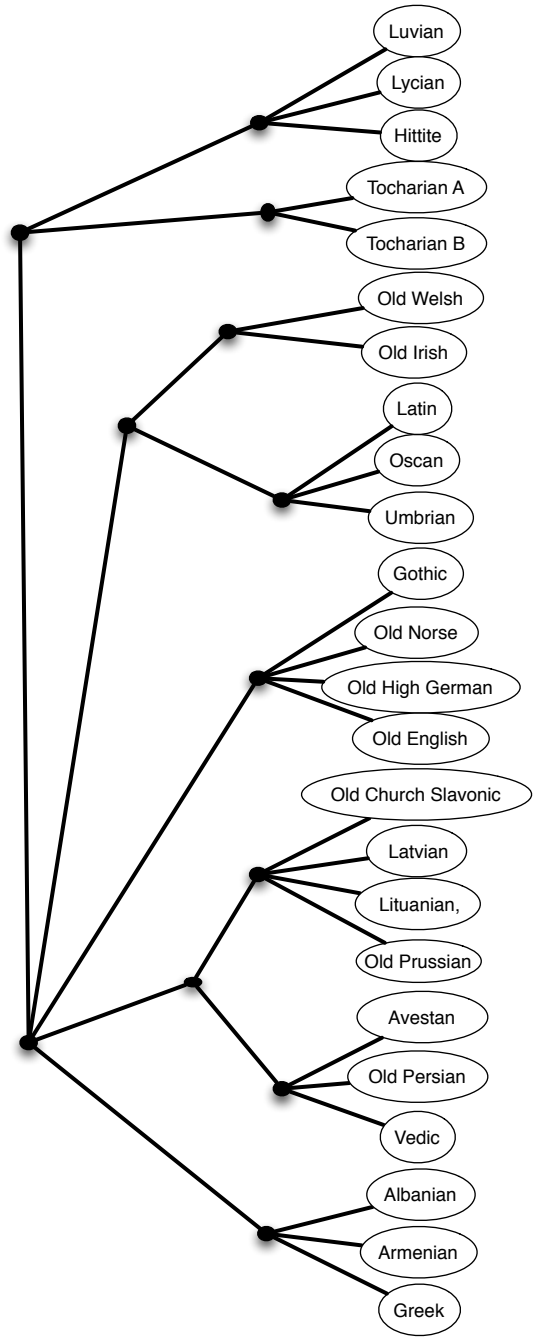
4.3 Comparative Indo-European Data Corpus

The dataset consists of 84 Swadesh 200-word lists representing contemporary Indo-European languages (Dyen et al., 1992). We used only the most reliable cognate sets numbered between 002 and 099, which contain forms that are cognate with each other and not cognate with the forms belonging to other sets. Each cognate set is treated as a separate character. For grouping purposes, we divided the languages into the following ten groups: Celtic, Romance, Germanic, Baltic, Slavic, Indic, Greek, Armenian, Iranian, and Albanian.

As a further simplifying step, we identified all cases where the same language group contains multiple cognate sets representing the same Swadesh meaning. For example, consider the words for 'neck': Nepali *manto* is cognate with Irish *muineal*, while Gujarati *gerden* is cognate with Macedonian *vrat*. In such cases, we removed all but one of the cognate sets involved, provided they are found in only two language families. In our example, we therefore remove the cognate set

[1] http://www.cs.rice.edu/~nakhleh/CPHL/

Figure 5: The tree given by our algorithm for the CPHL dataset, morphological and phonological characters only, with no language grouping.



shared between the Celtic and Indic families. This does not affect the total number of required borrowings.

Figure 6 shows the tree obtained when we ran our search on the Comparative Indo-European Data Corpus with language grouping. To construct this tree, our search found a minimum set of eight inconsistent cognates. We have analyzed these cognates, and while a few are inherited, most are either borrowings or annotation errors (e.g., Slovenian *jagat* 'to hunt' or Albanian *tuti* 'all').

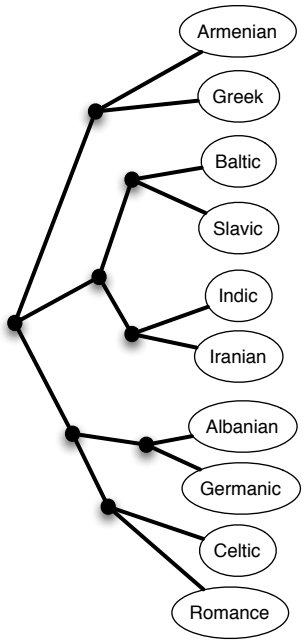
Although the consensus about the exact form of the Indo-European language phylogeny is yet to be reached in spite of many decades of research, our tree conforms to several well-established facts, including the affinity betwen Baltic and Slavic, Indic and Iranian, and the *Satem core*. However, there are some differences in comparison with the tree proposed by the CPHL project.

In Figure 6, the Albanian and Germanic languages are more closely grouped with the Celtic and Romance languages than the Baltic, Slavic, Indic, Greek and Armenian Languages. The opposite is true in the tree proposed by Nakhleh et al. (2005a). However, they note that the Germanic languages seem to have exhibited a large amount of borrowing compared to the other languages they considered, and mention that on their trees the position of Albanian is uncertain. If Germanic has undergone a substantial amount of non-tree-like evolution, then it is unsurprising that it is the major source of disagreement between our and their trees.

Our tree in Figure 6 also differs from our tree in Figure 5, in which a single vertex branches to many groups of languages. Another difference is the placement of Albanian.

Gray and Atkinson (2003) also built a phylogenetic tree for the Indo-European languages using the Comparative Indo-European Data Corpus. Their tree differs significantly from that proposed by Nakhleh et al. (2005a). The tree produced by Gray and Atkinson (2003) does not group Balto-Slavic with Indo-Iranian, while both our tree and the tree in (Nakhleh et al., 2005a) do. Gray and Atkinson (2003) instead groups Balto-Slavic with the Germanic, Romance, and Celtic language families. The placement of Albanian is different in all three trees. All three trees group Baltic with Slavic, Armenian with Greek, and Indic with Iranian.

Figure 6: The tree given by our algorithm for the Comparative Indo-European Data Corpus with language grouping.



5 Tiered search

We noticed when running experiments that characters that are thought in the literature to be borrowed are often present in few language families, whereas characters that are thought to be ancestrally inherited are often present in a larger number of language families. We therefore devised a tiered version of our heuristic search. In this version, we first run the search on only the characters that are present at more than two language families. We find the minimum number of these characters that must be removed to result in a chordal graph, remove that minimum set from the overall dataset, and run a subsequent search on this set in which the only vertices that are allowed to be removed are present at exactly two language families. We finally concatenate the minimum set of vertices that this search finds with the minimum set found in the earlier search to produce our overall minimal set of borrowing events.

We tested tiered search on the Comparative Indo-European Data Corpus. Our results were negative. In particular, the resulting tree fails to group Baltic and Slavic together, which is universally accepted in historical linguistics. This suggests that the observation is not sufficiently general to improve the proposed method.

6 Future work

We plan to extend our research on several directions.

First, our heuristic search could likely be made more efficient, though not asymptotically. Apart from the careful selection of nodes to evaluate as noted in Observation 1, we perform no search tree pruning. There are likely choices of nodes to be evaluated that are strictly dominated by other nodes. For example, consider two vertices u, v in a chordless cycle C of length greater than three such that the only neighbors of u are in C, but v is also in other chordless cycles, and has neighbors outside C. Then the choice to remove v strictly dominates the choice to remove u.

Second, we plan to modify our search procedure to minimize the total number of borrowings. rather than the total number of borrowed characters. We have found that in our experiments that minimizing the later has always minimized the former, but this may not be the case in all datasets.

Finally, we intend to improve the speed of the heuristic search, which would enable us to run tests on larger and more inconsistent datasets.

7 Conclusion

We have proposed conservative Dollo-phylogeny as a model for linguistic phylogenetics. We devised and tested an algorithm for calculating the minimum number of inconsistent characters within a dataset over all possible phylogenetic trees without enumerating those trees. We tested this approach on three datasets with positive results.

The main advantage of this approach is its speed. All computations took very little time - on the order of seconds. Previous approaches have been much slower. The trees produced by our method are therefore useful not only in their own right, but also as a very rapid initial stage of a computation. One possible approach would be to quickly generate trees with our method, and then use them as input to a more exhaustive algorithm.

Our approach calculates the minimum number of characters that are inconsistent with CDP across all possible phylogenetic trees without actually considering these trees individually. The CDP model and our heuristic algorithm may be particularly useful in cases where the number of languages is large, or where not even a partial tree is known.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank Dr. Lorna Stewart for cooperation.

References

William Day, David Johnson, and David Sankoff. 1986. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33–42.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5).

James S. Farris. 1977. Phylogenetic analysis under Dollo's law. *Systematic Zoology*, 26(1):77–88.

Joseph Felsenstein. 2004. *Inferring Phylogenies*. Number 1. Sinauer Associates, Massachusetts, USA.

Fanica Gavril. 1974. The intersection graphs of subtrees in trees are exactly the chordal graphs. *Journal of Combinatorial Theory (B)*, 16:47–56.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, November.

John M. Lewis and Mihalis Yannakakis. 1980. The node-deletion problem for hereditary properties is NP-complete. *Journal of Computer and System Sciences*, 20:219–230.

James W. Minett and William S.-Y Wang. 2003. On detecting borrowing: Distance-based and character-based approaches. *Diachronica*, 20:289–331(43).

Luay Nakhleh, Don Ringe, and Tandy Warnow. 2005a. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language (Journal of the Linguistic Society of America)*, 81(2):382–420.

Luay Nakhleh, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005b. A comparison of phylogenetic reconstruction methods on an IE dataset. *The Transactions of the Philological Society*, 3(2):171 – 192.

Teresa Przytycka, George Davis, Nan Song, and Dannie Durand. 2006. Graph theoretical insights into evolution of multidomain proteins. *Journal of computational biology*, 13(2):351–363.

Donald J. Rose, R. Endre Tarjan, and George S. Leuker. 1976. Algorithmic aspects of vertex elimination on graphs. *SIAM Journal of Computing*, 5(2):266–283.