PHYLO-ASP: Phylogenetic Systematics with Answer Set Programming

Esra Erdem

Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul 34956, Turkey

Abstract. This note summarizes the use of Answer Set Programming to solve various computational problems to infer phylogenetic trees and phylogenetic networks, and discusses its applicability and effectiveness on some real taxa.

1 Introduction

Cladistics (or phylogenetic systematics), developed by Willi Hennig [1], is the study of evolutionary relations between species based on their shared traits. Represented diagrammatically, these relations can form a tree whose leaves represent the species, internal vertices represent their ancestors, and edges represent the genetic relationships between them. Such a tree is called a "phylogenetic tree" (or a "phylogeny"). We consider reconstruction of phylogenies as the first step of reconstructing the evolutionary history of a set of taxa (taxonomic units). The idea is then to reconstruct (temporal) phylogenetic networks, which also explain the contacts (or borrowings) between taxonomic units, from the reconstructed phylogenies.

We studied both steps using Answer Set Programming: the first step is studied in [2,3,4], and the second step is studied in [5,6]. We call our ASP-based approach to phylogenetic tree and phylogenetic network reconstruction as PHYLO-ASP. We illustrated the applicability and effectiveness of PHYLO-ASP for the historical analysis of languages, and to the historical analysis of parasite-host systems.

Histories of individual languages give us information from which we can infer principles of language change. This information is not only of interest to historical linguists but also of interest to archaeologists, human geneticists, physical anthropologists as well. For instance, an accurate reconstruction of the evolutionary history of certain languages can help us answer questions about human migrations, the time that certain artifacts were developed, when ancient people began to use horses in agriculture [7,8,9,10].

Parasites occur worldwide, causing malnutrition, sickness, and even sometimes the death of their hosts. Historical analysis of parasites gives us information on where they come from and when they first started infecting their hosts. The phylogenies of parasites, with the phylogenies of their hosts, and with the geographical distribution of their hosts, can be used to understand the changing dietary habits of a host species, to understand the structure and the history of ecosystems, and to identify the history of animal and human diseases. This information allows predictions about the age and duration of specific groups of animals of a particular region or period, identification of regions of evolutionary "hot spots" [11], and thus can be useful to assess the importance of specific habitats, geographic regions, and biotas—all the plant and animal life of a

E. Erdem, F. Lin, and T. Schaub (Eds.): LPNMR 2009, LNCS 5753, pp. 567–572, 2009.

particular region—and areas of critical genealogical and ecological diversity [12,11]. Identification of the most vulnerable members of a community by this way allows us to make more reliable predictions about the impacts of perturbations (natural or caused by humans) on ecosystem structure and stability [12].

With PHYLO-ASP, we studied evolutionary history of 7 Chinese dialects based on 15 lexical characters, and 24 Indo-European languages based on 248 lexical, 22 phonological and 12 morphological characters. Some of the phylogenetic trees and networks computed by PHYLO-ASP are plausible from the point of view of historical linguistics. We also studied evolutionary history of 9 species of *Alcataenia* (a tapeworm genus) based on their 15 morphological characters. Some of the phylogenetic trees and networks computed by PHYLO-ASP are plausible from the point of view of coevolution—the evolution of two or more interdependent species each adapting to changes in the other, and from the point of view of historical biogeography—the study of the geographic distribution of organisms.

This note summarizes the use of PHYLO-ASP to solve various computational problems related to the inference of phylogenetic trees and phylogenetic networks, and discusses its applicability and effectiveness on some real taxa.

2 Phylogeny Reconstruction

A *phylogenetic tree* (or *phylogeny*) for a set of taxa is a finite rooted binary tree $\langle V, E\rangle$ along with two finite sets I and S and a function f from $L \times I$ to S, where L is the set of leaves of the tree. The set L represents the given taxonomic units whereas the set V describes their ancestral units and the set E describes the genetic relationships between them. The elements of I are usually positive integers ("indices") that represent, intuitively, qualitative characters, and elements of S are possible states of these characters. The function f "labels" every leaf v by mapping every index i to the state $f(v, i)$ of the corresponding character in that taxonomic unit.

A character $i \in I$ is *compatible* with a phylogeny (V, E, L, I, S, f) if there exists a function $g : V \times \{i\} \to S$ such that

(C1) for every leaf v of the phylogeny, $g(v, i) = f(v, i)$;
(C2) for every $s \in S$, if the set $V_{is} = \{x \in V : \ g(x, i) = s\}$ is nonempty then the digraph $\langle V, E\rangle$ has a subgraph with the set V_{is} of vertices that is a rooted tree.

A character is *incompatible* with a phylogeny if it is not compatible with that phylogeny.

The computational problem we are interested in is, given the sets L, I, S, and the function f, to build a phylogeny (V, E, L, I, S, f) with the maximum number of compatible characters. This problem is called the *maximum compatibility problem*. It is NP-hard even when the characters are binary [13]. We solve the maximum compatibility problem, by means of the following decision problem: given sets L, I, S, a function f from $L \times I$ to S, and a nonnegative integer n, decide the existence of a phylogeny (V, E, L, I, S, f) with at most n incompatible characters. In [2,4], we describe this decision problem as an ASP program whose answer sets correspond to such phylogenies.

3 Phylogenetic Network Reconstruction

A contact between two taxonomic units can be represented by a horizontal edge added to a pictorial representation of a "temporal phylogeny"—a phylogeny along with a function τ from vertices of the phylogeny to real numbers denoting the times when these taxonomic units emerged (Fig. 1). The two endpoints of the edge are simultaneous "events" in the histories of these communities. An event can be represented by a pair $v{\uparrow}t$, where v is a vertex of the phylogeny and t is a real number.

A finite set C of contacts defines a *(temporal) phylogenetic network* $\langle V \cup V_C, E_C \rangle$—a digraph obtained from $T = \langle V, E \rangle$ by inserting the elements $v{\uparrow}t$ of the contacts from C as intermediate vertices and then adding every contact in C as a bidirectional edge. We say that a set C of contacts is *simple* if the endpoints of all lateral edges are different from the vertices of T, and each lateral edge subdivides an edge of T into exactly two edges.

About a simple set C of contacts (and about the corresponding phylogenetic network $\langle V \cup V_C, E_C \rangle$) we say that it is *perfect* if there exists a function $g : (V \cup V_C) \times I \to S$ such that the function g extends f from leaves to all internal nodes of the phylogenetic network, and that every state s of every character i could evolve from its original occurrence in some "root" (i.e., every character i is compatible with the phylogenetic network).

We are interested in the problem of turning a temporal phylogeny into a perfect phylogenetic network by adding a small number of simple contacts. For instance, given the phylogeny in Fig. 1(a), the single contact $\{B{\uparrow}1750, D{\uparrow}1750\}$ is a possible answer.

It is clear that the information included in a temporal phylogeny is not sufficient for determining the exact dates of the contacts that turn it into a perfect phylogenetic network. To make this idea precise, let us select for each $v \in V \setminus \{R\}$ a new symbol $v{\uparrow}$, and define the *summary* of a simple set C of contacts to be the result of replacing

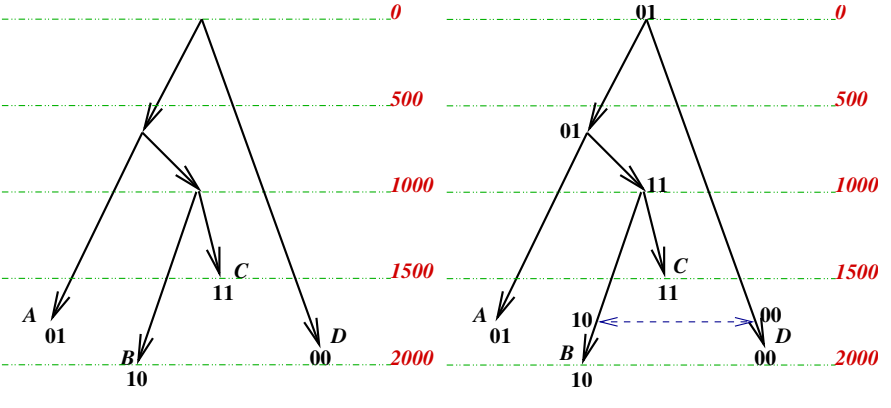

Fig. 1. A temporal phylogeny (a), and a perfect temporal network (b) with a lateral edge connecting $B{\uparrow}1750$ with $D{\uparrow}1750$

each element $v\uparrow t$ of every contact in C with $v\uparrow$. Thus summaries consist of 2-element subsets of the set $V\uparrow = \{v\uparrow : v \in V \setminus \{R\}\}$. For instance, the summary of the set of contacts of Fig. 1(b) is $\{\{B\uparrow, D\uparrow\}\}$.

An *IPSTN problem* (for "Increment to Perfect Simple Temporal Network") is defined by a phylogeny $\langle V, E, I, S, f\rangle$ and a function $v \mapsto (\tau_{min}(v), \tau_{max}(v))$ from the vertices of the phylogeny to open intervals. A *solution* to the problem is a set of 2-element subsets of $V\uparrow$ that is the summary of a perfect simple set of contacts for a temporal phylogeny $\langle V, E, I, S, f, \tau\rangle$ such that, for all $v \in V$, $\tau_{min}(v) < \tau(v) < \tau_{max}(v)$.

In [6], we describe IPSTN problem as an ASP program. We solve IPSTN problems in two steps: use an ASP system to compute summaries so that every character is compatible with the phylogenetic network, and then use a constraint programming system to check, for each of summary, whether the corresponding contact occurs within the given time intervals.

4 Experimental Results

We applied PHYLO-ASP to three sets of taxa: Chinese dialects, Indo-European languages, and Alcataenia (a tapeworm genus) species. For each taxa and a given integer n, first we computed all phylogenies with at most n incompatible characters, iteratively with a script as follows: at iteration i, compute the i'th phylogeny with the input program, and then add to the input program a constraint that prevents generation of the answer sets that describe the i'th phylogeny. After that, we identified the phylogenies that are plausible. For the Chinese dialects and Indo-European languages, the plausibility of phylogenies depends on the linguistics and archaeological evidence; for *Alcataenia*, the plausibility of the phylogeny we compute is dependent on the knowledge of host phylogeny (e.g., phylogeny of the seabird family *Alcidae*), chronology of the fossil record, and biogeographical evidence. Then, for each plausible phylogeny, we computed phylogenetic networks that require minimum number of lateral edges, and identified the plausible ones.

Experiments with Chinese dialects. We considered the Chinese dialects Xiang, Gan, Wu, Mandarin, Hakka, Min, and Yue. We used the dataset, originally gathered by Xu Tongqiang and processed by Wang Feng, described in [14]. In this dataset, there are 15 lexical characters; each character has 2–5 states.

After preprocessing this dataset, we computed 33 phylogenies with 6 incompatible characters, and we found out that there is no phylogeny with less than 6 incompatible characters. These phylogenies are presented in [4]. The sub-grouping of the Chinese dialects is not yet established. However, many specialists agree that there is a Northern group and a Southern group. That is, for the dialects we chose in our study, we would expect a (Wu, Mandarin, Gan, Xiang) Northern grouping and a (Hakka, Min) Southern grouping. (It is not clear which group Yue belongs to.) We identified 5 plausible phylogenies with respect to this hypothesis.

For each plausible phylogeny, we reconstructed phylogenetic networks. We observed that, among these phylogenies, two of them require at least 2 lateral edges (representing borrowings between Gan and Wu, and between (Mandarin, Wu) and Min) to turn into a plausible perfect phylogenetic network.

Experiments with Indo-European languages. We applied PHYLO-ASP to reconstruct the evolutionary history of the Indo-European language groups Balto-Slavic (BS), Italo-Celtic (IC), Greco-Armenian (GA), Anatolian (AN), Tocharian (TO), Indo-Iranian (IIR), Germanic (GE), and the language Albanian (AL). We used the dataset assembled by Don Ringe and Ann Taylor [15], with the advice of other specialist colleagues. There are 282 informative characters in this dataset, each with 2–22 states.

After preprocessing this dataset, we computed 45 phylogenies with at most 20 incompatible characters, taking into account the given domain-specific information (e.g., Anatolian is the outgroup for all the other subgroups, Albanian cannot be a sister of Indo-Iranian or Balto-Slavic). Out of these 45 phylogenies, 34 are identified by Don Ringe as plausible from the point of view of historical linguistics. These phylogenies are presented in [4]. The most plausible one with 16 incompatible characters is (AN, (TO, (IC, ((GE, AL), (GA, (IIR, BS)))))).

Based on this phylogeny, and given some time intervals for each node in the phylogeny, we reconstructed 3 plausible temporal phylogenetic networks with 3 lateral edges, taking also into account some domain-specific information (e.g., a contact between IC and BA is unlikely because the former was spoken in western Europe, while the Balts were probably confined to a fairly small area in northeastern Europe).

Experiments with Alcatenia species. We used PHYLO-ASP to reconstruct the evolutionary history of 9 species of *Alcataenia*—a tapeworm genus whose species live in alcid birds (puffins and their relatives): *A. Larina*, *A. Fraterculae*, *A. Atlantiensis*, *A. Cerorhincae*, *A. Pygmaeus*, *A. Armillaris*, *A. Longicervica*, *A. Meinertzhageni*, *A. Campylacantha*. We used the dataset described in [16]. In this dataset, there are 15 characters, each with 2–3 states.

After preprocessing this dataset, we computed 18 phylogenies, with 5 incompatible characters, for *Alcataenia*, taking into account some domain-specific information (e.g., the outgroup for all *Alcataenia* species is *A. Larina*). For the plausibility of the phylogenies for *Alcataenia*, we consider the phylogenies of its host *Alcidae* (a seabird family) and the geographical distributions of *Alcidae*. For instance, according to host and geographic distributions over the time, diversification of *Alcataenia* is associated with sequential colonization of puffins (parasitized by *A. Fraterculae* and *A. Cerorhincae*), razorbills (parasitized by *A. Atlantiensis*), auklets (parasitized by *A. Pygmaeus*), and murres (parasitized by *A. Armillaris*, *A. Longicervica*, and *A. Meinertzhageni*). This pattern of sequential colonization is supported by the phylogeny of *Alcidae* in [17]. Out of the 18 trees we computed, only two are consistent with this pattern. Each plausible tree needs 3 lateral edges to turn into a perfect phylogenetic network.

5 Conclusion

We have briefly described the use of ASP to reconstruct the evolutionary history of a set of taxonomic units (as in [5,3,6,2,4]), calling this ASP-based approach to phylogenetic systematics as PHYLO-ASP. We have discussed the applicability and effectiveness of PHYLO-ASP with three sets of taxa: Indo-European languages, Chinese dialects, and *Alcatenia* species. Our ongoing work involves extending PHYLO-ASP to analyze and compare phylogenetic trees and networks [18,19].

Acknowledgments. This work has involved, at various stages, close collaborations with Dan Brooks, Selim Erdoğan, Vladimir Lifschitz, Luay Nakhleh, James Minett, Don Ringe, Feng Wang. It has been partially supported by TUBITAK Grant 107E229.

References

1. Hennig, W.: Phylogenetic Systematics. University of Illinois Press (1966); by Davis, D.D., Zangerl, R.: Translated from Grundzuege einer Theorie der phylogenetischen Systematik (1950)
2. Brooks, D.R., Erdem, E., Minett, J.W., Ringe, D.: Character-based cladistics and answer set programming. In: Hermenegildo, M.V., Cabeza, D. (eds.) PADL 2004. LNCS, vol. 3350, pp. 37–51. Springer, Heidelberg (2005)
3. Erdem, E., Wang, F.: Reconstructing the evolutionary history of Chinese dialects. Accepted for presentation at the 39th International Conference on Sino-Tibetan Languages and Linguistics, ICSTLL 2006(2006)
4. Brooks, D.R., Erdem, E., Erdoğan, S.T., Minett, J.W., Ringe, D.: Inferring phylogenetic trees using answer set programming. Journal of Automated Reasoning 39(4), 471–511 (2007)
5. Erdem, E., Lifschitz, V., Nakhleh, L., Ringe, D.: Reconstructing the evolutionary history of Indo-European languages using answer set programming. In: Dahl, V., Wadler, P. (eds.) PADL 2003. LNCS, vol. 2562, pp. 160–176. Springer, Heidelberg (2003)
6. Erdem, E., Lifschitz, V., Ringe, D.: Temporal phylogenetic networks and logic programming. Theory and Practice of Logic Programming 6(5), 539–558 (2006)
7. Mair, V.H. (ed.): The Bronze Age and Early Iron Age Peoples of Eastern Central Asia. Institute for the Study of Man, Washington (1998)
8. Mallory, J.P.: In Search of the Indo-Europeans. Thames and Hudson, London (1989)
9. Roberts, R.G., Jones, R.M., Smith, M.A.: Thermoluminescence dating of a 50,000-year-old human occupation site in Northern Australia. Science 345, 153–156 (1990)
10. White, J.P., O'Connell, J.F.: A Prehistory of Australia, New Guinea, and Sahul. Academic Press, New York (1982)
11. Brooks, D.R., McLennan, D.A.: Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology. Univ. Chicago Press, Chicago (1991)
12. Brooks, D.R., Mayden, R.L., McLennan, D.A.: Phylogeny and biodiversity: Conserving our evolutionary legacy. Trends in Ecology and Evolution 7, 55–59 (1992)
13. Day, W.H.E., Sankoff, D.: Computational complexity of inferring phylogenies by compatibility. Systematic Zoology 35(2), 224–229 (1986)
14. Minett, J.W., Wang, W.S.Y.: On detecting borrowing: distance-based and character-based approaches. Diachronica 20(2), 289–330 (2003)
15. Ringe, D., Warnow, T., Taylor, A.: Indo-European and computational cladistics. Transactions of the Philological Society 100(1), 59–129 (2002)
16. Hoberg, E.P.: Congruent and synchronic patterns in biogeography and speciation among seabirds, pinnipeds, ans cestodes. J. Parasitology 78(4), 601–615 (1992)
17. Chandler, R.M.: Phylogenetic analysis of the alcids. PhD thesis, University of Kansas (1990)
18. Cakmak, D., Erdem, E., Erdogan, H.: Computing weighted solutions in answer set programming. In: Proc. of LPNMR (to appear, 2009)
19. Eiter, T., Erdem, E., Erdogan, H., Fink, M.: Finding similar or diverse solutions in answer set programming. In: Proc. of ICLP (to appear, 2009)