Incomplete Directed Perfect Phylogeny

Itsik Pe'er izik@math.tau.ac.il
Ron Shamir shamir@math.tau.ac.il
Roded Sharan roded@math.tau.ac.il
Department of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel

Keywords: perfect phylogeny, graph sandwich, SINE insertion

1 Introduction

Perfect phylogeny is one of the fundamental models for studying evolution. We investigate the following variant of the problem: The input is an $n \times m$ species-characters matrix. The characters are binary and directed, i.e., a species can only gain characters. The difference from standard perfect phylogeny is that for some species the state of some characters is unknown. The question is whether one can complete the missing states in a way admitting a perfect phylogeny. We call this problem *Incomplete Directed Perfect phylogeny* (IDP).

The problem arises in classical phylogenetic studies, when some states are missing or undetermined. Swofford's PAUP software package [4] provides an exponential solution to the problem by exhaustive search. Quite recently, a novel kind of genomic data has given rise to the same problem: Nikaido *et al.* [3] use inserted repetitive genomic elements, particularly SINEs, as a source of evolutionary information. The specific insertion events are identifiable by the flanking sequences on both sides of the insertion site. These insertions are assumed to be unique, irreversible events in evolution. However, the site and its flanking sequences may be lost when a large region of the genome which includes them is deleted. In that case we do not know whether an insertion had occurred in the missing site. One can model such data by assigning each locus a character, whose state is '1' if the SINE occurred in that locus, '0' if the locus is present but does not contain the SINE, and '?' if the locus is missing. An example of the problem input and solution is given in Figure 1.

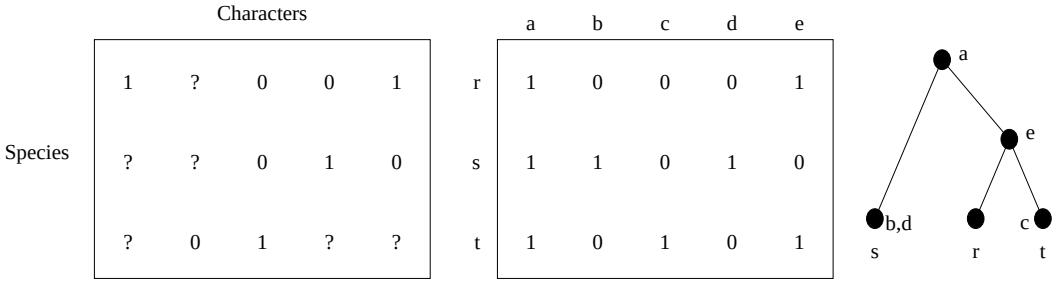

Figure 1: Left to right: An incomplete matrix $\mathcal{A}$, a completion $\mathcal{B}$ of $\mathcal{A}$, and a phylogenetic tree for $\mathcal{A}$ and $\mathcal{B}$. A character next to a node indicates that the state of the character is 1 on and below that node.

A generalization of IDP for non-binary characters was solved in $O(n^2m)$ time in [1]. We give a more elegant graph sandwich formulation of the problem, devise a faster, simple $\tilde{O}(nm)$ time algorithm, and analyze the generality of the solution it produces.

2 Method and Results

Let $S = \{s_1, \ldots, s_n\}$ be a species set, and let $C = \{c_1, \ldots, c_m\}$ be a character set. The input to IDP can be represented as an $n \times m$ species-characters matrix $\mathcal{A}$, where $a_{ij} \in \{0, ?, 1\}$. Construct the bipartite graph $G(\mathcal{A}) = (S, C, E)$ with $E = \{(s_i, c_j) : a_{ij} = 1\}$. An induced path of length four in $G(\mathcal{A})$ is called a Σ *subgraph* if it starts (and therefore ends) at a vertex corresponding to a species. A graph with no induced Σ subgraph is said to be Σ-*free*. In case the matrix $\mathcal{A}$ is complete (i.e., contains no ?-s), the following key observation can be shown: If $G(\mathcal{A})$ is connected and Σ-free, then it contains a *universal character*, i.e., a character that is adjacent to all species.

The pairwise compatibility theorem (cf. [2]), restated in terms of graph theory, says that a binary matrix $\mathcal{B}$ has a phylogenetic tree iff $G(\mathcal{B})$ is Σ-free. Denoting $E_x = \{(i, j) : a_{ij} = x\}$, for $x = 0, ?, 1$, this gives the following Σ-*free sandwich problem*, which is equivalent to IDP: Given a partition of $S \times C$ into three subsets: E_0 - the forbidden edges, E_1 - the mandatory edges, and $E_?$ - the optional edges, find a supergraph of (S, C, E_1) which is Σ-free and contains no forbidden edge.

Define the x-set of a character c to be $\{s : a_{sc} = x\}$. Define a *clade* of a tree T to be the leaf set of the subtree induced by some node of T. Our algorithm for IDP outputs a phylogenetic tree T for $\mathcal{A}$, represented as a list of its clades. It starts by discarding from $G(\mathcal{A})$ all characters whose 0-set or 1-set are empty, and by initializing the output tree to have the clade S, and all singleton clades. It then works iteratively as follows: In each step the algorithm examines some non-singleton connected component K of $G(\mathcal{A})$. It searches for semi-universal characters in K, i.e., characters whose 0-set does not intersect K. If no such character is found, the algorithm halts, declaring that the instance has no solution (by the key observation above). Otherwise, all semi-universal characters are removed from $G(\mathcal{A})$, and the species set of K is added as a clade of the output tree.

The algorithm can be naively implemented in $O(hnm)$ time, where $h \leq \min\{m, n\}$ denotes the height of the reconstructed tree. Using a dynamic data structure for maintaining the connected components of $G(\mathcal{A})$, we achieve a time bound of $O(nm + |E_1| \log^2(n + m) + h(n + m))$.

Our algorithm is guaranteed to find the *most general* phylogenetic tree if one exists, i.e., if the clade set of any other phylogenetic tree includes all clades reported. We can extend our algorithm to check whether the input instance has a most general solution. The following condition has to be verified, for each tree node: if three sons of this node induce clades S', S'' and S''', then S' remains connected, even when one removes from $G(\mathcal{A})$ all characters whose 0-sets do not intersect $S' \cup S''$. The complexity of the extended algorithm is $O(nm + d|E_1| \log^2(n + m) + h(n + m))$, where d denotes the maximum number of sons of a node in T.

References

[1] Benham, C., Kannan, S., Paterson, M. and Warnow, T., Hen's teeth and whale's feet, *J. Comput. Biol.*, 2(4):515–525, 1995.

[2] Meecham, C.A. and Estabrook, G. F., Compatibility methods in systematics, *Annual Review of Ecology and Systematics*, 16:431–446, 1985.

[3] Nikaido, M., Rooney, A. P. and Okada, N., Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales, *Proceedings of the National Academy of Science USA*, 96:10261–10266, 1999.

[4] Swofford, D.L., *PAUP, Phylogenetic Analysis Using Parsimony (and Other Methods)*, Sinaur Associates, Sunderland, Massachusetts, 1998. Version 4.