

A Framework for Representing Reticulate Evolution*

Mihaela Baroni, Charles Semple, and Mike Steel

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
mbaroni@ugal.ro, {c.semple, m.steel}@math.canterbury.ac.nz

Received November 12, 2003

AMS Subject Classification: 05C05, 92D15

Abstract. Acyclic directed graphs (ADGs) are increasingly being viewed as more appropriate for representing certain evolutionary relationships, particularly in biology, than rooted trees. In this paper, we develop a framework for the analysis of these graphs which we call *hybrid phylogenies*. We are particularly interested in the problem whereby one is given a set of phylogenetic trees and wishes to determine a hybrid phylogeny that ‘embeds’ each of these trees and which requires the smallest number of hybridisation events. We show that this quantity can be greatly reduced if additional species are involved, and investigate other combinatorial aspects of this and related questions.

Keywords: directed acyclic graph, reticulate evolution, hybrid species, subtree prune and regraft

1. Introduction

Creating a ‘tree of life’ has been a primary goal of systematic biology since Charles Darwin’s first sketch of an evolutionary tree in 1837. It has become an accepted dogma that such a tree would describe how all present-day species had evolved from a common ancestor. However, accumulating data suggest that evolution is more complex than this, because many species are mosaics of genes derived from different ancestors. This pattern may be the result of processes such as hybridisation (the formation of a species that contains genetic contributions from more than one ancestral species), a process that is widely recognised in certain plant and fish species. Nearly 20 years ago, Funk [8] cautioned “it is difficult to overemphasise the importance of hybridization and polyploidy in evolution.” Other mechanisms, such as the horizontal transfer of genes between species may also be important sources of reticulate (non-tree like) evolution particularly for deep divergences in the tree of life. The situation here has been recently summarised by Doolittle [6] who wrote that “molecular phylogeneticists will have failed to find the ‘true tree’, not because their methods are inadequate or because

* The authors thank the New Zealand Marsden Fund (UOC-MIS-005) for supporting this research.

they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree.”

To model reticulate evolution it seems increasingly appropriate to represent the evolution of the species under study with a directed graph, where the vertices correspond to extant and ancestral species, while each arc represents the transfer of genetic material from one species to another—for example, by hybridisation or horizontal gene transfer. This gives rise to several interesting mathematical and computational problems. One question is how best to represent and reconstruct these digraphs. To date much of the analysis in the biological literature has been somewhat ad-hoc. For example, starting from a tree, one can introduce additional arcs in a heuristic fashion to see if there is an improvement in the ‘fit’ to data. Such an approach was described by Legendre and Makarenkov [11] for inferring a reticulation network (‘reticulogram’) from a given distance matrix, and applied to examples from biogeography, population microevolution, and hybridisation. Other aspects of the problem of representing hybridisation in biology are discussed in [4, 9, 10, 13, 15, 17–19].

Another strategy for describing reticulate evolution has been to apply existing mathematical procedures that generate graphs (rather than trees) to biological data. Lapointe [9], reviewed four such approaches. These methods—pyramids [5], weak hierarchies [2], splitsgraphs [7] and reticulograms—were applied to the same data set and the results are compared. However, it is not clear that such general techniques for constructing graphs, often developed for quite different processes, are precisely the right tool for representing hybrid evolution.

In this paper, we take an alternative approach, developing a digraph representation that reflects directly the biological questions we consider. We call these digraphs, subject to simple constraints, ‘hybrid phylogenies’. In Section 2, we formally describe these phylogenies and identify an important subclass—the ‘regular’ hybrid phylogenies (these are naturally isomorphic to the cover digraph of their associated cluster system). By restricting our attention to regular hybrid phylogenies, we avoid many pathologies that can arise in the infinite set of possible hybrid phylogenies on a given set of extant species. Indeed, Section 4 shows for application purposes no generality is lost in confining ourselves to regular hybrid phylogenies.

One of the themes throughout this paper is to use this formalism to study a fundamental question of interest for biologists: given a collection of trees on sets of species that faithfully represent the (tree-like) evolution of different parts of various species genomes, we would like to know how these trees can be ‘displayed’ by a single hybrid phylogeny. In particular it is of interest to determine the smallest number of hybrid events that are required for the trees to be simultaneously displayed by a single hybrid phylogeny. This number then sets a lower bound on the degree of hybridisation that has occurred in the evolution of the species under consideration. Proposition 4.2 shows that the restriction to regular hybrid phylogenies does not change this minimum number.

In order to study these concepts it is first necessary to formalise the notion of what it means for a hybrid phylogeny to ‘display’ a rooted phylogenetic tree. We do this in Section 3 and show that, for any given collection \mathcal{P} of rooted phylogenetic trees, there is a canonical (and regular) hybrid phylogeny that displays each of the trees in this collection. This particular hybrid for when \mathcal{P} consists of two trees is considered further in Section 6. In general, this canonical hybrid exhibits more hybrid events than