Advances in Applied Mathematics 84 (2017) 34–46

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

A note on convex characters, Fibonacci numbers and exponential-time algorithms

Steven Kelk [*], Georgios Stamoulis

Department of Data Science and Knowledge Engineering (DKE), Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 19 August 2016
Accepted 20 October 2016
Available online xxxx

MSC:
05C30
05C85
05A15

Keywords:
Fibonacci numbers
Phylogenetics
Trees
Convexity
Counting
Algorithms

ABSTRACT

Phylogenetic trees are used to model evolution: leaves are labelled to represent contemporary species ("taxa") and interior vertices represent extinct ancestors. Informally, convex characters are measurements on the contemporary species in which the subset of species (both contemporary and extinct) that share a given state, forms a connected subtree. Given an unrooted, binary phylogenetic tree $\mathcal{T}$ on a set of $n \geq 2$ taxa, a closed (but fairly opaque) expression for the number of convex characters on $\mathcal{T}$ has been known since 1992, and this is independent of the exact topology of $\mathcal{T}$. In this note we prove that this number is actually equal to the $(2n-1)$th Fibonacci number. Next, we define $g_k(\mathcal{T})$ to be the number of convex characters on $\mathcal{T}$ in which each state appears on at least k taxa. We show that, somewhat curiously, $g_2(\mathcal{T})$ is also independent of the topology of $\mathcal{T}$, and is equal to the $(n-1)$th Fibonacci number. As we demonstrate, this topological neutrality subsequently breaks down for $k \geq 3$. However, we show that for each fixed $k \geq 1$, $g_k(\mathcal{T})$ can be computed in $O(n)$ time and the set of characters thus counted can be efficiently listed and sampled. We use these insights to give a simple but effective exact algorithm for the NP-hard *maximum parsimony distance* problem that runs in time $\Theta(\phi^n \cdot n^2)$, where $\phi \approx 1.618...$ is the golden ratio, and an exact algorithm which computes the *tree bisection and reconnection*

[*] Corresponding author.
E-mail addresses: steven.kelk@maastrichtuniversity.nl (S. Kelk), georgios.stamoulis@maastrichtuniversity.nl (G. Stamoulis).

http://dx.doi.org/10.1016/j.aam.2016.10.003

distance (equivalently, a *maximum agreement forest*) in time $\Theta(\phi^{2n} \cdot \text{poly}(n))$, where $\phi^2 \approx 2.619$.

1. Introduction

Phylogenetics is the science of accurately and efficiently inferring evolutionary trees given only information about contemporary species [12]. An important concept within phylogenetics is *convexity*. Essentially this captures the situation when, within a phylogenetic (i.e. evolutionary) tree, each biological state emerges exactly once: it should not emerge, die out, and then re-emerge. More concretely, given a phylogenetic tree and a set of states assigned to its leaves, can we assign states to the internal vertices of the tree such that each state forms a connected "island" within the tree? If this is possible, the assignment of states to the leaves is known as a *convex character*.

In this article we present a number of results concerning the enumeration of convex characters. In Section 2 we give formal definitions and describe relevant earlier work. In Section 3 we start by showing that an earlier result counting convex characters can be simplified to a term of the Fibonacci sequence. We then seek to count convex characters with the added restriction that each state should occur on at least k leaves, proving the somewhat surprising result that (as for $k = 1$) tree topology is irrelevant for $k = 2$, and that a formulation in terms of Fibonacci numbers is again possible. We give an explicit example showing that for $k \geq 3$ the topological neutrality breaks down. In Section 4 we show that for all k the size of the space can be counted in polynomial time and space using dynamic programming, which also permits listing and sampling uniformly at random, noting also that non-isomorphic trees can have exactly the same vector of space sizes (for $k = 1, 2, ...$). In Section 5 we give a number of algorithmic applications for NP-hard problems arising in phylogenetics that seek to quantify the dissimilarity of two phylogenetic trees. Finally, in Section 6 we briefly discuss a number of open problems arising from this work. The software associated with this article has been made publicly available.

2. Preliminaries

For general background on mathematical phylogenetics we refer to [5,12]. An *unrooted binary phylogenetic X-tree* is an undirected tree $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$ where every internal vertex has degree 3 and whose leaves are bijectively labelled by a set X, where X is often called the set of *taxa* (representing the contemporary species). We use n to denote $|X|$ and often simply write *tree* when this is clear from the context.

A *character* f on X is a surjective function $f : X \rightarrow \mathcal{C}$ for some set $\mathcal{C}$ of *states* (where a state represents some characteristic of the species e.g. number of legs). We say that