Molecular Phylogenetics and Evolution 107 (2017) 209–220

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Convex recoloring as an evolutionary marker

Zeev Frenkel [a], Yosef Kiat [b], Ido Izhaki [a], Sagi Snir [a,*]

[a] Department of Ecology and Evolutionary Biology, University of Haifa, Israel
[b] Israeli Bird Ringing Center, Society for the Protection of Nature in Israel, Israel

ARTICLE INFO

Article history:
Received 21 May 2016
Revised 16 October 2016
Accepted 25 October 2016
Available online 3 November 2016

Keywords:
Phylogenetics
Maximum parsimony
Character compatibility
Perfect phylogeny
Statistical significance
Supertree
Optimal convex recoloring cost

ABSTRACT

With the availability of enormous quantities of genetic data it has become common to construct very accurate trees describing the evolutionary history of the species under study, as well as every single gene of these species. These trees allow us to examine the evolutionary compliance of given markers (characters). A marker compliant with the history of the species investigated, has undergone mutations along the species tree branches, such that every subtree of that tree exhibits a different state. Convex recoloring (CR) uses combinatorial representation to measure the adequacy of a taxonomic classifier to a given tree. Despite its biological origins, research on CR has been almost exclusively dedicated to mathematical properties of the problem, or variants of it with little, if any, relationship to taxonomy. In this work we return to the origins of CR. We put CR in a statistical framework and introduce and learn the notion of the statistical significance of a character. We apply this measure to two data sets - Passerine birds and prokaryotes, and four examples. These examples demonstrate various applications of CR, from evolutionary relatedness, through lateral evolution, to supertree construction. The above study was done with a new software that we provide, containing algorithmic improvement with a graphical output of a (optimally) recolored tree.

Availability: A code implementing the features and a README is available at http://research.haifa.ac.il/ssagi/software/convexrecoloring.zip.

1. Introduction

The practice of constructing a tree depicting the evolutionary history of a set of organisms is nowadays common to almost every phylogenomic study - an area combining genomic data and techniques for the study of evolution (Eisen and Fraser, 2003; Delsuc et al., 2005). In particular, the deluge of the molecular data accumulating constantly, allows us to gauge the accuracy of the constructed trees. A character, genetic or morphological, classifies the species set into several character classes. If we consider each class as a different color, then every species is colored by the state of the character it possesses, and the given character induces a coloring over the tree leaves. We say that the coloring is *convex* on the given tree if every color class induces a clade or a subtree and these subtrees do not overlap (Moran and Snir, 2008) (or equivalently, do not intersect). Convexity is a desirable and natural property in classification. When a character is convex on a tree, it is denoted as *homoplasy free* meaning it displays no *reversals* or *convergence* (Zhang and Kumar, 1997). The well-founded and widespread phylogenetic approach *maximum parsimony* (Fitch, 1971) seeks a tree with minimal changes on its edges, summed over all input characters. A minimum can be obtained when a *perfect phylogeny* exists in which case each input character is homoplasy-free on that phylogeny (Fernandez-Baca, 2001). Such a tree not necessarily exists, and even finding it is computationally intractable (Bodlaender et al., 1992). In the above setting, the characters are given and assumed to be reliable, and a plausible tree is sought. In other settings, the tree is also given, along with the characters, but one or more characters are not convex on that tree. In this case, we may question about the reliability of that tree.

Alternatively, in a setting where the tree provides enough confidence, the question shifts to the reliability of the input characters. Moreover, we may wonder if the character under examination has evolutionary traces or is influenced by other factors such as environment or simply randomness. In both cases, questioning the tree while assuming character reliability or questioning the character evolutionary meaningfulness, we look for the *recoloring distance* that counts the minimum number of tree nodes we need to recolor in order to arrive at convexity. This value indicates the level of disagreement between the tree and the coloring. The notion of the recoloring distance was coined in Moran and Snir (2008) where the problem, *convex recoloring* (CR), was defined and studied for several types of trees and input colorings. Despite its biological

* Corresponding author.

http://dx.doi.org/10.1016/j.ympev.2016.10.018
1055-7903/

origin, due to its mathematical cleanliness, mainly combinatorial/algorithmic aspects of the problem and its derivatives, that have little if at all biological relevance, were studied. These include extensions to certain graph types rather than a tree, specific input colorings, constrained recoloring schemes, and alike (see e.g. Kanj and Kratsch, 2009; Kammer and Tholey, 2012; Campêlo et al., 2013 and references therein, but see also Matsen, 2015 for a classification oriented study).

In this work we bring back the high level theory of CR down to the biological ground in several aspects. For a taxonomist, it would be desirable to determine quantitatively and statistically, the relevance of a character (i.e. any classification) to the tree at hand. The recoloring distance is an absolute, context-less, number. We therefore introduce the notion of a *coloring significance*, indicating how likely we are to see, a coloring of this distance or less, by chance on the given tree. In the Results section we demonstrate the use of the coloring significance measure by applying CR to several examples. First, in order to obtain an intuition regarding this measure we show a simulation study. The results reveal that the recoloring distance is more structured than expected. Next, using two data sets, we demonstrate the various uses of CR as an evolutionary marker. The first data set is over eighty Passerine birds, and the second is over a hundred prokaryotes, with few colorings (characters) for each data set. The results obtained concern not only questions of phylogeny/character reliability, but also intensity of non tree-like activity in prokaryotes and the power of supertree methods.

Importantly, we provide a software that implements the features we describe in this work. To this respect, in the Method section we describe an algorithmic improvement to the algorithm presented in Moran and Snir (2008). The improvement is achieved by reducing the average number of colors checked at a node. We do not give an asymptotic analysis for this improvement but do provide rigorous proof for its correctness. We are aware that since the appearance of the algorithm of Moran and Snir (2008), there have been further improvements (e.g. Bar-Yehuda et al., 2008) to that first algorithm, and there might be other algorithms with better complexity than the one presented here. However a basic property of this algorithm, which to the best of our knowledge was not used before, is a *local view* that allows a dynamic calculation of the set of *candidate colors* of each tree node. Accordingly, we believe that the algorithmic improvements provided here, accompanied with more fundamental theoretical improvements to CR, viewing it as a fixed parameter tractable problem (Bodlaender et al., 2011), will allow application of CR to data sets of orders of thousands of species and hundreds of colors.

2. Results

We now show four examples for the application of convex recoloring to synthetic and real data. The first one is a simple example based on random colorings of a binary tree, demonstrating the distribution of optimal convex recoloring cost in one simple case. The other three are applications to real biological examples of colored trees where the colorings represent a different classification each time. In each case we compute the optimal recoloring and its associated p-value, signifying how much the given coloring complies with the evolutionary history of the given species set (that is also given as input, and is represented by the tree topology).

2.1. Example 1: Statistical distribution of the recoloring distance

Our first example shows how the recoloring distance distributes for a given tree size and number of colors. We constructed a set of random binary trees with 50 leaves. Next, we randomly and uniformly colored the tree leaves by 4 colors (no uncolored leafs, all internal nodes are uncolored). This is simply done by choosing for every leaf each color with probability 1/4. Therefore, the trees obtained are different in topology and also by the proportions between color sets. For each of these trees a convex recoloring was calculated. The distribution of cost of recoloring is presented in Fig. 1(a). We note that a naive upper bound to the expected value of this statistic, is the value of $3n/4$ where n is the number of leaves. This is achieved by recoloring all the leaves with the most common color. As this must have at least $n/4$, the bound is trivially obtained. However, as we see in the figure, a much smaller value (from $n/2$ to $3n/5$) is usually obtained, signifying existence of a more profound structure in this question than that naive bound. Notwithstanding, a more precise bound is not trivial to obtain and is beyond the scope of this work. Distribution of colors frequencies on the resulted convex trees is presented in Fig. 1(b). The results are divided into three cases (three bar charts in the figure) representing cases in which the most common color had (i) below 25 members (Blue bars), (ii) between 25 and 28 members (Brown bars), and (iii) above 28 members (Green bars). As shown, this difference in the prevalence of the most common color, affects minimally over the distribution of the final colors, where the most common color colors around 70% of the leaves. We note that as there are many (possibly even exponentially many) optimal recolorings, this distribution might be biased according to the strategy employed by the algorithm. One may observe that in a tree, every color is preserved at least by a single leaf as this does not violate convexity of the tree. This observation is explained by the three short bars in the right of Fig. 1(b).

2.2. Example 2: Birds moult strategies

In this example compatibility of adult/juvenile moult strategy of birds with their evolutionary history was examined. We took a tree over 80 bird taxa representing 29 of the 46 Passerine families (Treplin et al., 2008). The leaves of this phylogeny were classified by their main moult strategies in adult/juvenile life stages as described in Jenni and Winkler (1994), Cramp et al. (1993), and Ginn and Melville (1983). Such characterization was made only for 43 of these genus and species and was expressed by one, two or even three of three observed moult strategy types: "Summer complete/summer partial", "Summer complete/summer complete", and "Winter complete/winter complete". Such characterization induces the following coloring of phylogenetic tree's leafs: leafs corresponding to non-characterized species and species characterized by more than one strategy type - uncolored; leafs corresponding to species characterized by only one moult strategy type are colored by Blue, Red and Green (26, 7 and 4 leafs respectively). Based on our program we found that this coloring is not convex: $P_{opt} = 8$, p-value = 0.26 (see Fig. 2). Excluding the green color results in non-convex coloring with $P_{opt} = 5$, p-value = 0.46. Unifying colors Red and Blue (in the initial coloring) results in $P_{opt} = 3$, p-value = 1.0. The latter means the following. After unifications of Red and Blue, we are left with two colors - Red/Blue and Green - where the Green comprises of only 4 members, that are dispersed. A cost of 3 means that in order to arrive at convexity we must uncolor all but one of the Green leaves. As shown in previous section (Section 2.1), any tree recoloring retains at least one leaf of any color class intact. The latter implies that this is not only the minimum cost possible, rather also the maximum cost for the given configuration of 4 Green leaves. Moreover, since any other, random or not, input coloring with 4 Green leaves cannot achieve a cost higher than that (i.e. a cost greater than 3), all colorings attain this (3) or smaller cost, explaining the p-value of 1 of that result. The