



Primate molecular phylogenetics in a genomic era

Nelson Ting*, Kirstin N. Sterner

Department of Anthropology, 308 Condon Hall, 1218 University of Oregon Eugene, OR 97403, USA
Institute of Ecology and Evolution, 335 Pacific Hall, 5289 University of Oregon Eugene, OR 97403, USA

ARTICLE INFO

Article history:

Available online 30 August 2012

Keywords:

Molecular phylogenetics
Phylogenomics
Multispecies coalescent
Coalescent theory
Morris Goodman
Primate systematics

ABSTRACT

A primary objective of molecular phylogenetics is to use molecular data to elucidate the evolutionary history of living organisms. Dr. Morris Goodman founded the journal *Molecular Phylogenetics and Evolution* as a forum where scientists could further our knowledge about the tree of life, and he recognized that the inference of species trees is a first and fundamental step to addressing many important evolutionary questions. In particular, Dr. Goodman was interested in obtaining a complete picture of the primate species tree in order to provide an evolutionary context for the study of human adaptations. A number of recent studies use multi-locus datasets to infer well-resolved and well-supported primate phylogenetic trees using consensus approaches (e.g., supermatrices). It is therefore tempting to assume that we have a complete picture of the primate tree, especially above the species level. However, recent theoretical and empirical work in the field of molecular phylogenetics demonstrates that consensus methods might provide a false sense of support at certain nodes. In this brief review we discuss the current state of primate molecular phylogenetics and highlight the importance of exploring the use of coalescent-based analyses that have the potential to better utilize information contained in multi-locus data.

Published by Elsevier Inc.

1. Introduction

Fifty years ago, Dr. Morris Goodman published a series of papers investigating the molecular systematics of the living primates (Goodman, 1961, 1962a,b, 1963a,b). These papers were among the very first studies in the field of molecular phylogenetics. Using serological data and protein electrophoresis, he inferred the evolutionary branching order of the lemurs, lorises, galagos, New World monkeys, Old World monkeys, and apes, as well as relationships among certain species within those groups. Furthermore, he proposed a clade-based primate classification that placed humans and African apes in the same family (Hominidae) to the exclusion of orangutans (Pongidae) – a controversial arrangement that met strong resistance at the time [note: this is *prior* to the publication of Willie Hennig's *Phylogenetic Systematics* (1966) in English]. Dr. Goodman's primary motivation for his primate molecular systematic work was to gain a better understanding of our place in the natural world. In honor of his many contributions we thought now would be an appropriate time, on the 50th anniversary of his seminal papers, to assess the field of primate molecular phylogenetics in the age of genomics – in particular the use of consensus vs. coalescent-based approaches when analyzing large multi-locus datasets.

2. Analyzing multi-locus data using a single locus framework

One objective of molecular phylogenetics is to provide comprehensive and well-supported phylogenies that reflect species relationships. While this objective has remained constant throughout the last 50 years, the ability to interrogate different types of molecular data and the methodologies used to analyze these data has changed considerably. Earlier studies of molecular evolution inferred species relationships utilizing immunological and hybridization techniques. Although these studies made exciting contributions, particularly with regard to our close genetic relationship to chimpanzees and gorillas (e.g., Goodman, 1961, 1962a,b, 1963a,b), they offered only crude measures of genetic relatedness. Aided by technical advances in molecular biology, later studies were able to overcome this limitation by directly sequencing polypeptide chains and DNA, and from these data infer rates of sequence evolution, gene trees, and divergence times (e.g., as discussed in Goodman, 1996). The traditional molecular phylogenetic framework employed by such studies infers a gene tree and uses it as a hypothesis for the species tree, and infers divergence times across the gene tree through the use of a fossil calibration point and assumption of a molecular clock.

More recently the field of molecular phylogenetics has shifted from single locus studies to multi-locus studies that capture large segments of the nuclear genome. In order to apply the traditional molecular phylogenetic approach to large multi-locus datasets, multiple regions of the genome are concatenated into a supermatrix for analysis (reviewed in de Queiroz and Gatesy, 2007), or

* Corresponding author at: 308 Condon Hall, Department of Anthropology, 335 Pacific Hall, 5289 University of Oregon Eugene, OR 97403, USA. Fax: +1 541 346 0668.

E-mail address: nting@uoregon.edu (N. Ting).

multiple loci are analyzed separately and the resulting trees are combined to form a consensus tree (reviewed in Bininda-Emonds, 2004). While it has been known for some time that gene trees can be incongruent with one another and the species tree (e.g., see, Goodman et al., 1979), these “supermatrix” and “supertree” methods (we will from here on refer to these as “consensus methods”) assume that capturing the most common phylogenetic signal or tree will produce a gene tree that accurately represents species relationships. Within this framework, gene tree incongruence is viewed as an obstacle in the search for true species relationships.

The consensus method approach has been used extensively in recent studies to infer primate phylogenetic relationships from multi-locus data (including Horvath et al., 2008; Wildman et al., 2009; Jameson et al., 2011; Perelman et al., 2011; Wang et al., 2012). For example, Perelman and colleagues (2011) included as many as 191 primate species across 54 genes and 34,941 orthologous base pairs in their study and Jameson and colleagues (2011) sampled eight primate species across more than 1000 genes and one million orthologous base pairs. As a result, we now have a near fully resolved, well-supported, and taxonomically complete genus-level phylogenetic tree for the living primates (Perelman et al., 2011). But is this phylogeny an accurate representation of primate species relationships?

Over the past few years, it has become apparent that consensus methods have theoretical limitations and drawbacks when it comes to the analysis of large multi-locus datasets (Degnan and Rosenberg, 2009; Edwards, 2009). For example, obtaining a false phylogenetic signal through the use of supermatrix methods can occur if substantial amounts of sequence in the dataset are from linked parts of the genome and/or have faster rates of evolution than other parts, thus swamping out potential true phylogenetic signal found at other loci. Results from the supermatrix method are also highly influenced by the selection of alleles in the concatenation process, which is often not taken into account (Weisrock et al., 2012). Furthermore, results from consensus approaches can be confounded by gene tree incongruence due to past hybridization and can fail to uncover such subtleties in the evolutionary history of species. Lastly, consensus methods may lead to very well-supported and well-resolved relationships that are not concordant with the species tree. This can happen at nodes with short branch lengths and/or large ancestral effective population sizes because such nodes are expected to contain gene trees that are frequently discordant from the species tree due to incomplete lineage sorting (reviewed in Degnan and Rosenberg (2009) and Edwards (2009)). In such circumstances, the addition of sequence data in combination with the use of consensus methods could provide an increasingly false sense of confidence in the inferred relationships.

Consensus methods for inferring divergence times across a species tree face similar limitations. Although numerous improvements in inferring divergence times have been developed over the years, including the use of multiple fossil calibration points, allowing for error around calibration points, and incorporating rate heterogeneity across a tree, there remain inconsistencies between the results of primate divergence date studies that are not fully understood. For example, we do not know what characteristics make a sequence dataset ideal for divergence date analyses. Although it stands to reason that longer sequences will produce more accurate dates because they allow for better estimates of mutation rates, obtaining such sequences (i.e., tens of thousands of base pairs) requires concatenating many genomic regions that may not have the same history and/or mutation rates. The implications of this “averaging” approach are not entirely clear as these methods are meant to infer divergence times within a single gene tree. Furthermore consensus methods do not inform us on the timing of actual speciation events because gene tree divergence al-

ways predates species tree divergence by a time period that depends on ancestral effective population size, which consensus methods do not take into account.

3. Emergence of a new paradigm in molecular phylogenetics

The theoretical disadvantages of using consensus methods to analyze large multi-locus datasets are driven by a general lack of consideration for the population processes that underlie species relationships. Specifically, by treating the gene tree incongruence caused by various population processes as an artifact to be ignored rather than an important component of species histories, both supermatrix and supertree methods fail to utilize large genomic datasets to their full potential. Continued research into coalescent theory has led to the development of a framework that better realizes this potential for inferring species histories.

Coalescent theory refers to a retrospective model in population genetics that traces all alleles of a gene shared by individuals in a population to a single ancestral copy. The resulting tree-shaped genealogy shows ancestor–descendant relationships among alleles back in time and is known as “the coalescent.” Most commonly attributed to Kingman, coalescent theory was originally described as a set of probability models that demonstrated the time it would take for this process to occur in an idealized population based on genetic drift and population size (Kingman, 1982). After noting the difference between gene trees and species trees in early molecular phylogenetic studies (e.g., Goodman et al., 1979), molecular evolutionary biologists realized that this discrepancy can occur because of the coalescent process (Gillespie and Langley, 1979; Tajima, 1983; Hudson, 1983; Pamilo and Nei, 1988; see also references in Felsenstein, 2004). Through this work, some researchers started to view species trees as statistical distributions of genealogical coalescents, or a “cloud of gene histories” (Wu, 1991; Maddison, 1997; see also references in Felsenstein, 2004). It was also realized that, because coalescent theory is grounded in probability, direct inference of the species tree could be attained by calculating the likelihood that a set of gene tree topologies might exist within a given species tree topology given branch lengths and ancestral population sizes. This theoretical framework is now commonly referred to as the “multi-species coalescent” (Degnan and Rosenberg, 2009). With the collection of multi-locus datasets from non-model organisms becoming more common and an increase in access to computational power, there has been a recent proliferation of programs that use this framework for phylogenetic analysis, each of which implements the theory in different ways (see Liu et al., 2009 and Knowles, 2009 for reviews of some of these programs). In addition to directly inferring the species tree, many of these programs are also capable of inferring population divergence times (as opposed to gene tree divergence times) and estimating ancestral effective population sizes (e.g., BEST and *BEAST; Liu, 2008 and Heled and Drummond, 2010, respectively). Therefore, in theory these approaches hold a substantial advantage to supermatrix and supertree methods.

3.1. Performance of species tree inference within a coalescent framework

Despite growing support for coalescent-based approaches for species tree inference, their implementation has brought mixed results. In comparison to the supermatrix approach, simulation-based studies have shown that the coalescent approach is more accurate in inferring the species tree when there is a high degree of gene tree incongruence (Edwards et al., 2007; Kubatko and Degnan, 2007). However, empirical studies have shown that coalescent approaches often contain lower levels of support and/or resolution than consensus approaches (e.g., Leache, 2010; Townsend et al., 2011;

Download English Version:

<https://daneshyari.com/en/article/2834085>

Download Persian Version:

<https://daneshyari.com/article/2834085>

[Daneshyari.com](https://daneshyari.com)