Molecular Phylogenetics and Evolution 67 (2013) 494–508

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

Searching for the optimal data partitioning strategy in mitochondrial phylogenomics: A phylogeny of Acridoidea (Insecta: Orthoptera: Caelifera) as a case study

James R. Leavitt [a], Kevin D. Hiatt [a], Michael F. Whiting [a], Hojun Song [a,b,*]

[a] Department of Biology and M.L. Bean Museum, Brigham Young University, Provo, UT 84602, USA
[b] Department of Biology, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Article history:
Received 1 May 2012
Revised 10 February 2013
Accepted 15 February 2013
Available online 27 February 2013

Keywords:
Phylogeny
Mitochondria
Phylogenomics
Data partition
Orthoptera
Caelifera

ABSTRACT

One of the main challenges in analyzing multi-locus phylogenomic data is to find an optimal data partitioning strategy to account for variable evolutionary histories of different loci for any given dataset. Although a number of studies have addressed the issue of data partitioning in a Bayesian phylogenetic framework, such studies in a maximum likelihood framework are comparatively lacking. Furthermore, a rigorous statistical exploration of possible data partitioning schemes has not been applied to mitochondrial genome (mtgenome) data, which provide a complex, but manageable platform for addressing various challenges in analyzing phylogenomic data. In this study, we investigate the issue of data partitioning in the maximum likelihood framework in the context of the mitochondrial phylogenomics of an orthopteran superfamily Acridoidea (Orthoptera: Caelifera). The present study analyzes 34 terminals representing all 8 superfamilies within Caelifera, which includes newly sequenced partial or complete mtgenomes for 11 families. Using a new partition-selection method implemented in the software PartitionFinder, we compare a large number of data partitioning schemes in an attempt to identify the most effective method of analyzing the mtgenome data. We find that the best-fit partitioning scheme selected by PartitionFinder is superior to any *a priori* schemes commonly utilized in mitochondrial phylogenomics. We also show that over-partitioning is often detrimental to phylogenetic reconstruction. A comparative analysis of mtgenome structures finds that the tRNA gene rearrangement between cytochrome *c* oxidase subunit II and ATP synthase protein 8 does not occur in the most basal caeliferan lineage Tridactyloidea, suggesting that this gene rearrangement must have evolved at least in the common ancestor of Tetrigoidea and Acridomorpha. We find that mtgenome data contain sufficient phylogenetic information to broadly resolve the relationships across Acridomorpha and Acridoidea.

1. Introduction

As sequencing technologies advance at a remarkably rapid rate, the gap between the ability to generate data and the ability to analyze the data in a phylogenetic framework is also increasing rapidly (Delsuc et al., 2005; Philippe et al., 2005, 2004). Mitochondrial phylogenomics, the use of complete mitochondrial genomes (mtgenomes) in phylogenetics, provides a manageable platform for addressing various challenges in analyzing phylogenomic data (Song et al., 2010). A typical insect mtgenome contains 37 genes in a coding region: 13 protein-coding genes (PCGs), two ribosomal RNA genes (16S and 12S) and 22 transfer RNA genes (Wolstenhome, 1992). A number of recent studies have explored various ways of effectively analyzing mtgenome data for inferring phylogenetic relationships among insects (Cameron et al., 2006a, 2006b, 2007, 2009; Castro and Dowton, 2007; Dowton et al., 2009a; Fenn et al., 2008; Pons et al., 2010; Sheffield et al., 2009; Song et al., 2010). Specifically, the issues of gene exclusion (Cameron et al., 2006b; Nardi et al., 2003), data recoding (Cameron et al., 2007; Fenn et al., 2008), data partitioning (Fenn et al., 2008), base compositional bias (Sheffield et al., 2009; Song et al., 2010), and rate heterogeneity (Pons et al., 2010; Song et al., 2010) have been addressed in the context of insect mitochondrial phylogenomics.

Cameron et al. (2004) was the first to suggest that the inclusion of all available data from mtgenomes improved the resolution and support and Cameron et al. (2007) found that biologically realistic data partitioning would produce the best results. Fenn et al. (2008) echoed the same sentiment but warned that over-partitioning of the mtgenome might be more detrimental to the analysis than beneficial. However, there has not been a clear recommendation on how to decide which partitioning method is the best for

* Corresponding author. Address: Department of Biology, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816-2368, USA. Fax: +1 407 823 5769.
E-mail address: song@ucf.edu (H. Song).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.02.019

mitochondrial phylogenomics in general although a number of studies have tackled this very issue in multi-locus analyses of nuclear genes (Brandley et al., 2005; Brown and Lemmon, 2007; Li et al., 2008; McGuire et al., 2007). In deciding the optimal partitioning strategy, one typically compares several datasets differing in partitioning schemes using methods such as Bayes factor (Brandley et al., 2005; Brown and Lemmon, 2007; Nylander et al., 2004), cluster analysis (Li et al., 2008), or the Akaike Information Criterion (AIC) (McGuire et al., 2007). However, a general consensus appears to be that the optimal partitioning strategy depends on both taxon and character sampling and it is difficult to justify a certain partitioning strategy for a particular dataset *a priori*. Recently, Lanfear et al. (2012) developed a method that could overcome this issue by statistically comparing across numerous partitioning schemes for any given sequence data and selecting the best-fit scheme, which was implemented in the program PartitionFinder. They found that typical *a priori* partitioning schemes, such as partitioning by gene or codon position, were often not the most appropriate way of partitioning data across various empirical datasets that they examined (Lanfear et al., 2012). According to Lanfear et al. (2012), the number of possible partitioning schemes follows a relationship known as a Bell number (Bell, 1934). For mitochondrial genome data, this number would be $B_{37} = 5.28 \times 10^{31}$ if 37 genes are treated as individual partitions, and could increase to $B_{63} = 8.25 \times 10^{63}$ if protein-coding genes are further partitioned by codon position. To our knowledge, this new method has not been fully examined in mitochondrial phylogenomics.

Acridoidea (grasshoppers and locusts) is the largest superfamily in the insect order Orthoptera with over 7600 described species and eleven families: Acrididae, Charilaidae, Dericorythidae, Lathiceridae, Lentulidae, Lithidiidae, Ommexechidae, Pamphagidae, Pyrgacrididae, Romaleidae, and Tristiridae. Acridoidea is one of six currently recognized superfamilies that are grasshopper-like in morphology and have been traditionally grouped together and collectively referred to as the Acridomorpha (Dirsh, 1975; Song, 2010). Recent molecular studies (Fenn et al., 2008; Flook et al., 1999, 2000; Flook and Rowell, 1997; Matt et al., 2008) have consistently found this group to be monophyletic although the phylogenetic relationships among the superfamilies remain contentious (Eades, 2000; Flook and Rowell, 1997; Song, 2010). Acridoidea has been considered a monophyletic group based on the morphology of male genitalia and the lack of a basioccipital slit, among other characters (Amédégnato, 1974; Eades, 2000; Kevan, 1982; Roberts, 1941), but the phylogenetic relationships among families within this clade are not well understood (Song, 2010). This is mainly because the higher-level classification of Acridoidea (and Acridomorpha) has been largely influenced by the interpretation of internal male phallic structures, which has not been consistent among different taxonomists (Amédégnato, 1976; Dirsh, 1956; Eades, 2000; Roberts, 1941). Even the latest synthesis on the phylogeny of Acridomorpha (Eades, 2000) depicted the internal relationships of Acridoidea to be largely unresolved. Therefore, this study represents a unique opportunity to reassess the phylogeny of Acridoidea based on a different character system.

The migratory locust, *Locusta migratoria*, was the first hemimetabolous insect to have its mtgenome completely sequenced (Flook et al., 1995a), and the follow-up study (Flook et al., 1995b) suggested that grasshoppers exhibit a unique gene rearrangement within the mtgenome. In grasshoppers, the order of tRNA genes between cytochrome *c* oxidase subunit II (COII) and ATP synthase protein 8 (ATP8) is reversed from the ancestral insect arrangement, inferred from *Drosophila yakuba* (Clary and Wolstenhome, 1985), which exhibits tRNA-Lys preceding tRNA-Asp on the J strand. Flook et al. (1995b) hypothesized that this rearrangement was a unique condition for Caelifera based on an investigation of 5 orthopteran species, and Fenn et al. (2008) supported this hypothesis based on an analysis of 8 orthopteran mtgenomes. Sheffield et al. (2010) recently suggested that this gene rearrangement is a synapomorphy for a subgroup within the orthopteran suborder Caelifera instead of the suborder as a whole. There has been a rapid increase in the number of complete mtgenomes of Acridomorpha available on GenBank (Ding et al., 2007; Erler et al., 2010; Fenn et al., 2008; Liu and Huang, 2008; Ma et al., 2009; Sheffield et al., 2010; Zhang and Huang, 2008), as well as from our ongoing study of orthopteran mtgenomes. Taken all together, grasshoppers represent an excellent system for addressing the various challenges of analyzing mtgenome data in a phylogenetic framework, as well as for understanding the evolution of mtgenomes.

In this study we explore various partitioning schemes using PartitionFinder in the context of the mtgenome phylogeny of the grasshopper superfamily Acridoidea (Orthoptera: Caelifera) in a maximum likelihood framework. We specifically address the following questions: (i) What is the optimal data partitioning strategy in analyzing mtgenome data in a maximum likelihood framework for Acridoidea?; (ii) What are the patterns of mtgenome evolution in Acridoidea in terms of gene rearrangement, anti-codon conservation, and start and stop codon usage?; and (iii) What are the higher-level phylogenetic relationships within Acridoidea and Acridomorpha and how do these compare with previous hypotheses?

2. Materials and methods

2.1. Taxon sampling

We included nine of the 11 extant families within Acridoidea (Table 1). The two omitted families, Dericorythidae and Lathiceridae, which are found in the Middle East and the Namibian desert, respectively, are rare and we were not able to obtain DNA-grade tissues. For Acrididae, which is the largest family within Acridoidea, we included 16 terminals representing 8 subfamilies. For the remaining 8 families, we included one representative per family, except Pamphagidae, for which we included two taxa. Additionally, we included at least one representative for each of the acridomorph superfamilies (Eumastacoidea, Tanaoceroidea, Pneumoroidea, Trigonopterygoidea, and Pyrgomorphoidea) in order to determine the phylogenetic placement of Acridoidea as well as to test the monophyly of Acridoidea within Acridomorpha. For outgroups, we included one member of Tridactyloidea and Tetrigoidea, both of which have been traditionally considered basal lineages within Caelifera. Altogether, our analysis included 34 terminals representing all 8 superfamilies within Caelifera.

For this study, we generated seven complete and four partial mtgenomes representing six superfamilies, 11 families. Tissue samples used for data generation were either collected by the authors or provided by collaborators. The specimens used in this analysis were preserved in 100% ethanol and vouchered to the −80 °C cryofacility in the Insect Genomic Collection at Brigham Young University (BYU-IGC). The voucher information is presented in Supplementary Table S1. The remaining data were obtained from GenBank.

2.2. Character sampling

DNA was extracted from tissues from hind femur or thorax using the Qiagen DNeasy Tissue Extraction Kit (Valencia, California, USA) following the animal tissue protocol. The resulting DNA extracts were vouchered at the BYU-IGC. Mtgenome data were generated using a primer-walking technique. Initial long-range polymerase chain reactions (PCRs) were performed using both universal mtgenome primers (Simon et al., 2006) as well as Orthoptera specific primers, which were the same as those used by