Molecular Phylogenetics and Evolution 55 (2010) 1177–1182

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Short Communication

Identification of new molecular markers for assembling the eukaryotic tree of life

Yonas I. Tekle[a,*], Jessica R. Grant[a], Alexandra M. Kovner[a], Jeffrey P. Townsend[b], Laura A. Katz[a]

[a] Department of Biological Sciences, Smith College, Northampton, MA, USA
[b] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

Article history:
Received 29 July 2009
Revised 8 February 2010
Accepted 9 March 2010
Available online 17 March 2010

Keywords:
Phylogeny
New molecular markers
Eukaryotes
KOG
Tree of life
Phylogenetic informativeness

ABSTRACT

Six eukaryotic supergroups have been proposed based on both morphological and molecular data. However, some of these supergroups are contentious and the deep relationships among them are poorly resolved. This is due to a limited number of morphological characters and few molecular markers in current use. The lack of resolution in most multigene analyses, including phylogenomic analyses, necessitates a search for additional, appropriate molecular markers to enable targeted sampling of taxa in key phylogenetic positions. We evaluated the phylogenetic signal of 860 proteins obtained from the Clusters of Orthologous Groups of proteins (COGs) database. We report a total of 17 markers that resulted in well-resolved topologies that are congruent with well-established components of the eukaryotic tree. To establish their utility, we designed universal degenerate primers for six markers, some of which showed promising results in unicellular eukaryotes. Finally, we present phylogenetic informativeness profiles for seven selected markers, revealing that the markers contain phylogenetic signal that spans the whole tree including the deeper branches.

Published by Elsevier Inc.

1. Introduction

Assembling the eukaryotic tree of life is one of the most formidable tasks in evolutionary biology. This difficulty is mainly attributable to the tremendous diversity among eukaryotes coupled with the paucity of comparable morphological characters. Eukaryotes include the relatively well-studied macroscopic lineages such as plants, animals, and fungi, plus a vast diversity of microscopic lineages, many of which have yet to be sampled for molecular characters. Based on cellular complexity, more than 70 eukaryotic lineages can be identified. Additionally, there are more than 100 taxa of uncertain taxonomic affinity (Patterson, 1999). Our understanding of how many of these lineages are monophyletic and how the rest are interrelated is based primarily on morphology, and is in many cases tentative at best (Taylor, 1994, 1999; Patterson, 1999). More recently, molecular systematics has resolved some branches on the eukaryotic tree of life, frequently confirming those clades that are defined by clear ultrastructural identities (e.g. Gajadhar et al., 1991; Yoon et al., 2008). Other molecular systematics studies are beginning to decipher the placement of highly controversial taxa (Tekle et al., 2007; Yoon et al., 2008), and have resulted in the classification of the diverse microbial and macroscopic eukaryotes into six putative supergroups (e.g. Adl et al., 2005). However, the support for most of the supergroups is questionable (Parfrey et al., 2006), and interrelationships among them are poorly resolved (e.g. Baldauf et al., 2000; Yoon et al., 2008).

A few universal molecular markers that have been successfully employed in elucidating relationships within eukaryotes include rRNAs and nuclear protein-coding genes (e.g. EF-1 alpha, alpha-tubulin, beta-tubulin, Actin, RPB1, HSP 70, HSP90, RNA polymerase, myosin), as well as a number of conserved genes encoded in the chloroplast and mitochondrial genomes. However, many of the nuclear protein-coding markers have been problematic. For instance, lateral gene transfer (LGT) has clouded inferences from EF-1 alpha (Keeling and Inagaki, 2004) and alpha-tubulin (Simpson et al., 2008), and other genes have been plagued by heterotachy among eukaryotes (Philippe et al., 2005). Furthermore, markers that reside in organelles such as plastids and mitochondria are problematic for universal use because not all eukaryotes contain these organelles (e.g. Tekle et al., 2009). Despite continual advances in computational algorithms that accommodate diverse molecular evolutionary processes, none of the current markers, individually or in combination, have been able to provide robust solutions for deep-level relationships among eukaryotes. Recent multigene analyses of eukaryotes with increased taxonomic sampling demonstrated that the commonly well-sampled markers (ribosomal genes, cytoskeletal actin, alpha- and beta-tubulins) have limited power in resolving deep nodes of the eukaryotic tree of life (Baldauf et al., 2000; Yoon et al., 2008; Hampl et al., 2009). Our

[*] Correspondence to: Yonas I. Tekle, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA. Fax: +1 203 785 4782.
E-mail address: yonastekle@yale.edu (Y.I. Tekle).

1055-7903/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.ympev.2010.03.010

recent work also suggests the need for identification of additional molecular markers, in combination with increased taxonomic sampling, to resolve deep relationships within eukaryotes (Yoon et al., 2008).

Advances in molecular techniques have led to the proliferation of phylogenomic data such as Expressed Sequence Tags (EST) or completed genome projects. Phylogenomic studies incorporating large genomic data sets are providing some insights towards deciphering interrelationships of eukaryotes (Bapteste et al., 2002; Burki et al., 2007; Hackett et al., 2007; Hampl et al., 2009). However, these studies suffer from limited taxonomic sampling due to the high cost associated with generating such large datasets and produce matrices with large quantities of missing data due to their untargeted gene sampling. Hence, searching for additional appropriate markers for directed PCR-based studies is timely and will facilitate a cost effective way to capture the taxonomic breadth of eukaryotic diversity and surmount inferential challenges that arise due to limited sampling (e.g. long-branch attraction; Felsenstein, 1978). The utility of such a directed approach was demonstrated by a recent phylogenomic study of yeast (fungi), in which 20 appropriate markers were shown to provide comparable resolution and support to that achieved by phylogenomic data (106 markers; Rokas et al., 2003).

The search for new, universal molecular markers suitable for inferring ancient relationships is a challenging task. An ideal molecular marker for eukaryotes should be present among all lineages, not prone to LGT, conserved yet with desirable variability, and fairly large in size (e.g. Woese, 1987). Identification of new markers has recently been aided by numerous genome projects that are producing a plethora of sequence data. In this study, we report on the identification of a number of nuclear protein-coding genes that can serve as molecular markers to elucidate relationships among diverse eukaryotes. To identify these loci, we searched within the Clusters of Orthologous Groups for eukaryotic complete genomes (KOG) database applying NCBI data mining tools to find orthologous proteins (http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi?phy=127). KOG is a subset of Clusters of Orthologous Groups of proteins (COGs) (Tatusov et al., 1997) that includes proteins from seven eukaryotic genomes (*Saccharomyces cerevisiae, Schizosaccharomyces pombe, Encephalitozoon cuniculi*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*) to the exclusion of prokaryotes. The COG/KOG database is an ideal source for identifying molecular markers for phylogenetic purposes, because the foundation of the database is based on evolutionary classification of orthologous proteins in completed genomes (Tatusov et al., 2003). In addition to KOG, we used other databases (ST 3, Supplementary material online) of completed or nearly completed genomes of eukaryotes to search for additional homologous proteins. We also predict the selected markers' utility by profiling their phylogenetic informativeness (Townsend, 2007).

2. Methods

We assessed 860 proteins listed on the KOG database (as of January, 2007) for their phylogenetic content following three steps: identifying markers, inspecting alignment, and evaluating phylogenetic signal. First, we identified single copy markers or markers with few paralogs. This identification was based on the information provided on the KOG page for each protein. Subsequently, we examined the Neighbor-Joining (NJ) tree provided for each KOG member for preliminary assessment of topology and deep paralogy. This evaluation allowed us to exclude those proteins represented by numerous divergent paralogs.

In a second step, we searched for homologous proteins in the Entrez/NCBI using the BLINK search option to blast each selected KOG member from one of the seven species (see above). The BLINK result provided useful information to further assess the suitability of any KOG member such as: the presence of protein in other non-eukaryotic domains (Bacteria and Archaea); the representation/availability of data for other eukaryotes; and a preliminarily alignment to assess conserved alignable length of a protein and regions for primer design. Proteins that were not represented in other eukaryotic lineages or that showed limited signal (judged by the resolution of NJ tree) were excluded from our list. Additionally, ubiquitous KOG members that were present and conserved in bacteria and archaea were excluded due to risk of contamination (upon DNA extraction and PCR amplification) as well as to lower the potential for identification of a marker that would be compromised by lateral gene transfer (LGT). We also searched the literature to eliminate any selected markers for which there was any history of LGT.

The last step was evaluation of phylogenetic signal of the selected markers by compiling sequences from different databases, generating alignments and building phylogenetic trees. Sequences were aligned using Clustal X (Thompson et al., 1997) under the default parameters. Trees were first inferred using NJ in PAUP* 4.0b10 (Swofford, 1998), then using RaxML (Stamatakis et al., 2005a,b) as implemented in the CIPRES cluster using the WAG + Gamma + F model of sequence evolution. Bootstrap support (BS) was evaluated using the online version of RaxML BlackBox with 100 replicates (Stamatakis et al., 2008). We constructed phylogenies for a total of 57 markers. Their phylogenetic content was evaluated based on the presence of well-established clades with morphological (ultrastructural) identities as well as recapitulation of well-established relationships within these clades. These clades included euglenozoa (kinetoplastida + euglenids), stramenopiles, plants, animals (animals + choanoflagellates), and fungi (fungi + microsporidia). In cases where representatives were sufficiently numerous, we looked for more inclusive groups such as Alveolata, Opisthokonta, and 'Amoebozoa' (Table 1). GenBank accession numbers for the seven selected markers (see below) are provided in Supplementary Table 2 and all data analyzed are available from the authors upon request.

To evaluate the predicted phylogenetic utility of our top seven markers across time, we calculated the net and per site informativeness profiles as in Townsend (2007), based on rates for each character estimated by Rate4Site (Mayrose et al., 2004) and applied to an uncalibrated ultrametric tree inferred by application of r8s (Sanderson, 2003) to the concatenated tree for all seven loci.

3. Results and discussion

A total of 17 markers were selected based on our initial criteria (Table 1). These markers were further divided into two tiers. The first-tier included seven promising markers for which we were able to design universal primers (except for gamma tubulin) for preliminary laboratory experiments (ST 1). Among these markers, PCR analyses for three markers (TFIIH, U5 snRNP, and CCT3) showed promising results in some unicellular eukaryotes. Thorough phylogenetic analyses of each of the seven first-tier markers are reported in supplementary figures (SF1–SF7, Supplementary material online). The phylogenies from the separate (SF1–SF7) and concatenated seven markers (Fig. 1) generally corroborate morphological data (e.g. Patterson, 1999; Table 1) and previous molecular studies (e.g. Yoon et al., 2008).

The phylogenetic informativeness profiles indicated that these seven markers retain useful phylogenetic signal spanning the whole tree including the deeper nodes. The two longest genes, U5 snRNP and UBA1, show the highest informativeness both at shallow and deep node levels (Fig. 2). They are also significantly longer than the other five markers (2026 aa and 873 aa,