

1. Department of Biological Sciences, Smith College, Northampton, MA & Program in Organismic and Evolutionary Biology, UMass-Amherst, MA
2. Department of Biological Sciences & Roy J. Carver Center for Comparative Genomics, University of Iowa, IA
3. American Type Culture Collection, Manassas, VA
4. Section of Ecology, Behavior and Evolution, University of California - San Diego, CA
5. Josephine Bay Paul Center in Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA

AIMS

The aims of this study are to use coordinated multigene phylogenetic analyses to:

1. test specific hypotheses regarding eukaryotic evolution, and
2. elucidate a robust scaffold for the eukaryotic tree of life

To meet these aims, we will:

- characterize DNA sequences of nine genes from at least 200, predominantly free-living microbial eukaryotes
- sample well-circumscribed clades of eukaryotes & taxa of as yet unknown affinities
- analyze data by combining existing approaches with newly developed methods for partitioning multigene data

INTRODUCTION

The core objective of this study is to create a robust phylogeny for eukaryotes through analyses of multigene genealogies of microbial eukaryotes. Although the majority of major eukaryotic lineages are microbial, relationships among many groups remain unresolved. Much of this uncertainty arises from limited and patchy molecular data that exist for most protist lineages. To achieve our objective, we will isolate and sequence a set of nine genes from at least 200 phylogenetically-diverse protist species. This study will provide a dramatic increase in sampling of these understudied organisms, and will lay the foundation for a comprehensive tree of eukaryotic life.

Eukaryotes, cells with nuclei, existed exclusively as microorganisms for 0.5 – 1.5 billion years before the evolution of plants, animals, and fungi. These microbial eukaryotes, or protists, are a diverse assemblage of organisms that encompass the phylogenetic breadth of eukaryotic evolution. Protists are characterized by numerous innovations in cell biology (e.g. multiple acquisitions of plastids), play an essential role in ecosystems (e.g. carbon fixation in marine systems), and some are causative agents of prevalent infectious diseases (e.g. malaria) that impact the social and economic fortunes of many countries. Yet, a major gap in our knowledge of the tree of life resides in the uncertain relationships of divergent microbial eukaryotes (Fig. 1). Clearly, knowledge of the positions of protists is key to understanding the origins of eukaryotes, and where the ancestries of plants, animals and fungi lie within these microbial groups.

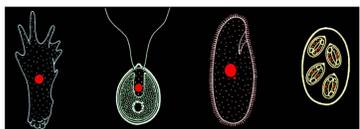


Figure 1: Traditional view of protists: amoebae, flagellates, ciliates, sporozoans. This perspective significantly underestimates the evolutionary and morphological diversity of these taxa.

PREVIOUS MOLECULAR STUDIES OF EUKARYOTIC MICROBES

Early molecular phylogenies relied on comparisons of rDNAs and suggested that the eukaryotic tree of life consisted of basal microbial lineages plus a ‘crown’ (Fig 2A). The conflict among single-gene genealogies led to the ‘Big Bang’ hypothesis (Fig 2B), which states that extant eukaryotes are descendants of a rapid radiation that occurred ~ 1 BYA.

To address the inconsistency among individual genes, several research groups have shifted their attention to phylogenetic studies based upon analyses of combined data (Fig. 2C). These combined phylogenies have identified potential super-clades of protists (e.g. the sister status of Euglenozoa + Heterolobosea (BALDAUF et al. 2000)). However, taxon representation in many of these analyses is sparse and many deep nodes are not well supported.

A major remaining question is where the root lies in the eukaryotic tree of life. Based on the distribution of a single gene fusion, the eukaryotic tree was recently rooted to generate three major clades: (1) Opisthokonts (fungi, animals, and related microbes), (2) Amoebozoa (slime molds, lobose and related amoebae), and (3) all remaining eukaryotes (Fig. 2D). Whereas this analysis is certainly an intriguing initial step, substantially more data are required to test this hypothesis.

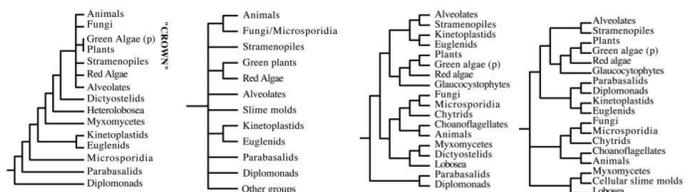


Figure 2: Alternative models for the scaffolding of eukaryotic diversity. (p) = polyphyletic.

UNIQUE CHALLENGES OF EUKARYOTIC MICROBES

A combination of features in microbial eukaryotes will create challenges in analyses:

- Considerable rate heterogeneity among lineages
- Potentially large and complex genomes
- Possibility of lateral gene transfers and/or symbioses
- Evolutionary changes in genetic code
- Modeling nucleotide, codon & amino acid evolution over long time scales

EXAMPLES OF HYPOTHESES

We will evaluate numerous evolutionary hypotheses through a combination of gene and taxon sampling. These hypotheses are not independent as many questions are intertwined. For example, resolving the phylogenetic position of particular unplaced flagellates or amoebae may not consolidate but rather disrupt groupings that have emerged from more limited analyses.

Hypothesis

‘Amoebae’



Specific Questions

The phylogenetic position of many amoebae is unknown, in part due to few morphological characters. Our questions include: What are relationships among the many amoeboid taxa of uncertain status? What is the status of recently-described clades including Gymnamoebia, and Archamoebae (CAVALIER-SMITH & CHAO 2003b; FAHRNI et al. 2003)?

‘Cercozoa’



The ‘Cercozoa’ lack a defining characteristic, and contain an extremely diverse set of species including both flagellates (e.g. Cercomonads) and amoebae (e.g. euglyphids) (ARCHIBALD et al. 2003; BERNEY and PAWLOWSKI 2003; CAVALIER-SMITH and CHAO 2003a; LONGET et al. 2003). How do multigene phylogenies reconstruct relationships within and among members of this group?

‘Chromalveolates’



Is the ‘Chromalveolate’ hypothesis, which posits the monophyly of the ‘chromist’ and alveolate protists (CAVALIER-SMITH 2002a; FAST et al. 2001), confirmed by nuclear gene trees? In particular, is the branching order within the Chromista (cryptophytes sister to haptophytes + stramenopiles) that emerged from plastid trees and gene loss data (NOZAKI et al. 2003; YOON et al. 2002) corroborated by multigene analyses?

OVERALL:



How are the major groups of eukaryotes related to one another? Which groups represent the deepest-branching eukaryotes? How do single-gene trees compare to the multigene phylogenies?

OUTREACH

This study will invigorate protist research while answering questions about the eukaryotic tree of life:

- **“Protist Diversity” workshop:** We will offer a workshop on the collection and identification (by light microscopy) of protists. Students will also be exposed to interpretation of ultrastructural data.
- **Graduate and post-doctoral training:** At least 3 postdoctoral fellows and 3 graduate students will be integrated into the proposed research. These participants will be trained in both molecular techniques and basic microscopy.
- **Undergraduate training:** PIs will seek REU supplements to train undergraduates, with a continued commitment to recruiting students from traditionally underrepresented groups.
- **Micro*scope:** This project will expand micro*scope, a web-based tool for exploring eukaryotic diversity (<http://www.mbl.edu/microscope>). We will develop micro*scope into a central resource for information & education on microbial diversity, through its use of the Universal Biological Indexer and Organizer, micro*scope offers a mechanism for making taxon-specific links with other ATOL sites.
- **Other:** Our project will inform decisions about expansion of protists at the ATCC & other collections.

RESPONSIBILITIES OF PIs

Laura Katz (LK), Smith College & University of Massachusetts

- management of this collaborative project, including coordinating DNA acquisition, assessment, storage and transfer, overseeing regular communications and organizing annual meetings
- collection of 1/3 of molecular data

Debashish Bhattacharya (DB), University of Iowa

- collection of 1/3 of molecular data
- development of appropriate databases for data organization and analyses.

Donald Burgess (DEB), American Type Culture Collection

- supervising protist culture and quality assurance,
- provide DNA samples for characterization by sequencing, PCR etc.

John Huelsenbeck (JH), University of California – San Diego

- provide guidance on phylogenetic analyses using existing approaches
- develop novel partitioning methods for multigene data

John Logsdon (JL), University of Iowa

- collection of 1/3 of molecular data
- support in database development

David J. Patterson (DP), Marine Biological Laboratories

- develop micro*scope (www.mbl.edu/microscope) into a resource to will link data to other relevant internet sites.
- development of a microscopy workshop, which will take place at MBL in years 2 and 4.

ANTICIPATED OUTCOMES

The resulting comprehensive phylogeny of eukaryotes is essential for:

- 1) interpreting the origins and diversification of eukaryotic cells
- 2) unifying the universal tree of life that includes both prokaryotes and eukaryotes,
- 3) understanding the multiple origins of multicellular eukaryotes

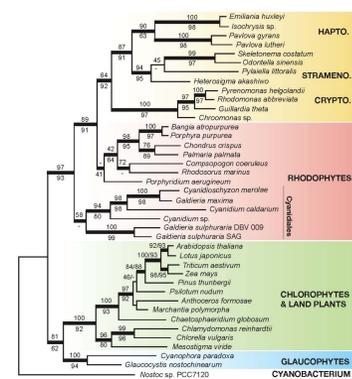


Figure 3: Example of the power of a multigene approach: a global tree of plastids (Yoon & Bhattacharya, unpub.)

Inferred from ML analysis (JTT + gamma model) of combined plastid proteins psaA, psaB, psbA, psbC, and psbD. The monophyletic chromist plastids are in brown and the rhodophyte donor of these plastids are in red. Results of a ML (100 replicates) and an unweighted MP (2,000 replicates) bootstrap analysis are above and below the branches, respectively. Thick branches represent >95% Bayesian posterior probability.

SPECIFIC METHODS

Analyses of multiple genes: DNA (and, in some cases, RNA) will be prepared mainly at ATCC using standard protocols, except that high nuclease activity in some protists requires modifications during extraction to ensure rapid inactivation of nucleases. We will characterize approximately one kilobase (kb) from 9 primary genes and do pilot studies on 4 secondary genes (Table 2). We will rely upon degenerate PCR to amplify genes, and we will clone PCR products into plasmid cloning vectors. We will fully sequence 2-3 clones per gene per taxon plus additional single reads from 4-5 clones to detect paralogs. We will also use RFLP surveys to search for additional paralogs.

Phylogenetic tree reconstruction: Sequences will be aligned using automated multisequence alignment algorithms such as Clustal W (THOMPSON et al. 1994), and resulting alignments will be manually edited. To assess the impact of alignment, phylogenies will be constructed using alignments generated with two sets of gap and gap length penalties. To reconstruct the gene phylogenies, we will analyze sequences using primarily maximum parsimony (MP) and maximum likelihood (ML) approaches. For example, unweighted maximum parsimony analysis will be performed with Paup*. Also, neighbor-joining analyses will use maximum likelihood distances calculated with the WAG (or other appropriate) model in Treezup v. 5.0. Finally, we will use the PROML method in PHYLIP (v3.6a3) and we will carry out Bayesian analysis of the data using different substitution models such as JTT, mtREV (ADACHI and HASEGAWA 1996), and WAG (WHELAN and GOLDMAN 2001).

We will generate trees from individual genes/proteins and on concatenated data sets; we will use a ‘total evidence’ approach (e.g. BALDAUF 1999; REGIER and SHULTZ 1997). We will also assess the extent of phylogenetic incongruence between different data partitions in multigene analyses (BULL et al. 1993), and evaluate potential cases of lateral gene transfers. Finally, we will explore the supertree approach to create a consensus phylogeny that incorporates two overlapping groups for which we may not have complete multi-gene data.

Data management: The sequence data will primarily be maintained on a server at the University of Iowa with weekly back-ups to protect against data loss. Project members will be able to access the data through a secure connection to a web site that will be created for the project. Sequence data will be submitted to GenBank as both individual sequences and alignments. Topologies will be imported into TreeBase (<http://www.treebase.org>) along with supporting alignments and documentation.

Table 2: 1st and 2nd tier genes

Primary	Function/Role	HSP90	heat shock protein - 90 kD
SSU-rDNA	small subunit rDNA	cox2	mitochondrial apocytochrome b
EF-1α	translation elongation factor		
α-tubulin	cytoskeletal protein	Secondary	Function/Role
β-tubulin	cytoskeletal protein	cox2	Mitochondrial cytochrome oxidase II
actin	cytoskeletal protein	RPB2	RNA polymerase II, 2 nd largest subunit
RPB1	RNA polymerase II, largest subunit	RAD51	eukaryotic recA homolog
HSP70	heat shock protein - 70 kD	V-ATPase	Vacuolar ATPase

ADACHI, J., & M. HASEGAWA, 1996 Instability of quartet analyses of molecular sequence data by the maximum likelihood method. Mol. Phy. Evol. 6: 72-76.
 ARCHIBALD, J.M., D. LONGET, J. PAWLOWSKI & P. KEELING, 2003 A novel polyubiquitin structure in Cercozoa & Foraminifera. Mol. Biol. Evol. 20: 62-66.
 BALDAUF, S.L., 1999 A search for the origins of animals and fungi: Comparing and combining molecular data. American Naturalist 154: S178-S188.
 BALDAUF, S.L., A.J. ROGER, I. WENK, SIEFER, & W.F. DOOLITTLE, 2000 A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290: 972-977.
 BERNEY, C., & J. PAWLOWSKI, 2003 Revised small subunit rRNA analysis provides further evidence that foraminifera are related to Cercozoa. J. Mol. Evol. 57: S120-S127.
 BULL, J.J., J.P. HUELSENBECK, C.W. CUNNINGHAM, D.L. SWOFFORD & P.J. WADDELL, 1993 Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42: 384-397.
 CAVALIER-SMITH, T., 2002 Chloroplast evolution: Secondary symbioses and multiple losses. Curr. Biol. 12: R62-R64.
 CAVALIER-SMITH, T., & E.E.Y. CHAO, 2003a Phylogeny and classification of phylum Cercozoa (Protozoa). Protist 154: 341-358.
 CAVALIER-SMITH, T., & E.E.Y. CHAO, 2003b Phylogeny of chlozoa, apusozoa, and other protozoa and early eukaryote megaevolution. J. Mol. Evol. 56: 540-563.
 SÖGIN, M.L., J.H. GUNDER, et al., 1989 Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. Science 243: 75-77.
 STECHMANN, A., & T. CAVALIER-SMITH, 2002 Rooting the eukaryote tree by using a derived gene fusion. Science 297: 89-91.
 THOMPSON, J.D., D.G. HIGGINS & T.J. GIBSON, 1994 Clustal W - improving the sensitivity of progressive multiple sequence alignment. Nucleic Acids Res. 22: 4673-4680.
 WHELAN, S., & N. GOLDMAN, 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18: 691-699.
 YOON, H.S., J.D. HACKETT, G. PINTO & D. BHATTACHARYA, 2002 The single, ancient origin of chromist plastids. Proc. Natl. Acad. Sci. U. S. A. 99: 15507-15512.