Function/Role

**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

**UNIQUE CHALLENGES OF EUKARYOTIC MICROBES**

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

- Considerate rate heterogeneity among lines
- Potentially large and complex genomes
- Possibility of lateral gene transfers and/or symbiosis
- Evolutionary genetic change in codon, amino acid sequences over long time spans

**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 interwoven. For example:

- The evolutionary position of a particular unplaced flagellate or amoeba may not consolidate but rather disrupt groups that have emerged from more limited analyses.

**PREVIOUS MOLECULAR STUDIES OF EUKARYOTIC MICROBES**

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

**OUTLOOK**

We will evaluate numerous evolutionary hypotheses through a combination of gene and taxon sampling. These hypotheses are not independent as many questions are interwoven. For example:

- The evolutionary position of a particular unplaced flagellate or amoeba may not consolidate but rather disrupt groups that have emerged from more limited analyses.

**SPECIAL METHODS**

Analyses of multiple genes (DNA, and in some cases, RNA) will be performed mainly at ATCC using standard techniques, except that high nucleotide content in some proteins requires modifications during extraction to ensure rapid maximisation of nucleases. We will characterise approximately one kilobase (kb) from 9 primary genes and do pilot studies on 6 secondary genes (Table 2). We will rely upon degenPCR to amplify genes, and we will clone PCR products into plasmid cloning vectors. We will fully delete genes per plasmid plus additional single reads from 6-3 clones to detect parasites. We will also use RFLP survey to search for additional parasites.

Phylogenetic tree reconstruction: Sequences will be aligned using automated multiple-sequence 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 with multiple models generated with two sets of gap and gap length penalties. To reconstitute the gene phylogenies, we will analyze sequences using primarily maximum parsimony and maximum likelihood methods. For example, unweighted maximum parsimony analysis will be performed with Paup*. Also, neighbor-joining analysis will use maximum likelihood distances calculated with the JTT (or appropriate) model in StepaRea v. 3.0. Finally, we will use the PhyML method in PHYML v. 1.6 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 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 multiple analyses (BRELL et al., 1993), and evaluate potential cases of lateral gene transfers. Finally, we will explore the utility of the alphabet in the context of a phylogenetic concept that incorporates two overlapping groups for which we may not have complete multigene data.

Data management: The sequence data will primarily be maintained on a server at the University of Iowa with weekly backups to protect against data loss. Protein data will be incorporated into the repository and a web site that will be granted to the project sequence data will be submitted to GenBank as both individual sequence and alignments. Table 2 will be imported into Taxtree3d (www-taxtree.org) along with supporting alignments and documentation.

**Table 2: 1st and 2nd tier genes**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Tiers</th>
<th>Species</th>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-ATPase</td>
<td>1st tier</td>
<td>eukaryotic protein</td>
<td>ATP synthesis</td>
<td>ATP synthase complex</td>
</tr>
<tr>
<td>V-ATPase</td>
<td>2nd tier</td>
<td>eukaryotic protein</td>
<td>ATP synthesis</td>
<td>ATP synthase complex</td>
</tr>
<tr>
<td>EF-1</td>
<td>1st tier</td>
<td>eukaryotic protein</td>
<td>Translation initiation</td>
<td>Translation initiation factor</td>
</tr>
<tr>
<td>RAD51</td>
<td>1st tier</td>
<td>eukaryotic protein</td>
<td>DNA repair</td>
<td>Homologous recombination</td>
</tr>
<tr>
<td>RPB2</td>
<td>1st tier</td>
<td>eukaryotic protein</td>
<td>Nuclear transcription</td>
<td>RNA polymerase II</td>
</tr>
</tbody>
</table>

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