DEFINING DNA-BASED OPERATIONAL TAXONOMIC UNITS FOR MICROBIAL EUKARYOTE ECOLOGY

David A. Caron[1], Peter D. Countway[1], Pratik Savai[1], Rebecca J. Gast[2], Astrid Schnetzer[1], Stefanie D. Moorthi[1,3], Mark R. Dennett[2], Dawn M. Moran[2], and Adriane C. Jones[1]

Running Title: DNA-based OTUs for Protistan Ecology

[1]Department of Biological Sciences, University of Southern California, 3616 Trousdale Parkway, Los Angeles, CA 90089-0371

[2]Woods Hole Oceanographic Institution, Woods Hole, MA 02543

[3]Present address: Carl-von-Ossietzky Universität Oldenburg, ICBM-Terramare, Schleusenstr. 1, D-26382 Wilhelmshaven, Germany

ABSTRACT

DNA sequence information has been increasingly used in ecological research on microbial eukaryotes. Sequence-based approaches have included studies of the total diversity of selected ecosystems, the autecology of ecologically relevant species, and the identification and enumeration of species of interest to human health. It is still uncommon, however, to delineate protistan species based on their genetic signatures. The reluctance to assign species-level designations based on DNA sequences is partly a consequence of the limited amount of sequence information presently available for many free-living microbial eukaryotes, and partly the problematic nature and debate surrounding the microbial species concept. Despite the difficulties inherent in assigning species names to DNA sequences, there is a growing need to attach meaning to the burgeoning amount of sequence information entering the literature, and a growing desire to apply this information in ecological studies. We describe a computer-based tool that assigns DNA sequences from environmental databases to operational taxonomic units at approximate species-level distinctions. The approach provides a practical method for ecological studies of microbial eukaryotes (primarily protists) by enabling semi-automated analysis of large numbers of samples spanning great taxonomic breadth. Derivation of the algorithm was based on an analysis of complete small subunit ribosomal RNA (18S) gene sequences and partial gene sequences obtained from GenBank for morphologically described protistan species. The program was tested using environmental 18S data sets from two oceanic ecosystems. A total of 388 operational taxonomic units were observed among 2,207 sequences obtained from samples collected in the western North Atlantic and eastern North Pacific.

INTRODUCTION

Ecological studies of aquatic microbial eukaryotes require the identification and enumeration of an extremely wide taxonomic diversity of organisms. These assemblages are typically dominated by phototrophic and heterotrophic protists (microalgae and protozoa), but microscopic metazoa from a variety of animal phyla can also contribute significantly. The identification of protists in environmental samples is particularly difficult because most species have been defined morphologically (41, 85). Protistan identifications involve a wide variety of procedures for collection, preservation, specimen preparation and examination (34, 85), as well as many different taxonomic expertises. Very few studies have attempted to identify and enumerate all protistan taxa because of these complexities, making it difficult to evaluate ecological studies of protistan diversity, community structure and biogeochemical function.

The growing database of DNA sequence information for a wide spectrum of microbial eukaryotes offers the possibility for greatly improving our existing tools for studying the phylogeny and ecology of these organisms. Much of the initial impetus for the acquisition of rDNA sequence information for microbial eukaryotes in the 1980s and 90s arose from a desire to improve our understanding of the evolutionary relationships among these taxa, especially among the many protistan lineages (66, 73-75). That research provided significant insights into the evolution of eukaryotic organisms, and continues to facilitate the generation, testing and modification of numerous hypotheses on this topic (1, 7, 15, 39, 72).

A molecular taxonomy has several real or potential advantages for ecologists relative to traditional taxonomies: (1) the ability to apply it to a wide range of taxa including those possessing few distinctive morphological features; (2) applicability across all life stages of a species; (3) a lessened requirement for formal (i.e. morphological) taxonomic training; (4) a standardized approach for sample processing, interpretation and comparison across different studies; (5) the potential for automation of much of the processing of sample characterization; and (6) the ability to taxonomically characterize large numbers of samples that are typical of most ecological studies.

DNA sequence information has been used to establish distinctions among protistan species with few or variable morphological features (11, 14, 23, 50, 87), as an aid to

characterize lineages of minute protists which largely lack morphology (3-5) and to identify and study specific protistan taxa in complex natural assemblages using fluorescence in situ hybridization (FISH) and quantitative real-time PCR (qPCR) (37, 43, 71). This work has helped establish the spatiotemporal distributions of a number of ecological important species such as harmful algal species, and/or species of significance to human health (6, 12, 19, 32, 55, 63, 89).

Genetic approaches have also been extensively applied to assess the composition of natural assemblages of protists from a variety of ecosystems. These studies have reported lists of gene sequences representing a wide array of protistan lineages from freshwater environments, various oceanic ecosystems ranging from polar to tropical, anoxic ecosystems and deep-sea environments (20-22, 25, 29-31, 36, 44-47, 51, 54, 78, 80). Interpretation of the results of these investigations of protistan community composition and structure could be improved by a clearer understanding of how sequence information translates into taxonomic composition. Moreover, the effectiveness of statistical approaches to compare the structure of microbial communities are dependent on accurate accounting of the number of taxa in these assemblages (67).

Methods utilizing DNA sequences for deriving microbial operational taxonomic units (OTUs), and subsequently species richness from sequence data, are now appearing in the literature (68, 69). These approaches hold great promise for ecologists by providing potentially powerful tools for examining community composition. A molecular taxonomy has been received enthusiastically by many within the ecological community, but with skepticism by some. Proponents have openly campaigned for the development of a DNA taxonomy to augment extant taxonomic schemes for microbes that rely primarily on morphology or physiology (9, 65, 81). Skeptics have noted technical and conceptual problems with the approach, and have expressed concern that molecular taxonomies do not necessarily facilitate an understanding of the morphological, physiological and behaviorial characters of organisms (27, 64). For ecologists, an optimal situation might involve the use of genetic signatures (to facilitate sample analysis) combined with an understanding of how that information relates to morphology, physiology and behavior in order to understand the biogeography of functional traits, not just taxonomic entities (35).

Very little work has been attempted regarding the derivation of a practical, sequence-based protistan taxonomy for ecological research. Diversity studies to date have used a range of approaches and/or a range of sequence similarity values to create OTUs from eukaryotic sequence libraries with little consistency or justification in the choice of these values (see Discussion). This inconsistency has caused confusion in interpreting data and comparing data sets between different studies. Resistance among many researchers to infer protistan species from sequence information exists, in part, because the species concept for protists is problematic. Morphological features have traditionally been used for species descriptions but reproductive and physiological criteria, and more recently DNA sequences, have also been incorporated (53, 56). This combination of disparate characters for defining protistan species has complicated the process of extrapolating these descriptions directly to species definitions based solely on DNA sequences. Regrettably, the complicated taxonomic schemes presently in use for protists are particularly difficult to apply in ecological studies.

The application of DNA sequence to ecological studies cannot await a resolution to the debate over the protistan species concept, if that ever happens (16). A practical method, recognizing the present limitations of this approach, could significantly improve our ability to interpret the large sequence data sets now appearing in the literature. The goal of this study was to establish a practical, reproducible approach for the use of DNA sequence information for defining molecular operational taxonomic units (OTUs) for ecological studies of microeukaryotic organisms, focused primarily on protistan taxa.

We designed and tested a computer program (Microbial Eukaryote Species Assignment; MESA) to establish species-level operational taxonomic units from 18S rRNA sequence information. Sequence data obtained from GenBank for a wide variety of taxa were used for design and testing. The program was then applied to sequence data obtained from environmental samples collected from the western North Atlantic and eastern North Pacific. A total of 388 taxonomic OTUs were derived from the combined sequences libraries totaling 2,207 partial 18S sequences. Within this large database, only 54 out of 388 of the OTUs were common to both the Atlantic and Pacific sample sets. Rare taxa (OTUs with ≤ 2 clones) comprised the majority of OTUs.

MATERIALS AND METHODS

The overall logic behind the design and application of the automated program for calling OTUs for protists was as follows. Full-length, 18S sequences of 'well-defined' (i.e. morphologically-defined) protistan species were selected from GenBank. The species included multiple strains within a variety of species, and multiple species within a number of genera across a wide phylogenetic range. Automated, pair-wise alignments of all sequences were performed using ClustalW (84). Intra-specific sequence variability (multiple strains within a species) and inter-specific similarity (different species within a genus) were analyzed based on the ClustalW alignments. A logical, overall demarcation (% similarity) for differentiating among the sequences at approximately the species level was determined based on an analysis of the alignments and the GenBank species identifications. A program (Microbial Eukaryote Species Assignment; MESA) was then developed for calling operational taxonomic units from 18S sequences using the percent similarity derived above. Finally, the MESA program was applied to an environmental sequence database for assessing microbial eukaryote diversity.

Analysis of intra- and inter-species sequence similarity. The design of the protistan OTU-calling program used publicly available sequence information (GenBank) from morphologically defined protistan taxa to establish an appropriate level of sequence similarity for use in the program. Morphologically defined species were employed because the overall purpose was to establish a link between DNA sequence similarity/dissimilarity and species identity based on traditional taxonomic schemes for protists. A wide range of taxa was specifically selected including those with extensive morphological features as well as taxa whose morphologies are variable or nondescript (e.g. amoebae and minute, non-flagellated algae) where ultrastructure, physiology or behavior have been invoked to delineate species. Our logic was that the chosen species might represent classifications ranging from taxonomic 'lumpers' to 'splitters'. Full-length, 18S gene sequences were used because eukaryotic databases now possess sufficient numbers of these sequences to begin to allow meaningful comparisons.

Both intra-specific and inter-specific comparisons were conducted to develop a program that would call OTUs with approximately species-level resolution. Seventeen species encompassing a total of 211 sequences were used to examine intra-specific

sequence variability. The number of strains in each species varied from 4 to 56 (Table 1). Thirty-one genera were used to examine inter-specific sequence variability. The number of species within each genus varied from 3 to 36 (Table 2). Sequence similarity among taxa above genus level was not examined because the intent was to identify species-level distinctions, and it was assumed that sequence-to-sequence variability among species from different genera would be greater than the variability between congeners. The species employed in these analyses included amoebae, minute chlorophytes, euglenoids, kinetoplastids, dinoflagellates, ciliates, diplomonads, heterokonts (diatoms, chrysophytes) and prymnesiophytes. No attempt was made to equalize or normalize sample numbers across this diversity of species.

Primary read, full-length 18S sequences (in FASTA file format) were prepared for pair-wise alignments by trimming each sequence (if necessary) at the 5' and 3' ends using an automated method that read from the end of the sequence toward the center and removed base sequences that contained more than 5 Ns per 25 base pairs. This process did not affect the full-length sequences obtained from GenBank to test the MESA program, but it was necessary for the environmental sequence databases. Intra-species (strain-strain within a species) sequence variability was determined using pair-wise alignments of full-length 18S sequences for the 17 species examined (total of 2,712 pair-wise comparisons; for n strains within a species, the number of pair-wise alignments within each species = $n!/((n-2)! \times 2!)$) using ClustalW without additional manual alignment. Aligned sequences were truncated to remove any non-overlapping sequence at the ends of each gene pair. Gaps were assigned one mismatch for each base pair difference. Pair-wise alignments of full-length 18S sequences for 323 species distributed among 31 protistan genera were also examined to establish the level of sequence similarity appropriate for distinguishing different species within a genus. A total of 2,439 pair-wise alignments of congeners were obtained using ClustalW, as noted above.

The ClustalW alignments were not manually adjusted because our goal was the development of an approach that would allow the comparison of large sequence databases with minimal human assistance rather than obtaining truly phylogenetically-informative alignments. Similarity values were calculated from the total number of basepair mismatches on the overlapping fragments of two sequences for every pair-wise

intra- and inter-species comparison and similarity matrices constructed for these two data sets. Average similarity values were determined for each species for all strain-strain comparisons, and a then weighted average for intra-specific sequence similarity across all species was calculated. A similar analysis was performed for the pair-wise comparisons among congeneric species to obtain an average inter-species sequence similarity.

The distributions of intra- and inter-specific sequence variability were examined visually, and a similarity value chosen that minimized discrimination among strains within each species but maximized discrimination among species within each genus. The resulting similarity value was used in the design and application of the MESA program (95% sequence similarity; see Results).

Derivation of the MESA program. The algorithm for the MESA program is shown in Figure 1. An initial round of sequence comparisons was conducted to place all sequences into provisional OTUs (see Fig. 1, Formation). The first OTU was established by selecting the first sequence in a sequence file. The second sequence was compared to the first OTU sequence using the ClustalW alignment to determine sequence similarity. If the similarity value was ≥ 95%, then the sequences were placed together in OTU #1. If the similarity value was < 95%, then the second sequence formed a separate OTU (OTU#2). Each subsequent sequence was then compared to OTU#1. If the sequence similarity of the new sequence was < 95% with any of the sequences in OTU#1, then it was compared to the sequences in OTU#2, and so forth until each sequence was either placed into an existing OTU, or formed a separate OTU.

An optimization step was conducted once all sequences had been placed into provisional OTUs in order to determine the best possible placement of each sequence among the OTUs (see Fig. 1, Optimization). The average sequence similarity of each sequence to all other sequences within an OTU was determined, and compared to the average similarity of that sequence to sequences in all other OTUs. Any sequence that revealed a greater average similarity to the sequences in another OTU was moved into the OTU with which it had the greater average similarity.

Finally, a condensation step was conducted to determine whether any two OTUs possessed overall average similarity that warranted the condensation of the two OTUs into a single OTU (see Fig. 1, Condensation). Average sequence similarities among the

sequences in two OTUs were compared for every pair of OTUs. If the average similarities were ≥ 95%, the two OTUs were condensed into a single OTU.

Testing the reliability of the MESA program. An initial test of the OTU-calling program was conducted using two replicate 18S clone libraries constructed from a single water sample obtained in the western North Atlantic. The purpose of this exercise was to test how closely OTUs were called from two clone libraries constructed independently from the same water sample. The replication of the cloning/sequencing approach employed for environmental samples was an inherent component of the evaluation.

Water collection, sample processing, cloning and sequencing protocols are detailed in Countway et al. (21). Briefly, water was collected using Niskin bottles from the subsurface euphotic zone at a station along the U.S. continental shelf (36˚21'N, 75˚14'W), and pooled to create a single sample. The sample was prefiltered through a 200 μm Nitex screen to remove most metazoa and filtered onto a 47 mm glass fiber GF/F filter (Whatman International, Ltd., Florham Park, NJ). DNA was released from cells using 1 ml lysis buffer (100 mM Tris [pH 8], 40 mM EDTA [pH 8], 100 mM NaCl, 1% SDS) at 70˚C with bead beating (0.5 mm zircon beads), followed by 1% CTAB (hexadecyl-trimethyl ammonium bromide, Sigma), and then extracted in phenol-chloroform and precipitated with isopropanol (33).

The resulting DNA was divided into two aliquots, and each aliquot was used in independent PCR reactions. Full-length 18S genes were amplified from the genomic DNA extracts using universal eukaryotic primers Euk-A (5'-AACCTGGTTGATCCTGCCAGT-3') and Euk-B (5'-GATCCTTCTGCAGGTTCACCTAC-3') (52). Amplicons of the appropriate length were excised, gel-purified, ligated into plasmids using the pGEM-T Easy Vector kit (Promega), and used to transform Electro10Blue electrocompetent cells (Stratagene) using procedures outlined in Countway et al. (21). DNA sequencing was carried out on a Beckman-Coulter CEQ8000 automated DNA sequencer (Fullerton, CA) according to manufacturer's specifications. A single sequencing read was performed using Euk-570F (5'-GTAATTCCAGCTCCAATAGC-3') (88). The sequences obtained ranged from 400 to 700 bp in length. The resulting partial sequences were checked for chimeric sequences using Ribosomal Database Project (RDP) Chimera Check

(http://rdp8.cme.msu.edu/html/), possible chimeric sequences were eliminated, and the remaining sequences analyzed in pair-wise alignments and placed into OTUs according to procedures outlined above. The lengths of the aligned sequences used for the estimation of the percent similarity value varied because of the variable read lengths.

Application of the MESA program to a large environmental dataset. The ability of the OTU-calling program to handle a large environmental data set was examined by applying it to a database of 2,207 partial sequences derived from previously published data from samples collected in the North Atlantic (980 sequences) (21) and from a study site in the coastal, eastern North Pacific (1,407 sequences). The latter data set was comprised of clone libraries constructed from water samples collected on a single date at 1, 20, 42, 150, 500 and 880 m at the San Pedro Ocean Time Series (SPOTS) station located mid-way between Santa Catalina Island and the U.S. mainland in the San Pedro Channel (33˚33'N, 118˚24'W). The location is the site of an ongoing Microbial Observatory, and a complete analysis of the data set will be presented elsewhere (GenBank accession numbers XXXXXXXX- XXXXXXXX). Sample collection and processing, and DNA extraction, amplification, cloning and sequencing were conducted as described above and in Countway et al. (18). Seawater used in the study was prefiltered through 200 µm screening and particulate material was collected on GF/F glass fiber filters (Whatman International Ltd., Florham Park, NJ). Sequencing of the libraries was conducted using Euk-570F (5'-GTAATTCCAGCTCCAATAGC-3') to provide compatibility with the North Atlantic data. Libraries from individual depths contained 137-257 sequences, but sequence information from all depths for the Pacific samples was combined for the present analysis.

The resulting partial sequences from the combined North Atlantic and North Pacific samples were processed as a single data set, placed into OTUs using the MESA program, and then separated according to sampling site. Taxonomic information pertaining to the 50 most abundant OTUs was obtained using BLAST (2) against the NCBI (8) and ARB (48) databases. Searches were conducted using all the occupants of each OTU.

RESULTS

Construction and evalution of the MESA program. The analysis of intra-species sequence variability indicated a high level of sequence similarity across the 211 strains within the 17 species examined (Fig. 2). Overall, sequence similarity between strains of the same species was high with an average of 98% similarity for all 2,712 pair-wise comparisons (Table 1), although a small percentage of the comparisons yielded relatively low values. A total of 89% of the strain-strain comparisons were placed within the same OTU by the MESA program using a demarcation of 95% sequence similarity. Most of the 11% of the comparisons that had similarity values less than 95% were contributed by a single amoeba species (*Acanthamoeba lenticulata*). That species yielded a particularly low overall average value in the pairwise matches (85%) relative to all other species.

The results of the intra-generic, inter-specific comparisons were less decisive than the intra-specific comparisons with respect to a similarity value that clearly demarcated species (Fig. 3). Seventy-eight percent of the inter-species pair-wise alignments exhibited ≤ 95% sequence similarity, while 22% of pairs showed > 95% similarity (that is, 22% of the time different species were placed in the same OTU). The overall sequence similarity among species within the same genus was 87% for all 2,439 pair-wise comparisons; Table 2). The congeneric species producing the highest sequence similarity values were observed in the genera *Tetrahymena*, *Leishmania* and *Nannochloropsis*.

The efficacy of the MESA program was also examined using partial sequences (approximately 600 bp of the 18S gene, beginning at 570F) of the same species and strains employed in the full-length sequence analysis described above. This analysis was conducted to determine if the program would produce results with partial sequences that were similar to our findings with full-length sequences. Many of the sequences appearing in GenBank that have been generated from environmental 18S clone libraries have been partial sequences, using 570F or a nearby primer for sequencing (20, 25, 42, 44, 47, 90). The results of the analysis using partial sequences and full-length sequences were virtually identical. The overall weighted average for intra-species similarity for the partial 18S sequences was the same as for the full-length sequences (98%). The inter-

species comparison yielded a value of 87% for all 2,439 pair-wise alignments using partial sequences, compared to a value of 90% for the full-length sequences.

Based on the analyses above, a sequence similarity value of 95% was chosen for use in the MESA program to provide approximately species-level distinctions among 18S sequences of protists. This value represented a compromise between identifying multiple strains of a single species as a single OTU on the one hand, and separating congeneric species into separate OTUs on the other hand.

Analysis of replicate clone libraries. A total of 357 partial sequences were obtained for the two replicate clone libraries from the North Atlantic sample. The libraries were combined for OTU calling and then separated for comparison. Application of these sequences to the MESA program yielded 51 and 61 OTUs for the two libraries (Fig. 4). The general shapes of the rank abundance curves for each library were similar. Twenty-four of the OTUs were observed in both clone libraries, while 64 OTUs were unique to either library. The 24 OTUs observed in both libraries were among the most abundant OTUs in the combined data set. That is, the MESA program yielded 'common' taxa that were observed in both libraries and at approximately similar relative abundances with a few exceptions (Fig. 4C). The presence of many 'rare' OTUs (OTUs represented by a single sequence) that were unique to either library was not surprising given the relatively low number of clones that were sequenced from each library and the potentially large number of these sequence types in natural samples.

Analysis of North Atlantic and North Pacific environmental clone libraries. A total of 2,207 partial sequences were analyzed using the MESA program from the combined North Atlantic and North Pacific clone libraries (Fig. 5). These sequences yielded a total of 388 OTUs using a sequence similarity value of ≥95%. The rank abundance curve for these OTUs revealed a relatively small number of OTUs (18% of total) that were composed of 5 or more sequences, while a large number of OTUs contained only one or two sequences.

Most of the OTUs observed in the combined data set were present in libraries obtained at one site or the other but not both (Table 3). Only 54 of the OTUs (14%) were observed in clone libraries constructed from samples from both study sites. A wide diversity of taxonomic groups was represented within the overall data set. Among the 50

most abundant OTUs from the combined data sets, twelve of these OTUs were metazoa (mostly copepods), while 38 returned best sequence matches that identified them as protistan taxa (Table 4). A substantial number of the latter sequences (17 of 38) showed closest phylogenetic affinity with unclassified alveolate taxa. Approximately one half of the 50 most abundant OTUs were observed in the data set from either the North Atlantic site or the North Pacific site, but not both.

The effect of the choice of the similarity value employed in the MESA program on the number of OTUs estimated from the environmental data set was examined by processing the 2,207 sequences using a range of similarity values (Fig. 6). The choice of the value dramatically affected the number of OTUs constructed by the program, particularly within the range of similarity values that have been generally employed in protistan molecular diversity studies. For example, increasing the similarity value from 95% to 99% resulted in a 2.5-fold increase in the number of OTUs among the sequences in the combined data set (from 388 to 956 OTUs).

DISCUSSION

Towards a DNA taxonomy. The development of a DNA taxonomy for microbial eukaryotes would provide a much needed tool for ecological studies of natural microbial communities, but the impediments to this goal include both technical and conceptual problems. Technical problems include potential artifacts pertaining to DNA extraction and amplication, cloning, sequencing and sequence manipulation. These issues will undoubtedly lessen at our present pace of biotechnological and computational advance. Conceptual issues are more problematic, as is the choice of gene(s) used in these taxonomies. For example, some species possess multiple RNA gene copies with somewhat different base pair compositions (70). These different sequences could conceivably produce multiple OTUs from a single specimen if the differences are sufficiently large, although these instances appear to be relatively rare. Similarly, the use of rapidly evolving genes or intergenic spacer regions might result in the creation of multiple OTUs for individuals that would be grouped into a single species using other criteria (60).

The algorithm and specific levels of similarity described here were developed using 18S gene sequences for a broad range of taxa. We chose the 18S gene because a substantial amount of information is available for this gene in public databases. However, the heterogeneous rates of evolution that have been noted for this gene (13) may make this gene less useful for some taxonomic groups. Another gene might prove more advantageous for those taxa, and this approach can easily be adapted as the databases for other genes expand. The use of ecologically relevant genes as the basis for a molecular taxonomy might aid reconciliation of molecular and traditional taxonomic schemes. Indeed, we anticipate that future protistan molecular taxonomies may involve the use of specific genes for specific taxa, or the use of multiple genes much the way multiple gene phylogenies are presently employed to yield intergrated perspectives on the evolutionary history of microbial taxa (24, 38, 49, 57). The MESA program presented here has not yet been applied in this way, but it provides a conceptual template for these adaptations.

More significantly, the debate regarding what constitutes a protistan species makes the reconciliation of traditional and molecular taxonomies a difficult task. The morphological species concept that dominates protistan taxonomy has been challenged by some investigators as inadequate because it sometimes fails to differentiate physiologically or sexually distinct entities within identical or nearly identical morphotypes (10, 17, 58, 65). Taxonomists continue to debate the integration of the morphological species concept, the biological species concept and the ecological species concept for describing protistan species and their global distributions (16). Within this debate, the decision by protistan ecologists to incorporate DNA analysis as one more observational and experimental tool is a pragmatic one, as is the conceptual approach presented here. It is impossible to formulate and test hypotheses on the environmental factors determining the distributions of ecologically- and commercially-important taxonomic entities without rapid, reliable means to determine the presence and abundances of those entities. Thus, new tools are required to allow large-scale studies of the ecology of these species that would not be possible using cumbersome, time-consuming and often inaccurate morphology-based taxonomies.

Establishing a protistan OTU-calling program. Extant approaches for establishing operational taxonomic units for microbial taxa are typically based on evolutionary distances (68). Such approaches afford the potential to derive a truly phylogeny-based taxonomy, but they are difficult to apply in ecological research because the computations typically require manual adjustments to multiple sequence alignments to improve automated alignments provided by programs such as ClustalW (83). Unfortunately, this is counterproductive when dealing with potentially 1000s of sequences that are often required in ecological studies. It would require an enormous amount of preparatory work and considerable training, slowing the processing time for a data set. This situation may improve in the future as algorithms for sequence alignment improve (28, 62). Given the present state of these programs, however, a specific objective of this study was to develop a program that could be applied in a 'hands-off' fashion to facilitate the rapid processing of large data sets characteristic of ecological studies.

Our objective for developing the MESA program was to establish practical guidelines for establishing protistan OTUs. The protistan MESA program does not provide a phylogeny-based taxonomy, nor does it attempt to resolve the controversial and difficult issue of the 'species concept' for protistan taxa. We have merely used protistan species with traditional identifications to provide information on setting the demarcation between taxonomic units to yield approximately species-level distinctions for use in ecological research. Therefore, consecutively called OTUs do not necessarily share a close phylogenetic relationship, because the manner in which the program handles gaps and variable regions is not necessarily appropriate for phylogenetic analysis. The information obtained in the analysis of intra-species and inter-species variability (Fig. 2, 3) assisted in selecting the level of sequence similarity used in the present analysis. The similarity value can be altered to permit more or less stringent formation of OTUs. Once formed, the sequences within each OTU can be analyzed by BLAST to determine the closest phylogenetic affiliation of that unit.

The three step process by which the program creates protistan OTUs entails an initial assignment, an optimization of the placement of sequences within OTUs following their initial establishment, and finally a test for condensation of some of the OTUs that exhibit strong similarity (Fig. 1). The final step is an attempt to prevent the creation of artificial

'microdiversity' by generating OTUs that show little distinction from one another. The latter situation was particularly important for highly populated OTUs but generally did not affect the existence of OTUs composed of only one or two sequences. The condensation step was also somewhat affected by the size of the database (i.e. number of sequences being compared). That is, as more sequences are added to a database, the potential for some degree of condensation increases. For example, the North Atlantic data set yielded 165 OTUs when analyzed alone, but 160 OTUs when analyzed in combination with the North Pacific data set. Based on these characters of the program, we believe that the program provides a conservative estimate of microbial eukaryotic taxa in a sample.

Choosing a similarity value for demarcating OTUs. The absolute number of OTUs obtained from the large environmental clone libraries examined in this study was critically dependent on the value of sequence similarity employed to distinguish among taxonomic units (Fig. 6). We chose a similarity value (95%) that was lower than values of 97-99% that have been employed in most studies. However, it is important to remember that our similarities result from automated, pair-wise alignments without manual adjustment through a hypervariable region of the 18S rRNA gene. Manual alignment would have undoubtedly increased the level of similarity between many pairs of sequences. As noted above, omission of a manual alignment step was a conscious decision to allow a 'hands-off' procedure for processing very large data sets such as the one depicted in Figure 5. This is a highly desirable approach for ecological investigations because of the large number of sequences that typically are processed in these studies, and will facilitate direct comparisons between different investigations.

A similarity value of 95% was chosen purposefully as a conservative estimator of species richness in a natural sample. This decidedly conservative choice probably masked considerable physiological diversity within some OTUs. This is supported by the observation that congeneric species were placed in a single OTU in 22% of the 2,439 pair-wise comparisons, and the congeners with the highest sequence similarity values were from the genera *Tetrahymena*, *Leishmania* and *Nannochloropsis*. Interestingly, these three genera include species whose taxonomic descriptions represent deviations from purely morphological descriptions. Mating type compatibility has been employed

to separate morphologically indistinguishable species of *Tetrahymena* (17, 58), and DNA sequence information has been used to differentiate between *Leishmania* strains that vary in their etiology (86, 87). The genus *Nannochloropsis* contains minute algae for which few morphological features are visible at the level of light microscopy, and for which physiological/biochemical characters have been employed to supplement morphological features. It is not surprising, therefore, that a comparison of 18S sequences for species within these three genera might not agree with similarity values obtained for other protistan species. These genera exemplify the present state of confusion regarding the species concept for protists. Some researchers might consider these latter distinctions strain-strain variability or 'ecotypes' within morphospecies, while others would confer species status (16). Nonetheless, refinement of the approach described here, and specifically the use of taxon-specific similarity values, could bring the outcome of the MESA program more in line with the accepted taxonomic distinctions for various protistan lineages. We anticipate that future iterations of the MESA program might enact an initial basic grouping based on one level of sequence similarity, and then apply a different level of similarity (or a different gene) that is more informative for each group.

The overall average similarity value for intra-species, pair-wise comparisons using the protistan sequences retrieved from GenBank in this study was 98% among all strains examined in 17 species (Fig. 2, Table 1). Only *A. lenticulata* yielded a substantially lower average similarity value for strain-strain comparisons (85%), with the next lowest value observed for *Euglena gracilis* (93%). Acanthamoebae are notoriously difficult to identify by morphological criteria, and *A. lenticulata* contains distinct clinical and genetic 'types' that may represent cryptic species. Therefore, a similarity value of 98% might be more appropriate for species-level distinctions among most protists. For the data examined in the present analysis using the MESA program, use of a 98% similarity value for calling OTUs resulted in a 1.7-fold greater number of OTUs, while a value of 99% yielded a 2.5-fold greater number (Fig. 6).

The use of a similarity value greater than 95% for the inter-species analysis carried out with full-length 18S sequences from GenBank resulted in better agreement between the number of OTUs called by the program and the number of congeneric species retrieved from GenBank (Fig. 3). However, use of a similarity value greater than 95% in

the MESA program also rapidly increased the frequency of placing strains of a single species into multiple OTUs in the intra-species comparisons (Fig. 2). A value of 95% was chosen in the present study because it represented a relatively conservative value for demarcating protistan OTUs in environmental sequence databases. An adjustment to the threshold value can be easily accommodated in the program, and the use of a range of similarity values may provide interesting insights into the microbial eukaryotic diversity present in a sample.

Molecular diversity studies of natural protistan assemblages conducted to date have employed a variety of methods, and a variety of sequence similarity values for calculating the number of OTUs in 18S clone libraries. One approach has been to generate RFLP patterns from the full-length 18S genes in a clone library and then group the clones into taxonomic units based on their RFLP patterns (25, 42). This method has been employed to reduce the number of clones that need to be sequenced. A more common approach has been to partially sequence and align a large number of clones and apply a specific sequence similarity value to group sequences into OTUs. A range of similarity values of 95-98% has generally been employed (20, 21, 59, 77, 79). Worden (90) examined OTU-calling at four different values of sequence similarity (range of ≥ 96-99), Doherty et al., (26) used a range of 88 to 99%, and Jeon et al. (40) examined the number of OTUs across a wide range of similarity values (50-99%). We are aware of no analysis of the type that we have conducted here to provide rationale to the specific value employed. The justifications for these values in other studies have often been omitted, but the implication is that they approximate species-level distinctions. This level of discrimination seems vaguely linked to empirical observations that SSU rDNA (16S) sequences of bacterial species differ by values on the order of 1-2% for well-aligned sequences. Our results provide the first analysis of 18S sequences for species of protists that have been identified by traditional approaches.

Calling OTUs for the North Atlantic and North Pacific clone libraries. The application of the MESA program to a large environmental clone library allowed automated processing of 2,207 sequences contained in the data set. The rank abundance curve generated from that data set is indicative of the results of the program (Fig. 5). The generation of the matrix of pair-wise alignments consumed the majority of the processing

time and is, of course, highly dependent on the total number of sequences and processor speed. Calling OTUs required comparatively little time. The program yielded 388 OTUs from the combined Atlantic/Pacific sequence database at approximately the level of morphospecies. A significant number of the OTUs most closely matched metazoan taxa in the BLAST analysis, particularly copepods (Arthropoda). Prescreening samples through 200 μm Nitex mesh did not remove these species. This result indicates that future iterations of the MESA program for 18S environmental libraries must take into account appropriate demarcations for metazoan taxa as well as protists if the approach is to have general applicability for ecological studies of microbial eukyarotes.

The shape of the rank abundance curve of OTUs generated by the program indicated the presence of a very large number of 'rare' OTUs (OTUs comprised of one or two clones) in the combined data set (Fig. 5). A relatively low percentage of OTUs (14%; 54 out of 388) were observed at both study sites. It is not uncommon in comparing environmental clone libraries from different locales that 'rare' taxa constitute the majority of the OTUs, and that these rare taxa tend to be different at different locales (61). This finding may indicate the existence of endemism among protistan species, but it is important to note that approximately half of the 50 most common phylotypes were observed at both oceanic sites within the limited databases generated in the present study. The presence of different rare taxa in the North Atlantic and North Pacific samples may simply indicate that there is very high local species richness for microbial communities, and severe undersampling at a given site cannot accurately assess the presence/absence of rare taxa. In addition, differences in environmental conditions and sampling depths presumably resulted in differences in relative abundance among the taxa at the two sites. It has been reported that minor changes in environmental conditions during bottle incubations resulted in rapid changes in the protistan assemblage at the North Atlantic site (21). The inability of other molecular diversity studies to attain sampling saturation supports this conjecture (20, 21, 47, 76, 91).

Finally, it is noteworthy that the use of 95% sequence similarity in the MESA program generated 388 unique OTUs from 2,207 sequences. Application of a higher value resulted in substantially more unique OTUs. The overall conclusion from this finding is that, if protistan taxonomists generally accept the incorporation of

physiological and behavioral data into the present morphological species concept employed for these taxa, then estimates of species richness of natural protistan assemblages could be dramatically higher then predicted by the use of a similarity value of 95%. A molecular taxonomy holds the most promise for ecologists to deal with this staggering diversity of forms and functions.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance of the captains and crews of the R/V Endeavor and R/V Seawatch for technical assistance with sample collection. We thank D. Beaudoin, J.M. Rose, R. Schaffner and M. Travao for assistance with the collection and processing of the samples from the North Atlantic and North Pacific, and K.B. Heidelberg for helpful comments on the manuscript. Support for this manuscript was provided by National Science Foundation grants MCB-0732066, MCB-0703159 and OCE-0550829 and a grant from the Gordon and Betty Moore Foundation. The MESA program described in this study is available for download on-line at http://www.XXXXXXXXXXXXXX.

REFERENCES

1. **Adl, S. M., A. G. B. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. Brugerolle, R. A. Fensome, S. Fredericq, T. Y. James, S. Karpov, P. Kugrens, J. Krug, C. E. Lane, L. A. Lewis, J. Lodge, D. H. Lynn, D. G. Mann, R. M. McCourt, L. Mendoza, O. Moestrup, S. E. Mozley-Standridge, T. Nerad, C. A. Shearer, A. V. Smirnov, F. W. Spiegel, and M. F. J. R. Taylor.** 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. Journal of Eukaryotic Microbiology **52:**399-451.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:**3389-3402.
3. **Andersen, R. A., and J. C. Bailey.** 2002. Phylogenetic analysis of 32 strains of Vaucheria (Xanthophyceae) using the rbcL gene and its two flanking spacer regions. Journal Of Phycology **38:**583-592.
4. **Andersen, R. A., R. W. Brett, D. Potter, and J. P. Sexton.** 1998. Phylogeny of the Eustigmatophyceae based upon 18S rDNA, with emphasis on *Nannochloropsis*. Protist **149:**61-74.
5. **Andersen, R. A., Y. Van de Peer, D. Potter, J. P. Sexton, M. Kawachi, and T. LaJeunesse.** 1999. Phylogenetic analysis of the SSU rRNA from members of the Chrysophyceae. Protist **150:**71-84.
6. **Audemard, C., K. S. Reece, and E. M. Burreson.** 2004. Real-time PCR for detection and quantification of the protistan parasite *Perkinsus marinus* in environmental waters. Applied and Environmental Microbiology **70:**6611-6618.
7. **Baldauf, S. L.** 2003. The Deep Roots of Eukaryotes. Science **300:**1703-1706.
8. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler.** 2004. GenBank update. Nucleic Acids Research **32:**D23-D26.
9. **Blaxter, M. L.** 2004. The promise of a DNA taxonomy. Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences **359:**669-679.
10. **Boenigk, J., S. Jost, T. Stoeck, and T. Garstecki.** 2006. Differential thermal adaptation of clonal strains of a protist morphospecies originating from different climatic zones. Environmental Microbiology **9:**593-602.

11. **Boenigk, J., K. Pfandl, P. Stadler, and A. Chatzinotas.** 2005. High diversity of the 'Spumella-like' flagellates: an investigation based on the SSU rRNA gene sequences of isolates from habitats located in six different geographic regions. Environmental Microbiology **7:**685-697.
12. **Bowers, H. A., T. Tengs, H. B. Glasgow, Jr., J. M. Burkholder, P. A. Rublee, and D. W. Oldach.** 2000. Development of real-time PCR assays for rapid detection of *Pfiesteria piscicida* and related dinoflagellates. Applied and Environmental Microbiology **66:**4641-4648.
13. **Brinkmann, H., M. van der Giezen, and Y. Zhou.** 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology **54:**743-757.
14. **Brown, S., and J. F. De Jonckheere.** 1999. A reevaluation of the amoeba genus *Vahlkampfia* based on SSUrDNA sequences. European Journal Of Protistology **35:**49-54.
15. **Burki, F., K. Shalchian-Tabrizi, Å. Skjaeveland, S. I. Nikolaev, K. S. Jakobsen, and J. Pawlowski.** 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One **8:**E790.
16. **Caron, D. A.** 2009. Protistan biogeography: why all the fuss? Journal of Eukaryotic Microbiology **56:**105-112.
17. **Coleman, A. W.** 2002. Microbial eukaryote species. Science **297:**337.
18. **Countway, P. D.** 2005. Molecular ecology of marine protistan assemblages. University of Southern California, Los Angeles.
19. **Countway, P. D., and D. A. Caron.** 2006. Abundance and distribution of *Ostreococcus* sp. in the San Pedro Channel, California (USA) revealed by qPCR. Applied and Environmental Microbiology **72:**2496-2506.
20. **Countway, P. D., R. J. Gast, M. R. Dennett, P. Savai, J. M. Rose, and D. A. Caron.** 2007. Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western N. Atlantic (Sargasso Sea and Gulf Stream). Environmental Microbiology **9:**1219-1232.

21. **Countway, P. D., R. J. Gast, P. Savai, and D. A. Caron.** 2005. Protistan diversity estimates based on 18S rDNA from seawater incubations in the western North Atlantic. Journal of Eukaryotic Microbiology **52:**95-106.
22. **Dawson, S. C., and N. R. Pace.** 2002. Novel kingdom-level eukaryotic diversity in anoxic environments. Proceedings of the National Academy of Sciences of the United States of America **99:**8324-8329.
23. **De Jonckheere, J. F.** 2004. Molecular definition and the ubiquity of species of the genus *Naegleria*. Protist **155:**89-103.
24. **Dewhirst, F. E., Z. Shen, M. S. Scimeca, L. N. Stokes, T. Boumenna, T. C. Chen, B. J. Paster, and J. G. Fox.** 2005. Discordant 16S and 23S rRNA gene phylogenies for the genus *Helicobacter*: implications for phylogenetic inference and systematics. Applied and Environmental Microbiology **187:**6106-6118.
25. **Diez, B., C. Pedros-Alio, and R. Massana.** 2001. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Applied and Environmental Microbiology **67:**2932-2941.
26. **Doherty, B. A. Costas, G. B. McManus, and L. A. Katz.** 2007. Culture independent assessment of planktonic ciliate diversity in coastal northwest Atlantic waters. Aquatic Microbial Ecology **48:**141-154.
27. **Ebach, M. C., and C. Holdrege.** 2005. More Taxonomy, not DNA barcoding. Bioscience **55:**822-823.
28. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32:**1792-1797.
29. **Edgcomb, V. P., D. T. Kysela, A. Teske, A. D. Gomez, and M. L. Sogin.** 2002. Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. Proceedings Of The National Academy Of Sciences Of The United States Of America **99:**7658-7662.
30. **Fawley, M. J., K. P. Fawley, and M. A. Buchheim.** 2004. Molecular diversity among communities of freshwater microchlorophytes. Microbial Ecology **48:**489-499.

31. **Fawley, M. W., K. P. Fawley, and H. A. Owen.** 2005. Diversity and ecology of small coccoid green algae from Lake Itasca, Minnesota, USA, including *Meyerella planktonica*, gen. et sp nov. Phycologia **44:**35-48.
32. **Galluzzi, L., A. Penna, E. Bertozzini, M. Vila, E. Garces, and M. Magnani.** 2004. Development of real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (dinoflagellate). Applied and Environmental Microbiology **70:**1199-1206.
33. **Gast, R. J., M. R. Dennett, and D. A. Caron.** 2004. Characterization of protistan assemblages in the Ross Sea, Antarctica by denaturing gradient gel electrophoresis. Applied and Environmental Microbiology **70:**2028-2037.
34. **Gifford, D. J., and D. A. Caron.** 1999. Sampling, preservation, enumeration and biomass of marine protozooplankton, p. 193-221, ICES zooplankton methodology manual. Academic Press, London.
35. **Green, J. L., B. J. M. Bohannan, and R. J. Whitaker.** 2008. Microbial biogeography: from taxonomy to traits. Science **320:**1039-.
36. **Guillou, L., W. Eikrem, M. J. Chretiennot-Dinet, F. Le Gall, R. Massana, K. Romari, C. Pedros-Alio, and D. Vaulot.** 2004. Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. Protist **155:**193-214.
37. **Handy, S. M., K. J. Coyne, K. J. Portune, E. Demir, M. A. Doblin, C. E. Hare, S. C. Cary, and D. A. Hutchins.** 2005. Evaluating vertical migration behavior of harmful raphidophytes in the Delaware Inland Bays utilizing quantitative PCR. Aquatic Microbial Ecology **40:**121-132.
38. **Harper, J. T., E. Waandersm, and P. J. Keeling.** 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. International Journal of Systematic and Evolutionary Microbiology **55:**587-496.
39. **Hoppenrath, M., and B. S. Leander.** 2006. Ebriid phylogeny and the expansion of the Cercozoa. Protist **157:**279-290.
40. **Jeon, S.-O., J. Bunge, T. Stoeck, K. J.-A. Barger, S.-H. Hong, and S. S. Epstein.** 2006. Synthetic statistical approach reveals a high degree of richness of microbial

eukaryotes in an anoxic water column. Applied and Environmental Microbiology **72:**6578-6583.

41. **Lee, J. J., G. F. Leedale, and P. Bradbury.** 2000. An illustrated guide to the protozoa. Allen Press, Inc., Lawrence.
42. **Lefranc, M., A. Thénot, C. Lepere, and D. Debroas.** 2005. Genetic diversity of small eukaryotes in lakes differing by their trophic status. Applied and Environmental Microbiology **71:**5935-5942.
43. **Lim, E. L.** 1996. Molecular identification of nanoplanktonic protists based on small subunit ribosomal RNA gene sequences for ecological studies. Journal of Eukaryotic Microbiology **43:**101-106.
44. **López_Garcia, P., A. Vereshchaka, and D. Moreira.** 2007. Eukaryotic diversity associated with carbonates and fluid-seawater interface in Lost City hydrothermal field. Environmental Microbiology **9:**546-554.
45. **López-García, P., H. Philippe, F. Gail, and D. Moreira.** 2003. Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. Proceedings Of The National Academy Of Sciences Of The United States Of America **100:**697-702.
46. **López-García, P., F. Rodríguez-Valera, C. Pedrós-Alió, and D. Moreira.** 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature **409:**603-607.
47. **Lovejoy, C., R. Massana, and C. Pedros-Alio.** 2006. Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. Applied and Environmental Microbiology **72:**3085-3095.
48. **Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüßmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.-H. Schleifer.** 2004. ARB: a software environment for sequence data. Nucleic Acids Research **32:**1363-1371.

49. **Martínez-Murcia, A. J., L. Soler, M. J. Saavedra, M. R. Chacón, J. Guarro, E. Stackebrandt, and M. J. Figueras.** 2005. Phenotypic, genotypic, and phylogenetic discrepancies to differentiate *Aeromonas salmonicida* from *Aeromonas bestiarum*. International Microbiology **8:**259-269.
50. **Maslov, D. A., S. J. Westenberger, X. Xu, D. A. Campbell, and N. R. Sturm.** 2007. Discovery and barcoding by analysis of spliced leader RNA gene sequences of new isolates of Trypanosomatidae from Heteroptera in Costa Rica and Ecuador. Journal of Eukaryotic Microbiology **54:**57-65.
51. **Massana, R., L. Guillou, B. Diez, and C. Pedros-Alio.** 2002. Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. Applied and Environmental Microbiology **68:**4554-4558.
52. **Medlin, L., H. J. Elwood, S. Stickel, and M. L. Sogin.** 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene **71:**491-499.
53. **Modeo, L., G. Petroni, G. Rosati, and D. J. S. Montagnes.** 2003. A multidisciplinary approach to describe protists: redescriptions of Novistrombidium testaceum and Strombidium inclinatum Montagnes, Taylor and Lynn 1990 (Ciliophora, Oligotrichia). Journal of Eukaryotic Microbiology **50:**175-189.
54. **Moon-van der Staay, S. Y., R. De Wachter, and D. Vaulot.** 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature **409:**607-610.
55. **Moorthi, S. D., P. D. Countway, B. A. Stauffer, and D. A. Caron.** 2006. Use of quantitative real-time PCR to investigate the dynamics of the red tide dinoflagellate Lingulodinium polyedrum. Microbial Ecology **52:**136-150.
56. **Moran, D. M., O. R. Anderson, M. R. Dennett, D. A. Caron, and R. J. Gast.** 2007. A description of seven Antarctic marine Gymnamoebae including a new species and a new genus: *Platyamoeba contorta* n. sp. and *Vermistella antarctica* n. gen. n. sp. Journal of Eukaryotic Microbiology **54:**169-183.
57. **Moreira, D., S. von der Heyden, D. Bass, P. López-Garcia, E. Chao, and T. Cavalier-Smith.** 2007. Global eukaryote phylogeny: combined small- and large-

subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. Molecular Phylogenetics and Evolution **44:**255-266.

58. **Nanney, D. L.** 1999. When is a rose?: the kinds of tetrahymenas, p. 93-118. *In* R. W. Wilson (ed.), Species: new interdisciplinary essays. MIT Press, Cambridge, MA.
59. **O'Brien, H. E., J. L. Parrent, J. A. Jackson, J.-M. Moncalvo, and R. Vilgalys.** 2005. Fungal community analysis by large-scale sequencing of environmental samples. Applied and Environmental Microbiology **71:**5544-5550.
60. **O'Mahony, E. M., W. T. Tay, and R. J. Paxton.** 2007. Multiple rRNA variants in a single spore of the microsporidian *Nosema bombi*. Journal of Eukaryotic Microbiology **54:**103-109.
61. **Pedrós-Alió, C.** 2006. Microbial diversity: can it be determined? Trends in Microbiology **14:**257-263.
62. **Poirot, O., E. O'Toole, and C. Notredame.** 2003. Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Research **31:**3503-3506.
63. **Popels, L. C., S. C. Cary, D. A. Hutchines, R. Forbes, F. Pustizzi, C. J. Gobler, and K. J. Coyne.** 2003. The use of quantitative polymerase chain reaction for the detection and enumeration of the harmful alga *Aureococcus anophagefferens* in environmental samples along the United States East Coast. Limnology and Oceanography **48:**92-102.
64. **Rubinoff, D., S. Cameron, and K. Will.** 2006. Genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. Journal of Heredity **97:**581-594.
65. **Scheckenbach, F., C. Wylezich, A. P. Mylnikov, M. Weitere, and H. Arndt.** 2006. Molecular comparsions of freshwater and marine isolates of the same morphospecies of heterotrophic flagellates. Applied and Environmental Microbiology **72:**6638-6643.
66. **Schlegel, M.** 1994. Molecular phylogeny of eukaryotes. Trends in Ecology and Evolution **9:**330-335.
67. **Schloss, P. D.** 2008. Evaluating different approaches that test whether microbial communities have the same structure. The ISME Journal **2:**265-275.

68. **Schloss, P. D., and J. Handelsman.** 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology **71:**1501-1506.
69. **Schloss, P. D., and J. Handelsman.** 2006. Introducing SONS, a tool for Operational Taxonomic unit-based comparisons of microbial community memberships and structures. Applied and Environmental Microbiology **72:**67773-6779.
70. **Scholin, C. A., D. M. Anderson, and M. L. Sogin.** 1993. Two distinct small-subunit ribosomal RNA genes in the North American toxic dinoflagellate *Alexandrium fundyense* (Dinophyceae). Journal of Phycology **29:**209-216.
71. **Scholin, C. A., K. R. Buck, T. Britschgi, G. Cangelosi, and F. P. Chavez.** 1996. Identification of *Pseudo-nitzschia australis* (Bacillariophyceae) using rRNA-targeted probes in whole cell and sandwich hybridization formats. Phycologia **35:**190-197.
72. **Simpson, A. G. B., and A. J. Roger.** 2004. The real 'kingdoms' of eukaryotes. Current Biology **14:**R693-696.
73. **Sogin, M. L.** 1991. Early evolution and the origin of eukaryotes. Current Opinion in Genetics and Development **1:**457-463.
74. **Sogin, M. L.** 1989. Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. American Zoologist **29:**487-499.
75. **Sogin, M. L., H. J. Elwood, and J. H. Gunderson.** 1986. Evolutionary diversity of eukaryotic small-subunit rRNA genes. Proceedings of the National Academy of Science **83:**1383-1387.
76. **Stoeck, T., and S. Epstein.** 2003. Novel eukaryotic lineages inferred from small-subunit rRNA analyses of oxygen-depleted marine environments. Applied and Environmental Microbiology **69:**2657-2663.
77. **Stoeck, T., B. hayward, G. T. Taylor, R. Varela, and S. S. Epstein.** 2006. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. Protist **157:**31-43.
78. **Stoeck, T., G. T. Taylor, and S. S. Epstein.** 2003. Novel eukaryotes from the permanently anoxic Cariaco Basin (Caribbean sea). Applied and Environmental Microbiology **69:**5656-5663.

79. **Stoeck, T., A. Zuendorf, A. Behnke, and H.-W. Breiner.** 2007. A molecular approach to identify active microbes in environmental eukaryote clone libraries. Microbial Ecology **53:**328-339.
80. **Takishita, K., H. Miyake, M. Kawato, and T. Maruyama.** 2005. Genetic diversity of microbial eukaryotes in anoxic sediment around fumaroles on a submarine caldera floor based on the small-subunit rDNA phylogeny. Extremophiles **9:**185-196.
81. **Tautz, D., P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler.** 2003. A plea for DNA taxonomy. Trends in Ecology and Evolution **18:**70-74.
82. **Tekle, Y. I., L. Wegener Parfrey, and L. A. Katz.** 2009. Molecular data are transforming hypotheses on the origin and diversification of eukaryotes. BioScience **59:**471-481.
83. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22:**4673-4680.
84. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research **22:**4673-4680.
85. **Tomas, C. R.** 1997. Identifying marine phytoplankton. Academic Press, San Diego.
86. **Uliana, S. R. B., M. H. T. Affonso, E. P. Camargo, and L. M. Floeter-Winter.** 1991. *Leishmania*: genus identification based on a specific sequence of the 18S ribosomal RNA sequence. Experimental Parasitology **72:**157-163.
87. **Uliana, S. R. B., K. Nelson, S. M. Beverley, E. P. Camargo, and L. M. Floeter-Winter.** 1994. Discrimination amongst *Leishmania* by polymerase chain reaction and hybridization with small subunit ribosomal DNA derived oligonucleotides. Journal of Eukaryotic Microbiology **41:**324-401.
88. **Weekers, P. H. H., R. J. Gast, P. A. Fuerst, and T. J. Byers.** 1994. Sequence variations in small-subunit ribosomal RNAs of *Hartmannella vermiformis* and their phylogenetic implications. Molecular Biology and Evolution **11:**684-690.

89. **Whipps, C. M., and M. L. Kent.** 2006. Phylogeography of the cosmopolitan marine parasite *Kudoa thyrsites* (Myxozoa: Myxosporea). Journal of Eukaryotic Microbiology **53:**364-373.
90. **Worden, A. Z.** 2006. Picoeukaryote diversity in coastal waters of the Pacific Ocean. Aquatic Microbial Ecology **43:**165-175.
91. **Zuendorf, A., J. Bunge, A. Behnke, K. J.-A. Barger, and T. Stoeck.** 2006. Diversity estimates of microbial eukaryotes below the chemocline of the anoxic Mariager Fjord, Denmark. FEMS Microbiology Ecology **58:**476-491.

Figure Legends

Figure 1. Flow diagram of the Microbial Eukaryote Species Assignment algorithm for calling protistan OTUs (Operational Taxonomic Units) using full-length 18S rRNA gene sequences.

Figure 2. Cumulative intra-specific sequence similarity determined at the 95% similarity level for 211 full-length small subunit ribosomal RNA gene sequences distributed among 17 species (see Table 1; total pairwise comparisons = 2,712).

Figure 3. Cumulative intra-generic, inter-specific sequence similarity determined at the 95% similarity level for 323 full-length small subunit ribosomal RNA gene sequences distributed among 31 genera (see Table 2; total pairwise comparisons = 2,439).

Figure 4. OTU calling on replicate clone libraries. The clone libraries (A,B) were established from DNA subsamples taken from the same water sample collected from the North Atlantic. Libraries were constructed independently, but the sequences were combined into a single data set for OTU calling (C). Note different axes for A, B and C. OTU rank order differs from panel to panel, with the overlap of common OTUs shown in (C).

Figure 5. OTUs established for sequences obtained from combined environmental clone libraries constructed for samples collected from the western North Atlantic and eastern North Pacific.

Figure 6. Effect of the choice of sequence similarity value on the number of OTUs estimated from sequences obtained from combined environmental clone libraries collected in the western North Atlantic and eastern North Pacific.

Table 1. List of 17 species whose complete 18S sequences were obtained from GenBank and used to examine intra-species sequence variability within full-length small subunit ribosomal RNA (18S) genes.

Species	Number of Strains	Average % Intra-species Similarity
Acanthamoeba castellanii	12	0.965
Acanthamoeba lenticulata	12	0.849
Alexandrium catenella	16	0.999
Alexandrium tamarense	37	0.982
Chlamydomonas noctigama	6	0.995
Entamoeba histolytica	4	0.990
Euglena gracilis	6	0.931
Euglena mutabilis	6	0.951
Euplotes aediculatus	4	0.999
Euplotes vannus	4	0.987
Giardia intestinalis	9	0.975
Gymnodinium beii	5	0.997
Nannochloropsis gaditana	10	0.999
Phaeocystis globosa	8	0.996
Plasmodium knowlesi	12	0.970
Thalassiosira rotula	4	0.994
Trypanosoma cruzi	56	0.984
Total: 17 Species	211	Average = 0.980

Table 2. List of 31 genera employed to examine inter-species sequence variability within full-length small subunit ribosomal RNA (18S) genes. Species in the genera noted by italics (*) were employed in the intra-species comparison in Table 1.

Genus	Number of Species	Average % Inter-species Similarity
*Acanthamoeba**	20	0.830
*Alexandrium**	13	0.946
Amphidinium	8	0.886
Bodo	7	0.852
Chaetoceros	5	0.932
*Chlamydomonas**	26	0.942
Chrysochromulina	8	0.960
Cryptomonas	8	0.821
Dinophysis	5	0.985
*Entamoeba**	12	0.760
*Euglena**	29	0.704
*Euplotes**	13	0.913
*Giardia**	4	0.895
*Gymnodinium**	6	0.953
Gyrodinium	10	0.954
Leishmania	5	0.996
Mallomonas	9	0.963
*Nannochloropsis**	6	0.988
Oxytricha	4	0.947
Paramecium	13	0.936
Paraphysomonas	6	0.914
*Phaeocystis**	5	0.973
*Plasmodium**	12	0.865
Prorocentrum	9	0.931
Pyramimonas	6	0.974

Table 2 (continued).

Genus	Number of Species	Average % Inter-species Similarity
Scrippsiella	3	0.983
Synura	6	0.959
Tetrahymena	17	0.988
*Thalassiosira**	8	0.950
Tintinnopsis	4	0.944
*Trypanosoma**	36	0.878
Total: 31	323	Average = 0.870

Table 3. Summary of microbial eukaryote OTU distributions among Pacific and Atlantic clone libraries, indicating the numbers of OTUs that were unique to either Pacific or Atlantic libraries or present in libraries from both oceans. OTUs are organized according to Teckle et al. (82). Uncl. = Unclassified; Unres. = Unresolved

Supergroup	OTU Category	Total OTU	Unique to Pacific database	Unique to Atlantic database	Combined Pacific/Atlantic databases
'Rhizaria'	Polycystinean	11	11	0	0
	Acantharean	9	9	0	0
	Sticholonchid	6	6	0	0
	Cercozoan	5	2	3	0
'Chromalveolata'	Stramenopile	51	22	21	8
	Ciliate	38	11	16	11
	Dinoflagellate	26	16	7	3
	Apicomplexan	1	0	1	0
	Haptophyte	4	2	1	1
	Cryptophyte	2	0	1	1
	Group I Alveolate	21	19	1	1
	Group II Alveolate	48	30	11	7
	Uncl. Alveolate	76	51	15	10
	Perkinsean	1	1	0	0
'Plantae'	Chlorophyte	11	4	5	2
	Rhodophyte	3	1	2	0
	Streptophyte	2	2	0	0
'Excavata'	Euglenozoan	15	12	3	0
'Opisthokonta'	Arthropod	27	10	12	5
	Cnidarian	5	2	2	1
	Ctenophore	3	2	0	1
	Echinodermate	1	0	1	0
	Urochordate	4	2	1	1
	Choanoflagellate	5	4	1	0
	Fungi	5	3	0	2

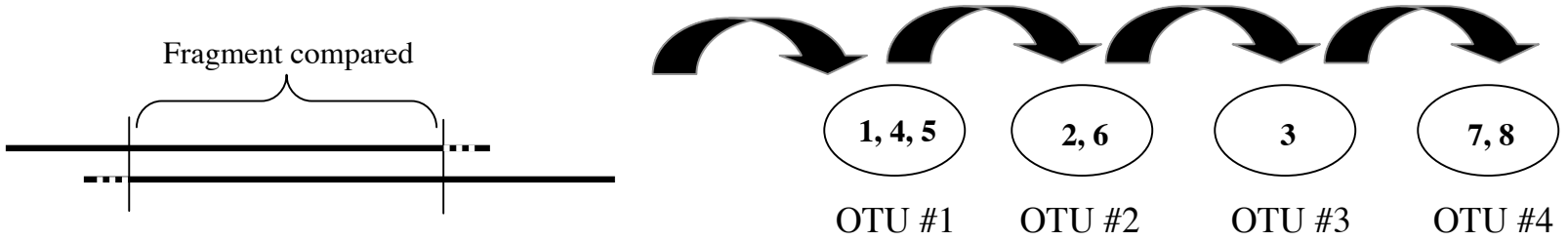
Table 3. (continued)

Supergroup	OTU Category	Total OTU	Unique to Pacific database	Unique to Atlantic database	Combined Pacific/Atlantic databases
	Uncl. Metazoan	1	0	1	0
Unres. lineages	Cryothecomonad	1	0	1	0
	Ichthyosporean	1	1	0	0
Unknown	Uncl. Eukaryote	5	5	0	0
TOTAL		388	228	106	54

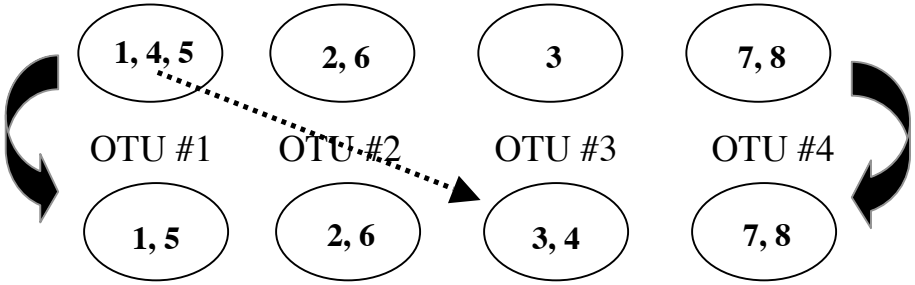
Table 4. Taxonomic groupings of the most abundant OTUs in the North Atlantic and North Pacific dataset. Alv. = Alveolate; Uncl. = Unclassified; Rhiz. = Rhizaria; Chrom. = Chromalveolata; Plan. = Plantae; Opist. = Opisthokonta.

Rank	Taxonomic Group	Rank	Taxonomic Group
1	Arthropod (Opist.)	26	Ciliate (Chrom.)
2	Dinoflagellate (Chrom.)	27	Chlorophyte (Plan.)
3	Cnidarian (Opist.)	28	Arthropod (Opist.)
4	Ciliate (Chrom.)	29	Ciliate (Chrom.)
5	Ciliate (Chrom.)	30	Stramenopile (Chrom.)
6	Group II Alv. (Chrom.)	31	Dinoflagellate (Chrom.)
7	Group II Alv. (Chrom.)	32	Group II Alv. (Chrom.)
8	Arthropod (Opist.)	33	Group II Alv. (Chrom.)
9	Ctenophore (Opist.)	34	Group II Alv. (Chrom.)
10	Acantharean (Rhiz.)	35	Stramenopile (Chrom.)
11	Group II Alv. (Chrom.)	36	Ciliate (Chrom.)
12	Arthropod (Opist.)	37	Uncl. Alv. (Chrom.)
13	Polycystinean (Rhiz.)	38	Group I Alv. (Chrom.)
14	Uncl. Alv. (Chrom.)	39	Urochordate (Opist.)
15	Ciliate (Chrom.)	40	Ciliate (Chrom.)
16	Arthropod (Opist.)	41	Group I Alv. (Chrom.)
17	Uncl. Alv. (Chrom.)	42	Arthropod (Opist.)
18	Uncl. Alv. (Chrom.)	43	Ciliate (Chrom.)
19	Group I Alv. (Chrom.)	44	Arthropod (Opist.)
20	Chlorophyte (Plan.)	45	Stramenopile (Chrom.)
21	Arthropod (Opist.)	46	Uncl. Alv. (Chrom.)
22	Uncl. Alv. (Chrom.)	47	Ciliate (Chrom.)
23	Chlorophyte (Plan.)	48	Uncl. Alv. (Chrom.)
24	Haptophyte (Chrom.)	49	Sticholonchid (Rhiz.)
25	Arthropod (Opist.)	50	Group I Alv. (Chrom.)

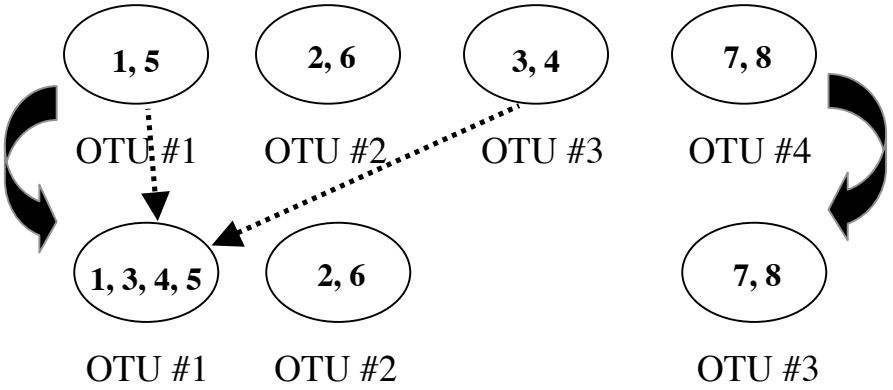
• **Formation:** Align and truncate sequences. Begin OTU assignment. Group each sequence with existing OTU at ≥ 95% similarity, or form a new OTU.

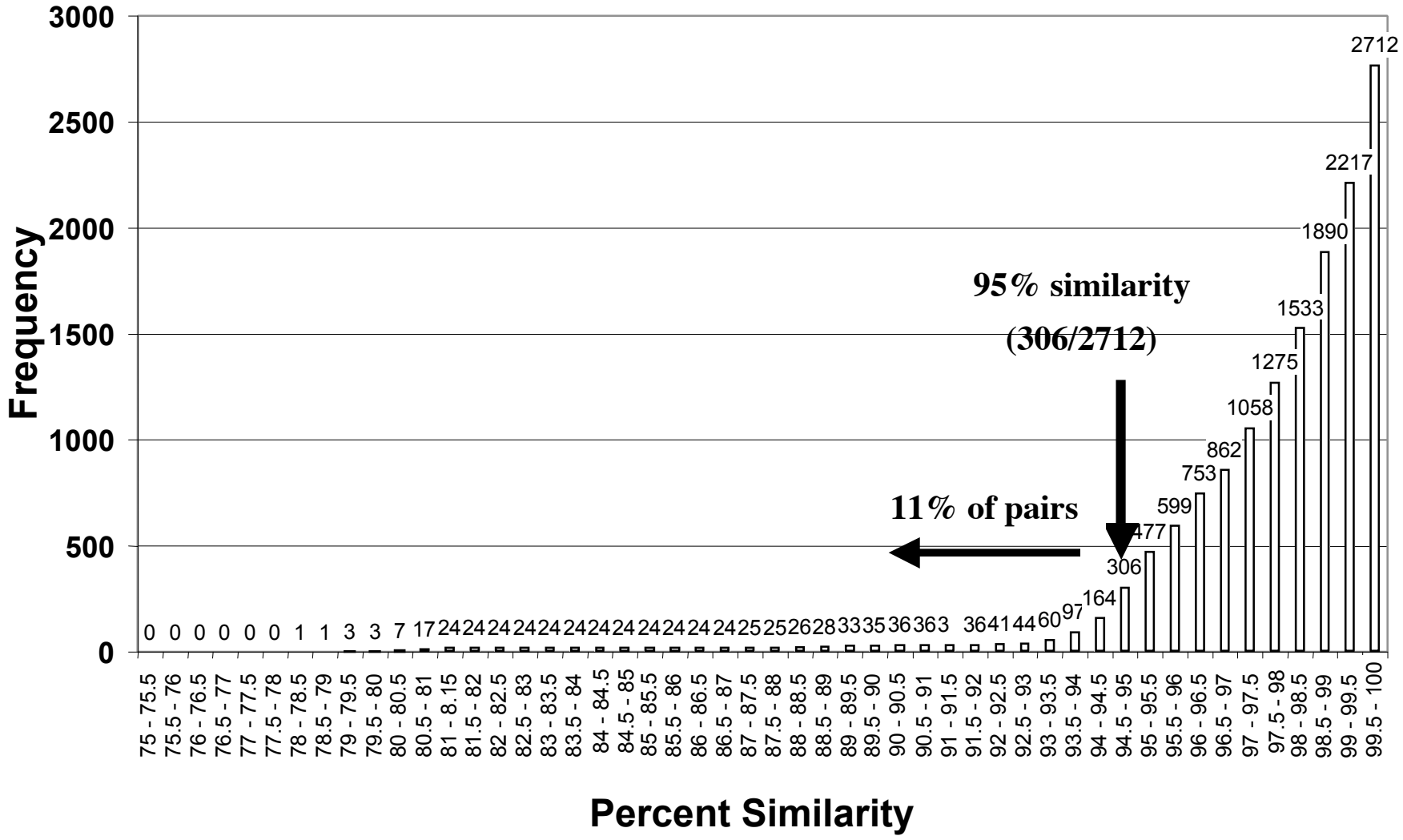


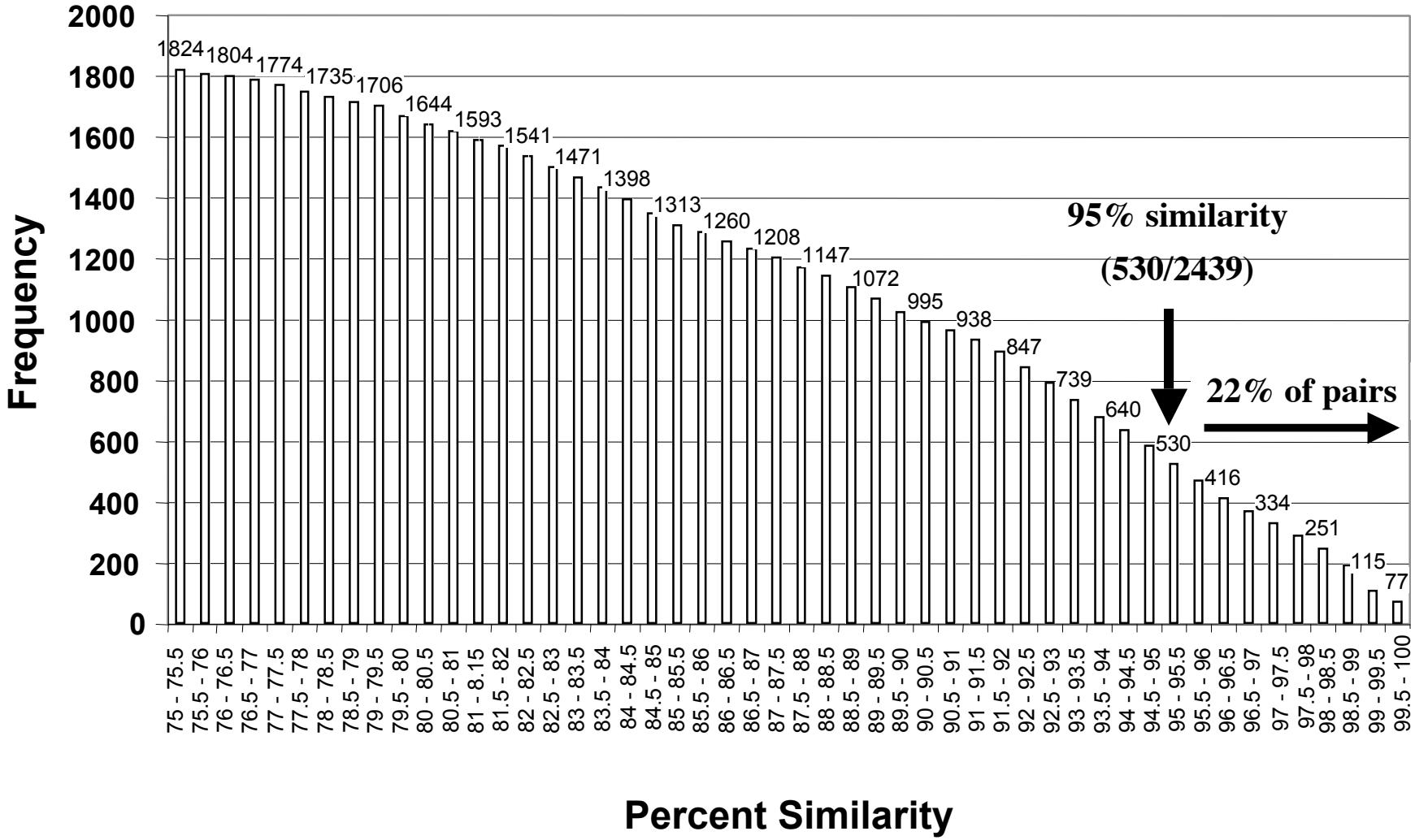
• **Optimization**: Compare each sequence in a given OTU with the sequences from all other OTUs. Redistribute to a different OTU based on highest average similarity.

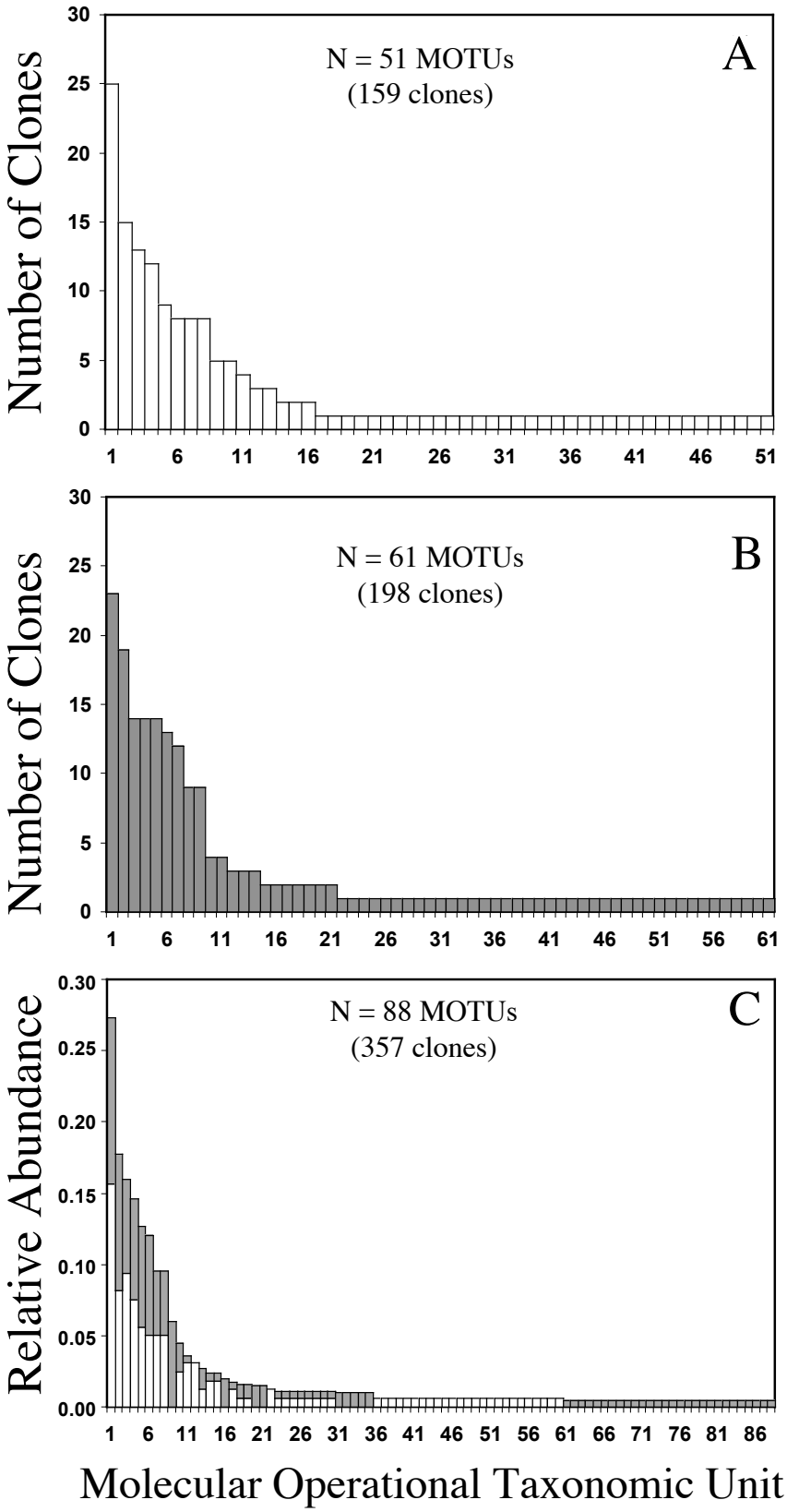


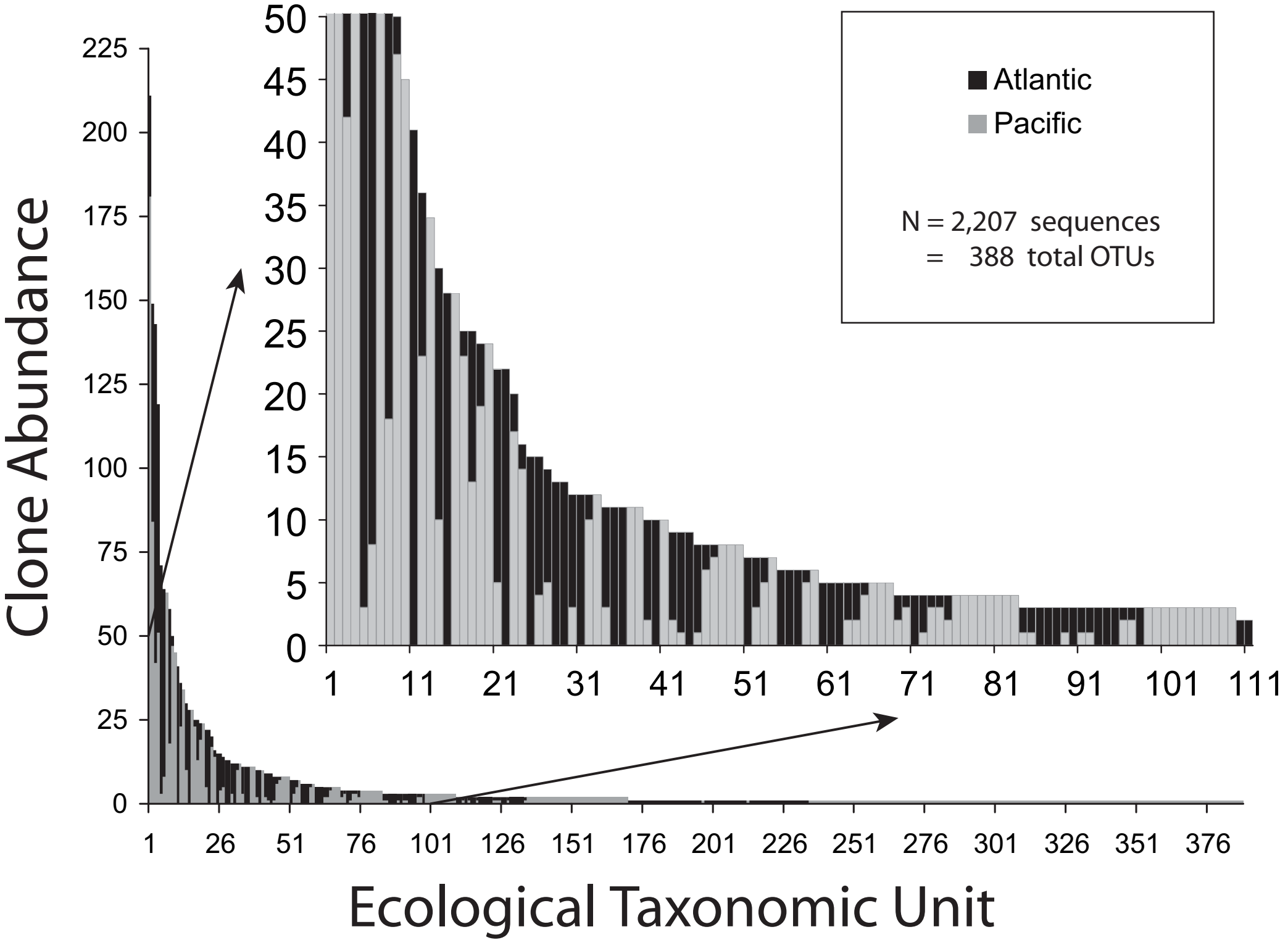
• **Condensation**: Determine the average pair-wise sequence similarity between sequences in pairs of OTUs. If overall average sequence similarity between two OTUs is ≥ 95%, merge OTUs.

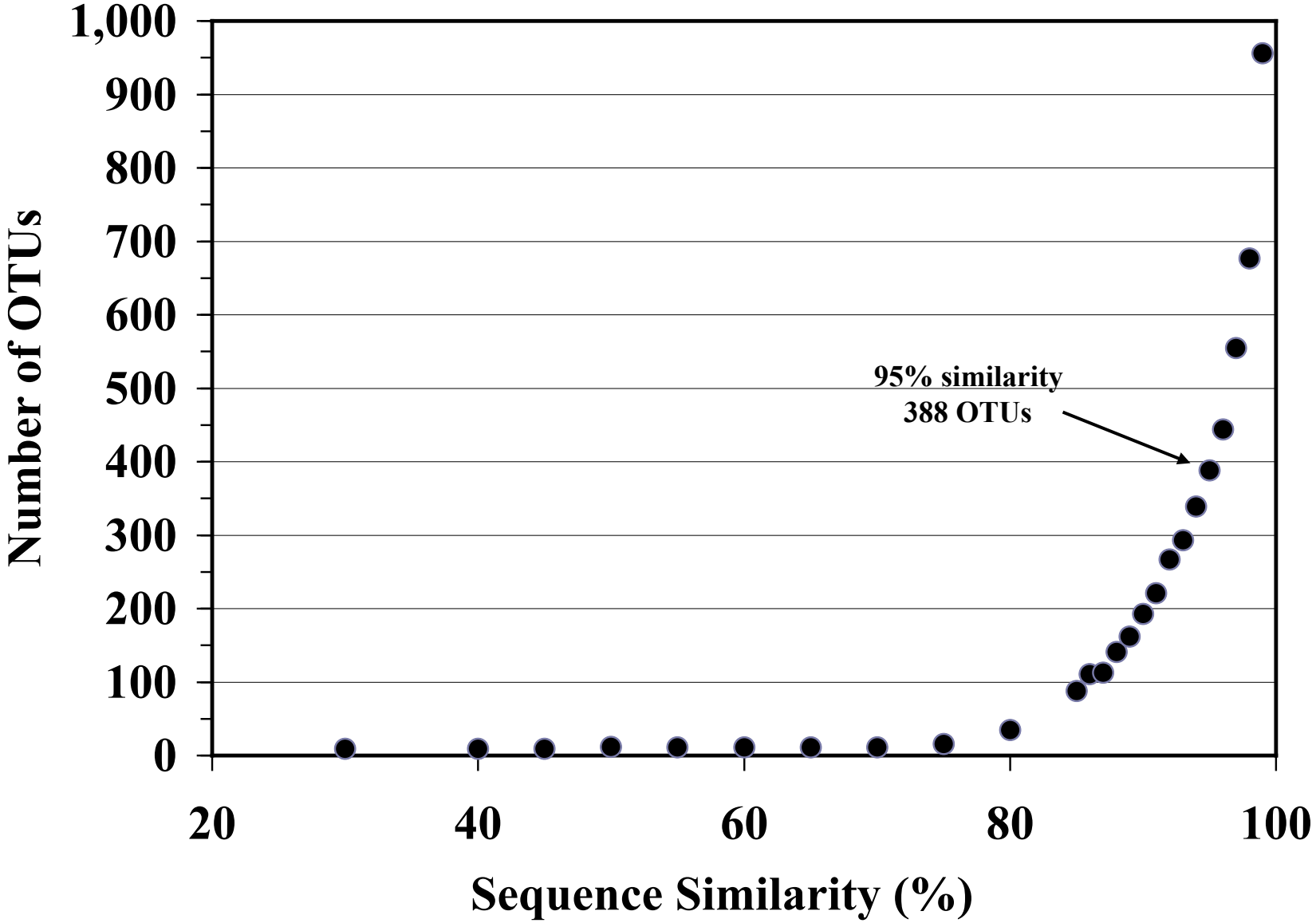



Frequency
3000
2500
2000
1500
1000
500
0
95% similarity
(306/2712)
11% of pairs
0 0 0 0 0 0 1 1 3 3 7 17 24 24 24 24 24 24 24 24 24 24 24 24 24 25 25 26 28 33 35 36 36 3 36 41 44 60 97 164 306 477 599 753 862 1058 1275 1533 1890 2217 2712
75 - 75.5
75.5 - 76
76 - 76.5
76.5 - 77
77 - 77.5
77.5 - 78
78 - 78.5
78.5 - 79
79 - 79.5
79.5 - 80
80 - 80.5
80.5 - 81
81 - 8.15
81.5 - 82
82 - 82.5
82.5 - 83
83 - 83.5
83.5 - 84
84 - 84.5
84.5 - 85
85 - 85.5
85.5 - 86
86 - 86.5
86.5 - 87
87 - 87.5
87.5 - 88
88 - 88.5
88.5 - 89
89 - 89.5
89.5 - 90
90 - 90.5
90.5 - 91
91 - 91.5
91.5 - 92
92 - 92.5
92.5 - 93
93 - 93.5
93.5 - 94
94 - 94.5
94.5 - 95
95 - 95.5
95.5 - 96
96 - 96.5
96.5 - 97
97 - 97.5
97.5 - 98
98 - 98.5
98.5 - 99
99 - 99.5
99.5 - 100
Percent Similarity


Frequency
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1824
1804
1774
1735
1706
1644
1593
1541
1471
1398
1313
1260
1208
1147
1072
995
938
847
739
640
530
416
334
251
115
77
95% similarity
(530/2439)
22% of pairs
75 - 75.5
75.5 - 76
76 - 76.5
76.5 - 77
77 - 77.5
77.5 - 78
78 - 78.5
78.5 - 79
79 - 79.5
79.5 - 80
80 - 80.5
80.5 - 81
81 - 8.15
81.5 - 82
82 - 82.5
82.5 - 83
83 - 83.5
83.5 - 84
84 - 84.5
84.5 - 85
85 - 85.5
85.5 - 86
86 - 86.5
86.5 - 87
87 - 87.5
87.5 - 88
88 - 88.5
88.5 - 89
89 - 89.5
89.5 - 90
90 - 90.5
90.5 - 91
91 - 91.5
91.5 - 92
92 - 92.5
92.5 - 93
93 - 93.5
93.5 - 94
94 - 94.5
94.5 - 95
95 - 95.5
95.5 - 96
96 - 96.5
96.5 - 97
97 - 97.5
97.5 - 98
98 - 98.5
98.5 - 99
99 - 99.5
99.5 - 100
Percent Similarity


A
N = 51 MOTUs
(159 clones)
Number of Clones
B
N = 61 MOTUs
(198 clones)
Number of Clones
C
N = 88 MOTUs
(357 clones)
Relative Abundance
Molecular Operational Taxonomic Unit


Clone Abundance
225
200
175
150
125
100
75
50
25
0
1
26
51
76
101
126
151
176
201
226
251
276
301
326
351
376
50
45
40
35
30
25
20
15
10
5
0
1
11
21
31
41
51
61
71
81
91
101
111
Atlantic
Pacific
N = 2,207 sequences
= 388 total OTUs
Ecological Taxonomic Unit


1,000
900
800
700
600
500
400
300
200
100
0
Number of OTUs
95% similarity
388 OTUs
20
40
60
80
100
Sequence Similarity (%)