Assembling the Tree of Life:
Research Needs in Phylogenetics and Phyloinformatics

Report from a Workshop
Yale University, July 28 and 29, 2000

Sponsored by the National Science Foundation
(Proposal 0089975 to the American Museum of Natural History)

EXECUTIVE SUMMARY

Science and society would benefit enormously from detailed and well-supported knowledge of the Tree of Life (TOL). At present, however, we know relatively little about the phylogenetic relationships of most of the species on Earth, or even among many of the major branches of the Tree. Fortunately, we have reached a turning point. Owing to fundamental theoretical advances, to the development of powerful analytical tools, and to the availability of major new sources of data, we now have the power to assemble the entire TOL. Much of this task can be accomplished within the next two decades, given sufficient vision, resources, and coordination.

Despite the potential to accomplish this task, and the great value of the product, this goal will not be achieved if traditional approaches are maintained. The growth of phylogenetic knowledge is too haphazard, and the resources, including human resources, are too limited. Given the magnitude of the problem, and the size of the datasets that will be generated, current funding and infrastructure for research are clearly inadequate. A TOL initiative requires the development and widespread use of new tools to gather and analyze massive phylogenetic datasets. It also requires computational tools and infrastructure to archive the results of phylogenetic research in a form that will allow comparison and synthesis, and to render the results widely accessible and valuable to the user community.

A workshop was held at Yale University on July 28 and 29, 2000, to provide advice to the National Science Foundation on research and funding strategies that would significantly advance the assembly of the Tree of Life. A group of 28 participants, broadly representing the phylogenetics research community, made the following major recommendations:

1. Design and implement new mechanisms to advance a TOL initiative:

- Support Tree of Life Research Networks, or TOLNets, to coordinate multi-investigator phylogenetic research programs.

- Support Tree of Life Research Centers, or TOLCenters, to foster the integration and synergy of TOLNets and yield efficiencies in research and training.

- Support, through TOLNets and TOLCenters,

  (a) the assembly and analysis of high quality data, including molecular and morphological characters, on as many taxa as possible, including extant and extinct organisms.

  (b) the development and maintenance of the necessary technical staff, the infrastructure for research and informatics, and the collection, curation, and preservation of the specimens and materials from which data are obtained.

2. Expand support to develop new phylogenetic methods and tools to:

- Rapidly and efficiently obtain phylogenetically relevant evidence, and to better document and archive these data.
- Meet the challenge of increasingly large datasets, with special attention to the combination of datasets from disparate sources.
- Allow the assembly of "supertrees" from component trees with partially overlapping samples of taxa.
- Assess the robustness of phylogenetic hypotheses.
- Optimally represent and visualize phylogenetic knowledge, and navigate through the TOL and between it and other data resources.

3. Establish a "phyloinformatics" infrastructure to ensure:

- Development and long-term growth and maintenance of a comprehensive database of phylogenetic knowledge, including character information and trees.
- Interconnectivity and interoperability between the phylogenetic knowledgebase and a wide variety of other data resources to facilitate data mining.
- Coordination of phylogenetic and phyloinformatics research, and outreach to the user community and the public.

4. Sponsor two additional workshops to:

- Evaluate and refine the structure of the TOL initiative, and the research objectives and needs of such a program within NSF, with additional input from the coordinators of other relevant programs.
- Explore in depth the research and infrastructural needs of the TOL phyloinformatics program, with additional input from the computer science and bioinformatics communities.

Table of Contents

I. Introduction

A. The Tree of Life as a Megascience Question

Modern evolutionary biology, with its concern for understanding the history of life, began with the publication in 1859 of Darwin's *On the Origin of Species*. In that volume Darwin's sole illustration was a "phylogenetic tree" depicting the hierarchical relationships of 15 hypothetical species. Despite the power of this imagery, for the next 100 years surprisingly few attempts were made to represent the evolutionary relationships of groups of organisms in the form of a branching diagram, or tree, in part because it often seemed too speculative and the methods for building trees were primarily narrative, not analytical. During the last several decades, however, following the introduction of new theory and methods in the form of "phylogenetic systematics," or "cladistics," as well as the unprecedented increase in relevant character information, especially derived from DNA sequences, knowledge of phylogenetic relationships has expanded tremendously. We now have the power to assemble the Tree of Life (**TOL**).

Nevertheless, resolving the TOL is among the most complex scientific problems facing biology, presenting challenges that far exceed sequencing the human genome. Darwin's diagram depicted relationships among 15 species, but the entire TOL is almost unimaginably vast. Presently, about 1.75 million species of organisms have been discovered and described, but it is estimated that millions or perhaps tens of millions remain to be discovered. This represents an extraordinary

challenge for resolving the TOL, since for every N species on the tree, there are N-1 genealogical groups of species, or clades, to be discovered.

To date we estimate that approximately 50,000 species have been placed in some form of phylogenetic tree, and a substantial percentage of the relationships that have been depicted are not yet well supported. The scale of the task confronting systematic biologists is therefore enormous. By all measures, however, we are making extraordinary progress. After all, phylogenetic analysis is almost entirely a phenomenon of the last three decades. Moreover, owing to new methods of analysis and an explosion of new sources of data, systematic biologists are now poised to assemble a detailed picture of the TOL within a reasonable time frame. The genetics community was in a very similar position a decade ago when the possibility of sequencing the entire human genome was seriously entertained. The time is right to assemble the TOL.
The societal need for a TOL is great and urgent, but a satisfactory understanding is not likely to emerge if we continue to pursue the problem in a piece-meal fashion. Accomplishing this task, and making the results of general use, will require new and innovative approaches to systematics research as well as new tools for the manipulation and dissemination of phylogenetic information. Articulating this new research initiative was the goal of the Yale Workshop.

B. Scientific and Societal Benefits from a Tree of Life

Organizing biological knowledge according to phylogenetic relationships has become increasingly important for many segments of science and society. People profit immensely from knowledge of phylogenetic relationships, and much of that benefit follows from the inferences that can be derived from a detailed chronicle of lineage branching events and character changes along the branches of the tree. Among many applications, knowledge of phylogenetic relationships has:

- provided a basis for all comparative studies in basic and applied biology by

    a. establishing which similarities were maintained from a common ancestor and which arose through convergent evolution
    b. inferring the direction of character change and adaptation
    c. providing a basis for analyzing rates of diversification and character change
    d. permitting more rational choice of experimental systems
    e. establishing the statistical independence of compared characteristics

- permitted the identification and place of origin of emerging diseases and their vectors (e.g., Hanta virus, Nipah virus, West Nile virus, HIV),

- enabled the reconstruction of the epidemiological history of disease transmission (e.g., HIV), as well as predicted locations of future viral outbreaks (Hanta virus)

- guided analyses of vaccine efficiency (e.g., in meningitis studies)

- provided new approaches to forensic analysis

- guided the search for new pharmaceuticals or biotechnological products

- permitted the identification of invasive pest species, their hosts and geographic origin, as well as potential biological control agents

- been used to reconstruct the history of functional changes in gene and protein sequences that are linked to patterns of development or to disease (e.g., cancer, color blindness)

- provided the comparative framework for bioinformatic knowledgebases (e.g., Genbank)

The use of phylogenetic information and methods is now pervasive throughout the biological sciences, and there is no question that society has gained significantly from our current, yet relatively incomplete, picture of the TOL. However, it is imperative to have increased knowledge of phylogenetic relationships to advance comparative developmental biology, physiology, and biochemistry, and to enable society to meet the challenges facing human health, threats to agriculture and forestry from invasive species and disease, and the management of our natural resources. Perhaps most importantly, without a substantial improvement in our knowledge of the TOL, managing, understanding, and manipulating biological information held in numerous databases worldwide, including the burgeoning information from genomic sciences, will become increasingly difficult and inefficient.

C. The Yale Workshop on the Tree of Life

Historical Context. In the early 1990's, with support from the National Science Foundation, a consortium of systematic biologists established Systematics Agenda 2000 as a mechanism to survey the research priorities and societal importance of systematic biology. This broad initiative within the systematics community defined three general goals, or missions, of systematics. In addition to inventorying life on Earth, the second and third missions —to gather and analyze comparative data across all organisms in order to construct a phylogenetic history of life, and then to use the hierarchical information contained in those phylogenetic hypotheses to create predictive information systems— comprise the core of the research initiative considered at the Yale Workshop.

Purpose and Goals of the Workshop. The objective of the Yale Workshop was to make recommendations to the National Science Foundation regarding the organization of a research initiative capable of assembling the TOL within a reasonable time frame, such as within 10-15 years, and also to consider how information from this research effort might be made available electronically to a global user community. This task was facilitated by addressing the following major topics:

1. The identification of current impediments to assembling the TOL over the short-term using conventional approaches.

2. The identification of institutional changes within the systematics community that would accelerate assembly of the TOL.

3. The identification of new approaches that would facilitate large-scale increases in phylogenetically informative data.

4. The identification of improvements in methods of data analysis to support a TOL initiative.

5. The identification of the bioinformatic challenges implied by a TOL initiative and new approaches to their solution.

Participants. The 28 Workshop participants are listed at the end of this report. These individuals were selected to represent the breadth of relevant disciplines (phylogenetic theory, molecular and morphological systematics, evolutionary biology, paleontology, genomics, bioinformatics) as well as major groups of organisms (bacteria, fungi, plants, invertebrate and vertebrate animals). Discussions took place in a series of plenary and breakout sessions. Deliberations and recommendations of the Workshop are discussed in the following sections of this report.

II. Impediments to Discovering the Tree of Life

The Workshop participants identified a number of impediments to assembling the TOL within a 10-15 year time frame assuming that current research approaches were maintained.

A. Research Strategies, Organization, and Human Resources

- The growth of phylogenetic knowledge for all major groups of organisms is too incremental and uncoordinated across research groups.

Although the individual investigator-driven approach has steadily improved our knowledge of the TOL, this research strategy alone will not meet the goal of a substantial representation of the TOL within 10-15 years. The problem is analogous to sequencing the human genome using only a series of single investigator grants; without large-scale cooperation and focus, the effort would likely be significantly incomplete to this day. Gathering phylogenetically informative character data is time-consuming and labor-intensive, even for relatively "small" and well-known taxonomic groups such as most major lineages of vertebrate animals. For diverse taxa, such as many invertebrate animal groups, the immensity of this effort scales upward.

Looked at from a TOL perspective, current research appears idiosyncratic. Individual investigators use different character systems or sets of taxa, generally with little coordination among themselves, thus there are many gaps in taxonomic coverage or character-systems across laboratories working on the same major group. In other cases there has been substantial duplication of effort. Many laboratories, for example, have sequenced the same genes for the same taxa. These situations impede building a TOL expeditiously and with maximal supporting evidence.

- Current funding opportunities within granting agencies, including NSF, encourage investigators to think small, whereas assembling the TOL requires the systematics community to think big.

As already noted, assembling the TOL is a megascience question. There are very few major groups of organisms whose relationships can be investigated adequately without extensive taxon sampling and large datasets (see below). Most systematists, however, are forced to restrict the scope of their investigations because grants for phylogenetics typically have funding restrictions and short funding periods. This contributes to incremental growth as well as phylogenetic results that are less robust than they would be if projects were more comprehensive in scope. Although there are some mechanisms in place at NSF to encourage collaboration [e.g., Research Coordination Networks (RCN); Integrated Research Challenges in Environmental Biology (IRC-EB)], these are still insufficient and underutilized for TOL activities.

• Human resources within phylogenetic research institutions and laboratories are inadequate to complete the TOL over a short to medium time-frame.

Although technologies of data acquisition and analysis continue to improve, phylogenetic research is still labor intensive. Much basic phylogenetic research is carried out by graduate students, postdoctoral fellows, and technicians. Yet, funding levels to support such positions within the individual investigator grant format are inadequate. These funding levels also have negative implications for training the next generation of systematic biologists.

B. Data Acquisition and Methods of Analysis

• Character evidence within and among groups is highly variable with respect to kind (molecules, morphology, etc.), quantity, and quality, which hinders broad-scale comparisons.

Many phylogenetic studies use only one kind of character data and either ignore or reject other sources of phylogenetic evidence that may be available. This has resulted in apparent conflicts among different sources of data. While we recognize the value of investigating phylogenetic signal in different datasets, each kind of data presents difficulties and ambiguities of analysis and interpretation. Well resolved phylogenetic relationships will depend ultimately on consilience across all relevant datasets.

• Phylogenetic analyses typically are inadequate with respect to the sample of relevant taxa.

Systematists now recognize a variety of problems associated with taxon sampling, and it is broadly appreciated that inadequate sampling can lead to spurious yet apparently well-supported results. For many major lineages the numbers of relevant taxa are in the tens of thousands. This constitutes a severe technical problem given current approaches to data capture and analysis. Other impediments include lack of access to appropriate specimens and inadequate human resources needed to increase sampling. In many major lineages we still know too little about species-level diversity. It is now abundantly clear, for example, that much of microbial diversity has not yet been discovered, which severely compromises our ability to infer the deepest branches of the TOL.

• Phylogenetic analyses typically are inadequate with respect to character data.

Most phylogenetic studies include too few characters to confidently resolve many relationships. Most systematic questions of large scope—for example, across the basal lineages of life, of arthropods, or even of birds and mammals—have been difficult to resolve, and will require much additional data. In the case of molecular data the more difficult problems may require tens of thousands of base pairs of sequence, rather than a few thousands, to achieve adequate resolution. Moreover, character data are sometimes inappropriate for the taxonomic level being investigated and provide little support for relationships. Thus, additional sources of evidence will be needed to resolve relationships with greater certainty and to account for both deep and more recent branching events.

• Current approaches to gathering character data are inadequate to generate the large comparative datasets needed to resolve the TOL.

If it is correct that large amounts of data are necessary to resolve the TOL, then current approaches to data gathering can be judged as inadequate for generating the character information

that will be necessary over the research time-frame considered in this report. New methods for gathering large amounts of comparative morphological and sequence data across a large number of taxa must be made broadly available to the systematics community.

- Current methods of data analysis are unable to meet the challenge of vastly increased taxon and character sampling.

A wide variety of algorithms are currently used for phylogenetic analysis of taxon-by-character data matrices. The pros and cons of these phylogenetic methods have received considerable attention, and it is clear that all methods are challenged in finding optimal (under a variety of criteria) solutions as the size of the problem increases dramatically. Participants in the Workshop agreed that at least for datasets uncomplicated by large amounts of missing data or high levels of homoplasy (convergent evolution, reversal, etc.), some current methods can provide adequate resolution for trees containing up to about 1000 taxa. Methods dependent on more complicated models of character-state change have difficulties finding solutions for far fewer taxa. Increasingly large datasets will require new computational infrastructure, as even the best desktop stand-alone computers will have difficulties resolving large trees within reasonable time periods. These observations raise serious theoretical and practical issues that need to be addressed if a TOL initiative is to be successful.

- Conceptual and computational challenges to combining and linking trees based on subsets of taxa and characters must be solved if the TOL is to be assembled across all groups.

Although large portions of the TOL may someday be reconstructed using a single data set of homologous character data, the vast diversity of taxa implies that a global representation of the TOL will be based on combining results from separate phylogenetic analyses, perhaps based on different types of data and overlapping in only a few taxa. Construction of combined trees, or supertrees, has generated considerable interest in recent years, yet we currently lack adequate methods and tools to generate such trees.

C. Phylogenetic Information Management and Dissemination

Phylogenetic knowledge must be rendered accessible and useful to the researchers, institutions, and government agencies who need it. Moreover, such knowledge must be in a form that it can serve as a framework to make bioinformatic databases more efficient and interactive. Major impediments include the following:

- Scientific results from the vast majority of phylogenetic research, including the original data and the resulting trees, are not being preserved in a readily accessible form.

This is due to the lack of adequate infrastructure and human resources to store this information, and to insufficient commitment to databasing such knowledge within the systematics community as well as the potential users of phylogenetic knowledge. A major problem is that the systematic community itself has generally failed to require the consistent archiving of phylogenetic results. An informal survey of the major systematics journals revealed that many phylogenetic data matrices are not available either in print or in some digital form, and that a very small percentage

of studies reported deposition of data in any publicly accessible database. In short, many phylogenetic studies are not replicable, and the data are effectively lost.

At present there are two major databases devoted primarily to the storage of phylogenetic knowledge: TreeBASE, http://www.herbaria.harvard.edu/treebase, and Tree of Life, http://phylogeny.arizona.edu/tree/phylogeny.html. Of these, the mission of TreeBASE is to store published data matrices and trees. Although TreeBASE now includes information on some 14,000 taxa, this is a small portion of the taxa that have already been included in some formal analysis, and a tiny percentage of the entire TOL. A major impediment identified by the Workshop was that as a TOL initiative scales upward, TreeBASE would be quickly overwhelmed with data (assuming that as part of this effort, systematists would consistently submit their results). The magnitude of the informatics problem is far larger if we also consider the inclusion of data on the specimens from which the character information was obtained. And, just as it is crucial to archive digitally the sequence alignments upon which phylogenetic analyses are based, morphological data require digital storage of images and of "synonymy tables" to ensure comparability of characters across different studies. If the TOL initiative is to succeed it will be crucial to create and support the necessary infrastructure to store phylogenetic information and render it accessible to the user community.

- Methods to compare and synthesize phylogenetic hypotheses derived from different data sources and analytical techniques are inadequate to meet the future needs of scientists and bioinformatics.

Differences in underlying data type (e.g., sequences, morphology) and in methods of analysis (e.g., distance, maximum likelihood, or maximum parsimony) present special difficulties for comparison and synthesis of phylogenetic evidence and trees. An information system using phylogenetic results as an organizing framework must be able to integrate congruent and incongruent hypotheses of relationships derived from a variety of sources in order to make data mining maximally efficient. Furthermore, since phylogenetic hypotheses frequently vary in their empirical support, information systems must be able to clearly reflect ambiguities and degrees of confidence in relationships.

- The potential for phylogenetic knowledge to provide a central organizing framework for biological information cannot currently be fulfilled due to insufficient infrastructure and computational tools.

Hierarchical classification systems reflecting phylogenetic relationships are predictive tools for organizing and exploring comparative biological data of all kinds. Phylogenetic knowledge must be linked to, and interoperate with, other biological databases, such as those for genetic data, species-, specimen-, and taxon-level information, as well as other biological and geophysical data. Biological data associated with trees must be accessible and analyzable in ways that generate entirely new information associations, linkages, and hypotheses. In short, an interactive knowledgebase of phylogenetic information should be able to harness the predictive power of phylogeny to expand comparative biological research for use by science and society.

As noted, we now lack comprehensive databases of phylogenetic knowledge, and the computational research to implement necessary tree comparisons is in its beginning stages. Linkages among databases are still inadequate, and the computational capabilities to employ hierarchical information to search those databases have not been developed. Molecular genetic databases, such as GenBank and EMBL, are housed, curated, and supported by computational research in centralized facilities, with global support. Comparable infrastructure does not exist to

support phylogenetic knowledge, its use, and the activities that would be enabled with the implementation of interoperable hierarchical data mining.

D. The Need for a TOL Initiative

Because of the scope and complexity of the scientific challenges to resolving the TOL and making those results available to a broad user community, the Workshop participants strongly and unanimously agreed that a significant new research initiative is required. Current programs at the NSF, and at other funding agencies that have research portfolios concerned with phylogenetic knowledge, are currently not organized to insure the assembly of a TOL over the next few decades. If ongoing programs and approaches are left unchanged, the benefits of a comprehensive understanding of phylogenetic relationships may be realized only in the distant future. Strategies and recommendations to implement a new research initiative on the TOL are discussed in the next three sections of this report.

III. Resolving the Tree of Life: A Research Initiative in Phylogenetic Analysis

Because of the complex research problems that a TOL program must investigate, we are proposing the creation of a new research initiative in phylogenetic analysis that (a) calls for institutional changes in the way research efforts are currently designed and implemented, (b) new research to increase substantially the amount of phylogenetically informative evidence available to systematists to resolve relationships, and (c) research to improve significantly systematists' ability to analyze ever larger `datasets`. Research needs related to data storage, management, and dissemination will be discussed in the following section.

A. Recommendations for Changes in Research Strategies, Organization, and Design

As noted, phylogenetic research on the major groups of organisms is too incremental and uncoordinated. Moreover, the emphasis of funding agencies on relatively narrowly defined individual-investigator grants hinders researchers from addressing large-scale phylogenetic questions with dense taxon and character sampling. Moreover, it is difficult for single research laboratories to investigate those large-scale questions in the detail required when grants are awarded for relatively short time periods and resources are limited. To overcome these impediments, the Workshop makes two recommendations designed to create a more coordinated and focused research effort that will accelerate progress on the TOL.

Recommendation 1. NSF should establish systematics research awards that support multi-investigator and coordinated phylogenetic research. We further propose that each multi-investigator research group be called a Tree of Life Research Network (TOLNet).

The Workshop identified at least two general categories of TOLNets, but others might be considered:

1. TOLNets that are taxon-oriented and designed to increase knowledge of organismal relationships using large amounts of phylogenetically informative character data.

2. TOLNets whose research is primarily theoretical, methodological, or computational.

The majority of basic systematic research needed to build the TOL would be accomplished by the taxon-oriented TOLNets, yet because of the impediments we have previously identified, research on theory and methods should also be supported. TOLNets will address larger and more complex research objectives than those typical of individual investigator proposals, and we envision TOLNet awards may require substantially larger budgets and have longer award periods (e.g., four to five years) than typical systematics grants now funded.

Guidelines should be established for research at TOLNets that increase the likelihood of significant progress toward resolving the TOL. Workshop participants identified the following criteria, among others that might be considered, as being important for achieving the objectives of a TOL initiative:

1. A taxon-oriented TOLNet proposal should be designed to resolve the relationships of a major group of organisms and address complex phylogenetic problems that have been resistant to solution using past approaches. Methods-oriented proposals should likewise address large-scale questions.

2. A case should be made why a collaborative effort is needed to solve the problem at hand.

3. TOLNet proposals should demonstrate why they will result in economies of scale and effort. In principle, TOLNet collaborations should undertake larger research questions more efficiently and more cost effectively than if that question were being addressed by individual investigator awards.

4. Phylogenetic research within TOLNets that is taxon-oriented should be expected to employ large taxon samples, large character datasets, and to integrate different kinds of data to the maximum extent possible.

5. TOLNet proposals should seek to partition research questions among investigators to avoid duplication of effort and maximize synergy.

6. TOLNet proposals should show how training and educational objectives can be realized beyond what might be expected from individual-investigator grants.

7. TOLNet proposals should discuss how they will foster collaboration with individual investigators who are not formal members of the TOLNet to facilitate research and avoid duplication of effort.

8. TOLNet proposals should seek to integrate their research with that of other existing, or proposed, TOLNets in order to advance the global TOL effort and avoid duplication.

The Workshop participants believe that if the TOLNet concept is implemented, research on the TOL can be measurably accelerated and the quality of the resulting research significantly improved over current approaches. We want to emphasize strongly, however, that we do not see TOLNet awards as substituting for individual investigator proposals undertaking phylogenetic research. There are numerous and important phylogenetic questions that would be best studied within the framework of individual-investigator research awards rather than through TOLNets.

Individual-investigator awards would not preclude formal collaborative work with TOLNets, where relevant.

Recommendation 2. NSF should establish a new form of phylogenetics research award that supports the formation of Tree of Life Research Centers (TOLCenters) whose function is to foster the integration and synergy of TOLNets to advance efficiencies in research and training.

The participants of the Workshop agreed that TOLNets may not embody all the necessary elements to complete the TOL and may not always lead to economies and efficiencies of scale. We can envision, for example, instances in which TOLNets might involve redundancies in their need for large, expensive equipment or infrastructure. Examples might be TOLNets that intersect at the same institution —such as could easily occur at large universities or natural history museums or botanical gardens having major support for systematics programs— or at two or more nearby institutions having investigators in the same TOLNet. The redundancies could also extend to overlaps in training programs. A number of intersecting TOLNets might require training programs that would best be coordinated rather than implemented individually.

We can envision three important criteria, among others that might be developed, for establishing a TOLCenter:

1. TOLCenter activities would be expected to build on, and integrate activities of, the participating TOLNet PIs, not duplicate them. Thus, TOLCenters are tightly linked to TOLNets doing taxon-oriented or theoretical/methodological research.

2. TOLCenter activities would be expected to create economies and efficiencies of scale among TOLNets with respect to infrastructure, shared research, and training. Examples of this might be shared high throughput data gathering or computational facilities, among others.

3. TOLCenters would also be expected to facilitate phylogenetic research of individual investigators who may not be formal members of a TOLNet.

B. Recommendations for Increasing Phylogenetically Informative Data

Participants in the Workshop agreed that to achieve a solid understanding of the TOL within a 10-15 year time-frame, systematists will need to develop methods and technologies to increase taxon and character sampling dramatically.

1. Taxon sampling

Recommendation 3. Phylogenetic research associated with the TOL initiative should endeavor to sample as many relevant taxa as possible, including, when possible, both fossil and Recent organisms.

The Tree of Life is an expression of the phylogenetic relationships of all known taxa, living and extinct. Given this, current samples of taxa, for all major groups of organisms, are inadequate to provide a satisfactory picture of the TOL, and Workshop participants agreed that efforts must improve significantly if the TOL initiative is to be successful.

Denser taxon sampling will result in phylogenetic hypotheses that are more comprehensive, more useful to those seeking phylogenetic information, and less likely to contain results that are

artifacts of inadequate sampling. Several groups of organisms, most notably bacteria, but also single-celled eukaryotes and fungi, are so poorly known that it is likely that many new species, representing entire clades, have yet to be discovered. Construction of the Tree of Life will therefore require augmented support for exploration and discovery of new species. Furthermore, in assembling the TOL it is critical to appreciate that over 99% of species that have ever lived are extinct. Some groups of extant organisms are even better known from fossil species. Paleontologists continue to discover fossils of incredible completeness and new records of biological structures that would not previously have been predicted to fossilize. Examples include stomachs and hearts in dinosaurs, organelles of insects, feather keratins, the flowers of many angiosperms, and billion year old cells undergoing mitosis. Because phylogenetic hypotheses are sensitive to the sample of species included in an analysis, the vast taxonomic storehouse in the fossil record will in many cases improve trees derived from the small samples afforded by living species. In addition, fossils may record intermediate forms that enable systematists to see connections among disparate morphologies that might otherwise be hidden by subsequent evolutionary change.

Recommendation 4. TOL research projects must provide for the appropriate collection, curation, and preservation of taxonomic specimens and associated materials, such as tissues, DNA samples, etc., and establish guidelines and techniques to ensure data reliability and accessibility.

The Workshop participants emphasize the critical importance scientific inventory will have for a TOL initiative. To meet the needs of the TOL project and the demands it will place on gathering various kinds of systematic character data, inventory efforts will need to be increased. This is most obviously the case for microscopic organisms, but extends broadly across the TOL. Ongoing programs supporting inventory projects should anticipate the needs of a TOL and strongly encourage a range of preservational techniques that will facilitate extracting morphological and molecular character data in the future

A major limitation of much current work on molecular systematics is that the data, particularly gene sequences, are often not linked to original samples or specimens. For a TOL project to be successful, this linkage is essential. Thus, mechanisms must be put in place to use materials (e.g., tissues, seeds, cell lines, DNA extracts, etc.) associated with voucher specimens housed in a recognized collection and to maintain the identify and long-term integrity of these materials

Storage and curation of samples requires space, infrastructure, and human resources, and funding for these activities is a necessary part of the TOL project. Although many collections-based institutions have facilities for storage of tissues and DNA, others do not. Some institutions with facilities, such as zoos, do not have professional systematists on their staffs, and the reliability of tissue samples from specimens is sometimes in doubt because vouchers have not been properly identified and housed. Many molecular investigators undertaking TOL activities house tissues and samples in their own laboratories, generally without proper curation or links to vouchers. It may be appropriate to designate and fund TOL repositories. This would allow researchers not affiliated with an institution having proper collection care to have access to storage capabilities.

2. Character sampling

The heart of phylogenetic research is data acquisition. Many controversies in phylogenetic analysis are debates over evidence, and frequently it is the lack of phylogenetically informative character data that leads to lack of phylogenetic resolution. Addressing impediments to the

acquisition of character evidence necessary for satisfactory resolution of the TOL was a central concern of the workshop, and resulted in a series of recommendations.

Recommendation 5. Phylogenetic analyses associated with the TOL initiative should take into account all relevant and available evidence, including both molecular and morphological characters.

Many phylogenetic studies have employed either morphological or molecular data, not both. One major function of the TOLNets recommended above would be to encourage studies to gather and integrate large amounts of both kinds of data to achieve phylogenetic results having maximal character support. Developing high throughput methods to gather more data, both molecular and morphological, will be key to meeting the TOL objectives.

a. Morphological character data

In this Report "morphological" is used in its broadest sense, meaning heritable phenotypic data. Character data obtained from soft tissues, ultrastructure, physiology, biochemistry, and behavior are potentially relevant for phylogeny reconstruction, and multiple phenotypic datasets can be acquired for many taxa, in parallel with molecular datasets. Morphological data can provide critical evidence on relationships, and on patterns of phylogenetic diversification, often because the rate of morphological evolution can be independent of the rate at which many molecules evolve (e.g., in an adaptive radiation key morphological synapomorphies may arise without concomitant substitutions in molecules commonly used for phylogeny reconstruction). Morphological data also provide the primary link between fossil and recent organisms. Furthermore, it is based on these data, that keys and other identification guides render branches of the tree accessible to systematists, other biologists, conservationists, resource managers, and the public.

There is also much that we do not understand about how genotype is translated into phenotype via developmental pathways. It is becoming increasingly apparent that this becomes more understandable if morphological data have been investigated and scored for phylogenetic analysis.

Recommendation 6. TOL research utilizing morphological data must have adequate resources for support staff, and for the development and use of new data acquisition and interpretation technologies.

For high quality morphological research the great expense has been, and continues to be, acquisition and maintenance of highly trained support staff needed to execute the research. The slow rate of morphological data collection often results more from the lack of trained personnel than from methodological obstacles. For many morphological systematists, perhaps especially paleontologists, scientific preparators and illustrators are crucial. Addressing personnel needs will greatly increase productivity and meet TOL objectives.

A second critical step for morphological research is to document visual observations and communicate these to the scientific community. It is impractical to expect that future researchers will repeat the morphological data collection of previous workers. To the extent possible, taxa included in the TOL should have one or more digital images associated with them. All investigators must be able to access and evaluate raw data to make appropriate comparisons, and part of achieving this will be the development of "synonomy tables" across major groups as well

as methods of creating "super matrices" across major groups of organisms. New imaging methods such as CT scanning, collection of 3-D data, and sophisticated pattern recognition imaging systems should be explored as part of the TOL data collection initiative. Approaches such as CT scanning are increasingly important to avoid destructive sampling of irreplaceable museum specimens, and rare and endangered species. Digitization of classical 2-D images and 3-D images would also be an important contribution to making morphological data available to an international community of systematists as well as users of systematic information.

b. Molecular character data

The explosion of molecular sequence data and their application to deciphering relationships of organisms has transformed TOL research. We anticipate molecular data playing an ever increasing role in phylogenetics, especially in groups such as microbes for which morphological data are limited or difficult to obtain. With this rich source of data, however, have come a host of procedural problems. In part because of the relative ease with which sequences can be obtained and the ready availability of algorithms to produce phylogenetic trees, molecular studies are now often undertaken by investigators with insufficient training in the systematics of the group under consideration. Many studies use inadequate data and taxon samples. Furthermore, too little attention has been paid to data reliability, to the taxonomic status of the samples used, to vouchers specimens housed in systematics institutions, or to legal issues surrounding the collection and exchange of specimens or tissues.

Recommendation 7. In order to meet TOL demands new methods of DNA extraction, particularly employing automation, need to be developed and made available to practicing systematists. Likewise, technologies for high throughput sequencing need to be modified and extended for comparative analyses and made available to the phylogenetics community.

The ability to analyze large numbers of samples, spanning the entire TOL, would be greatly facilitated by improvements in techniques to quickly isolate DNA from a wide variety of tissues. Moreover, it is highly desirable to isolate bio-molecules in such a way that they can be re-analyzed by other researchers. Research is needed to determine which isolation and storage methods are best for enabling future analysis. Possibilities include freezing of materials, making large insert libraries of DNA, making cDNA libraries, or immortalizing cells.

Because extraction procedures may need to be repeated, new methods are required for non- or minimally-destructive sampling. For many organisms, especially microorganisms, these concerns are not trivial: in sequencing whole genomes, for example, the entire specimen is utilized, and no "voucher" may exist for an important set of sequences. Thus, it will be of great value to develop highly sensitivity methods (e.g., the ability to analyze single DNA molecules) which would limit the destruction of samples as well as allow the analysis of species for which samples are limited (e.g., single celled organisms which have not been cultured).

High throughput sequencing will be a key element in building the TOL. New efficiencies are required to increase the generation of sequence data by orders of magnitude over the output from most individual laboratories. Technologies exist within the genomics community that may be appropriate for this task but this machinery has not been adapted for the comparative analyses required by systematists. The modification of such technologies to suit the needs of the TOL initiative needs immediate attention, and should be a topic for more detailed discussion in a subsequent workshop.

Workshop discussions suggested that the ability to obtain comparative data from a wide range of taxa will be a greater bottleneck than will the sequencing itself. Methods currently available to target and isolate homologous sequence should be improved and automated as much as possible. For example, PCR with degenerate primers is frequently used to clone homologous genes from a wide range of taxa. However, if this is to be done on tens of thousands of species, for dozens or hundreds of genes, we will need more automated procedures for primer design, optimization of PCR conditions, and cloning and sequencing of PCR products. Methods that will improve the ability to clone and sequence homologous genes from large numbers of species need to be developed. These could include targeted cloning methods in which homologous regions are preferentially cloned directly into a DNA library, and targeted genome sequencing (e.g., to identify baseline genes for particular taxa).

We note, finally, that in the future, non-sequence molecular data may also be useful for the TOL. Examples of such data include gene expression patterns, gene order, and presence and absence of biochemical pathways. Methods to obtain and analyze large amounts of such data must also be encouraged.

C. Recommendations for Improving Data Analysis and Tree Assembly

Among our most important recommendations for resolving the TOL over the short-term include significantly increasing the size of datasets, both in terms of taxa and character data. Yet, as these increase, particularly the former, impediments to data analysis also increase. Evaluating each of the possible trees of a very large taxon sample to find the best-fit tree for the data, using any criterion, is regarded as effectively impossible, thus heuristic solutions are required. But finding satisfactory heuristic solutions has also proven to be a major challenge. The Workshop devoted considerable discussion to these and other issues of data analysis as they relate to a TOL initiative and agreed that much more research is needed.

Recommendation 8. More research should be undertaken to improve heuristic search strategies using algorithms designed to meet the demands made by increasingly large datasets.

The history of quantitative phylogenetic analysis shows that the limits of tree building algorithms have continued to expand as a result of research effort. We are now entering an era in which datasets are becoming significantly larger, and it can be anticipated that a TOL initiative will entail larger datasets still. The last several years have seen remarkable advancements in algorithms for tree searching, but ensuring that this trend continues is critical.

Some approaches to tree building, especially using molecular sequence data, are faster at tree searching than others, and speed depends on the underlying assumptions describing character transformation, strategies to search tree space, and criteria for optimization. Additional research on tree searching is critically important and efforts should be made to foster new approaches that might represent fundamentally new ways of addressing the challenges presented by large datasets.

Recommendation 9. Improvements need to be made through research and experimentation on parallel processing and the technologies required to make these systems widely available to the systematic community.

In order to analyze the ever-larger datasets generated in TOL research, significant advances will be required in both the algorithms and hardware used by tree search software. The Workshop

agreed that significant improvements need to be made in the computational hardware that is available to investigators undertaking TOL research. The overall conclusion of the meeting was that emphasis should not be on the use of supercomputers but rather on other high-performance computing hardware, possibly including parallel processing. Many noted that access times to supercomputers can be so long as to make many contemporary desktop machines as effective for undertaking research. Recent advances in low-cost LINUX-based cluster computers have brought supercomputer level resources within the reach of many investigators.

Although this hardware is well understood and widely available, the software is not. Only a few phylogenetic analysis packages are currently available that can operate on such parallel clusters (e.g., fastDNAml, MALIGN, POY, and most recently PAUP*), and such efforts are in their infancy. Furthermore, the use of even these few parallel tree search programs requires significant effort on the part of the user. The setup and execution of such mediation packages as MPI an PVM must be understood as well as the specific options of the programs which interact with them. The parallel algorithms required should be highly scaleable and the code portable. This will allow investigators to operate on a diversity of hardware configurations and dimension.

Hence, research on parallel algorithms should be fostered, as well as purchases of computer equipment (such as clusters of workstations that can be configured to work as a parallel machine). Development of user-friendly software (perhaps written in Java and running inside a web browser) for interacting with fast, parallel software should be encouraged. "Screen saver" style of parallelism used by the SETI project might be a possibility, but currently it is viewed as difficult to implement for the specific problems generally presented by phylogenetics.

Recommendation 10. Research is needed to develop more efficient algorithms and procedures to align and combine character data.

Assembly of the TOL will require the combination of multiple sources of data. Not only does this involve the combination of sequence data and morphological data, but also the great multiplicity of genetic loci which can now be sequenced.

Many of the loci sequenced today show significant length variation. Future efforts may involve more non-coding DNA, hence the situation is likely to become more pressing. Methods to align sequence data (e.g. CLUSTALL, TreeAlign, MALIGN) are available but limited. Each of these multiple alignment programs use heuristic solutions which break down when sufficient length variation is found. Superior methods and implementations are required, possibly using parallel algorithms.

Recommendation 11. More research is required to characterize more fully concepts of tree support and robustness.

Phylogenetic analyses should be accompanied by an evaluation of group support (i.e., some measure of evaluation of the quantity and quality of the evidence). Many methods have been proposed to assess tree support, but research has shown that virtually all these measures have limitations. As datasets become ever larger and more complex in terms of missing or inapplicable data for subsets of taxa, as different kinds of data are combined, or as underlying assumptions, or models, of character change become increasingly intricate, it is more difficult to assign meaning to the concept of support and to quantify it.

The Workshop concluded that much more research on tree support is needed. For example, attention might be focused on issues such as sensitivity to missing data, to alternative alignments, and inapplicable characters across taxa, or on methods for elucidating conflicts across trees and within the data set itself. Moreover, there is a need to develop support measures that are computationally efficient for large datasets. Markov chain Monte Carlo (MCMC) bayesian methods were suggested as another promising avenue of research on nodal support that is still in its infancy.

Recommendation 12. Increased research is needed on the assembly of supertrees that realistically reflect the topological patterns and underlying empirical support expressed by trees being combined.

Ideally, the TOL will be assembled by the analysis of large amounts of comparable data across taxa spanning the entire tree. In actuality, the process of assembling the tree will include linking together trees generated by separate analyses into what is termed a "supertree." Recent research has shown, however, that combining topologies is a complex computational problem.

There are other issues aside from understanding how multiple topologies might be combined into a supertree. Little research has been devoted, for example, to how topologies might be combined so as to reflect the underlying support of different clades, especially when conflicts are present across topologies and datasets.

IV. Interpreting and Using the Tree of Life: A Research Initiative in Phyloinformatics

One fundamental concern of the TOL initiative is phylogenetic analysis and its goal of assembling a tree depicting phylogenetic relationships among very many of the Earth's known species, representing all major branches of the tree. A second, and no less important, aim of the TOL initiative is to compile and make this phylogenetic knowledge of value to users. Deriving concrete benefits from the TOL project depends critically on the development of a new research initiative we call "phyloinformatics."

Our vision of phyloinformatics is very broad. At the most basic level, it will be essential to archive and database the character data underlying phylogenetic analyses and the phylogenetic hypotheses (trees) resulting from these analyses. As noted, this is currently the mission of the TreeBASE database, but the latter contains only a small proportion of published phylogenetic work and will be inadequate in its present form to support the TOL project. Also critical are connections to the voucher specimens from which data were originally obtained.

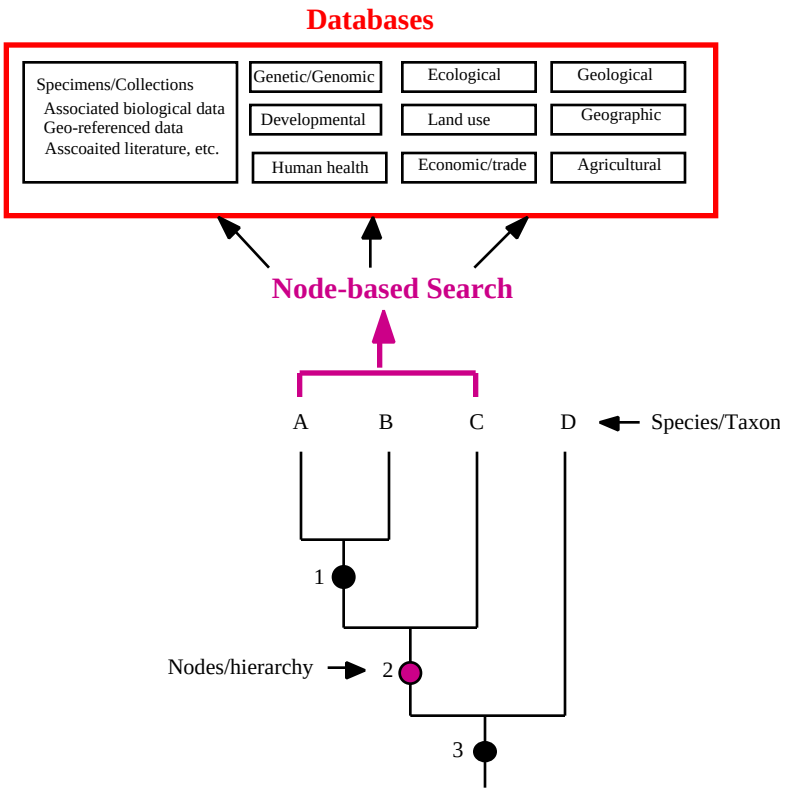

Figure 1. Phyloinformatics: a conceptual framework. In addition to archiving trees, data, and having the ability to manipulate these, a phyloinformatics infrastructure will use the hierarchy implied by trees to conduct node-based searches across many kinds of databases. A search using node 2, for example, would simultaneously retrieve information from databases for taxa A, B, and C, and would combine and synthesize those data in ways not currently possible. Searches with this power and sophistication would use the predictive capabilities of phylogenetic trees and serve to integrate biological information in new ways.

The phyloinformatics research initiative must also include the development of new tools to synthesize and navigate the data and trees in the phylogenetic knowledgebase, and new ways of visualizing trees and the information associated with them, including ready access to information on robustness of alternative phylogenetic hypotheses. Further, phyloinformatics will be an entirely new way of using the hierarchical information implied by phylogenetic hypotheses to investigate and synthesize information from many other kinds of data resources (Figure 1). Tree-oriented navigation and data mining will enable the users of biological information to bring together comparative information in ways impossible with traditional search strategies, thus actually creating new associations and new knowledge, and facilitating predictivity. Finally, phyloinformatics will spawn a host of new connections to educational institutions, to students, the media, and the public at large. These would highlight the TOL discovery process, the results of the project, and the general significance of these results. To a considerable extent the current Tree of Life project (http://phylogeny.arizona.edu/tree/phylogeny.html) serves this purpose, but it is far from complete and presently is inadequately prepared to cope with the anticipated increases in knowledge.

Phyloinformatics is a critically important concept, because it is the mechanism for connecting the results of phylogenetic research with other biological information and making those results of available to science and society. During the Yale Workshop a series of concrete recommendations were developed regarding the need for this new research program and the

infrastructure and human resources to support it. These recommendations are presented below, but it was also recognized that such a large initiative requires much more attention, with input from a wider spectrum of individuals with expertise in the relevant computer sciences and experience with other biological databasing efforts. We therefore strongly recommend that an additional workshop be devoted specifically to the phyloinformatics needs of the TOL initiative.

Recommendation 13. Creating a comprehensive database of phylogenetic knowledge requires research to develop better mechanisms for data capture and data mining, and to find optimal ways to represent and visualize phylogenetic knowledge.

A comprehensive database of phylogenetic knowledge that includes both character data and trees will be critical to the success of the TOL initiative. Although we now have some experience with such a database, through the TreeBASE and Tree of Life efforts, there are a host of difficult conceptual and practical issues that will need to be addressed to move this databasing effort forward in a way that will best serve the needs of the entire TOL project. Development of a new data model will require renewed attention to how best to represent both data matrices and phylogenetic trees (e.g., whether as objects, decomposed into fields representing each node, etc.). Here a critical problem is the representation of alternative phylogenetic hypotheses, derived from the same dataset and analysis, or from different datasets (e.g., morphology versus gene sequences; alternative gene trees) and analyses (e.g., maximum parsimony versus maximum likelihood).

The success of such a databasing effort depends critically on the efficiency of data entry (including retrospective data capture, and entry of new data, perhaps as a corequisite of publication) and curation (including relevant annotation of data for interoperability with other data resources; see below). Much additional attention is needed to the design of appropriate software for such tasks. Mechanisms to navigate through the trees in such a database have been explored in connection with the development of TreeBASE (e.g., the concept of "surfing" a "neighborhood" of trees in “tree space”), but much additional research is needed to develop and implement new and improved strategies.

Recommendation 14. Assembling and interpreting the TOL will require increased attention to the theory and practice of synthesizing "supertrees,” to the representation of ambiguity and alternative hypotheses, and to the visualization of information associated with trees.

Assembly of the entire TOL will, at some level, depend on the synthesis of data and results from individual studies based on a more limited sample of taxa. This task raises a variety of theoretical and methodological questions that will need to be addressed if the goals of the TOL initiative are to be realized. In recent years some attention has been devoted to the problem of assembling a "supertree" so as to best represent the information contained in a set of underlying source trees that differ to a greater or lesser extent in the sample of taxa they include. Much more research is needed in this area,

Directly related to this task is the visualization of very large phylogenetic trees and the information associated with them. As noted earlier, mechanisms are needed to take into account and display levels of support or ambiguity surrounding particular phylogenetic hypotheses. Also critical will be the development of new tools for the display of character information associated with trees. To some extent this can build on previous efforts such as MacClade and WinClada, but as datasets and trees are increasingly combined (e.g., via construction of supertrees), more attention to the issue of synonymy of characters across different studies will be required (e.g., see discussion above of the development of synonymy tables for morphological characteristics).

Recommendation 15. Linkages need to be designed and implemented between the hierarchy contained in the phylogenetic knowledgebase and a wide variety of other data resources by creating the search and data mining tools to synthesize information across those databases.

The value of the TOL will not be realized unless the results are effectively connected to other digital data resources such as those major databases now well established in the molecular biological community, especially to archive and explore molecular sequence data (e.g., GenBank, EMBL). Enormous efforts are now underway in the taxonomic community to database taxonomic names (e.g., International Plant Names Index, Species 2000, the Integrated Taxonomic Information System, and others) and data associated with museum specimens (e.g. the Ocean Biogeographic Information System, Global Butterfly Information System, and others). This work will be promoted and facilitated when the Global Biodiversity Information Facility (GBIF) is formally instituted early in 2001. The first priority of GBIF will be to complete an Electronic Catalog of Names of Known Organisms. Other work areas are to develop search engines and other interoperability tools, to convene technical advisory groups on data standards and other topics, and to coordinate among local and regional information networks such as the North American Biodiversity Information Network (NABIN) and the National Biological Information Infrastructure (NBII), among others.

It is imperative that these efforts be linked with the phylogenetic knowledgebase because it is through these links that the power of phylogenetic knowledge to organize and integrate disparate information will be realized. At the same time, these other databases will contribute to the construction and searching of the TOL. Reaping the ultimate benefits of the TOL initiative will require very significant attention to the appropriate annotation of data, so as to maximize the potential of these linkages for the user community.

Recommendation 16. The phylogenetic knowledgebase must be made comprehensible and valuable at a variety of levels, to have the broadest possible impact for science, education, and the rest of society.

The Tree of Life web project is a very important start in the right direction, as this provides summary information of great use to teachers, students, and the public, including digital images, references to the relevant literature, web linkages, and so forth. Tree of Life "treehouses" attempt to reach an even younger audience. Such efforts need to be greatly expanded, which will entail research on the most effective strategies and greater attention to the design of user interfaces and tools to display and utilize trees. Here we envision especially creative intersections with research mentioned above on the visualization of the data associated with trees, and on the exploration and connection of disparate data using phylogenetic relationships as a guide. Ready access to the latest representation of the TOL will have an important impact on teaching at all levels, as well as the representation of nature in public museums, in the media, on the web, and elsewhere.

Recommendation 17. A phyloinformatics infrastructure must be established to insure the development and long-term growth of a phylogenetic knowledgebase, coordination of informatics research, linking of phylogenetic hierarchies with other biological databases, and the public outreach required by the TOL initiative.

Phylogenetic databasing efforts are not now supported at any significant level. If the goals of the TOL initiative are to be realized, this situation must change dramatically and as quickly as possible. Major new commitments are required to support this activity, which in scope exceeds the major databasing efforts of the molecular genetics community, such as GenBank and EMBL. The needs of the TOL initiative will be best served through the establishment of a phyloinformatics facility, with primary responsibility for the long term development, growth, and

maintenance of the phylogenetic knowledgebase, as well as conducting research that will produce ever-improving software. Establishment of a phyloinformatics facility would also include support for the necessary computer scientists, information technology personnel, and scientific curatorial staff. Because the phyloinformatics-oriented research projects interdigitate in so many ways, these efforts might benefit greatly from the coordination that a facility would provide. It should also be noted that there was interest on the part of some Workshop participants for a facility whose function would also include coordination of the TOLNets and TOLCenters and facilitation of activities across the TOL initiative. This issue requires further attention at the follow-up workshops.

V. A National Initiative to Assemble the Tree of Life

The recommendations enumerated above, when taken together, encompass the elements of a National Tree of Life Research Initiative (Figure 2). The Research Component includes the numerous Tree of Life Research Networks (TOLNets), comprising the coordinated taxon-focused and methodology-focused research groups whose empirical work will be responsible for building the hierarchy of the TOL, as well as a variable number of TOLCenters that will synergize and build on the research programs of multiple TOLNets.

We are proposing a research program that attempts to maximize individual-investigator creativity and independence, yet at the same time fosters formal and informal collaboration to achieve the focused and intense research effort that will be required if significant progress on the TOL is to be achieved. We suggest the need for some flexibility within this organizational framework. TOLCenters, for example, might provide infrastructure that would foster research of systematists having specialized requirements, such as centralized sequencing centers for microbial phylogenetics.

In addition to the Research Component, the National Tree of Life Research Initiative also includes a Phyloinformatics Infrastructure which would spearhead the databasing effort, informatics research, and public outreach functions of the TOL. As noted already, considerable thought will have to be given to this infrastructure, but the Workshop participants were unanimous in their belief that such an infrastructure is critically necessary if the research efforts of thousands of scientists are to be made available to people everywhere. It may be that this infrastructure could also facilitate coordination and support more broadly across a TOL research effort.

Finally, scientific research has the purpose to expand the frontiers of knowledge and to use that knowledge to serve the well-being of people everywhere. Systematics has a demonstrated importance to society, and those contributions will be expanded enormously with increased understanding of the Tree of Life and the power it brings to organizing biological information. The research initiative we propose is an ambitious effort but one that will surely generate many tangible benefits for science and society.

Organizational Plan
National Tree of Life Research Initiative

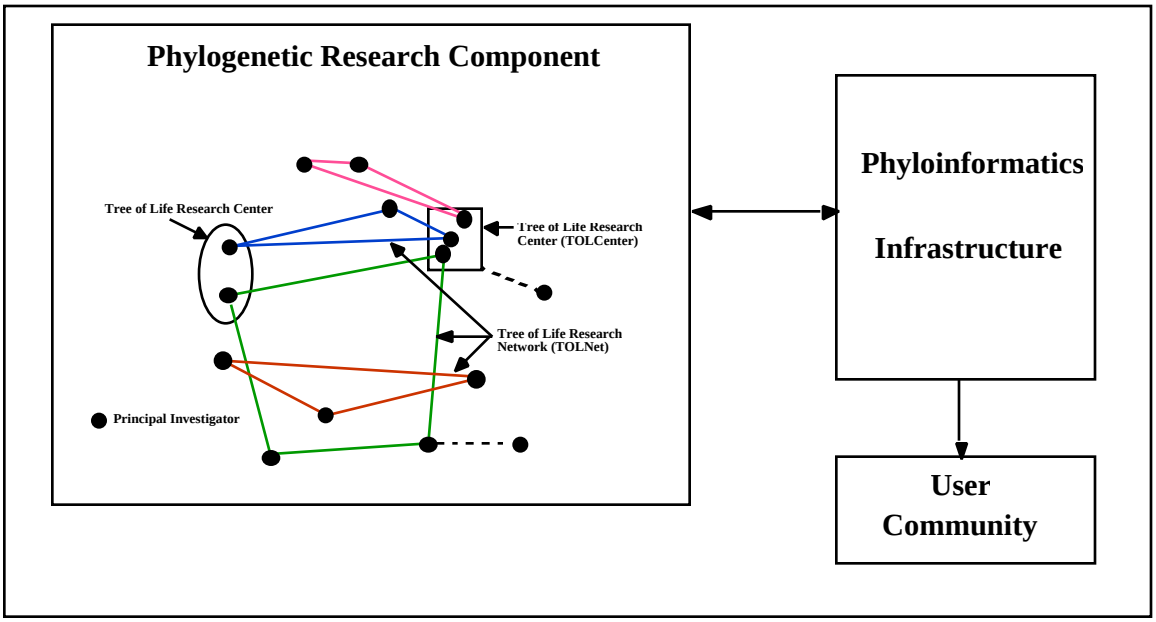

Figure 2. An organizational plan for a National Tree of Life Research Initiative. The initiative would be composed of two core program elements. The first is a core element of research; the second would be a core infrastructure for phyloinformatics. The research component consists of formal Tree of Life Networks (TOLNets) and Tree of Life Research Centers (TOLCenters). Individual TOLNets (solid colored lines) would consist of two or more principal investigators sharing a cooperative research program (grant) in taxon-based phylogenetic research, or focused on theoretical and methodological problems in phylogenetic analysis, or phyloinformatics. Some PIs (dashed lines) might not be formal members of a TOLNet but may have collaborative research with one or more members of a network or with a TOLCenter. Tree of Life Research Centers might be established at single institutions where multiple TOLNet PIs are located (rectangle) or among local or regional institutions housing TOLNet PIs (oval). A Phyloinformatics Infrastructure would store the results of phylogenetic research, undertake informatics research to support TOL activities, and ensure optimal linkages to multiple databases to serve the user community.

VI. Participants of the Workshop

Meredith Blackwell, Louisiana State University, Baton Rouge, LA
Judy Blake, The Jackson Laboratory, Bar Harbor, ME
Joel Cracraft (co-organizer), American Museum of Natural History, New York, NY
Rob DeSalle, American Museum of Natural History, New York., NY
Michael Donoghue (co-organizer), Yale University, New Haven, CT
Jonathan Eisen, The Institute for Genomic Research, Rockville, MD
Douglas Futuyma, State University New York, Stony Brook, NY
Jacques Gauthier, Yale University, New Haven, CT
Pablo Goloboff, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
David Hillis, University of Texas, Austin, TX
Darlene Judd, Oregon State University, Corvallis, OR
Junhyong Kim, Yale University, New Haven, CT
Meredith Lane, Academy of Natural Sciences, Philadelphia, PA
Paul Lewis, University of Connecticut, Storrs, CT
Diana Lipscomb, George Washington University, Washington, DC
Francois Lutzoni, The Field Museum, Chicago, IL
Wayne Maddison, University of Arizona, Tucson, AZ
Brent Mishler, University of California, Berkeley, CA
Maureen O'Leary, State University of New York, Stony Brook, NY
Jeffrey Palmer, Indiana University, Bloomington, IN
Kathleen Pryer, The Field Museum, Chicago, IL
Michael Sanderson, University California, Davis, CA
Pam Soltis, Washington State University, Pullman, WA
James Staley, University of Washington, Seattle, WA
Naomi Ward, Louisianna State University, Baton Rouge, LA
Quentin Wheeler, Cornell University, Ithaca, NY
Ward Wheeler, American Museum of Natural History, New York, NY
Kevin White, Stanford University**,** Palo Alto, CA

NSF:

Mary McKitrick, Program Officer, Division of Environmental Biology
Terry Yates, Director, Division of Environmental Biology