Tree of Life Workshop III:

Developing the Technology and Infrastructure

Needed for Assembling the Tree of Life

Report of a workshop held at the
University of Texas, Austin, December 3 and 4, 2000

Table of Contents

I. The Tree of Life Problem *4*

II. Description of the Workshop and Participants *4*

Organizers 5
Participants 5
NSF Observers 6

III. Infrastructure and Community Coordination *7*

A. Tree of Life Networks and Hubs **8**
What is a TOLNet? 8
Priorities for Proposed TOLNets 8
Taxon-based proposals 8
Analysis-based proposals 9
What is a TOLHub? 9
Priorities for Proposed TOLHubs 10

B. Phyloinformatics and Coordination Infrastructure (PICI) **11**
Primary informatics and global archiving of phylogenetic knowledge 11
Development and delivery of user interface tools and metadata standards 12
Phyloinformatics research and development to support PICI 12
Coordinating services for TOLNets and TOLHubs 13
Training activities for Assembling the Tree of Life 13
Recommendations for establishing PICI 13
Location 14
Administration 14

IV. Specific Issues to be Addressed in Technology and Infrastructure Proposals *14*

A. Specimen and Data Collection **14**
Sampling of Taxa 15
Specimens 15
Specimen acquisition 15
Handling of specimens 15
Archiving specimens 16
Data 16
Character sampling strategies 16
Data collection 16
Archiving and distributing the data 17
Analytical considerations in database design 17
Summary 18

B. Data Analysis **19**
General issues involving data analysis 20
Support structures 23

C. Integrating Existing Knowledge in Systematics **23**

D. Technology Development and Transfer **25**
High-throughput data collection 25
Examples of steps in need of development 25
Molecular or genotypic characterization 25
Phenotypic studies 25
Specific steps currently in need of improvements 26
Morphology and other aspects of phenotype characterization 26
Extraction of DNA for molecular studies 26
Other areas of need for technological improvements or development 26

Control of and Credit for Research 27
D. Training 28
The Problem 28
Recommendations for Improving Training 28
Recommendations to Increase Recruitment 29
V. Coordination, Timing, and Implementation 29
A. Coordination with other efforts 29
B. Timing and Implementation 30

I. The Tree of Life Problem

Whether biologists are interested in describing the history of life, or in using evolutionary relationships in a comparative framework to analyze biological data from other fields, phylogenetic trees provide the basic framework for interpreting biological data. A complete description of the Tree of Life would give biologists the same kind of predictive power that chemists have from the Periodic Table of Elements, but on a much larger and more complex scale. The proposed project, "Assembling the Tree of Life," aims to expand this power of analysis for all biological taxa on Earth. Although the goal of describing the relationships among all known species is an ambitious idea, biologists are discovering that complete phylogenetic analyses of particular groups are providing unprecedented power for interpreting and understanding biological diversity. For instance, phylogenetic trees are being used increasingly within the health sciences to identify disease agents and predict disease outbreaks. Within comparative genomics, trees help us understand changes in gene structure and function across life's diversity and make sense of normal and abnormal patterns of development in different organisms. Within resource management, phylogenetic methods are being used as research tools to identify exotic invasive species that may disrupt ecosystems or compromise our food supply and to locate their point of origin. These example applications and many others depend on accurate and thorough estimates of the Tree of Life.

Although complete information on the Tree of Life would be enormously useful for biologists, there are a number of limitations that must be overcome before it can be estimated in detail. These challenges can be grouped in the areas of data collection and management, data analysis, and personnel training. The challenges of data collection include the collection and preparation of specimens, as well as the linking of specimens to existing phylogenetic knowledge. Data collection issues also include the rapid acquisition of morphological and/or molecular data from over 1.7 million species. Furthermore, even if we had the complete genomic sequences or morphological summaries of all 1.7 million known species, we could not produce a global estimate of evolutionary history with existing computational approaches. Therefore, the development of computer technology and methods of analysis is critical to the field of phylogenetics. There is also a need to develop and support programs to train the individuals who will collect the specimens, acquire the necessary data, and conduct the analyses, and a need to coordinate these activities across researchers and institutions.

II. Description of the Workshop and Participants

An NSF-sponsored workshop was held at the University of Texas in Austin, Texas on 3-4 December 2000. This was the third and final of three workshops held on the feasibility of and planning for Assembling the Tree of Life (ATOL). The first two meetings were held at Yale and the University of California-Davis. The Austin meeting was devoted to an

assessment of the development of infrastructure and new technologies that will be needed to assemble the Tree of Life.

The participants of the workshop were chosen to represent a wide diversity of taxonomic groups, research expertise in data collection and analysis, research approaches, technology development, and experience in the organization and administration of national centers. The following individuals participated in the workshop:

Organizers

David M. Hillis, University of Texas, Austin. Interests: Phylogenetic methods for large data sets; viral evolution; phylogenetic applications; metazoan systematics
Wayne Maddison, University of Arizona. Interests: Computational phylogenetics; phylogenetic software; spider systematics

Participants

Judith Blake, The Jackson Laboratory. Interests: Comparative genomics and genome informatics
Jeffrey Boore, Group Leader in Comparative Genomics, DOE Joint Genome Institute. Interests: Large scale comparative genomics
Stuart Brand, Co-chair, All Species Initiative. Interests: Applications of information technology to large biological problems
David C. Cannatella, University of Texas, Austin. Interests: Comparative analysis, vertebrate systematics and morphology
Joel Cracraft (co-organizer of Yale Workshop), American Museum of Natural History. Interests: avian systematics, large data sets
Michael Donoghue (co-organizer of Yale workshop), Yale University. Interests: plant systematics, large data sets
Jonathan Eisen, The Institute for Genomic Research. Interests: Bacterial phylogeny, genomics
John Huelsenbeck, University of Rochester. Interests: Statistical analysis of phylogenies; phylogenetic methods
Robert Jansen, University of Texas, Austin. Interests: Plant systematics; using rearrangements of genomes for phylogenetic analysis
Kevin Kelly, Editor, Wired Magazine; Co-chair, All Species Initiative. Interests: Applications of information technology to large biological problems
Junhyong Kim, Yale University. Interests: Phyloinformatics, biostatistics, and developmental evolution
Leonard Krishtalka, University of Kansas. Interests: Systematics, museum databases
Francois Lutzoni, Field Museum. Interests: Systematics of lichens, co-evolution
David M. Maddison, University of Arizona. Interests: Phylogenetic software and theory, insect systematics
Maureen O'Leary, State University of New York, Stony Brook. Interests: Mammalian systematics and vertebrate paleontology

Jim Reichman, University of Santa Barbara. Director, National Center for Ecological Analysis and Synthesis

Tim Rowe, University of Texas, Austin. Interests: Paleontology, morphological analysis, computed tomography, scientific visualization

Thomas Schmidt, Michigan State University. Interests: Microbial systematics and ecology

Chris Simon, University of Connecticut. Interests: Insect systematics

David Swofford, Smithsonian Institution, Laboratory of Molecular Systematics. Interests: Phylogenetic analysis (including software development)

Tandy Warnow, University of Texas, Austin. Interests: Computational phylogenetics, computation and visualization of large phylogenetic trees

Greg Wray, Duke University. Interests: Developmental evolution, echinoderm evolution

Anne Yoder, Northwestern University Medical School. Interests: Primate systematics, molecular evolution

Elizabeth Zimmer, Smithsonian Institution, Laboratory of Molecular Systematics. Interests: Plant molecular systematics

NSF Observers

Mathew Kane, James Rodman, Joann Roskoski, Grace Wyngaard, and Terry Yates.

III. Infrastructure and Community Coordination

Assembling the Tree of Life will depend on sharing expertise, infrastructure, and data in new ways. Both at this workshop and at the previous workshops, it was recognized that the success of ATOL depends on the contributions of research working groups and institutions, and a common informatics infrastructure. Thus, guiding our recommendations are the following principles:

- To attack a particular empirical or theoretical question, an important contribution to ATOL can be made by a **coordinated working group** that draws expertise from wherever it may be found, often scattered across institutions. Indeed, linking across institutions to build a sense of community effort is important to the success of ATOL. We envision many such working groups contributing to the effort.
- To foster cross-disciplinary interaction, breadth of training, and efficiency of sharing infrastructure, an important contribution to ATOL can be made by an institution with a **concentration of expertise and facilities** that span diverse but ATOL-related interests. We envision many such institutions contributing to the effort.
- To ensure efficient growth and use of the shared data bases that will be a primary output of this entire effort, there must be **oversight and coordination of the informatics aspects** of ATOL.

In the workshop, most of the discussion focused on one particular approach to realize the three principles, namely that each represent a separate funding competition or category of implementation: TOLNets, TOLHubs, and the PICI, respectively. A TOLNet would link scientists at various institutions and from various disciplines into a working group focusing on a particular problem. TOLHubs would represent institutions with a special concentration of, and commitment to, ATOL activities. The PICI ("Phyloinformatics and Coordination Infrastructure") would be a coordinating group or facility that would support and oversee shared data bases.

It is worthwhile emphasizing that there was strong agreement on the fundamental value of collaborative working groups, institutional contributions, and a shared informatics infrastructure. In this report, we will center our discussion around the TOLNet/TOLHub/PICI model, while realizing that this is but one of various possible implementations. For instance, another model would retain the PICI component as described below but have no clear distinction between TOLNets and TOLHubs, instead allowing for both net-like and hub-like activity to be part of any collaborative effort. A collaborative effort might have particular strength and breadth at one institution (i.e. have a hub-like aspect) while at the same time reach out to various institutions to draw in needed expertise (i.e. have a net-like aspect). Regardless of whether TOLNets and TOLHubs are distinct categories of collaborative efforts, or represent different emphases

that can be blended into a single collaboration, the arguments we give below regarding the types of efforts that need to be supported remain valid.

A. Tree of Life Networks and Hubs

What is a TOLNet?

Tree of Life Networks ("TOLNets") are envisioned as collaborations among multiple investigators, spanning different fields of expertise, that focus on a particular empirical or theoretical problem. Such collaborations are fundamental for advancing both data collection and analysis. These networks ideally should be avenues for training, synergy of ideas and information, as well as assemblage of data and analysis for the ATOL. Proposals submitted to a TOLNet competition should allow for an annual meeting of these people, even encourage inclusion of people not in the TOLNet, by hosting annual meetings. There should be a mechanism for an annual meeting to review progress and to promote coordination. All activities associated with the TOLNet should be coordinated with the PICI and the TOLHubs. We strongly recommend that the TOLNets should be funded on at least a five-year cycle.

TOLNets are the essential mechanisms for coordinating individual investigators from diverse fields of knowledge. Proposals for TOLNets can be organized in a variety of ways—they may be taxon-based or methods-based or may even be focused on databasing. Given the ambitious scope of the ATOL effort, all TOLNets should be broad-based. Strong proposals will have an educational component (postdocs, graduate students, undergraduates). Investigators are encouraged to contact program officers about the appropriateness of the scope of the proposed project.

Priorities for Proposed TOLNets

We here outline expectations for two possible categories of TOLNets, one focusing on taxa, the other on analyses.

Taxon-based proposals

A taxon-based TOLNet proposal would seek to reconstruct the phylogeny of a particular clade. The exact scope would vary, from the deepest branches of the Tree of Life to relationships within smaller groups of species. The following would be considered valuable components of a taxon-based TOLNet proposal:

- Ambitious scope, in the breadth region of the tree to be investigated and/or in the density of species sampled
- A fully-articulated strategic plan for coordination
- Established systematics expertise in the clade of interest
- Expertise in a wide variety of data types (e.g., genomics, behavior, morphology…)

- An articulated plan for sampling, including choice of species, specimen acquisition, and identification of specimens
- Adequate curation of material (e.g., deposition of vouchers where possible, establishment of frozen tissue cultures)
- Appropriate procedures for checks/quality control on acquired data
- Expertise in a wide variety of systematic data-analysis techniques
- Appropriate technical support (preparators, lab technicians, etc…)
- Collaboration with field biologists (in the broadest sense, including neontology and paleontology)
- A mechanism for dissemination of data/information to the community; this may include web based access to raw data and educational materials more generally defined
- Collaboration with experts in visualization or graphics. Such experts may include but are not limited to scientific illustrators, computer graphics experts, etc.
- Collaboration with computer scientists and statisticians for analyses
- Use and development of innovative technology

Analysis-based proposals

An analysis-based proposal would seek to develop theory and analytical methods important at various stages of reconstructing the Tree of Life. Appropriate topics would include investigation of optimal taxon-sampling strategies, automated quality control for sampled data, tree-selection criteria, search strategies for large data sets, combining heterogeneous data, synthesis and presentation of results through supertrees or visualizations, and databasing methods. The following would be considered valuable components of a TOLNet analysis-based proposal:

- Collaboration of biologists with expertise in analysis and theory, empirical biologists, computer scientists, and mathematicians
- Solving computational problems on real data sets in collaboration with biologists
- Broad coverage of topics
- A final product that is professional and widely usable; development and distribution of software
- Association with PICI to enable high performance computing
- Educational workshops for the systematics community, with emphasis on outreach to enable taxon-oriented specialists to learn new methods

What is a TOLHub?

A Tree of Life Hub ("TOLHub") is an institution (or consortium of physically proximate institutions) that provides a concentration of people and resources for research and training in diverse areas related to the ATOL effort. TOLHubs will have shared resources for generating and analyzing molecular and non-molecular data, as well as

person resources—researchers, technicians, interactors, and administrative personnel. A proposed TOLHub might be one, or a more usually a combination, of the following:

- An idea-incubator devoted to broad-ranging and innovative issues of data analysis and synthesis
- A major curation/collections center for a particular taxonomic group, or a repository for genetic resources, cultures, etc.
- A data factory, where molecular or other types of data are gathered using high-throughput automation
- A super-computer facility for phylogenetic analysis and metadata synthesis
- An informatics center for management of genomic and/or collections-based data and databases

By virtue of their geographic concentration of people and resources, TOLHubs can make unique contributions to the ATOL effort by providing:

- Supplemental training of students, postdocs, and cross-disciplinary researchers in systematic, taxonomic, and theoretical methods. This training will enhance, but not replace, that provided in local institutions.
- A locus for research interactions among national and international scientists
- An administrative center for programs, such as workshops, mini-conferences, and visitor services
- An economy of scale for research activities

TOLHubs are not expected to be independent of TOLNets. TOLHubs provide foci with which TOLNets may interact, providing value-added resources and opportunities for TOLNets. In this way, a TOLHub can synergize the activities of individual TOLNets and provide opportunities for individual researchers. Given that a TOLHub may have one or a few thematic foci, and that a TOLNet may involve annual meetings and shared resources, there is not a clear distinction between them. Indeed, as mentioned in the introduction to this section III, it is possible that TOLNets and TOLHubs will be end points on a continuum, and a single collaborative proposal may blend elements of both.

Priorities for Proposed TOLHubs

Proposals will be accepted from groups of researchers to base a TOLHub at a lead institution (which may represent a consortium). The following would be considered valuable components of a TOLHub proposal:

- A demonstrated international record of excellence in systematics research at the institution(s) in a variety of techniques, subdisciplines, and/or taxa
- A proven track record of training graduate students and postdocs in modern systematics approaches and methods among established PIs
- Demonstration of the presence of, or the commitment to establish, substantial support facilities, such as conference facilities, computational hardware and

support personnel, high-throughput sequencing technology, image-capture hardware, or modern collection facilities
- A proven track record of hosting visitors (national and international) and provision of services related to systematics research

B. Phyloinformatics and Coordination Infrastructure (PICI)

To meet the increasing demand for phylogenetic research, information gathering and dissemination must be managed more effectively and research must be coordinated to make it maximally useful to society. Participants at the Austin Workshop strongly endorsed the conclusions of the two previous Tree of Life workshops that a phyloinformatics infrastructure (here referred to as the Phyloinformatics and Coordination Infrastructure, or PICI) should be created as a crucial element of any ATOL effort. As reiterated by the Austin participants, the potential usefulness of ATOL research to basic and applied biology will depend on having phylogenetic results and the underlying data archived, easily accessible to the user community, and in a form that new associations and interpretations among the data will lead to innovative scientific conclusions. This will require a phyloinformatics infrastructure with sophisticated database and informatics equipment and personnel as well as research capabilities to build the new generation of software and analytical tools that will be required to manipulate phylogenetic data and information more efficiently and make the results available to the global user community.

Moreover, if phylogenetic research is to be undertaken efficiently and effectively, and if the results of research on the Tree of Life are to be delivered quickly to agencies and institutions that need them, as well as to the general public, a synthesis and coordination mechanism will be required. The Workshop concluded that for efficiencies and economies of scale, this coordination mechanism would be most effective if integrated with the phyloinformatics infrastructure.

We therefore recommend that a Phyloinformatics and Coordination Infrastructure (PICI) be established with the following functions and responsibilities:

Primary informatics and global archiving of phylogenetic knowledge

The primary informatics responsibility of PICI will be to archive phylogenetic results and data, including phylogenetic trees and character data of all types. These would be deposited into PICI by the global community of systematics researchers, and, once there, the information would be stored and maintained in such a way as to be retrievable by any potential user.

An additional crucial informatics function of PICI would be to capture retrospectively phylogenetic results and information available in the printed literature or housed in independent electronic databases. The Workshop recommends that this activity be

undertaken intensively for at least five years and after that on a maintenance basis, at a level appropriate and necessary.

At the same time, a primary informatics function will be to develop a database of higher taxon names that will be required for any sophisticated query and search functions of the phylogenetic database. Several global initiatives (such as Species 2000) are underway to create databases of species names, but comparable efforts on the names of groups of organisms are scattered and not coordinated. PICI will develop a list of formal taxonomic names and their synonyms and link those to vernacular (common) names.

Perhaps the most important informatics function of PICI in the future will be the development of capabilities for phylogenetic (node-based) driven queries and data mining of biological databases in order to permit prediction and comparative inferences across biological data. The Austin workshop strongly endorses the vision for this capability developed in the previous two workshops.

Development and delivery of user interface tools and metadata standards

Research results and syntheses from the ATOL effort should be made available to the user community as soon as possible. Attention must be paid to developing interfaces that will serve the needs of a broad user community, encompassing levels from school children to basic and applied biologists. Whereas basic retrieval of phylogenetic data and results can be made available relatively quickly, more sophisticated and integrative and synthetic interpretations will require significant research activities within the biological and computer science communities. In order to foster a range of creative approaches, we see this research developing through the activities of individual investigators and TOLNets/TOLHubs, as well as at PICI.

A high-priority endeavor will be the development of metadata standards, along with an effective query language and efficient tools for data submission. These activities will require a community of systematists and computer information specialists beyond those housed at PICI.

Phyloinformatics research and development to support PICI

As noted in the Davis Workshop report, PICI will require a substantial commitment to onsite research activities that will support the ATOL effort in general, and PICI functions in particular. Some of these activities should include research on the development of a phylogenetic query language, visualization of large trees and collections of trees, methods to combine smaller sets of trees into large "supertrees," resolution and display of phylogenetic ambiguity and conflict, and development of strategies to handle synonymy and resolve conflicts among names for species and higher taxa. Although we recommend these research activities also take place within TOLNets or at TOLHubs, we believe it is essential to have informatics research at PICI to support its ongoing needs.

All three workshops recognized the paramount importance of using the reconstructed Tree of Life as the conceptual foundation for node-based searches of biological databases. Realizing this important function will require significant new research, some of which could appropriately take place at PICI.

Coordinating services for TOLNets and TOLHubs

Assembling the Tree of Life is a mega-science research initiative. It will require significant coordination and common support functions. The PICI is the most logical locus for undertaking these essential general functions that are applicable across all the TOLNets and TOLHubs, although we envision PICI as not providing a top-down coordination for the ATOL effort but as a provider of coordinating services. Some coordinating services that PICI might facilitate would include sharing methods of data capture and analysis, helping TOLNets incorporate new technologies, facilitating common approaches to problems and avoiding duplication of effort, integrating studies that incorporate shared portions of the Tree of Life, and sharing informatics functions such as metadata standards, training, and data storage.

PICI would sponsor and host working groups and workshops on specific research projects pertaining to the ATOL effort and would disseminate their results to the entire community. It is expected that PICI would thus serve as a catalyst for additional research activity.

In addition PICI would be expected to be the internet and WWW gateway to the entire ATOL effort, and provide some support for these activities at the level of the TOLNets and TOLHubs. PICI would also host coordination meetings among TOLNet and TOLHub directors and personnel.

Training activities for Assembling the Tree of Life

Because of its research and coordination functions, PICI could also serve as a point of training at several different levels. Graduate and postgraduate phyloinformatics training would be particularly appropriate. PICI would also be an appropriate location for training support personnel of TOLNets and TOLHubs in subjects such as informatics, web development, and possibly data analysis.

Recommendations for establishing PICI

The Workshop also took up the issue of the physical location and structure of PICI and makes the following recommendations:

Location

Like other national infrastructures, we envision PICI being located at an institution having, in this case, an established track record and ongoing support for phylogenetic and systematics research. Whereas we endorse an institutional affiliation, at the same time we suggest that PICI might best operate quasi-independently of the host institution so as to be seen as providing free and open access and input to the entire community. This could be assured by the appointment of an independent Director and Advisory Board. We also suggest that, if possible, PICI be sited off-campus so as to foster an independent atmosphere and a sense of "ownership" on the part of systematists everywhere.

Another aspect of location considered important by the participants of the Workshop was accessibility and cost-effectiveness. Because of the large number of visitors likely to attend meetings and research and training activities at PICI, we consider it important that PICI be sited at a location with easy year-round access. In addition, to be cost-effective PICI should be located where transportation and accommodation costs are reasonable in aggregate.

Administration

The Austin Workshop participants recommended the appointment of a full-time Director whose background indicates he/she will have the experience, knowledge, and vision to make PICI a success over time. Ideally, that Director would have an affiliation through a staff appointment with the host institution. The Director would be answerable to an Advisory Board (broadly representative of the ATOL effort) and to the National Science Foundation.

In addition to the Director, other key positions should be an Associate Director and administrative support staff; a Director of Informatics with a staff to oversee prospective and retrospective data capture and storage, and database systems development; a Director of Computing who might oversee staff dedicated to systems administration, metadata systems development, user interface development, and others. Graduate and postgraduate positions would also be required.

IV. Specific Issues to be Addressed in Technology and Infrastructure Proposals

A. Specimen and Data Collection

There are three sources of data for the ATOL effort: 1) retrospective information from existing studies, 2) data collected as part of the ATOL effort, and 3) data from other phylogenetics projects that are independent of the ATOL effort. This section addresses

specifically those data collected as part of the ATOL effort. Retrospective data will be deposited and integrated at the PICI and so is not considered specifically in this section.

TOLNets and TOLHubs are the primary mechanisms for the generation of specimen and character data needed to build the tree for all life on Earth. Additional discussion of taxon-based proposals is included in the TOLNet section of this document.

Sampling of Taxa

Proposals will need to address how the selection of taxa will enhance the ATOL effort. The goal of the ATOL is to discover as much of the Tree of Life as possible. So, proposals will be judged primarily on their quality, but secondarily as to how they contribute to the completion of the Tree of Life. After the first several years, calls for proposals may encourage research for specific taxa/clades/ that are underrepresented in the larger initiative. This will be facilitated by the development of the PICI and of TOLNet and TOLHub resources.

One of the aspects of the developing initiative will be consensus building as to the standard for coverage and the sense of when a node is finished.

Taxon-specific issues should be addressed in any TOLNet proposal. For example, for bacteria, visualization may be important but archiving may be problematic (e.g., symbionts).

Specimens

Specimen acquisition

Specimens are the primary source of data for the ATOL effort. Proposals must address how specimens will be acquired. For some taxa, significant new collecting may be needed. Indeed, collecting and identification of new specimens may be one of the rate-limiting steps of the ATOL effort for poorly sampled mega-diverse groups. Cooperation with other biodiversity efforts at an international level will be important.

A core set of data will be submitted to the PICI for each specimen. The data standards for this information will be determined in conjunction with the PICI staff, and should include voucher accession IDS, geoposition of collection site, species name, date of collection. This set of 'minimal' data would not preclude the submission of complete specimen information with the voucher .

Handling of specimens

Specimens must be identified in a timely manner, to allow for removal of redundancies and identification of gaps in sampling. The issue of high throughput analysis of specimens to select representative specimens will be common to most proposals. The use

of innovative approaches, for instance the use of artificial intelligence for pattern recognition or rapid field identification methods, should be encouraged. New species must be formally described according to community standards.

Specimens must be computer-catalogued in the institution that will house them. In addition, the specimen data must be accessible to the PICI via the WWW.

Archiving specimens

A critical part of these efforts will be additional support for maintenance of collections at museums, universities, and culture collections, all of which will experience a significant influx of new material during the ATOL effort. This support will include database activities, curatorial positions, and physical infrastructure.

The ATOL project will not maintain centralized collections. However, management and dispersal of specimen and collection information will be a central component of the PICI.

Data

Character sampling strategies

The selection of character sets for analysis, including gene sequences, will be developed by TOLNets in coordination with the relevant TOLHub. There will be a dynamic tension between centralized vs. decentralized decisions regarding the selection of characters, quality control standards, and sampling strategies. In general, individual TOLNets would have the responsibility of selecting characters, measurement criteria, quality control standards, etc. in a clade-specific manner. However, the PICI can provide guidance, and, in particular, reviews of a TOLNet proposal may encourage specific character analysis for the purpose of integrating different data sets / taxa over a common set of information.

We envision, therefore, a mixed model of centralized and distributed input into project design at any level. Additionally, different TOLNet proposals may include components for character sampling that are best accomplished at a TOLHub site.

Data collection

Data collected in the course of the ATOL effort will be used both for phylogeny estimation and for comparative analyses of phenotypic and genotypic evolution. What data are gathered and how it is collected will depend on a particular study in a particular taxon.

Certain kinds of data collection would benefit from the establishment of one or several facilities for high-throughput analysis. This will be important for methods that rely on expensive equipment (as with x-ray scanners), or where efficiencies of scale are possible (as in genome centers).

High-throughput data collection will clearly be important for nucleotide sequences and other genetic data. Dedicated and well-equipped facilities exist at several scales, from departmental or university-level sequencing centers to national genomics centers (e.g., DOE Joint Genome Institute). A significant investment in infrastructure is probably not necessary for these activities.

High-throughput data collection will also be important for morphological and other types of phenotypic data, including fossil material. The most likely high-tech applications include x-ray scanners, confocal microscopy, environmental scanning electron microscopes, and 3D surface scanners. Other important applications include digital photography, sound collection, and georeferencing of localities and collection sites. Some appropriate facilities exist in many institutions and universities, but others are currently available at only a few places (e.g., an x-ray scanner with trained support staff). Thus, it will be necessary to invest in facilities that can facilitate the collection of geographic, and morphological and other phenotypic data.

Archiving and distributing the data

Data must be stored so as to be retrievable and interpretable. In response to the needs of TOLNets, the PICI will coordinate the development of data standards, and of tools for the deposition and updating of character and geographical data. The deposition of data in the PICI does not preclude the development of other forms of public data access.

It is essential that data be made available to other members of each TOLnet as it is collected. However, these data will also need to be subject to quality control measures, to augmentation, and to potential re-characterization in response to acquisition of new specimens and new types of analyses.

Information about the primary data (metadata) must be coordinated with the TOLHubs and with PICI. This will include information about versions of data sets, when they were last updated, what kinds of data they contain, and which taxa they cover.

Analytical considerations in database design

The research and coordination effort that designs database standards, and the individual research projects that contribute to the ATOL effort, must both result in data that satisfy the needs of the analyses. There are several possible models for the structure of the character and geographic databases that could be entertained. In serving as the raw material for phylogenetic analyses, an important criterion is flexibility, which could be achieved under more than one database model. New methods for estimating phylogenetic trees are increasing at a rapid rate, and the data must be structured in a way that facilitates the analysis as new methods become available and estimation models are refined. Moreover, there are differences in opinion as to the most appropriate methods of analysis. The database must therefore provide the ability for investigators to obtain the necessary primary data. It should always be possible to edit or re-analyze the data at any

level of the hierarchy and regenerate the full analysis (subject to this change) in some semi-automated way. The flexibility of the database not only allows it to be agnostic with respect to the choice of methods, but also to facilitate the development of newer and better methods as our knowledge of the problem improves.

In addition to providing the necessary flexibility, the database design must address the following issues:

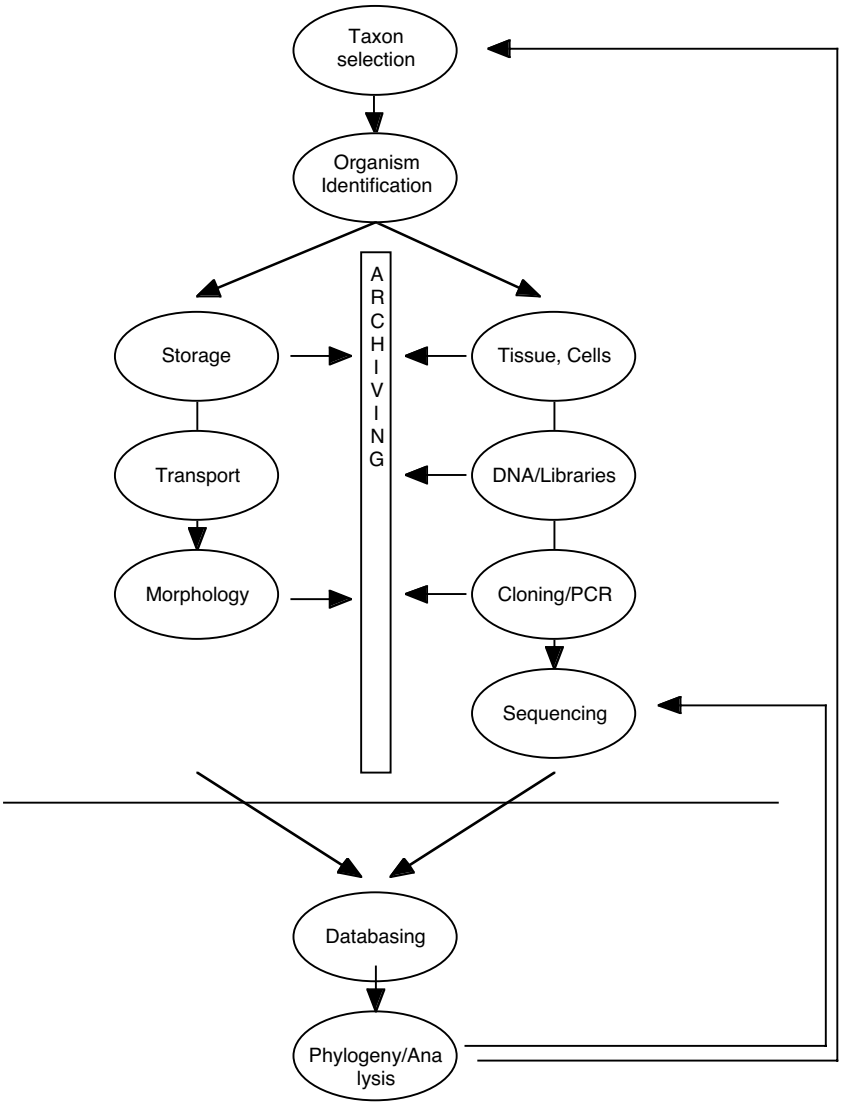
- A mechanism must be provided for assessing the reliability of individual data items, as well as a method for easily omitting subsets of substandard data. For example, DNA sequences could be coded with reliability values of base calls for individual sites or regions; morphological data could be scored as to whether the data were culled from the literature or verified through direct observation).
- The database must permit all relevant data for a set of taxa to be easily input to phylogenetic analysis software (e.g., all available DNA sequence data from a variety of genes, morphological and other phenotypic data relevant to a specific phylogenetic problem).
- Appropriate metadata standards must be established so that adequate information about data and analyses is an integral component of the database. These metadata standards will include the conditions under which the data were collected and detailed description of the character data (e.g., documentation in the form of images for morphological characters, protocols used for amplification and sequencing, references for previously published information).
- In addition to the primary data and the metadata, the database must also store information about the results of phylogenetic analyses (trees, support values, and other relevant statistics) as well as the methods used to generate them.

Summary

We discussed here issues concerning specimen and character selection, sampling, analysis and archiving. Many decisions about these issues will follow existing community standards. The aspects of specimen acquisition and data collection that are particularly affected by the ATOL effort are those relating to centralized high-throughput data collection, and the imposition of ATOL standards beyond the needs of a particular proposal. The objective is to keep the larger ATOL goal in mind, to leverage technological development elsewhere, and to support special technological development within the ATOL community that will expedite and standardize the handling of specimen and character data.

B. Data Analysis

The challenge for the data-analysis component of the ATOL effort is to take character data sampled from a vast number of individual taxa and propose a phylogenetic tree that best represents the evolutionary history of these organisms. The problem can be summarized in the following figure:



Using character data sampled from individual organisms, find a phylogenetic tree that best represents the evolutionary history of life

Each of the arrows in the figure represents an inference step for which there is a certain level of uncertainty (in addition to the uncertainty inherent in the raw data); these uncertainties must be propagated as we move from the raw data to the final estimate. It is critical that the final estimate not only provides a useful representation of our accumulated knowledge but also honestly represents the uncertainty associated with this

estimate. There are many unsolved problems associated with phylogenetic analysis. The ATOL effort must cope with these, and must also address a number of new challenges for phylogenetic theory and data analysis. In the following we outline some general considerations that should guide the interplay between the data collection/archival efforts and the analysis component of the project. We outline some of the unique problems posed by the ATOL effort, and make recommendations for the infrastructure necessary for addressing these issues.

General issues involving data analysis

One approach to inferring a Tree of Life would involve a single phylogenetic analysis that includes all available data from all taxa sampled. However, such an analysis is impractical for a number of reasons; compatible data will not generally be available across life and the sheer size of the phylogenetic tree poses computational problems that will be insurmountable for decades, if not forever. It is clear, then, that a "supertree" approach will at some level be necessary as an approximate surrogate to a "supermatrix" approach. Nonetheless, maximizing the size of the component subtrees will probably provide the most reliable inferences. Therefore, even the subtree analyses will be, in many cases, much larger than the largest problems yet tackled by phylogeneticists.

Specific issues that will need to be addressed include:

- If a supertree approach is involved in estimating the Tree of Life, its methods require further development. First, the uncertainty inherent in each of the component subtrees must be accommodated. New methods for accomplishing this must be developed, and the methods for assessing the reliability of the individual subtrees must be refined. Second, some overlap between component subtrees is a necessary requirement of the approach. However, it is unclear at present what the minimal amount of overlap might be, nor what the optimal tradeoff between subtree size and degree of overlap among subtrees should be. Because this aspect of the analysis is so critical to the quality of the final result, this should be an area of intensive investigation.
- While the goal of the ATOL effort is to produce a single tree representing the history of life, this may not be feasible for a number of reasons. First, processes such as horizontal gene transfer, hybridization, and lineage-sorting either cause the pattern of evolution to be non-tree-like or obscure the tree structure. Second, even when the underlying history is treelike, considerable uncertainty about the tree structure cannot be avoided. This raises interesting and important issues. First, we must ask whether it is reasonable to provide a single, highly resolved tree to the scientific community without indicating which areas of the tree are ambiguous or poorly supported. Providing this kind of information in a concise and intuitive format represents a line of investigation that must involve statisticians (who fundamentally address issues of uncertainty) as well as computer scientists interested in the visualization of complex data structures and in the compact representation of uncertainty. Second, methods that explicitly

accommodate processes that cause non-treelike structure or obscure a treelike pattern must be developed. For instance, coalescence theory provides a basis for accommodating multiple gene trees imbedded within a single species tree.

- Most methods of phylogenetic analysis involve choosing an objective function or optimality criterion that allows any tree to be scored, and a method for selecting trees that are optimal under that criterion. Because the choice of an optimality criterion can, in at least some cases, have a dramatic impact on the inferred tree, it is critical to explore the properties of these criteria and the suitability of their assumptions. In fact, for many methods that explicitly incorporate assumptions about the evolutionary process, we need to know more about the robustness of these methods to violation of their assumptions and to develop models with less restrictive assumptions where possible.
- The choice of taxa to include in a phylogenetic analysis can have a strong impact on the reliability of the final result. Paradoxically, it is often easier to achieve high accuracy in simulation studies by including large numbers of densely sampled taxa even though the number of possible outcomes grows exponentially with the number of included taxa. However, dense taxon sampling cannot be assumed to represent a panacea for the problem of inferring phylogenies reliably. Even for the complete true Tree of Life, there will exist some regions that are sparsely branching or that contain long, undivided branches, which are particularly problematic for some optimality criteria. Emphasis should be placed on developing strategies for identifying areas of the tree where reliability would be enhanced by improved taxon sampling. The identification of methods that are less sensitive to taxon-sampling issues should be encouraged.
- One of the most active areas of current research by advocates of criterion-based methods is the discovery of new algorithms for searching tree space for optimal solutions. The ATOL effort, which will involve the analysis of data sets that are orders of magnitude larger than current analyses, will necessitate the development of more powerful methods of tree-searching. Although promising new developments are occurring (e.g., simulated annealing and genetic algorithms, perturbation-based heuristics, Markov-chain Monte Carlo), much work remains to be done in improving these methods and the discovery of radically new approaches does not seem out of the question. This is an area where intensive collaboration among mathematicians, computer scientists, and biologists will be necessary to develop exciting new technologies while maintaining relevance to the most important biological issues at hand.
- The discovery and refinement of new search algorithms must be accompanied by the development of good software for implementing these methods, which in many cases will require intensive interaction between biologists and computer scientists (and possibly engineers, if the implementation of the new methods can be enhanced by the development of special-purpose hardware). We anticipate that many of the new methods will involve new and innovative strategies for exploiting parallel computer architectures. In fact, the optimal level of parallelism is not obvious. Many problems can best be solved on relatively inexpensive computer clusters, whereas some problems will require a more fine-

grained approach using shared memory and tight processor integration. Biologists have much to gain from the experience of computer scientists involved in this active area of research.

- One of the least satisfying aspects of current phylogenetic practice is the need to separate the steps of sequence alignment and tree estimation. A much more desirable strategy involves estimating a phylogeny that treats insertion, deletion, and substitution events directly rather than requiring a prior alignment in which gaps are used as placeholders. There are a number of reasons why the current approaches have been suboptimal. First, traditional methods typically must treat gaps as "missing data" and do not make use of the phylogenetic information provided by gaps. Other strategies for making use of indel information are limited by the inappropriateness of treating gaps spanning more than one nucleotide position as independent events. Secondly, even for an alignment that does not contain gaps, computer programs effectively treat the alignment as an observation when in fact the alignment is one step removed from the true observations—the sequences themselves. Ideally, the tree building process should accommodate the uncertainty in the alignment process when inferring phylogenies in a manner that is not *ad hoc*. Although some progress has been made in this area (particularly for the parsimony criterion), this represents the most difficult computational challenge facing any attempt to estimate extremely large phylogenies such as the entire Tree of Life, and its solution will require the concerted efforts of biologists, computer scientists, statisticians, and other scientists.
- Although most of the analytical issues involve traditional sources of information such as macromolecular sequences and morphology, a number of additional sources of data are becoming available for use in phylogenetic information. We need to develop new ways of incorporating "nonconventional" sources of data such as genome organization (e.g., gene order, duplications), secondary structure, and gene function (including expression/microarray data) into phylogeny estimation.
- Not only can new sources of information be incorporated into phylogenetic estimates, we need to continue research into ways of combining diverse data sets (morphology, sequence data, and other sources of data mentioned above) into a single analysis. This approach has been advocated as ideal (or even mandatory) by many, but it brings a new set of challenges when applied to projects as large and complex as the ATOL effort. For example, character sets will only partially overlap for many of the taxa included in analyses of data sets across the more diverse lineages of the Tree of Life. It is unclear how the analysis of matrices containing many missing elements will affect the reliability of the estimates. This is an area where simulation studies might be particularly useful in evaluating methods.

In addition to the data-analysis issues discussed above, there is another suite of problems more oriented toward the application of the reconstructed Tree of Life to study other aspects of evolution (e.g., biogeography, character evolution, epidemiology, genome organization and change). Much work remains to be done in this area, as many of the

new methods do not account for uncertainty about the trees or make unreasonable assumptions about the evolutionary process.

Support structures

One of the greatest impediments to making bold discoveries in data analysis is the lack of communication among theoretical biologists on the one hand and computer scientists, mathematicians, and statisticians on the other. It is often the case that one investigator has either clearly formulated a problem or made considerable advances in solving a problem, but the critical information for a complete solution resides with another theoretician with largely non-overlapping interests and training. Bringing these people together in an environment that promotes long-term and close interaction could in turn promote the development of exciting new methods for data analysis relevant to the ATOL effort. Often solutions to complex problems are found but these solutions are not transferred to the potential consumers in the form of usable software, or lack an adequate description necessary for implementation by others. Finally, many of the methods that are developed in the course of the ATOL effort will undoubtedly be very complex. There must be a mechanism for transferring information about the biological assumptions made by new methods to the user-community in such a way that individual biologists can judge for themselves the appropriateness and utility of any new method of analysis for their own work.

Two factors contribute heavily to the situation described above. First, current administrative structures at universities and other research institutions tend to isolate biologists and non-biologists who have much to gain by knowing each other better. Within an institution, these individuals are nearly always housed in different departments in different buildings. Furthermore, even if this separation were overcome, most single institutions lack the critical mass of scientists with complementary skills and interests necessary to achieve the synergy that is needed for solution of these complex problems. The second factor is simply a "language barrier" to effective communication between disciplines. These language differences range from fundamentally different styles of communicating ideas within disciplines to matters as basic as the terms used to describe parts of trees (e.g., edges versus branches, binary versus dichotomous, and leaves versus tips).

C. Integrating Existing Knowledge in Systematics

Systematics and taxonomy are not new fields, having existed for more than 200 years. There are two reasons the ATOL effort must make use of, and integrate into, this existing knowledge base. First, systematics and taxonomy provide us with information about species already discovered and catalogued, character systems and methods already studied, and phylogenetic hypotheses that can provide a basis and guide our further sampling. Even if we currently have only a fraction of the species discovered, of the characters sampled, and of the analyses performed that we would need to resolve

adequately the Tree of Life, the existing knowledge is an indispensable beginning. Second, existing biological data of all sorts, from the molecular to the ecological and evolutionary, have been stored by reference to the contemporaneous names and classification of the organisms studied. The Tree of Life that arises from the ATOL effort must be linked to those data. To do so, it must be possible to translate between the reconstructed Tree of Life and the existing classifications.

Integrating the ATOL effort with existing systematics knowledge may be particularly needed in these respects:

- As noted under "Specimens and Data Collection," specimens need to be identified to be used effectively in the ATOL effort. Even if identification may be, conceptually, no different from placement by phylogenetic analysis, preliminary identification to currently recognized taxa (whether species or clades) will often be an important filtering procedure in specimen sampling schemes. In some groups identification will be easily accomplished, in others not, especially where species vastly outnumber specialists.
- A major contribution of our current knowledge will be to guide taxon sampling. Species chosen for sampling can be distributed throughout our current preliminary phylogenetic arrangement in whatever way is optimal according to our theories of taxon sampling. In addition, knowledge necessary for acquiring fresh specimens will be embedded in systematics works, including geographic and phenological distributions and even collecting techniques.
- As noted under the description of the PICI, in developing and implementing the databases it will be necessary to create a database of higher taxon names and their synonyms, so that queries can accommodate, as well as possible, varied and possibly changing naming conventions. The research required to design such a database will involve both computer science and biology, but the implementation of the database will deal very much with existing biological literature. To answer the question "to what does and did the name Insecta apply?" requires knowledge of the group itself and its literature, including issues of characters and phylogenetic interpretations.

Satisfying each of these needs for integration will depend on existing systematics knowledge. Published monographs, classifications and phylogenetic works may be sufficient in many groups, but in other groups the best knowledge will reside only in the unpublished wisdom of aging systematists. In that regard, NSF's PEET program and efforts of that sort may be vital to the efficiency of the ATOL effort in many groups of organisms. Even in those groups for which published works will suffice to guide us, there may need to be a non-trivial effort to gather and synthesize these works. The needed syntheses include not only the recapturing of published quantitative phylogenetics analyses discussed in the context of the mission of the PICI, but also the gathering of traditional classificatory treatments in those groups lacking recent phylogenetic work.

There will be tradeoffs in addressing existing systematics knowledge, because we need to make use of existing expertise without becoming bogged down in fine details of

synonymies and nomenclature, for example. It is clear that if we are to succeed with the ATOL effort, we must start fresh and focus on efficiency. Efficiency will dictate that we make effective use of existing knowledge.

D. Technology Development and Transfer

High-throughput data collection

The success of the Tree of Life project will depend in part on improvements in the ability to acquire phylogenetically informative characters for a large number of taxa. In some cases, these goals may be met by work at individual facilities. The attention of experts is required for several aspects of the ATOL effort and it is here where their attention is best devoted. Alternatively, the technology could be developed in centralized resources that would be accessible to scientists working on ATOL projects. Centralized resources have many potential advantages, including improved linking to curation and database facilities, economies of scale, and the ability to change rapidly to new technologies. While centralized facilities should not be a requirement of any particular project, if such resources are available (e.g., DNA sequencing), they should be used whenever feasible, scientifically valuable, and cost effective.

Examples of steps in need of development

Improvements that will be needed include: (1) minimization of cost; (2) maximization of speed; (3) increase in accuracy; and (4) increase of access to technology. The processes in most need of development depend in part on the type of data (e.g., genotypic vs. phenotypic), the taxa, and the nature of the study (e.g., fossil vs. extant species). We have outlined some of the steps in the gathering of data for different processes below.

Molecular or genotypic characterization

(1) Identification and collection of organisms
(2) Isolation of tissues or cells
(3) Isolation of DNA
(4) Cloning or amplification (e.g., PCR) of DNA, RNA, or individual genes.
(5) Sequencing or other analysis of molecules

Phenotypic studies

(1) Collection and identification of organisms
(2) Storage and stabilization of collections
(3) Analysis of selected traits and characters

Specific steps currently in need of improvements

Although all aspects of gathering of phylogenetically informative characters could use improvement, there are some areas in clear need of development that would have immediate benefits on the ATOL projects. Emphasis should be placed on those areas that are currently rate-limiting or very costly. These include the following:

Morphology and other aspects of phenotype characterization

Phenotypic characterization is an incredibly important component not only of phylogenetic analysis, but in species identification and even collection. Included are studies of ontogeny, behavior, physiology, etc. Such studies are currently quite expensive (per taxon), not readily archivable, and time consuming. Therefore, proposals that would likely produce improvements in the speed, accuracy, cost, and accessibility of morphological studies should be greatly encouraged. In addition, those methods that could be used or adapted to field setting would be very beneficial in species identification and collection and therefore may have a great impact on some of the rate-limiting steps in ATOL projects.

Extraction of DNA for molecular studies

Molecular characterization, such as DNA sequencing, will also likely be a major component of ATOL projects for many types of organisms. A major rate-limiting step in such studies is the extraction and availability of DNA from multiple taxa. The actual DNA sequencing step is already highly automated, relatively cheap and will continue to see improvements that will be driven by other fields of research. However, the ability to isolate DNA for the construction of libraries is still very cumbersome for many taxa and needs improvements. Proposals that seek to improve DNA isolation and library construction, as well as making such methods available to many people are to be encouraged. Methods to provide access to DNA, once isolated, to many researchers (e.g., library construction and archiving) will also be beneficial and should be encouraged.

Identification and collection of organisms and the isolation of tissues generally will require the work of experts, as will the analysis of either type of data. Other steps may benefit from automation and centralization. We encourage DNA extractions, PCR, and library construction to be done in a centralized location using automated, high-throughput techniques.

Other areas of need for technological improvements or development

1. Archiving primary or secondary materials (e.g., specimens and DNA).

2. Archiving information/databases. All information must be readily accessible. This responsibility would be part of the PICI.

3. Need for interactions across all steps in a process and with people developing technology for these steps. For example, tissue and cell extraction could be done in a one lab, but the DNA extraction could be done elsewhere. Those developing methods for rapid DNA isolation should be in regular communication with those developing methods for tissue storage and isolation.

4. Collaborations with industry. Many aspects of technology for the ATOL effort could be of great benefit to other areas of science and society.

5. Inclusion of those with current expertise in these areas. For example, museums have already developed methods for rapid curation and description of materials gathered in different field sites, as well as voucher methods for keeping track of samples. Proposals that would seek to adapt methods developed in other fields or areas of research for the ATOL effort should be encouraged. That is, don't reinvent the wheel; modify those wheels that are out there already.

6. Quality control of data gathering is going to need to be assessed. This will be of concern for both high throughput facilities as well as for individual research labs. The entire effort will suffer if the quality of the data or of the linking of data with proper identifications and voucher specimens, is not of the highest level. Therefore, all proposals should address quality control issues.

Control of and Credit for Research

For the ATOL effort to work, it will be necessary to make sure that those scientists that are involved get adequately credited for their contributions and are able to maintain adequate control of their materials and data. Although making data and materials accessible to as many people as possible will be important, some accommodation may be necessary to allow scientists to retain primary access to data prior to publication. For example, if there were a project to sequence all the rRNA genes from a group of bacteria, a scientist might be allowed to send samples to a central sequencing facility without being required to immediately release the sequence data to the web. Some middle ground may be necessary to ensure full participation of all scientists.

Whole genome projects are underway by several organizations that will be provide data complementary to the ATOL effort. Every effort should be made to coordinate these projects and to target data that will be of mutual benefit. For example, phylogeneticists could well advise on taxon selection for whole genome sequencing and genome scientists could provide data for analysis of genome level features at certain critical nodes of the phylogeny of life.

D. Training

The Problem

There currently are not enough trained people to complete the Tree of Life in a timely manner. Many taxonomists are retired or nearing retirement. Many species may go extinct before they are discovered and studied. Students (and faculty) need to be broadly trained in specimen collection and field work, morphological techniques (e.g. scanning EM; confocal computed tomographic imaging), molecular techniques, analytical methods, evolution of development, and web-based presentation (key construction, species description, and information distribution). Training must change through time to keep up with latest techniques. New categories of funding are needed specifically for ATOL students and faculty to travel among TOLHubs and TOLNets and to attend field courses. Training is needed to broaden perspectives on biology. Students need to be trained in the "ologies" (e.g. mammalogy, entomology) to replace organism-based courses that have been eliminated from many universities. Rapidly developing new technologies mean that faculty also need retraining and time to read, think and learn.

Recommendations for Improving Training

ATOL should offer grants for undergraduate students, graduate students, postdocs and faculty in the following areas:

TOLHub grants for travel/per diem to visit TOLHubs for several weeks, several months, or a year or more. Courses could be held at TOLHubs with local staff and visiting faculty. The hubs would have the technologies/equipment too expensive to be duplicated at each TOLNet node. Students could spend time at TOLHubs to learn theory, analysis, and molecular/morphological techniques.

Workshops and Field Courses: We need more undergraduate programs like the Smithsonian Summer undergrad research course and graduate programs like the Woods Hole Summer molecular systematics course. Because the collection of specimens is expected to be a bottleneck step in the ATOL, we also need field techniques training courses. A new course could be developed or we could use existing courses at various biology field stations. For example, Organization for Tropical Studies courses could be designed to include more systematics and biodiversity training along with the standard broad training in ecology and evolutionary biology. An intensive field or lab program would serve to develop a network of ATOL students and faculty who join a community of scientists with a common history.

TOLNet travel/per diem grants for students and faculty to facilitate travel among nodes of the TOLNet to which they belong.

Special research fellowships for undergraduate and graduate students not already at appropriate ATOL institutions. These would provide stipends to complete a degree or to work for a semester or two with experts on a particular group of organisms.

Special postdoctoral fellowships for ATOL participants would provide support for some of the most skilled and productive members of our systematics community.

Faculty Development Grants for faculty sabbaticals and short term learning experiences would provide valuable time to think and opportunities to learn new methods. We need to include scientists from computer science and other disciplines relevant to the project. Time is also needed to develop courses and curricula for ATOL training at individual nodes of the nets.

Recommendations to Increase Recruitment

- Money is needed for web development. We need to create "training opportunities" and "careers in systematic biology" information on the ATOL website. This page would include links to relevant institutions, investigators, and societies around the world.

- The courses and grants for undergrads mentioned above can be advertised on the web page and in relevant journals and will help recruit new students into systematic biology.

V. Coordination, Timing, and Implementation

Two general issues are considered here: the need for immediate coordination activities and the nature of the initial implementation steps.

A. Coordination with other efforts

Even if the ATOL effort becomes a national priority, the general effort of reconstructing the Tree of Life will not be restricted to the United States. It must be an international effort, because residing beyond the boundaries of the country are both substantial expertise in biological systematics and the bulk of the diversity of species to be sampled and studied. The ATOL effort must be, and be viewed as, an international effort with scientists from around the world participating.

To ensure that the systematics community worldwide accepts and commits to the broader enterprise, it is vital that an effort be made to reach out internationally. This might take the form of email and other low-effort means to invite and communicate, but it would probably be more effective if there were a more tangible commitment to international involvement, such as the sponsorship or hosting of an international meeting.

In addition to involving the systematics community at large, it will be important to coordinate with other institutions and organizations that could make valuable

contributions to the goals of the ATOL effort, by supplying direct contributions to the biological studies, expertise in managing an effective big science project, and assistance communicating its results. These could include:

- Organizations and people interested in biological diversity (NGOs, public and private sector). For example: All Species Initiative, Convention on Biological Diversity (CBD), Conservation International, World Wildlife Fund, Mellon, Sloan and/or MacArthur Foundations, CSIRO, directors of major international museums and/or botanical gardens, heads of key scientific societies (SSB, ASC, Hennig Society, AIBS), key scientists from mega-diverse countries (i.e., from Latin America, Africa, Asia, Australasia), well-respected international systematists.
- Managers of big projects from public and private sector. For example, in addition to those mentioned above: NASA Space Program, NIST/DARPA, LIGO, NCEAS, SDSC, NCBI, NIH big equipment/facilities institute/department, Earthquake monitoring project, Lee Hood's operation/dept at U Wash, Human Genome Project, Arabidopsis Genome project, LTER network, CELERA, systems engineers and information management specialists from private sector .
- Representatives of relevant agencies that might go in for crosscut funding—for example: NRC, NIH, DOE, USGS, Smithsonian, USDA, ATCC.
- Journalists and science writers, e.g. from *Science, Nature, New York Times, Wall Street Journal, Fortune*.

B. Timing and Implementation

Both RFPs for a PICI and for supporting research contributing to the ATOL effort (e.g., TOLNets/TOLHubs) would be essential at the outset of this effort; the RFPs should be issued simultaneously.

The PICI is essential immediately in order to be prepared to capture data as soon as it starts to be gathered, and to begin recapturing already published data into an accessible database. Included in this is not only setting up the hardware and software for data bases, but designing data storage standards so that data are captured in appropriate ways. In addition, it is vital that the systematics community as a whole become involved in the ATOL effort as soon as possible, and the PICI will facilitate that as it coordinates the designing of standards and implements a shared resource (the databases).

RFPs for proposals to do empirical and methodological studies for the ATOL effort should also begin immediately, because the size of the task at hand demands beginning quickly. The reason to begin research efforts of all kinds simultaneously (methodological and empirical, from sampling through analysis and databasing) is that, as we undertake this major effort at a scale not previously tried, we will need reciprocal illumination between methods and data to ensure efficient progress. Initial efforts should leverage existing expertise, and enhance its effectiveness. For those questions for which existing expertise is distributed at various institutions, research following the TOLNet model

should begin quickly to network this expertise for efficiencies of effort and economies of scale.

The Workshop recognized the value of networking scientists at various institutions, and the value of enhancing research and training at institutions strong in ATOL-related activities. As noted in the introductory remarks to section III, those two values might or might not be realized as two separate competitions (TOLNet and TOLHub respectively). Three possibilities regarding timing of RFP's emerged in the workshop:

- RFP for TOLNets first at the same time as the initial RFP for the PICI, then RFP's for TOLHubs later, once TOLNets have established a track record
- Simultaneous RFP's for TOLHubs and TOLNets at the same time as the initial RFP for the PICI
- No distinction made between TOLNets and TOLHubs, and thus there would be a single RFP for ATOL research (whether the research proposed was to be answered via a Net-like or Hub-like organizational model), simultaneous with the initial RFP for the PICI

Regardless of the structure of the competitions, it was agreed that the research answering ATOL questions needs to begin as soon as possible. The Tree of Life is critically important for the entire field of biology, and its rapid and accurate resolution should be an international research priority.