ASSEMBLING THE TREE OF LIFE

HARNESSING LIFE'S HISTORY TO BENEFIT SCIENCE AND SOCIETY
This document is based on three National Science Foundation Tree of Life Workshops held in 1999 and 2000 at Yale University, the University of California Davis, and the University of Texas Austin. See http://research.amnh.org/biodiversity/center/features/tol.html for full reports of each workshop.

Editorial Committee: Joel Cracraft, American Museum of Natural History
Michael Donoghue, Yale University
Jerry Dragoo, University of New Mexico
David Hillis, University of Texas
Terry Yates, University of New Mexico

Participants:
Meredith Blackwell, Louisiana State University
Judy Blake, The Jackson Laboratory
Rob DeSalle, American Museum of Natural History
Jonathan Eisen, The Institute for Genomic Research
Douglas Futuyma, University of Michigan
Jacques Gauthier, Yale University
Pablo Goloboff, University of Tucuman, Argentina
Darlene Judd, Oregon State University
Junhyong Kim, Yale University
Meredith Lane, Academy of Natural Sciences
Paul Lewis, University of Connecticut
Diana Lipscomb, George Washington University
Francois Lutzoni, The Field Museum
Wayne Maddison, University of Arizona
Brent Mishler, University of California Berkeley
Maureen O’Leary, State University of New York Stony Brook
Jeffrey Palmer, Indiana University
Kathleen Pryer, Duke University
Michael Sanderson, University of California Davis
Pam Solotis, University of Florida
James Staley, University of Washington
Naomi Ward, Louisiana State University
Quentin Wheeler, Cornell University
Ward Wheeler, American Museum of Natural History
Kevin P. White, Yale University
Brad Shaffer, University of California Davis
Gavin Naylor, Iowa State University
Peter Cranston, University of California Davis
Chuck Delwiche, University of Maryland
Tom Bruns, University of California Berkeley
Dirk Redecker, University of California Berkeley
Roderick D. M. Page, University of Glasgow
David Maddison, University of Arizona
Emilia Martins, University of Oregon
Bill Piel, University of Leiden
Tim Lilburn, Michigan State University
Michael Freeston, Kings College, Aberdeen
Jim Gannon, Parabon Computation
Susanne Chambers, Parabon Computation
Hasan Jamil, Mississippi State University
Peter Karp, SRI International

NSF Observers: Matt Kane, Mary McKitrick, James Rodman, Joann Roskoski, Terry Yates, and Grace Wyngaard.

This document is an outgrowth of discussions and the reports of the three workshops, which were supported by a grant (DEB-0089975) from the National Science Foundation to the American Museum of Natural History. Any opinions, findings, conclusions, or recommendations expressed in this document are those of the participants, and do not necessarily represent the official view, opinions, or policy of the National Science Foundation.
One of the most profound ideas to emerge from modern science is the realization that all life, from the smallest microorganism to the largest vertebrate, is connected through genetic relatedness on a vast evolutionary tree. This Tree of Life provides the framework for much of our modern understanding of biology because it reveals the diversity of life as well as the historical basis for similarity and differences among organisms. Increased knowledge of phylogenetic relationships will improve human health, push the frontiers of comparative developmental biology, meet threats to agriculture and forestry from invasive species and pests, and improve management of our natural resources. Perhaps most important, without substantial growth in our knowledge of the Tree of Life, it will become increasingly difficult and inefficient to manage, understand, and manipulate biological information held in numerous databases worldwide, including the burgeoning information from the genomic sciences.

The hierarchical arrangement of this phylogeny provides a comparative and predictive framework for all fundamental and applied biology. Our understanding of the Tree of Life has advanced rapidly over the past decade fueled by enormous progress in the fields of genomics and information technology. Now, with new theoretical and technical innovations that cut across many areas of scientific research, systematic biologists are poised to develop a comprehensive understanding of life’s history that will advance all biology and provide enormous benefits to society.

The Tree of Life depicts the evolutionary relationships of Earth’s taxonomic diversity — including all living and extinct forms — over the past 3.5 billion years of its existence.

One of the most profound ideas to emerge from modern science is the realization that all life, from the smallest microorganism to the largest vertebrate, is connected through genetic relatedness on a vast evolutionary tree. This Tree of Life provides the framework for much of our modern understanding of biology because it reveals the diversity of life as well as the historical basis for similarity and differences among organisms. Increased knowledge of phylogenetic relationships will improve human health, push the frontiers of comparative developmental biology, meet threats to agriculture and forestry from invasive species and pests, and improve management of our natural resources. Perhaps most important, without substantial growth in our knowledge of the Tree of Life, it will become increasingly difficult and inefficient to manage, understand, and manipulate biological information held in numerous databases worldwide, including the burgeoning information from the genomic sciences.

Resolving the Tree of Life is unquestionably among the most complex scientific problems facing biology and presents challenges much greater than sequencing the human genome. The entire Tree of Life is almost unimaginably vast. Although 1.75 million species of organisms have been discovered and described, it is estimated that tens of millions remain to be discovered. Placing these species on the Tree calls for increased amounts of information about each, as well as new analytical tools to build our understanding of the relationships among them.

At present, we know relatively little about the phylogenetic relationships of Earth’s species or even among many of the major branches of the Tree. Only 60 or 70 thousand species have been studied and even for these the data are far from complete. The lack of a universal tree is severely hampering progress in many areas of science and limiting the ability of society to address critical problems affecting human health and environmental management.

Nevertheless, we have reached a turning point. The convergence of three important developments—conceptual and methodological advances in phylogenetic analysis, the rise of comparative genomics with its vast quantities of data, and rapid advances in information technology and processing—have now made possible the construction of a robust Tree of Life depicting the genealogical relationships of all known species.

Although many scientific challenges still remain, they also represent opportunities to advance integrative solutions across numerous scientific disciplines. The size and complexity of this endeavor will require vision, sufficient human resources, and coordination and collaboration at an international level.

Yet, assembling an accurate universal tree depicting the relationships of all life on Earth, from microbes to mammals, holds enormous potential and value for society. It is imperative that we begin now.
The Tree of Life: Benefits to Society through Phylogenetic Research

Using phylogenetic analysis to discover new life forms for biotechnology

Phylogenetic analysis is playing a major role in discovering and identifying new life forms that could yield many new benefits for human health and biotechnology. Many microorganisms, including bacteria and fungi, cannot be cultivated and studied directly in the laboratory, thus the principle road to discovery is to isolate their DNA from samples collected from marine or freshwater environments or from soils. The DNA samples are then sequenced and compared in phylogenetic analyses with the sequences of previously described organisms. This has led to major new discoveries.

For several decades microbiologists have been searching for new bacteria in extreme environments such as hot springs or marine hydrothermal vents. The thermal springs of Yellowstone National Park have yielded a host of new and important bacterial species, many of which were identified using phylogenetic analysis of DNA sequences.

The most famous bacterium from Yellowstone is *Thermus aquaticus*. An enzyme derived from this species — DNA Taq polymerase — powers a process called the polymerase chain reaction (PCR), which is used in thousands of laboratories to make large amounts of DNA for sequencing. This discovery led to the creation of a major new biotechnological industry and has revolutionized medical diagnostics, forensics, and other biological sciences. Many microorganisms in extreme environments may yield innovative products for biotechnology.

Fungi — an unknown world revealed by phylogenetic analysis

Fungi are among the most ecologically important organisms. By feeding on dead or decaying organic material, fungi help recycle nutrients through ecosystems. Additionally, fungi are important economically as foods and as biotechnological sources for medicines, insecticides, herbicides, and many other products.

About 200,000 species of fungi are known, but there may be millions more to be discovered because most are extremely small and found in poorly studied habitats such as soils. Increasingly, phylogenetic analysis is being used to discover new microfungi through isolation and sequencing of DNA. Biological studies on these new species hold great promise for developing novel natural products.
Knowledge of the phylogenetic relationships of the world’s crop plants and domestic animals is crucial for understanding their origins and for their continued improvement. Over the past few years new discoveries have been made on the relationships of the major cereal grains, especially wheat, rice, and corn, as well as other crops such as tomatoes. Nevertheless, the origins of many other important sources of food and fiber upon which large segments of the human population depend remain uncertain. Identifying the close relatives of these organisms is crucial from the standpoint of potential genetic improvements. The task is urgent because these wild relatives may become extinct due to habitat loss before they can be studied.

Cultivated maize (or corn, *Zea mays mays*) is of enormous economic value, and its continued improvement is of paramount importance for sustaining the human population. Theories on the relationships of corn to other grasses have varied widely, owing to the highly unusual structure of the familiar fruiting stalk or cob. Attention has focused on several grasses from Mexico and Guatemala known as “teosintes,” although the exact relationships among these plants has remained unclear until recently. Armed with new molecular and computer techniques, scientists have identified two teosinte lineages as the closest relatives of cultivated maize. This phylogenetic knowledge has helped trace the origin of cultivated maize to 9,000 years ago in the highlands of Mexico and to identify the precise genetic changes responsible for the highly modified corn cob and other unique features of the modern cultivars.

The Tree of Life is essential for interpreting the origin and subsequent modification of developmental genes called HOX genes during the history of animals (above). These genes help specify the development of the animal body plan and thus are the subject of intense scientific study within fundamental and applied biology. Early animals such as sponges (Silicea and Calcarea) appear to lack a HOX gene complex. A small set of genes is present in the more advanced jellyfish and sea anemones (Cnideria), but the HOX gene cluster becomes even more elaborated in bilaterally symmetrical organisms. Developmental biologists could not fully understand these genetic and developmental changes without the information afforded by the Tree of Life, which provides a comparative framework for deciphering the genomic milieu underlying developmental processes. As knowledge of relationships improves, new insights into how genes and their functional interactions have changed over time will be possible. This, in turn, will lead to greater knowledge about normal and abnormal development, which will contribute to improvements in human health.
Using phylogenetic knowledge to discover snakebite antivenins

Australia has more poisonous snakes than any other continent, and many people die from snakebites each year. Developing effective antivenins is thus a high priority, but little is known about the venins of most species. Phylogenetic analysis is helping with this task because venin properties correlate strongly with evolutionary relationships. Although the red-bellied black snake looks very different from the king brown, it is actually closely related and can be treated with the same antivenin. Conversely, the western brown looks very similar to the king brown, but it is only distantly related and thus responds best to different antivenin. The phylogeny is also predictive: the recent demonstration that the poorly-known bardick is closely related to the death adder (orange lineage) predicts that the former is also highly dangerous and might respond to widely-available death adder antivenin.

Using phylogenetic analysis to protect ecosystems from invasive species

Invasive species are a worldwide problem. The United States spends about $137 billion annually to control or mitigate the effects of exotic species, thus identifying potential invasives in ecosystems is of crucial economic and ecological importance. Phylogenetic analysis is an important tool in this effort.

An exotic alga was recently found at several locations along the California coast and scientists used phylogenetic analysis of DNA sequences to identify the alga as *Caulerpa taxifolia*. Although the alga is thought to be native to Australia where it is a natural part of the ecosystem, the California alga proved to be closely related to a highly invasive strain that has caused extensive ecological damage in the Mediterranean. This discovery pointed to the need for an immediate eradication program before damage to the coastal ecosystem of California could take place.
The Tree of Life and Human Health

Understanding how organisms, as well as their genes and gene products, are related to one another has become a powerful tool for identifying disease organisms, tracing the history of infections, and predicting disease outbreaks.

Identifying Emerging Diseases: West Nile Virus

When an encephalitis-like viral infection emerged in people living in the New York region in 1999, it was first suspected to be the St. Louis encephalitis virus. Transmitted by mosquitoes, the virus was simultaneously found to be associated with a high mortality in wild and domesticated birds. It has currently spread as far west as California and has resulted in numerous human deaths.

In two separate studies, health workers used phylogenetic analysis to identify viral isolates from mosquitoes and birds as a new outbreak of the West Nile virus rather than St. Louis encephalitis. The viral tree (right) demonstrated that the New York isolate was most closely related to one found in dead birds in Israel, East Africa, and eastern Europe. This knowledge provided health officials with key information about the basic biology of the virus that was needed for diagnosis and predicting its spread. Such knowledge was critical in preventing human and animal infection.

Several species of mosquitoes, including the recently invasive tiger mosquito, transmit West Nile virus from bird and mammal hosts to human populations. Understanding the phylogenetic relationships of disease vectors such as mosquitoes is important in tracing the origin and spread of diseases.

Disease Transmission and Medical Forensics

The use of phylogenetic analysis has become a powerful tool within the health sciences for understanding the history of disease transmission.

One of the earliest applications of phylogenetic methods to disease transmission and medical forensics investigated whether a Florida dentist infected with the HIV virus had transmitted the disease to his patients following invasive dental procedures. Once it was suspected the dentist might be the source of infection for several HIV-positive patients, other patients were tested. The sequences of viral envelope genes of those testing positive, along with the genes of local controls (HIV-positive nonpatients of the dentist), were determined and a phylogenetic analysis was performed on all the sequences (left). The gene sequence of the dentist clustered with those from his patients, thus supporting the original hypothesis. At the same time, the tree also showed that other patients (F and D on the tree) as well as the local controls had not contracted HIV from the dentist.
In the spring of 1993, a previously undescribed disease emerged in the American Southwest, killing 10 people during a 10-week period from April to June. For the first few weeks, the mortality rate was 70%. There was no cure, no successful medications or treatment, and the disease agent (virus, bacterium, or toxin) was completely unknown. It was soon discovered, however, that the disease-causing agent was an unknown hantavirus, later named Sin Nombre virus, and that its reservoir was the common deer mouse.

Analyses of published phylogenies for hantaviruses reveal that the new virus was related to Old World strains but differed in many ways, suggesting it was not likely to have been introduced recently from the Old World. When phylogenetic trees of the viral genes were compared with those of rodents related to deer mice, a high degree of concordance in the pattern of branching was discovered, suggesting a long co-evolutionary history in the Americas (right). This finding, moreover, led to the prediction that most, if not all, major mice lineages would be found to have a closely associated hantavirus.

Since the discovery of Sin Nombre virus, some 25 additional hantaviruses have been described from the New World alone, totally consistent with the prediction of the tree. Fourteen of these newly recognized hantaviruses have been detected in countries throughout Central and South America, and many of these viruses are pathogenic for humans. The phylogeny predicts that yet unknown hantaviruses are likely present in other rodent hosts in North and South America and elsewhere in the world, especially in poorly studied regions of Africa and Asia.

The documentation of a co-evolutionary relationship between hantaviruses and their rodent hosts has triggered similar investigations of other emerging pathogens, including those that cause African and South American hemorrhagic fevers. These studies, in turn, are leading to the discovery of many new species of both viruses and their hosts, which increases our ability to diagnose and treat these diseases.
Phyloinformatics: A Comparative Framework for Exploring Biological Databases

Phyloinformatics

Knowledge of the Tree of Life has advanced considerably over the past 10 years, and accompanying this has been a near-exponential growth in the amount of data (genetic and otherwise) that have been used to build our current picture of life’s history. For society and science to reap the benefits of the powerful predictive capabilities of the Tree of Life, new technologies and analytical techniques will have to be developed to cope with our growing information about trees and their associated data. The emerging discipline being created to meet these new challenges is called phyloinformatics.

The systematics community’s vision of phyloinformatics is expansive, encompassing the development of new informatics tools to archive, synthesize, and retrieve the data used to build trees, and also of novel approaches for visualizing and searching trees and the information associated with them. Scientific innovations that will emerge from phyloinformatics will spawn a host of benefits to educational institutions, to students, and the public at large.

Natural history collections are essential for Tree of Life research

Specimen collections housed in museums, herbaria, and seed banks comprise a permanent, verifiable record of Earth’s biological diversity. Information about species — including genomic, anatomical, ecological, and other data — increases in reliability when it is linked to properly identified voucher specimens in collections. Thus, capturing the information associated with specimens in natural history collections is of crucial importance for building the Tree of Life.

Phyloinformatics and Predictivity

A phylogenetic tree for a group of species is derived from a comparison of their characteristics. This hierarchical pattern of similarity (for example, vertebrates share characters not found in other animals, mammals share similarities absent in other vertebrates, and primates share features lacking in other mammals) provides a powerful comparative framework for interpreting, manipulating, and predicting biological information about organisms.

A key promise of phyloinformatics is that it can provide the tools to exploit this hierarchy of relationships for data mining. Phyloinformatic data mining will use the hierarchy implied by trees to conduct node-based searches across many kinds of databases (figure below). A search using node 2, for example, would simultaneously retrieve information from databases for groups A, B, and C, and would combine and synthesize those data in ways not currently possible. Tree-oriented navigation and data mining having this power and sophistication make use of the predictive capabilities of phylogenetic trees to integrate biological information in new ways, and thus create new associations and new knowledge.

Phyloinformatics: A Vision for Comparative Biology

Databases

Node-based Search and Data Mining

Groups of Organisms

Specimens/Collections
Associated biological data
Geo-referenced data
Associated literature, etc.

Genetic/Genomic
Developmental
Human health
Economic/trade
Agricultural

Specimen/Collections
Geo-referenced data
Associated literature, etc.

Geological
Land use
Geographic

Node 1
Node 2
Node 3
Node 4

Node 2
Node 1
Node 3
Node 4
Building a Tree of Life for Earth’s species, living and extinct, is a megascience initiative that already has produced numerous benefits for science and society. Meeting the challenges of constructing this Tree will create unprecedented opportunities to foster interdisciplinary research across fields as diverse as genomics, computer science and engineering, informatics, mathematics, Earth sciences, developmental biology, and environmental biology.

The scientific problem of building the Tree of Life is enormous because life itself is so diverse and complex. Collecting the necessary genomic, anatomical, and behavioral data on all species with which to determine the Tree will require a significant commitment of human resources and technological innovation. Yet, the challenges can be met. Already, for example, the genomics database, GenBank, has information on nearly 100,000 species, and this database is growing exponentially, and will keep doing so as new sequencing technologies are developed. Having genomic information on all of Earth’s species is no longer a dream. Nongenomic data — anatomical, behavioral, biochemical, or physiological — have also been collected on many thousands of species, but if a comprehensive understanding of life’s complexity over time is to be developed, those data will have to be greatly expanded, databased, and then integrated with genomic information.

The burgeoning information used to construct the Tree of Life will have numerous applications within industry, agriculture, the health sciences, and in environmental management — but only if that information is organized and available. Science now has the ability to create huge databases, but much of this information is not organized and integrated in such a way as to create new knowledge. This must change if the full benefits of Tree of Life research are to be achieved.

The organizational and comparative power of phylogenetic relationships is creating the new field of phyloinformatics, which will harness phylogenetic knowledge to integrate and transform data held in isolated databases, thereby inventing new information and new knowledge. Realizing the broad vision of phyloinformatics will require innovative approaches to data mining, as well as to the visualization and interpretation of its results. The complex computational problems associated with constructing trees and manipulating different kinds of data are attracting the attention of computer scientists who are creating new algorithms for building trees and comparing data. The computational advances underlying phyloinformatics will contribute solutions to similar problems in other disciplines and will generate applications far beyond the field of phylogenetic research.

Assembling the Tree of Life is as grand and important a mission as scientists have ever tackled. Much like the mission to place a man on the moon, or to decipher the human genome, constructing a comprehensive Tree of Life will promote intellectual achievement across the sciences. Considering the significant benefits that have already flowed from current knowledge about life’s history, expanding the effort to build a comprehensive Tree of Life over the next several decades is vitally important, and it can be accomplished.