# Genomes, phylogeny, and evolutionary systems biology

**Mónica Medina***

Department of Evolutionary Genomics, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

**With the completion of the human genome and the growing number of diverse genomes being sequenced, a new age of evolutionary research is currently taking shape. The myriad of technological breakthroughs in biology that are leading to the unification of broad scientific fields such as molecular biology, biochemistry, physics, mathematics, and computer science are now known as systems biology. Here, I present an overview, with an emphasis on eukaryotes, of how the postgenomics era is adopting comparative approaches that go beyond comparisons among model organisms to shape the nascent field of evolutionary systems biology.**

postgenomics | eukaryotic genomics | network evolution

> Systems biology is in the eye of the beholder.
> Leroy Hood

Only in the last decade have we had access to nearly complete genomes of a diversity of organisms allowing for large-scale comparative analysis. The access to this immense amount of data is providing profound insight into the tree of life at all levels of divergence (Fig. 1*A*). It is thus not surprising that understanding phylogenetic relationships is a prevalent research goal among not only evolutionary biologists but also all scientists interested in the organization and function of the genome. New genome sequences and analysis methods are helping improve our understanding of phylogeny, and at the same time improved phylogenies and phylogenetic theory are generating a better understanding of genome evolution. Currently however, the level of genome sequencing for different branches of the tree of life is far from equivalent. Prokaryotic genome projects are abundant, mainly due to their small genome sizes, with >200 genomes already published and at least 500 currently in progress (www.genomesonline.org). In contrast, <300 eukaryotic genomes are either finished or in progress (www.genomesonline.org). Nevertheless, these data are starting to have a major impact on our understanding of eukaryotic evolution.

These new genomic data have informed our understanding of phylogenetic relationships, and the emerging consensus topologies are adding new insight to the small subunit ribosomal RNA phylogenies. For example, the topology of the ribosomal eukaryotic tree has been recently redrawn with the use of genomic signatures that place the root of all eukaryotic life between two newly uncovered major clades, Unikonts and Bikonts (Fig. 1*A*). Unikonts, which contain the heterotrophic groups Opisthokonta and the Amebozoa, share a derived three-gene fusion of enzyme-encoding genes in the pyrimidine synthesis pathway (1), whereas Bikonts, which contain the remaining eukaryotic clades, share another derived gene fusion between dihydrofolate reductase and thymidine synthase (2). All photosynthetic groups of primary and secondary plastid symbiotic origins are now thought to be within the Bikonts. Although the animal, fungal, and plant lineages are the most widely represented in terms of genome initiatives (Fig. 1 *B–D*), it is significant that multiple protistan genome projects have also been initiated by the interest of diverse scientific communities, including parasitologists (3), plant pathologists (4), oceanographers (5), and evolutionary biologists (www.biology.uiowa.edu/workshop).

As more whole-genome projects are being completed, postgenomic biology is also providing insight into the function of biological systems by the use of new high-throughput bioanalytical methods, information technology, and computational modeling. This new revolution in biology has become known as systems biology (6). In addition to shifting approaches to biological research from reductionist strategies to pathway- and system-level strategies (7), another paradigm is rapidly emerging, namely the use of phylogenetically based inference in systems biology. Before the genomic revolution, research questions were typically addressed within a single model organism, with only occasional comparative studies when similar information was available for another organism. These comparisons were made between distantly related taxa, and the evolutionary implications were rarely mentioned or taken into account. The increasing importance of comparative analysis is evident in the growing proportion of new prokaryotic genome projects that have been chosen primarily because of their phylogenetic relationship to model organisms, such as *Escherichia coli* and *Bacillus subtilis* and their corresponding related taxa. This same trend is occurring for eukaryotes. Some prominent examples are the multiple *Saccharomyces* genome projects and those of other ascomycote fungi, the several *Plasmodium* projects and other genome initiatives for apicomplexan taxa, the numerous *Caenorhabditis* and other nematode genome projects, the multiple *Drosophila* and arthropod genome projects, and the large number of primate and mammalian genome projects.

## Genomes and Phylogeny of Higher Eukaryotes

**Metazoa.** The sampling of the metazoan tree, and in particular of the chordate branch, was undertaken primarily due to the usefulness of the genomes in understanding human biology. However, this larger genomic dataset is already providing a powerful tool for comparative analysis and more accurate evolutionary inference. Deeper divergences in the Metazoan tree have become the target of major scrutiny due to the interest in comparative developmental genetics (Fig. 1*B*). Based on molecular phylogenies, the bilaterian phyla have been rearranged into three large clades, deuterostomes, lophotrochozoans, and ecdysozoans, these last two being sister taxa inside the protostome clade. At present, there is still debate regarding the placement of nematodes in the tree (i.e., the Ecdysozoa vs. Coelomata hypotheses) because analysis of genomic data

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, ''Systematics and the Origin of Species: On Ernst Mayr's 100th Anniversary,'' held December 16–18, 2004, at the Arnold and Mabel Beckman Center of the National Academies of Science and Engineering in Irvine, CA.

*Present address: School of Natural Sciences, University of California, P.O. Box 2039, Merced, CA 95344. E-mail: mmedina@ucmerced.edu.





www.pnas.org/cgi/doi/10.1073/pnas.0501984102

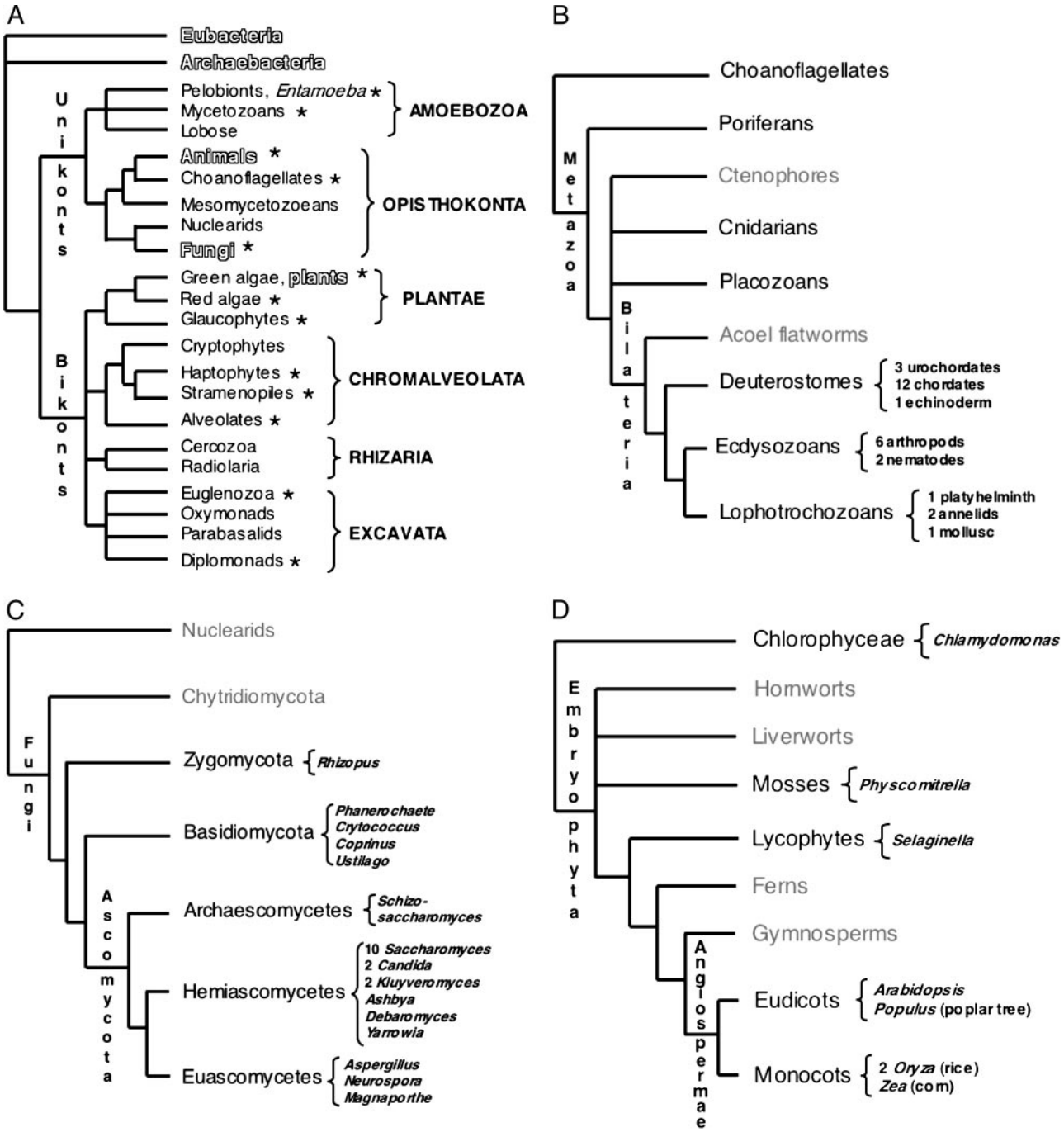


**Fig. 1.** Current consensus eukaryotic tree. (*A*) The large subclades within Unikonts and Bikonts are recovered by a combination of multiple gene phylogenies, EST data, and genomic level characters (1, 80, 81). Six major eukaryotic groups are now recognized although resolution within them is still lacking. The placement of the root is based on two gene fusion events (1, 2). Lineages where whole-genome projects are in progress are marked with asterisks. Lineages being studied by large postgenomic initiatives are shadowed. (*B*) Metazoan consensus phylogeny of major branches (82–84) and a conservative estimate of finished and ongoing genome projects (highlighted in black). (*C*) Fungal consensus phylogeny (13, 14) and estimate of ongoing genome projects (www.broad.mit.edu/annotation/fungi/fgi) (highlighted in black). (*D*) Consensus phylogeny of green plants (13, 20) and estimate of ongoing genome projects (highlighted in black).

currently challenges the placement of *Caenorhabditis elegans* as an ecdysozoan (8, 9).

In addition to the traditional developmental model organisms, genomes from unrepresented protostome (Annelida, Platyhelmintha, and Mollusca) and basal phyla are now being sequenced (Porifera, Placozoa, and Cnidaria) (www.jgi.doe.gov/sequencing/cspseqplans.html). Finally, another node in the tree of life that has gained recent interest is that of the choanoflagellates, a unicellular sister group to metazoans (10). Ribosomal phylogenies suggest that choanoflagellates are the most likely unicellular lineage to have shared common ancestry with the multicellular animals (11), but there are a few other unicellular protists that also fall out in this part of the tree in other gene phylogenies (12). A choanoflagellate genome project is now in progress, and multiple EST initiatives for unicellular opisthokont protists are also in place.

In summary, postgenomic research on the metazoans is advancing rapidly because of the large number of model organisms, e.g., *C. elegans* (nematode), *Drosophila melanogaster* (fruitfly), *Danio rerio* (zebrafish), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Homo sapiens* (human). On the other hand, sequencing metazoan genomes is a major technical challenge, because of higher level of complexity associated with multicellularity and tissue compartmentalization. These challenges are giving a leading role to the yeast and other unicellular systems described in the next section.

**Fungi.** The initial driving force behind the choice of genome projects in fungi was the prime status of yeast (*Saccharomyces cerevisiae*) as a model organism. Additionally, the relatively small genome size in other fungi has facilitated the explosion of numerous large scale sequencing projects (www.broad.mit.edu/annotation/fungi/fgi). Consensus phylogenies of fungi place the Chitridiomycota as the most basal lineage, followed by the Zygomycota, with Ascomycota and Basidiomycota as sister crown clades (13, 14). Ribosomal phylogenies suggest that the Nucleared amoeba are the likely unicellular sister group to Fungi (11, 15).

After the completion of *S. cerevisiae*, subsequent fungal genome projects were chosen within the Ascomycota (Fig. 1*C*) mainly based on phylogenetic proximity (within the Hemiascomycetes) (16–18), although now more distantly related taxa including additional model organisms such as *Neurospora crassa* and *Aspergillus nidulans* have also been sequenced. Basidiomycete genomes have been sequenced (19) or are in progress as well (www.broad.mit.edu/annotation/fungi/fgi). The combination of both *S. cerevisiae* as the best characterized unicellular eukaryote and the thorough comparative genomics allowed by the numerous fungal genome projects have made this branch of the eukaryotic tree an ideal target for validation and improvement of postgenomic approaches.

**Plantae.** Most of the species diversity of plants is represented in the crown group, the angiosperms, which encompasses the Monocots and the Eudicots. Consensus phylogenies place paraphyletic gymnosperms basal to angiosperms and ferns as a sister group to this clade (20) (Fig. 1*D*). The placement of some of the basal groups in the Embryophyta (hornworts, liverworts, and mosses) is still unresolved, although lycophytes are now considered the sister group to the clade containing ferns, gymnosperms, and angiosperms (13, 20) (Fig. 1*D*). Finally, multiple sources of evidence point to the green algae as the unicellular sister group to plants (reviewed in ref. 21) (Fig. 1*D*).

Genome projects for green plants have been hampered by the larger genome sizes of most members of this group. Nonetheless, the first draft Plantae genome published was from *Arabidopsis thaliana*, a flowering plant model organism (22). Genome drafts of two different rice strains (*Oryza sativa*) have been recently published (23, 24). This effort is now complemented by the completion of a unicellular alga (*Chlamydomonas reinhardtii*), the poplar tree (*Populus trichocarpa*), and partial genome data from corn (*Zea mays*), whereas two basal lineages, the moss *Physcomitrella patens* and the lycophyte *Selaginella moellendorffii*, will be sequenced this year (www.jgi.doe.gov/sequencing/cspseqplans.html). Thus, although more sparse than for metazoans and fungi, the Plantae branch of the eukaryotic tree is rapidly expanding in terms of genomic data. Agricultural interests will likely drive future choice of Plantae genomes to some degree, but decisions will also be influenced by phylogenetic implications as reflected in the recent choice of the *P. patens* and *S. moellendorffii* for genome sequencing.

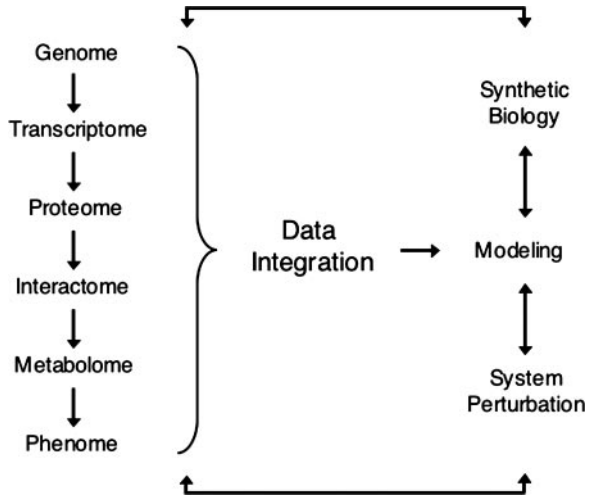


**Fig. 2.** Overview of systems biology. Hierarchical information from the genome (DNA) to the phenome (phenotype) is integrated to predict mathematical models. These models can then be tested by "synthetic biology" (*de novo* design of biological modules) and/or by system perturbations that generate a cycle of hypothesis-driven science (26, 27, 32).

## Evolutionary Systems Biology

With whole-genome data allowing reconstruction of more robust phylogenies for the major eukaryotic groups, new biological questions can now be addressed. Genomic and postgenomic data offer a new "global" view of the function of living systems across the tree of life. These new data suggest that biological systems (e.g., a cell) are composed of discrete "modules" of interacting components with different functions, and in turn these modules form biological networks that carry out the myriad functions of living systems (7). Multiple metabolic and regulatory networks are now being characterized in diverse organisms for which reasonably annotated genomes are available. Metabolites, being the end products of cellular regulatory networks, are one of the most directly accessible windows into the cell's dynamic phenotype (25).

Systems biology is a rapidly expanding field that integrates widely diverse areas of science such as physics, engineering, computer science, mathematics, and biology, toward the goal of elucidating the hierarchy of metabolic and regulatory systems in the cell, and ultimately leading to a predictive understanding of the cellular response to perturbations (26, 27) (Fig. 2). As the theoretical and experimental tools of systems biology rapidly advance, multiple fields are embracing systems biology approaches as a mainstream method of research. Because postgenomics research is taking place throughout the tree of life, comparative approaches are a way to combine data from many organisms to understand the evolution and function of biological systems from the gene to the organismal level. Therefore, systems biology can build on decades of theoretical work in evolutionary biology, and at the same time evolutionary biology can use systems approaches to go in new uncharted directions. For instance, although comparative genomics has benefited from a long tradition of theoretical work by molecular evolutionists (28), new datasets being provided by systems biology are allowing theoreticians new ways to study evolutionary processes (29).

Comparative studies can give insight into even the highest-level principles of life. For example, revolutionary findings in network theory have in part come from genomic data from a wide range of organisms, leading scientists to propose laws that seem to govern biological networks (30, 31). Different types of cellular networks (e.g., protein interaction and metabolic networks) seem to share properties with other complex abiotic

networks such as their "scale-free" nature and "small world" organization. In scale-free networks, a few nodes (hubs) have the largest number of connections to other nodes, whereas most of the nodes have just a few connections. This property is reflected in a power-law distribution. In practical terms, this relationship means that, in a protein interaction network, most proteins interact with a couple of others whereas a few proteins (hubs) interact with a large number, and that, in a metabolic network, a few molecules (hubs) participate in most reactions whereas the rest participate in one or two. The "small world" concept refers to the property of such spoke-and-hub networks that there is a small path length between nodes, just as in modern air travel where only a few flights connect any two cities in the world. This property means that a path of just a few interactions or reactions will connect almost any pair of molecules in the cell (29).

Additional levels in the hierarchy of biological networks and the interactions between them are now being characterized that will allow for integration of data and new theoretical predictions (32). Processes widely studied by evolutionary biologists such as selection, gene duplication, and neutral evolution are being examined in the context of network models as opposed to at the level of individual genes or molecules (33–37).

## Evolution of Biological Networks

**Transcriptional Networks.** High-throughput global gene expression approaches such as EST sequencing and microarrays are now common practice for functional assessment of the genome. The extensive microarray gene expression datasets available for model and non-model organisms are starting to be incorporated into a comparative approach to study transcriptome evolution at multiple levels of divergence. At lower levels of divergence, studies in organisms including fish (38), fruitfly (39, 40), and yeast (41) have now shown that extensive variation exists in the transcriptome in natural populations and that this variation is likely to be an important factor in organismal evolution. Transcriptome comparisons across several primate and mouse species, however, suggest that the majority of gene expression differences within and between species evolve in a selectively neutral or nearly neutral fashion (42). At intermediate levels of divergence, less information is available at present due to lack of genomic data. Although analytically challenging, the use of gene expression profiling by heterologous hybridization to a single species cDNA microarray is starting to be explored, potentially opening the door to comparative analyses of taxa as divergent as 200 mega-annum (Ma) (43). This application would be of great significance for the comparative study of non-model organisms that are only distantly related to an already sequenced species. At deep levels of divergence, coexpression of large aggregates of functionally related genes seems to be conserved across evolution. Two recent comparisons of the transcriptomes of several of the model organisms [*S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *H. sapiens* in one case (44), and these four plus *A. thaliana* and *E. coli* in the second case (45)] support the hypothesis that coexpression networks can be split into multiple components enriched for genes involved in similar functional processes. Some of these identified components can be unique to a certain clade, such as the signaling pathway and neuronal function components present only in metazoans in the four-species comparison (44). These cross-species comparisons promise to provide more information about coexpression network evolution as the transcriptomes of additional diverse lineages becomes available (46).

Central to postgenomic analysis is the accuracy of genome annotation. The degree of accuracy in which genomes are annotated is affected by the quality of sequence assembly, gene prediction, and functional annotation by both bioinformatics and experimental data. This relationship is particularly critical in genome projects of non-model organisms where little genetic work has been performed in the past. All these factors, combined with the lack of network information outside the model organisms, point to the tradeoff between a comprehensive systems analysis of a particular network within a well-studied organism, versus the historical perspective introduced by evolutionary conservation or divergence of systems through time in phylogenetic comparisons. Therefore, although only partial inference is possible at present, studies have already shown that the comparative approach to coexpression not only is giving insight into the universal rules that govern biological systems but also has practical implications by helping improve functional annotations of both model and non-model organisms (44, 45). Because comparative analyses of coexpression data from several model organisms have shown high levels of conservation between such divergent taxa as prokaryotes (*E. coli* and *B. subtilis*) (47), opisthokont eukaryotes (44), and even prokaryotes and eukaryotes (45), some efforts are now targeting the coupled evolution of regulatory networks and the transcriptome.

**Regulatory Networks.** The characterization of the transcriptome is only a fraction of the information needed to understand global cellular processes because gene expression is driven by the spatio-temporal localization of regulatory networks and details of specific protein–DNA and protein–protein interactions. Genome-wide efforts to characterize transcriptional regulatory networks have already been fruitful in model organisms like yeast (48) and *E. coli* (49). In multicellular organisms, fractions of the regulatory networks are being characterized for sea urchins (50), *Drosophila* (51), and mammals (52).

Transcription factors are regulatory proteins that influence the expression of specific genes. They work by binding to cis-regulatory elements (short and often degenerate sequence motifs frequently located upstream of genes) where they interact with the transcription apparatus to either enhance or repress gene expression. Even though identifying cis-regulatory elements in new genomes is an inherently difficult task due to their short sequence length and as yet unknown syntax, comparative approaches have been helpful. By aligning orthologous regions flanking a gene from multiple species, conserved noncoding sequence motifs can be distinguished. These evolutionary conserved motifs are then hypothesized to be potential functional elements. This method, called phylogenetic footprinting (53), has successfully been used to identify a limited number of regulatory regions in vertebrates (54, 55) and plants (56, 57). More sophisticated comparative approaches are starting to combine computational prediction and laboratory validation of regulatory networks. Coexpression data and known cis-regulatory elements from *S. cerevisiae* were used in a multispecies comparison of 13 published ascomycete genomes, finding multiple cases of regulatory conservation but also some cases of regulatory diversification (58). It has become apparent, however, that sequence conservation alone will not help identify all cis-regulatory elements by phylogenetic footprinting, and additional data and experimental approaches have to be integrated (59).

Gene expression can be regulated not only at transcriptional initiation but also at other levels, such as during mRNA editing, transport, or translation, and characterizing these interactions and their evolution is one of the many future challenges of systems biology (60). For example, comparative work on populations of yeast and fruitfly has recently shown that protein–protein interactions are negatively associated with evolutionary variation in gene expression (61). A comparative analysis of the *E. coli* and yeast regulatory networks has demonstrated that gene duplication has a key role in network evolution both in eukaryotes and prokaryotes (62). Finally, introducing concepts of network dynamics has revealed new topological changes in the regulatory network in yeast (63), an approach that, incorporated into a comparative framework, will eventually provide answers

to the evolution of morphological divergence in multicellular taxa (64).

**Protein Networks.** The proteome for several of the model organisms is now characterized, and this global scale information has been used to predict protein–protein interaction networks (interactomes) for *D. melanogaster* (65), *C. elegans* (66), and *S. cerevisiae* (67). Assuming some degree of evolutionary conservation, these data can also be used to transfer interactome annotations to genomes that have not been characterized experimentally. Comparisons across multiple species have shown conserved protein interactions that allow for initial drafts of protein–protein interaction maps of human (68) and *A. thaliana* (69). When formulating evolutionary hypotheses, however, attention to the phylogenetic relationships is necessary. For example, some of the conclusions from the analysis of the *C. elegans* interactome (63) are weakened by the incorrect assumption that plants (*A. thaliana*) and animals (*C. elegans*, *D. melanogaster*) are more closely related to each other than to yeast. Current phylogenies show that multicellularity has occurred independently in metazoans, fungi, and plants (Fig. 1*A*), and that unicellularity in yeasts is a derived rather than ancestral state (Fig. 1*C*).

**Metabolic Networks and "Ome" Data Integration.** The metabolome is made up of all of the low-molecular weight molecules (metabolites) present in a cell at a particular time point, and their levels can be regarded as the functional response of biological systems to genetic or environmental stimuli (25). Challenges faced in the global study of metabolites, such as their dynamic behavior and chemistry, are being addressed by emerging technologies such as liquid and gas chromatography mass spectrometry and NMR (70). Plant biologists have led in the application of these advances (71), and soon there will likely be large datasets for multiple plant and other eukaryotic species. Although high-throughput metabolome projects are just now being initiated, comparative analysis of 43 known metabolic networks has already shown that they seem to follow a power-law distribution (29, 30).

The integration of data from the different levels of cellular networks (transcriptome, regulome, interactome, and metabolome) is the next obvious step to identify patterns of network interactions in individual species and in multispecies comparisons (32, 72, 73). This integrative approach has already been fruitful in model organisms such as *C. elegans* (74) and *S. cerevisiae* (75).

It is clear that producing a large scale comparative systems biology analysis will have to involve the work of many research groups and that many challenges will need to be overcome. For example, rigorous standards will need to be established to facilitate the comparison of results from high-throughput "omic" analyses before we can make conclusive evolutionary inferences (76). A pioneer example is the ENCODE initiative, which aims to identify all functional elements in the human genome by using coordinated computational and experimental efforts in a multispecies framework (52).

Although we can already find global patterns of network evolution, in the future we should be able to look at trends and patterns in the evolution of biological systems within phylogenies. For instance, we should be able to look at how many of the biological network similarities are due to homoplasy as opposed to phylogenetic constraints due to common ancestry. Thus, by using the theoretical framework developed for the comparative method, phylogenetic information can allow only for improvement of evolutionary inference at the systems level.

Finally, to bring evolutionary systems biology to the highest level of biological organization, ecosystem-level factors have to be taken into consideration. To this end, the use of high-throughput approaches for the study of interactions among organisms and between organisms and their natural environments is engaging the interest of ecologists (77, 78).

## Historical Perspective

Darwin's theory of natural selection and, later on, the integrative nature of the modern synthesis consolidated the study of evolution as a solid discipline to address fundamental questions in biology. The scientific advances that allowed for the discovery of the structure of DNA and the development of molecular biology eventually led to large-scale whole-genome initiatives. This unfolding was a revolutionary moment in the scientific mentality of 20th century researchers, because it generated the integrative approaches of systems biology, which will most likely become the standard of 21st century biology. Organismal biologists have been thinking along these lines for the past few decades, advocating integrative and multidisciplinary approaches to evolutionary questions (79). Thus, bridging knowledge between evolutionary theory and systems biology will be only a natural process. Together, these approaches offer the promise to solve two of the ultimate questions in biology: the function of biological systems and an understanding of the evolution of life's diversity.

I thank Francisco Ayala, Jody Hey, and Walter Fitch for the invitation to participate in the Mayr colloquium. Pilar Francino and Paramvir Dehal provided thoughtful insight for both the seminar and the manuscript. I also thank Mike Colvin, Benoît Dayrat, Jodi Schwarz, Rick Baker, Kevin Helfenbein, and an anonymous reviewer for helpful comments on previous versions of the manuscript. Benoît Dayrat and Peter Brokstein helped with figure design. I thank Caturro Mejía for introducing me to Mayr's writings many years ago. This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, University of California, Lawrence Berkeley National Laboratory under contract DE-AC03-76SF00098. I also acknowledge support by National Science Foundation Grant OCE 0313708.

1. Stechmann, A. & Cavalier-Smith, T. (2003) *Curr. Biol.* **13,** R665–R666.
2. Stechmann, A. & Cavalier-Smith, T. (2002) *Science* **297,** 89–91.
3. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., *et al.* (2002) *Nature* **419,** 498–511.
4. Waugh, M., Hraber, P., Weller, J., Wu, Y., Chen, G., Inman, J., Kiphart, D. & Sobral, B. (2000) *Nucleic Acids Res.* **28,** 87–90.
5. Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., *et al.* (2004) *Science* **306,** 79–86.
6. Hood, L. (2003) *Mech. Ageing Dev.* **124,** 9–16.
7. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999) *Nature* **402,** C47–C52.
8. Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. (2004) *Genome Res.* **14,** 29–36.
9. Dopazo, H., Santoyo, J. & Dopazo, J. (2004) *Bioinformatics* **20,** Suppl. 1, I116–I121.
10. King, N. (2004) *Dev. Cell* **7,** 313–325.
11. Medina, M., Collins, A. G., Taylor, J. W., Valentine, J. W., Lipps, J. H., Amaral Zettler, L. A. & Sogin, M. L. (2003) *Int. J. Astrobiol.* **2,** 203–211.
12. Ruiz-Trillo, I., Inagaki, Y., Davis, L. A., Sperstad, S., Landfald, B. & Roger, A. J. (2004) *Curr. Biol.* **14,** R946–R947.
13. Hedges, S. B. (2002) *Nat. Rev. Genet.* **3,** 838–849.
14. Berbee, M. L. & Taylor, J. W. (1993) *Can. J. Bot.* **71,** 1114–1127.
15. Amaral Zettler, L. A., Nerad, T. A., O'Kelly, C. J. & Sogin, M. L. (2001) *J. Eukaryotic Microbiol.* **48,** 293–297.
16. Kellis, M., Birren, B. W. & Lander, E. S. (2004) *Nature* **428,** 617–624.
17. Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., *et al.* (2004) *Nature* **430,** 35–44.
18. Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., *et al.* (2004) *Science* **304,** 304–307.

19. Martinez, D., Larrondo, L. F., Putnam, N., Gelpke, M. D., Huang, K., Chapman, J., Helfenbein, K. G., Ramaiya, P., Detter, J. C., Larimer, F., *et al*. (2004) *Nat. Biotechnol.* **22,** 695–700.
20. Pryer, K. M., Schneider, H., Zimmer, E. A. & Ann Banks, J. (2002) *Trends Plant Sci.* **7,** 550–554.
21. Archibald, J. M. & Keeling, P. J. (2002) *Trends Genet.* **18,** 577–584.
22. *Arabidopsis* Genome Initiative (2000) *Nature* **408,** 796–815.
23. Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al*. (2002) *Science* **296,** 79–92.
24. Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., *et al*. (2002) *Science* **296,** 92–100.
25. Fiehn, O. (2002) *Plant Mol. Biol.* **48,** 155–171.
26. Ideker, T., Galitski, T. & Hood, L. (2001) *Annu. Rev. Genomics Hum. Genet.* **2,** 343–372.
27. Kitano, H. (2002) *Science* **295,** 1662–1664.
28. Wolfe, K. H. & Li, W. H. (2003) *Nat. Genet.* **33,** Suppl., 255–265.
29. Barabasi, A. L. & Oltvai, Z. N. (2004) *Nat. Rev. Genet.* **5,** 101–113.
30. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000) *Nature* **407,** 651–654.
31. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. (2002) *Science* **297,** 1551–1555.
32. Ge, H., Walhout, A. J. & Vidal, M. (2003) *Trends Genet.* **19,** 551–560.
33. Wagner, A. (September 30, 2003) *Sci. STKE*, 10.1126/stke.2003.202.pe41.
34. Wagner, A. (2003) *Proc. R. Soc. London Ser. B Biol. Sci.* **270,** 457–466.
35. van Noort, V., Snel, B. & Huynen, M. A. (2004) *EMBO Rep.* **5,** 280–284.
36. Wuchty, S. (2004) *Genome Res.* **14,** 1310–1314.
37. Hahn, M. W., Conant, G. C. & Wagner, A. (2004) *J. Mol. Evol.* **58,** 203–211.
38. Olesiak, M. J., Churcill, G. A. & Crawford, D. L. (2002) *Nat. Genet.* **32,** 261–266.
39. Meiklejohn, C. D., Parsch, J., Ranz, J. M. & Hartl, D. L. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9894–9899.
40. Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D. & Hartl, D. L. (2003) *Science* **300,** 1742–1745.
41. Townsend, J. P., Cavalieri, D. & Hartl, D. L. (2003) *Mol. Biol. Evol.* **20,** 955–963.
42. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. & Paabo, S. (2004) *PLoS Biol.* **2,** E132.
43. Renn, S. C., Aubin-Horth, N. & Hofmann, H. A. (2004) *BMC Genomics* **5,** 42.
44. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. (2003) *Science* **302,** 249–255.
45. Bergmann, S., Ihmels, J. & Barkai, N. (2004) *PLoS Biol.* **2,** E9.
46. Zhou, X. J. & Gibson, G. (2004) *Genome Biol.* **5,** 232.
47. Snel, B., van Noort, V. & Huynen, M. A. (2004) *Nucleic Acids Res.* **32,** 4725–4731.
48. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al*. (2002) *Science* **298,** 799–804.
49. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002) *Nat. Genet.* **31,** 64–68.
50. Davidson, E. H. (2001) *Genomic Regulatory Systems: Development and Evolution* (Academic, San Diego).
51. Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B. & Celniker, S. E. (2004) *Genome Biol.* **5,** R61.
52. ENCODE Project Consortium (2004) *Science* **306,** 636–640.
53. Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L. & Jones, R. T. (1988) *J. Mol. Biol.* **203,** 439–455.
54. Gumucio, D. L., Heilstedt-Williamson, H., Gray, T. A., Tarle, S. A., Shelton, D. A., Tagle, D. A., Slightom, J. L., Goodman, M. & Collins, F. S. (1992) *Mol. Cell. Biol.* **12,** 4919–4929.
55. Dermitzakis, E. T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C. & Antonarakis, S. E. (2003) *Science* **302,** 1033–1035.
56. Hong, R. L., Hamaguchi, L., Busch, M. A. & Weigel, D. (2003) *Plant Cell* **15,** 1296–1309.
57. Kaplinsky, N. J., Braun, D. M., Penterman, J., Goff, S. A. & Freeling, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6147–6151.
58. Gasch, A. P., Moses, A. M., Chiang, D. Y., Fraser, H. B., Berardini, M. & Eisen, M. B. (2004) *PLoS Biol.* **2,** e398.
59. Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., *et al*. (2005) *Genome Res.* **15,** 1–18.
60. Wei, G. H., Liu, D. P. & Liang, C. C. (2004) *Biochem. J.* **381,** 1–12.
61. Lemos, B., Meiklejohn, C. D. & Hartl, D. L. (2004) *Nat. Genet.* **36,** 1059–1060.
62. Teichmann, S. A. & Babu, M. M. (2004) *Nat. Genet.* **36,** 492–496.
63. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. & Gerstein, M. (2004) *Nature* **431,** 308–312.
64. Howard, M. L. & Davidson, E. H. (2004) *Dev. Biol.* **271,** 109–118.
65. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al*. (2003) *Science* **302,** 1727–1736.
66. Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., *et al*. (2004) *Science* **303,** 540–543.
67. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al*. (2000) *Nature* **403,** 623–627.
68. Lehner, B. & Fraser, A. G. (2004) *Genome Biol.* **5,** R63.
69. Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M. & Gerstein, M. (2004) *Genome Res.* **14,** 1107–1118.
70. Stitt, M. & Fernie, A. R. (2003) *Curr. Opin. Biotechnol.* **14,** 136–144.
71. Oksman-Caldentey, K. M., Inze, D. & Oresic, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 9949–9950.
72. Castrillo, J. I. & Oliver, S. G. (2004) *J. Biochem. Mol. Biol.* **37,** 93–106.
73. Papin, J. A., Reed, J. L. & Palsson, B. O. (2004) *Trends Biochem. Sci.* **29,** 641–647.
74. Walhout, A. J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K. C., Schetter, A. J., *et al*. (2002) *Curr. Biol.* **12,** 1952–1958.
75. Ge, H., Liu, Z., Church, G. M. & Vidal, M. (2001) *Nat. Genet.* **29,** 482–486.
76. Levesque, M. P. & Benfey, P. N. (2004) *Curr. Biol.* **14,** R179–80.
77. Benfey, P. N. (2004) *Dev. Cell* **7,** 329–330.
78. Thomas, M. A. & Klaper, R. (2004) *Trends Ecol. Evol.* **19,** 439–445.
79. Wake, M. L. (2003) *Integr. Comp. Biol.* **43,** 239–241.
80. Simpson, A. G. & Roger, A. J. (2004) *Curr. Biol.* **14,** R693–R696.
81. Bhattacharya, D., Yoon, H. S. & Hackett, J. D. (2004) *BioEssays* **26,** 50–60.
82. Adoutte, A., Balavoine, G., Lartillot, N. & de Rosa, R. (1999) *Trends Genet.* **15,** 104–108.
83. Medina, M., Collins, A. G., Silberman, J. D. & Sogin, M. L. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 9707–9712.
84. Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, U., Baguña, J. & Ruitort, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11246–11251.