Collins, Allen G. 1999. Molecules and evolutionary history. In: Springer D. and J. Scotchmoor (eds.) Evolution: investigating the evidence. Paleontological Society Papers, volume 3.

This publication by the Paleontological Society aims at providing high school teachers with readable reviews of a variety of topics related to evolutionary history.

PALEONTOLOGISTS LEARN and tell the history of life; it is our job. You might suspect that paleontologists spend most of their time studying fossils. While fossils are an important source of information for the paleontologist, other types of evidence can also tell us about biological history. For instance, the rocks themselves provide important information, especially about past climates. It makes perfect sense that organisms are more easily understood if you know the environment in which they lived. A third important source of information is all around us. The organisms alive today are the current products of the various processes of evolution that have been at work for more than three billion years. Organisms carry the legacy of their histories with them, in their anatomy, behavior, and genes. By studying and comparing living organisms, we learn about the past. Advances in technology have made the abundant historical information contained in biological molecules, chiefly genes and their RNA and protein products, easier to obtain. Thus, it is not too surprising to see today’s paleontologist setting about his or her business with a rock hammer in one hand and a pipettor in the other.

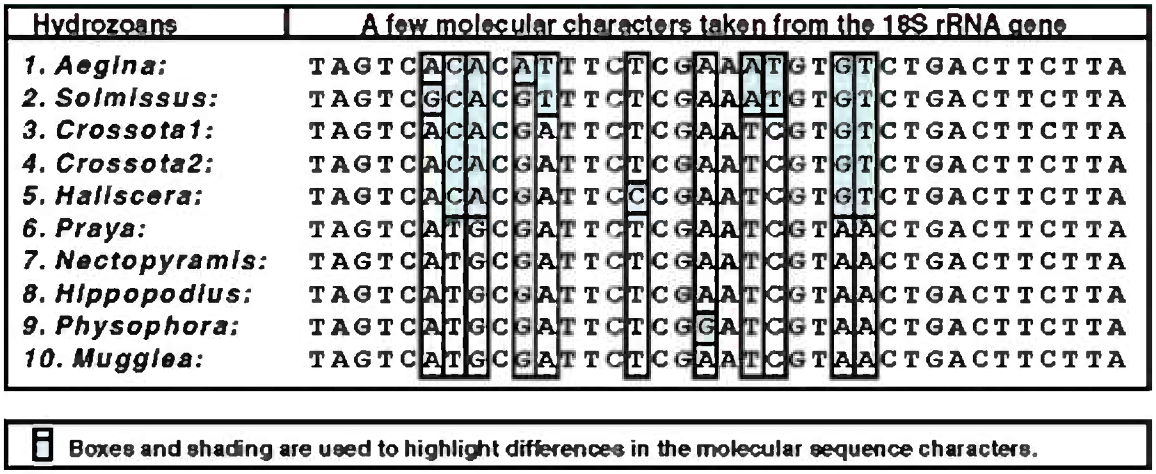
Many different things can be learned about the history of life from molecules. The most important lesson lies in their ability to unveil how organisms are related to one another, i.e. life’s phylogeny (see chapters in this volume that deal with phylogeny). Our understanding of evolutionary relationships has been revolutionized in a very short period, 20 years or so, spurred on by the study of molecules. For instance, molecules have shown us that there are two distinct types of prokaryotes (Archaea and Bacteria). Furthermore, you might be surprised to learn that you are more closely related to one type of prokaryote than to the other, as will be discussed later. While interesting on its own, phylogeny is also very useful. Biological questions are easier to figure out if you know something about the phylogeny of the organisms concerned. For instance, you might want to know how an adaptation such as insect wings came about. In this case, it would be helpful to compare the anatomy and behavior of insects to the anatomy and behavior of those organisms that are most closely related to insects. But, what organisms are most closely related to insects? Phylogeny guides the paleontologist to make comparisons that best unveil the answers to biological questions of all sorts.

Technological advances have put molecules within the grasp of historical biologists and molecules have proven useful for tasks other than phylogenetic reconstruction. As you will read below, some molecules have been used as clocks to date the origination of groups of organisms. Other molecules are revealing how body plans and structures of multicellular organisms are formed. These molecules hold the exciting promise of exposing how body plans first evolved. Molecules are even found in rocks, as fossils that mark the presence and/or activity of organisms in the past.

MOLECULES AND PHYLOGENY

How does molecular evidence reveal phylogeny?

In the late 1950s, even before the basics of protein synthesis had been worked out, it was becoming clear that molecules would be useful for reconstructing phylogeny. Francis Crick, co-discoverer of the structure of DNA, was probably among the first to make the connection between molecular sequences and phylogeny. He wrote, "sequences are the most delicate expression possible of the phenotype of an organism . . . vast amounts of evolutionary information may be hidden away within them" (Crick, 1958). Not long after that, during the early 1960s, two researchers presented a more formal explanation of how molecules can document biological history (Zuckerkandl and Pauling, 1962, 1965). They clarified that while not every biological molecule holds promise for aiding in phylogenetic reconstruction, proteins and nucleic acids do.



Hydrozoans	A few molecular characters taken from the 18S rRNA gene
1. *Aegina:*	TAGTCACACATTTCTCGAAATGTGTCTGACTTCTTA
2. *Solmissus:*	TAGTCGCACGTTTCTCGAAATGTGTCTGACTTCTTA
3. *Crossota1:*	TAGTCACACGATTCTCGAATCGTGTCTGACTTCTTA
4. *Crossota2:*	TAGTCACACGATTCTCGAATCGTGTCTGACTTCTTA
5. *Haliscera:*	TAGTCACACGATTCCCGAATCGTGTCTGACTTCTTA
6. *Praya:*	TAGTCATGCGATTCTCGAATCGTAACTGACTTCTTA
7. *Nectopyramis:*	TAGTCATGCGATTCTCGAATCGTAACTGACTTCTTA
8. *Hippopodius:*	TAGTCATGCGATTCTCGAATCGTAACTGACTTCTTA
9. *Physophora:*	TAGTCATGCGATTCTCGGATCGTAACTGACTTCTTA
10. *Mugglea:*	TAGTCATGCGATTCTCGAATCGTAACTGACTTCTTA

▯ Boxes and shading are used to highlight differences in the molecular sequence characters.

Figure 1. Patterns that can be seen in a few molecular characters taken from the 18S ribosomal RNA gene sequences of ten hydrozoan species (taken from the author's unpublished data). Differences in the molecular sequence characters are in boxes and highlighted with shading. From the patterns present in these data, one might recognize three groups of hydrozoans based on similarity of the sequences (group A: hydrozoans 1 and 2; group B: 3,4, and 5; group C: 6,7,8,9, and 10). Further, one might suppose that groups A and B inherited the characters that they uniquely share as a result of common history, suggesting that A and B are more closely related to each other than either is to Group C. Full phylogenetic analyses include so much data that the alternative grouping possibilities are immense. Computer programs are necessary to carry out phylogenetic analyses to completion.

DNA, RNA, and proteins have potential to reveal evolutionary relationships for three reasons. First, nucleic acids and proteins are composed of linear strings of numerous smaller parts, nucleotides and amino acids respectively. Each nucleotide or amino acid in a molecular sequence is a character, albeit not a very colorful one, that can be used to describe an organism. Second, these molecules are replicated from generation to generation, but not perfectly. Changes of various sorts in the genomes of all organisms happen all the time.

Some of these mutations are inherited by descendants. Finally, all living organisms on Earth share some history as a common lineage. That is, living organisms are all connected by ancestor-to-descendant relationships. Thus, by comparing the nucleotide and/or amino acid sequences of different organisms, it is possible to identify characters that two or more organisms share as a result of their common history. For example, Figure 1 shows a handful of molecular characters for ten hydrozoan jellyfish. By eyeballing these data, you might detect some patterns that possibly indicate shared history. Recognizing patterns such as these is the beginning of a phylogenetic analysis. However, the data and the alternative grouping possibilities are so numerous that computer programs, as described below, are needed to carry phylogenetic analyses to completion.

It should be emphasized that besides molecules, anatomical, physiological, behavioral, embryological and other variable and inherited characters are also useful for phylogenetic inference. Today's emphasis on using molecular characters (for phylogenetic analysis) is due in part to technological advances that have made it possible to gather numerous molecular characters inexpensively. Another reason that molecules are so commonly used is probably that they are fashionable. Fortunately, the current spate of molecular phylogenies is spurring on phylogenetic analyses based on non-molecular characters. Technological advances, for instance in image acquisition and analysis, are also making non-molecular characters more readily available. All types of data that have the potential to reveal phylogenetic history should be investigated.

Choosing the right molecule for revealing phylogeny.

Rates of molecular change are highly variable, from one gene to another, from one portion of a gene to another, from one lineage to another, and from one time in the past to another. Consequently, not every molecular sequence is useful for answering every question of phylogeny. A general balance must be struck between the age of the phylogenetic divergence being examined and the rate at which a given molecule evolves. For instance, if you were to compare sequences of a slowly evolving gene from organisms that shared a very recent common ancestor, you would find that the sequences are nearly identical. Insufficient historical information has accumulated in slowly evolving sequences to reveal the relationships of closely related species. Recent divergences are better investigated by using molecules that tend to evolve rapidly. Similarly, slowly evolving sequences are most useful for revealing divergences that occurred long ago. Molecules that evolve quickly are of little use for identifying ancient divergences because subsequent mutations destroy the evidence that certain molecular characters were shared by organisms as a result of common history. Revealing divergences of lineages that happened long ago over a brief period of time is especially challenging. In such cases, slowly evolving molecules may not have accumulated enough change during the period of rapid branching. On the other hand, quickly evolving molecules that did accumulate change during the period have subsequently experienced mutations that hide the ancient changes. The prescription for all difficult phylogenetic questions is additional data and fresh analyses.

Computer programs are a must for molecular analyses.

Programs and computers are necessary tools for figuring out evolutionary relationships. Phylogenetic questions are surprisingly complex and a great number of methods have been developed for attacking them (see Swofford et al., 1996 for a thorough review). Phylogenetic analyses often involve generating and evaluating different possible topologies (branching trees) by some standard measure. In these analyses, alternative topologies are scored based on a given standard. The "best" tree found is the one that has the optimal score. Finding the best tree is not always a simple task. As the number of taxa being considered increases, the number of possible tree arrangements that join the taxa increases dramatically. For example, while there are just 15 topologies possible for five taxa, this number increases to 10,395 for 8 taxa and to 34,459,425 for just 11 taxa. For analyses that contain a dozen or more taxa, it is computationally impossible to consider every tree. A variety of algorithms have been devised to evaluate only a fraction of the total number of trees while seeking to minimize the chance of missing the overall optimal tree.

Is there any assurance that molecular analyses work?

You might wonder if there is any way to check whether the techniques of phylogenetic reconstruction actually reveal evolutionary history. After all, it is impossible to go back in time and actually watch lineages diverging. Or, is it? Viruses evolve extremely rapidly. Hillis and colleagues (1992, 1994) used viruses that live in bacteria to experimentally manipulate phylogeny. They were able to create divergences in viral lineages by splitting colonies of their bacterial hosts. Thus, the actual phylogenetic history of the viruses was known. Subsequently, they determined molecular sequences for the viruses and attempted to reconstruct their phylogeny using typical techniques. Happily, their results matched the known branching history of the viruses. It is nice to know that the techniques that are being used to reconstruct phylogenies hold for viruses in the laboratory, even though they represent just a limited sample of evolution.

The results of molecular phylogenetic analyses.

Any phylogenetic tree, with the exception of the viral phylogenies mentioned above, is a hypothesis of evolutionary relationships. Therefore, phylogenies are not final results. Molecular data are used to test phylogenetic hypotheses derived from other types of data. Molecules and morphology often point to the same evolutionary relationships, but not always. Some molecular phylogenies have contradicted hypotheses based on morphology and have suggested new possibilities for how organisms are related. In turn, these hypotheses must be tested with other sets of data. Through this process, a coherent picture of life's phylogeny will emerge.

Consider the case of animals, plants, and fungi. Traditionally, plants and fungi were grouped together. Today, most fungal collections reside at botanical institutions as an historical consequence of this view. Later, as fungi became better characterized, they were placed as one of the five great kingdoms of life, on a par with plants and animals. Later still, as phylogenetic thinking was beginning to take hold, some morphological characteristics hinted that fungi and animals may be more closely related to each other (Cavalier-Smith, 1987). Not surprisingly, this phylogenetic hypothesis was tested with molecular sequences. Ribosomal RNA sequences corroborated the link between fungi and animals (Wainwright et al., 1993). Since then, evidence from several other genes that suggest that animals and fungi are more closely related to each other than either group is to plants has been reported (Baldauf and Palmer, 1993; Borchiellini et al., 1998). Unless or until contradictory information is brought

into view, it will be accepted that this hypothesis best represents the true evolutionary relationship of these three groups.

Some phylogenetic questions generate considerable controversy. One example includes the use of fossil DNA. Under circumstances where fossils are preserved without free water (examples would be extreme cold or in amber), DNA may not degrade. Recently, DNA from a fossil mammoth was extracted and an attempt was made to determine how it is related to the two living species of elephants, Asian and African (Ozawa et al., 1997). These researchers "confirmed" that the mammoth was more closely related to the Asian elephant than it is to the African elephant in accordance with morphological data. Less than a year later, a second group of researchers reported that fossil DNA that they had extracted from mammoth indicated the contrary (Noro et al., 1998). Disagreements, such as this, are sometimes used to conclude that molecular sequences are not good at revealing phylogeny. While they may be frustrating, contradictions are normal events in the progression of scientific knowledge.

Molecular phylogenies do not only focus on events of the distant past. Some have practical importance to our lives. Recently, a molecular phylogenetic analysis was used to suggest that the virus which causes AIDS in humans, HIV, is derived from a similar virus that exists harmlessly in chimpanzees (Gao et al., 1999). Moreover, the phylogenetic results were so robust (well-supported) that they allowed the researchers to strongly suggest that a specific subspecies of chimpanzee from western equatorial Africa is the host of the strain of HIV that causes AIDS in humans. Interestingly, chimpanzees are hunted for food in this region of Africa, providing a likely mechanism of cross-species transmission of the virus to humans.

Another recent phylogenetic study, which included chimpanzees and humans, attempted a new classification of primates based on molecular, morphological, and fossil data (Goodman et al., 1998). Among the interesting conclusions of this study (for humans anyway) was that humans and chimpanzees ought to be given the same generic name. This argument rests on the fact that the degree of sequence divergence between other primate species of the same genus is equivalent to or exceeds that observed for chimpanzees and humans. Our generic name, *Homo*, is older and thus has precedence over the generic name of chimpanzees, *Pan*. Had it gone the other way, just imagine Carolus Linnaeus, the type specimen of our species, rolling in his grave on learning that he had suddenly become *Pan sapiens*.

Molecular phylogenies and systematics.

Today, molecular phylogenies are prominent in systematics, a broad field that deals with the discovery, description, organization, and naming of life (See Carlson chapter). The value of systematics can hardly be doubted as we face the responsibility of preserving biodiversity in the face of explosive natural resource depletion and human population growth. Nevertheless, while the tasks of systematics appear to be rather straightforward, they are difficult to achieve. First among these difficulties is the sheer number of species on the planet. With tens of millions of species yet to be described, it could be argued that the work of systematists might never be completed. Other, more subtle difficulties come to light when one considers groups of species, and the task of placing these groups in meaningful hierarchies. This task has given rise to divisive debates among biologists that will not be discussed here. Instead, let's

consider a problem that current systematics is posing for biology teachers at all levels.

Because textbooks are not keeping up with our rapid gain in knowledge of evolutionary relationships, teachers are put in an unfortunate position. Teachers can hardly be expected to consult the primary literature to get the latest phylogenies or classifications. Our views of evolutionary relationships, and the classification schemes based on them, are changing so rapidly that textbooks are quickly outmoded. So, what are teachers to do? Here are a few recommendations.

1. Focus on the general utility of phylogenetic classification schemes. Phylogeny provides a natural and useful scheme for organizing life. By giving organisms names that correspond to evolutionary history, then learning names becomes equivalent to learning history.

2. Emphasize that phylogenetic classifications are not compatible with the hierarchical ranks of the Linnean system (e.g., phylum, kingdom, order, etc.). The levels in phylogenetic hierarchies are too numerous to categorize in this manner.

3. Stress hypotheses and hypothesis testing. The organization of life's groups in textbooks is not dogma. They are reasoned ideas and they are subject to further testing and revision. It is a valuable lesson for students to learn and accept that uncertainty exists, in their minds, as well as in the minds of teachers and scientists.

4. Finally, do not teach that molecules provide all of the answers in phylogeny. Molecules have played an important role in revolutionizing our understanding of phylogeny, but they have not given us all of the answers. Morphological analyses of phylogeny are equally important. These two types of data complement each other in the basic goal of recovering evolutionary relationships.

Molecular phylogenies and biogeography.

Biogeographers seek to comprehend the spatial distribution of organisms on the planet. History plays an important role in determining biogeographic patterns. Not surprisingly, molecular phylogenies have been useful in understanding the processes behind biogeographic patterns. (See Avise, 1994 for review.) A recently completed doctoral dissertation illustrates how a phylogeny is not only helpful, but also necessary for interpreting the geographic distributions of cowries (Meyer, 1998). Cowries are a group of marine snails that are largely associated with reef environments. Like many other groups of marine organisms, the number of cowrie species is much greater in the tropical western Pacific (TWP) than it is elsewhere. This pattern has been recognized for a long time, and many opinions have been offered to explain it. Among the competing hypotheses to explain the high species richness observed in the TWP, are the following:

1. Species were generated preferentially inside this region.

2. Species were generated on the periphery of this region where genetic isolation occurred, and subsequent migration back towards the center of the region causes its high diversity.

3. Species were generated in the Indian and Pacific oceans and range overlaps of closely related species caused the high diversity.

How can these ideas be tested? Branch points on a phylogeny represent events of speciation, and speciation is what generates new species. Thus, a necessary

first step to evaluating these hypotheses was to determine how cowrie species are related to each other. To this end, Meyer created a comprehensive phylogeny using the sequences of two genes (Meyer, 1998). He then mapped geographic distributions from living and fossil cowrie species onto this phylogeny. The resulting pattern was mosaic, but a clear picture began to emerge as he incorporated the geologic history of the region. What he found was that Hypothesis 3 is unlikely to be responsible for the high diversity seen in the TWP, because his phylogeny showed that very few closely related species have ranges that overlap in the TWP. Hypothesis 2 could also be ruled out as the primary explanation, especially over longer periods of geologic time (greater than three million years), since his phylogeny indicated that species living on the periphery of the TWP were typically ancient lineages that had remained isolated for many millions of years. Finally, Hypothesis 1, that species had been preferentially generated within the western Pacific appeared to be very likely. His phylogeny revealed that one particular group of cowries had diversified into many species in the relatively recent past. Sea level changes associated with ice ages during the last 2.5 million years have apparently isolated small basins in the TWP, providing an ideal setting for speciation to occur.

OTHER USES FOR MOLECULAR DATA

Dating divergences and the molecular clock.

When Zuckerkandl and Pauling articulated their ideas concerning the usefulness of molecules for documenting evolutionary history, they suggested that molecules might evolve at a rate constant enough to estimate when two lineages diverged (1962, 1965). Today, it is widely recognized that there is a general relationship between time and molecular divergence. However, the idea that molecules can be used as clocks to gauge the age that two lineages split remains controversial and complicated (Ayala, 1997). As noted above, rates of molecular change are highly variable, from one gene to another, from one portion of a gene to another, from one lineage to another, and from one time in the past to another. Thus, it is difficult to determine the expected range of errors for molecular clock estimates (see Hillis et al., 1996a). Nevertheless, molecular clocks are quite often used. Given the ever-growing amounts of molecular data available and the inherent interest in particular biological events of the past, especially the origination of major groups of organisms, increasingly sophisticated applications of molecular clocks are appearing in the professional literature.

When a molecular clock estimate is made for the origin of a group of organisms that has very little or no fossil record, arguments can be made about techniques and assumptions. The debates become far more interesting, however, when molecular clocks are used to date the origin of a group with a relatively robust fossil record. Fossils and molecular sequences are independent lines of evidence that ideally would corroborate each other. Molecular clock estimates for the origin of groups usually predate the earliest fossil evidence for the group. In a certain respect, one would expect this. The first fossil of a group would be a minimum age estimate of when that group first evolved. However, it is sometimes difficult to invoke such an explanation for the discrepancies observed between molecular clock estimates and first fossil estimates. Such is a case with a number of groups, e.g., primates, birds, mammals, animals, flowering plants, plants, etc.

As an example, consider the animals. Molecular clock estimates for the divergence of early animal lineages range from 1,500 to 700 million years before present (Runnegar, 1982; Wray et al., 1996; Nikoh et al., 1997; Gu, 1998; Ayala et al., 1998; Bromham et al., 1998). The oldest fossil evidence of animals is roughly 600 million years old (Brasier and McIlroy, 1998; Li et al., 1998). At present, there is a disparity of roughly 100 to 900 million years between the molecular clock estimates for when the major animal clades originated and the oldest fossil evidence that definitively demonstrates their existence. Two opposing possibilities could explain this disparity. There is either a hidden period of animal history or there is a systematic bias in the molecular clock estimates. For many paleontologists, it is hard to imagine an adequate explanation for the absence of animal fossils over hundreds of million years. The leading explanation offered by the proponents of molecular dates invokes the idea that animals were too small to be fossilized during this extended period. It can be countered that many fossils of small animals, while rare, are known. As for a bias in the molecular dates, little work has been done to address this possibility. Thus, we are currently in a state of partial ignorance concerning molecular clocks. A certain amount of caution in applying, interpreting, and evaluating molecular clocks is warranted.

A few words on the molecules behind development.

Have you ever wondered how animals and plants grow from a single cell to recognizable adult forms, or how the basic types or body plans of adult forms came to be in the first place? If so, you are not alone. A blooming area of molecular research deals with the genes that are involved in the development of multicellular organisms. Some genes have protein products that regulate the transcription of other genes. That is, they control when and where other genes will be turned on and off. Pathways of gene action exist because the expression of a single gene can set off an entire cascade of downstream effects. Studies that examine these regulatory gene cascades during the development of an organism are providing clues as to precisely how the basic parts of animals and plants are put together into cohesive functional individuals. Furthermore, as more and more regulatory genes of various types are identified in different organisms, it is becoming possible to infer what genetic components are responsible, at least in part, for the evolution of novel body plans. This goal is being achieved by comparing the existence and function of regulatory gene pathways in a phylogenetic context. (See Valentine chapter for more on this exciting topic.)

A Few Words on Fossil Molecules, Biomarkers.

Some organic molecules are preserved over geologic time and indicate or mark the presence and/or activity of organisms. These fossil molecules, or biomarkers, can be found in fossil shells, sedimentary rocks, and oil deposits. (Identification is usually accomplished with gas chromatography and mass spectrometry, should you want to look it up.) One nice thing about biomarkers is that they can be used to infer that an organism once lived at a particular place and time in the past, even in the absence of more traditional fossil evidence. For instance, a group of geologists studying oil deposits noted that some previously unidentified organic compounds were particularly abundant in their samples derived from source rocks of Vendian age, roughly 600 to 540 million years old (McCaffrey et al., 1994). The type of compounds they found, steranes, are derived from sterols, which are precursor lipid molecules present in eukaryote organisms. In particular, the sterols that would give rise to the

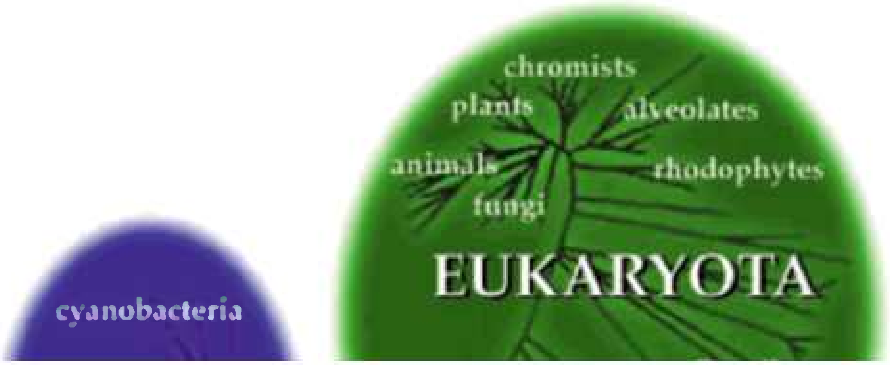
steranes they had discovered are a major component of a particular group of sponges. No fossil sponges had ever been described from Vendian rocks. Nevertheless, since oil deposits from the Vendian contained the novel steranes, the petroleum researchers predicted that sponges did indeed exist during this time. Subsequently, fossil sponges from the Vendian period have been described (Gehling and Rigby, 1996; Li et al., 1998).

Other biomarkers, preserved in fossil animal skeletons such as shells, have been used to infer ecological interactions of the past (CoBabe, in press_a, in press_b). Work of this sort is extremely promising because it strongly integrates ecology and evolution. For example, by determining the presence and ratio of certain biomarkers in a given snail shell, one can deduce whether the snail was an herbivore or a carnivore. Moreover, one can determine whether the snail was a generalized feeder or specialized in one food source. This information could be used to make connections between changes in diet and changes morphology or the environment through time. Biomarkers incorporated into skeletal material have also been used to infer the presence of intercellular chemosymbionts in animals. This valuable information about the lives of past organisms could not otherwise have been ascertained without the use of biomarkers.

Perhaps it should be mentioned briefly that not all biomarkers are molecules. Paleontological studies employ isotopes in a variety ways, sometimes as biomarkers. Just recently, isotopes have been used to infer that life existed at least 200 million years prior to the oldest fossil remains, which are known from strata dating to 3,500 million years (Rosing, 1999). These researchers looked at the ratio of two isotopes of carbon, carbon-12 and carbon-13, in 3,700 million year old strata, and found that the rocks were enriched in carbon-12. They inferred the presence of life because sediment formed on sea bottoms today is similarly enriched in carbon-12 in areas rich with bacterial plankton.

Rooting the Tree of Life, A Clever Use of Molecules.

Recall from the introduction that there are two distinct types of prokaryotes (Archaea and Bacteria). Beginning with the work of Woese, molecular studies with a number of different genes have shown that there are three fundamental divisions of life, the Archaea, Bacteria, and Eukaryota, as shown in Figure 2 (Woese et al., 1990). You and I are members of the Eukaryota; our cells have the organelles and nuclei to confirm this. Organelles and nuclei are not present in the other two groups. Does that mean that Archaea and Bacteria are more closely related to each other, or could it be that one of these two groups is more closely related to eukaryotes? It turns out that you (as a eukaryote) are more closely related to Archaea than to Bacteria.



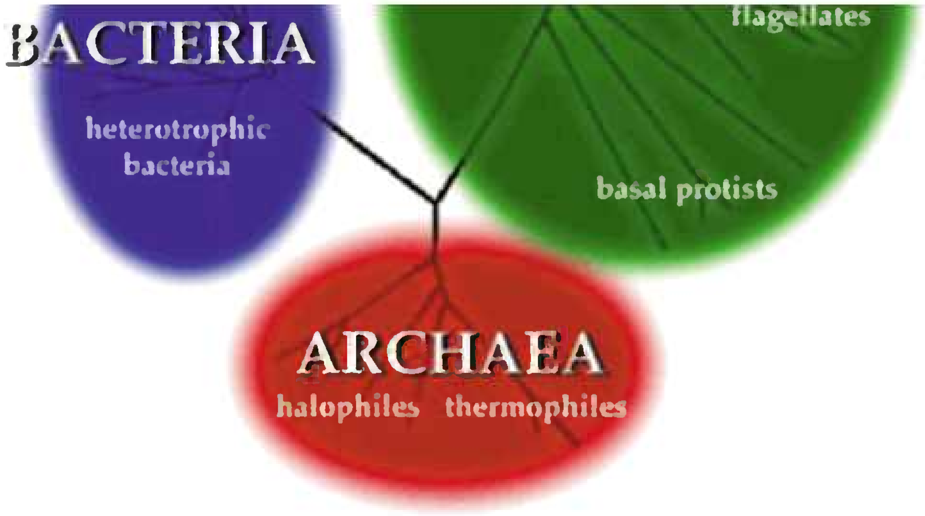


Figure 2. The three major divisions of life, Archaea, Bacteria, and Eukaryota (used with permission from the website of the University of California Museum of Paleontology: http://www.ucmp.berkeley.edu/alllife/threedomains.html).

The method by which this result was determined is not exactly straightforward, but it required an elegant and novel use of molecules, so it is worth outlining. The key to resolving the relationships of Archaea, Bacteria, and Eukaryota was to establish the root of the tree of life. The root of a phylogenetic tree is the place on the tree that represents the last common ancestor of all organisms being considered. Figure 3 shows how alternative possibilities for placing a root on the tree of life imply alternative relationships among Archaea, Bacteria, and Eukaryota. The general method for determining the root of phylogenetic trees is to use an outgroup (one or more organisms that are more distantly related to the organisms in question). Figure 4 shows how an outgroup is used to place a root on a phylogenetic tree. In the case of the tree of all life, however, there is no outgroup because the only possibilities would be non-living things, which are not related to life by definition. Rooting the tree of life was a conundrum until molecules were cleverly put to use to answer it.

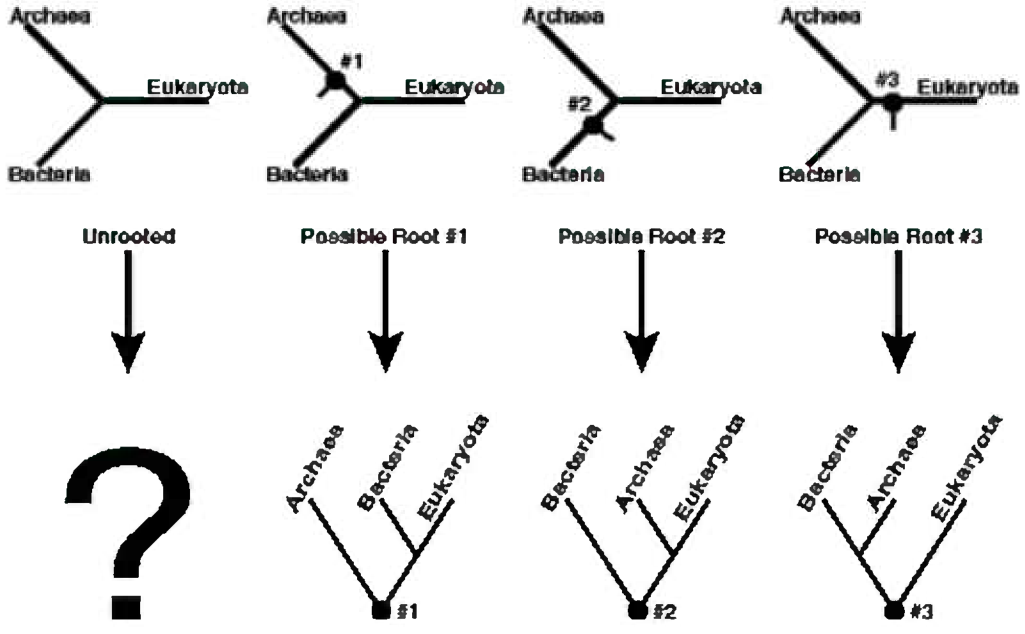


Figure 3. Three alternative possibilities for the root (the point which represents the last common ancestor) on the tree of life. Without a root, it is not possible to tell which two of the three groups (Archaea, Bacteria, and Eukaryota) is most closely related. The three alternative placements of the root on the tree, imply three separate possibilities for the evolutionary relationships of the three groups.

There is strong molecular evidence to support the assertion that Archaea and Eukaryota share a more recent common ancestor than either group does with the Bacteria (possibility #2 in Fig. 3). In order to understand how this relationship was determined, it is necessary to know a little about the process of gene duplication. Some mutations involve the duplication of portions of the genome, which may result in the creation of a redundant copy of a gene. After the duplication event, the two genes evolve separately. The first step to solving the root of life was to find a gene that was duplicated prior to the last common ancestor of everything alive today.

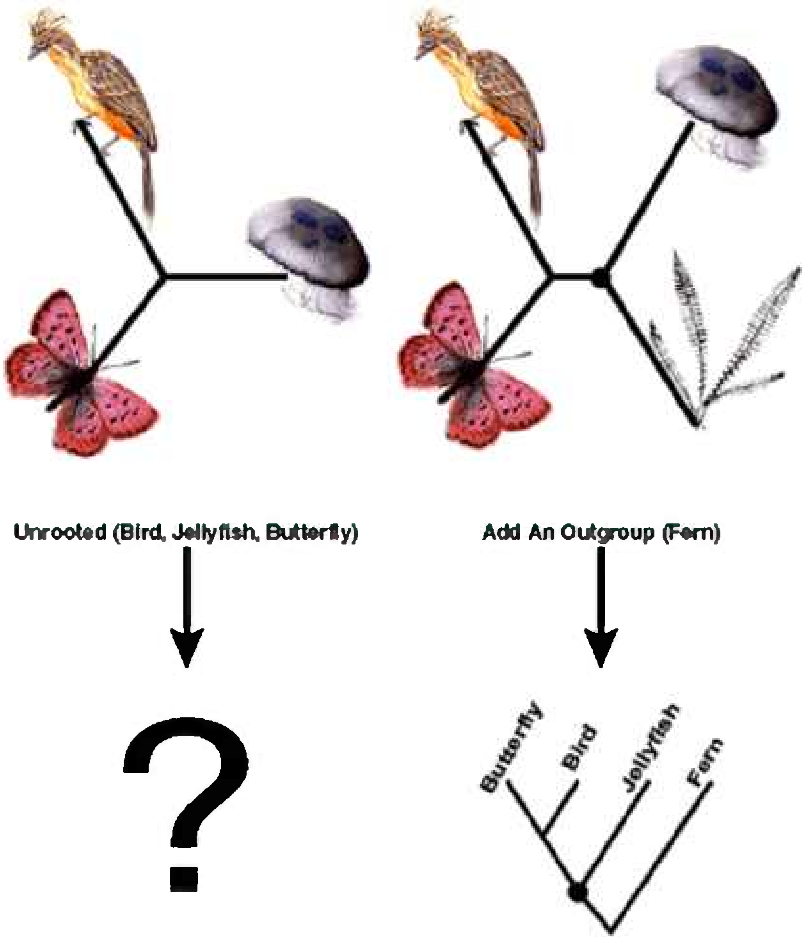


Figure 4. Illustration of how an outgroup, one or more organisms that is less closely related to the organisms under consideration, is normally used to root a branching diagram. Without an outgroup to root the tree, it is not possible to tell which two of the three organisms (bird, jellyfish, and butterfly) are most closely related. By including a fern as an outgroup to the analysis, the tree can be

rooted, and the relatedness between the bird, jellyfish, and butterfly is revealed.

Following along with Figure 5, let's step through the logic of how such a gene can resolve this puzzle. Consider a gene, GeneA, which was duplicated (A1 and A2) prior to the last common ancestor of everything alive today. Next, suppose the last common ancestor inherits the two forms of the gene, GeneA1 and GeneA2, and subsequently passes them on to all of its descendants. The two forms of GeneA, A1 and A2, will be present in any living organism (Fig 5A.).

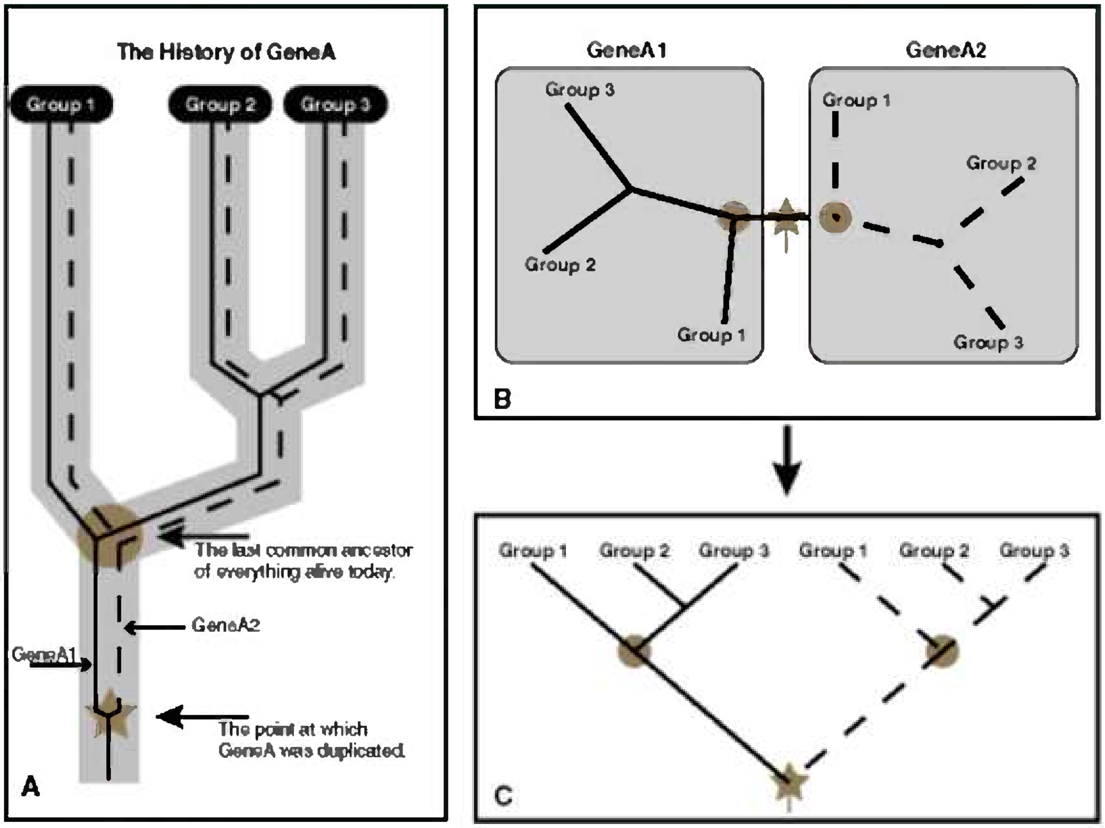


Figure 5. Illustration of how an anciently duplicated gene could be used to determine the root of the tree of life.
A. The true history of an anciently duplicated gene, GeneA. The duplication, denoted with a gray star, results in two separate genes, GeneA1 and GeneA2, that are passed on in the lineage that leads to the last common ancestor of all life, denoted by a gray circle. Through time, the two genes are passed on to all descendants of the last common ancestor, including everything alive today. A solid line traces the history of GeneA1, while a dashed line traces the history of GeneA2.
B. A tree reconstructed from the sequences of GeneA1 and GeneA2 that correctly matches the true history as shown in A. The GeneA1 sequences, connected by solid lines, cluster on one side, while the GeneA2 sequences, connected by dashed lines cluster on the other. The gene duplication, marked with a gray star, happened before the last common ancestor of life, denoted by two gray circles, existed.
C. By using the last common gene ancestor as the root of the tree, the relative relationships among the three groups of life is revealed.

If you were to build a tree using the sequences from Archaea, Bacteria, and Eukaryota for either GeneA1 or GeneA2, then the three groups would be revealed as distinct from one another. However, without a root there would be no way of knowing which two of the three groups shared the most recent common ancestor. All is not lost, however, because we know that GeneA1 and GeneA2 are related to each other. Further, we know that the gene that gave rise to GeneA1 and GeneA2 (denoted by a gray star in Fig. 5) predates the last common ancestor of all life (denoted by a gray circle in Fig. 5).

The next step is to build a gene tree using the sequences of both GeneA1 and GeneA2. Finally, we can reason that the root of this gene tree should be placed on the branch that connects the GeneA1 side of the tree to the GeneA2 side. This is because the root represents the common gene ancestor, which existed prior to the last common ancestor of the three groups of life.

It was through this ingenious method that two groups of scientists independently concluded that the Archaea and Eukaryota are more closely related to each other than either is to the Bacteria (Gogarten et al., 1989; Iwabe et al., 1989). Since then, several studies that have relied on different pairs of anciently duplicated gene sequences have all reached the same conclusion (Brown and Doolittle, 1995; Lawson et al., 1996; Gribaldo and Caqmmarano, 1998). It is difficult to imagine that Halobacterium, a salt-loving single-celled organism without organelles or a nucleus, shares more history with you than it does with true bacteria, but it appears to be true.

THE BASIC STEPS TO OBTAINING MOLECULAR SEQUENCE DATA

Using molecular sequence data to derive phylogeny has become a widespread practice because these data are reasonably easy to obtain. The relative ease of collecting molecular sequences is largely due to a key technological innovation, the polymerase chain reaction (PCR). When a biological problem is recognized that can be addressed by a molecular phylogeny, appropriate tissue samples must be gathered. After that, a few simple tricks are performed back in the laboratory to extract historical information from the molecules.

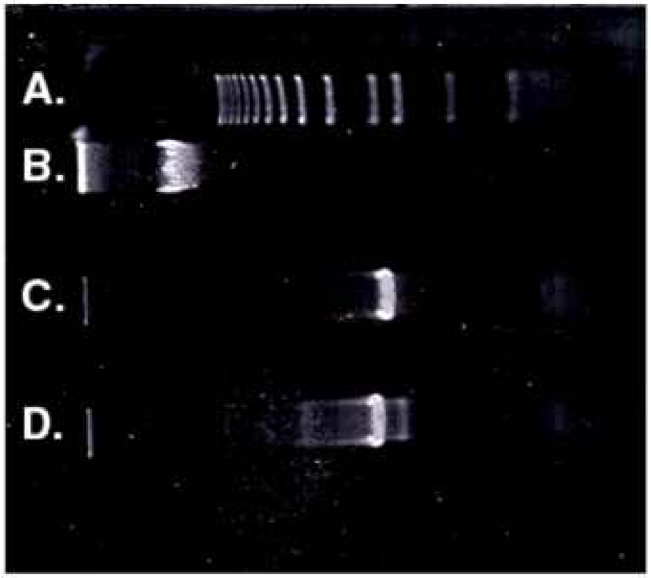
This section should provide a general understanding of how molecular sequence data are obtained and is not a guide to performing this type of work. There are numerous variants on each of the basic steps outlined here, including ones using household items that can be easily used in the classroom. Detailed descriptions of a bevy of molecular techniques useful to the systematist can be found in a comprehensive book edited by Hillis et al. (1996b).

Extracting nucleic acids from tissue.

In the laboratory, the first step to getting sequence data is to separate DNA from all other components of the organism’s tissue. (Under some circumstances, isolating RNA is the initial procedure, but the distinction is not important for this explanation.) DNA is extracted from tissue that is either fresh, frozen, or preserved. Preservation in 80 to 100 percent ethanol is common. Experience varies as to what method works best. A series of chemicals (some methods require noxious chemicals, while others do not) are added to finely ground or chopped tissue. These chemicals break down cell membranes and deactivate enzymes that destroy and fragment the DNA. At each step of the extraction, the researcher keeps track of whether the DNA is

dissolved in solution or precipitated out of solution. This allows portions of the mixture to be discarded while retaining the DNA. The goal is to have only DNA remaining in pure water.

Once extracted, DNA can be visualized by placing a tiny amount in a gel and applying a current. Since DNA is negatively charged, it will migrate towards a positive charge at a speed that is proportional to the size of the DNA fragment. Thus, after current has been applied for a period, smaller pieces of DNA will have moved farther through the gel than larger pieces. The next step is to stain the gel with a substance called ethidium bromide. Ethidium bromide sticks to DNA, and has the handy characteristic of fluorescing under UV light. Thus, you can take a photo of DNA in an ethidium bromide stained gel under UV light (Fig. 6). Having large pieces of extracted DNA maximizes your chances of successful PCR.



DNA fragments of known sizes for comparison

Large pieces of DNA extracted from tissues

Product of Successful PCR with one band of DNA corresponding to one gene

Less successful PCR results with two bands of DNA, one faint

Figure 6. Partial photograph of a gel containing DNA that has been stained with ethidium bromide. Samples were loaded at the left, and moved through the gel to the right when a current was applied.
A. A DNA ladder, a solution with DNA fragments of known length used for comparison to other DNA samples. The smaller fragments traveled farther through the gel to the right during the period that the current was applied.
B. DNA of high molecular weight, as is seen with successful extractions of DNA from tissue.
C. Results from a successful PCR. A single band of DNA that corresponds to the target gene has been amplified.
D. Results of a less successful PCR. A second faint band means that two pieces of DNA were preferentially amplified during the PCR. Parameters of the PCR reaction can subsequently be changed in order to increase the specificity of the reaction.

<u>Amplifying the target gene by PCR.</u>

The second step to procuring sequence data is to amplify (make a multitude of copies) a target gene (the single gene that you are investigating). This is accomplished through an ingenious technique called the polymerase chain reaction (PCR). PCR amplification consists of heating and cooling a mixture of one enzyme, two primers, many nucleotides (Gs, As, Ts, and Cs), extracted DNA, and a buffer solution with ions. The ions and buffer solution create the right ionic and chemical conditions for the magic to happen. The extracted DNA contains a few or more copies of your target gene. The nucleotides provide the necessary raw material. The two primers are short stretches of DNA (15 to 40 nucleotides long) that match each end of your target gene. The enzyme catalyzes the replication of DNA. If everything works correctly, the target gene is replicated millions of times in just a few hours. PCR sounds like a miracle, but the logic of it is relatively simple.

The PCR mix is put into a machine that cycles (25 to 35 times) through three temperatures. During the time that the PCR mix is at the first temperature, which is very high (92o C to 95o C), double-stranded DNA splits (denatures) into single strands. At the second temperature of the cycle, which is much cooler (37 o C to 60 o C), single strands of DNA bind (anneal) to complementary pieces of DNA. During this annealing stage, the primers preferentially find their match because they are much smaller than other pieces of DNA in the solution. During the third stage, DNA is replicated. This process occurs at 72 o C, the temperature at which the enzyme optimally extends new DNA strands by binding free nucleotides. This enzyme is active at 72 o C because it is derived from a heat loving bacterium. The discovery of this enzyme was an essential step in the development of PCR. Similarly acting enzymes from most organisms would be active at lower temperatures and destroyed by the high temperature needed for DNA denaturation. After the third stage, the cycle begins again. During each cycle, the concentration of the target gene increases, which enhances the efficiency of the annealing process in step two. In this way, the number of copies of the target gene grows exponentially. In order to check the results after the PCR is completed, a portion of the PCR product is run through a gel, stained with ethidium bromide, and photographed as described above (Fig. 6).

Deriving the sequence of nucleotides.

The final step to obtaining sequence data is to “read” the sequence of the target gene. This is a two-stage process. The first stage relies on the PCR technique in a modified form. This time the PCR solution includes some of the previously amplified gene as the starting DNA and a single primer that corresponds to one end of the gene. In addition, the solution includes special nucleotides, along with the normal ones. Recall that during the third stage of this PCR reaction, free nucleotides are being added to the end of growing strands of DNA. When one of the special nucleotides is incorporated into a growing strand, it stops the creation of the remainder of the strand. In addition, each of the special nucleotides is labeled so that it can be identified as a G, A, T, or C. The result of this type of PCR is a collection of partial target genes that end with a labeled nucleotide. The lengths of the partial genes range from one nucleotide to the full number of nucleotides in the target gene. Ideally, there are approximately equal numbers of partial genes at each length, all of which end with a labeled nucleotide.

The second stage in “reading” the sequence is to run the partial DNA strands through a gel. The fragments move through the gel at a speed that is proportional to their length. The first fragment to reach the end of the gel is one nucleotide long. It is followed by the fragment that is two nucleotides long, and so forth. Near the bottom end of the gel is a laser that shines on each fragment as it passes by. A sensor records the characteristic signal given off by the final nucleotide of the fragment; a G, A, T, or C. The sequence of the target gene is recorded nucleotide by nucleotide. The process is rarely perfect, so genes are usually sequenced in both directions and the complementary sequences are compared to check for discrepancies. The discrepancies are then resolved by visually inspecting digital images recorded by the sensor. Accuracy of molecular sequences is extremely important for later analyses. To paraphrase an esteemed colleague, you cannot make chicken salad from chicken “excrement”.

CONCLUSION

I hope I have shown that biological molecules are an enormous source of information about the history of life. Many biological questions have already been clarified using molecules (I have shared just a few), but countless questions still remain un-addressed. Current and future generations of scientists have a great deal of work ahead of them. But this is fortunate, because this type of work is incredibly fun. Many historical questions require time spent travelling and working in the field where other fascinating questions arise. Solving these questions is a challenging task, requiring creative and synthetic thinking. But it is a rewarding endeavor because these problems are tractable, all the more so given our growing knowledge of biological molecules. And so, we are able to share what we have learned about the history of life. In fact, being a historical biologist is so enjoyable that it is more like play than labor.

ACKNOWLEDGEMENTS

I thank a number of people who have reviewed and improved this manuscript, including J. A. Johnson, H. H. Hamilton, M. Stefanski, J. W. Valentine and two anonymous reviewers. As always, I am indebted to the University of California Museum of Paleontology for providing support. Finally, I am grateful to M. L. Sogin and J. D. Silberman who taught me how to pipette and think at the laboratory bench.

REFERENCES

Avise, J. C. 1994. Molecular Markers, Natural History and Evolution. Chapman & Hall, New York, 511 p.

Ayala, F. J. 1997. Vagaries of the molecular clock. Proceedings of the National Academy of Sciences, USA, 94:7776-7783.

Ayala, F. J., Rzhetsky, A., and F. J. Ayala. 1998. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. Proceedings of the National Academy of Sciences, USA, 95:606-611.

Baldauf, S. L., and J. D. Palmer. 1993. Animals and fungi are each other’s closest relatives: congruent evidence from multiple proteins. Proceedings of the National Academy of Sciences, USA, 90:11558-11562.

Borchiellini, C., Boury-Esnault, N., Vacelet, J., and Y. Le Parco. 1998. Phylogenetic analysis of the Hsp70 sequences reveals the monophyly of Metazoa and specific phylogenetic relationships between animals and fungi. Molecular biology and Evolution, 15:647-655.

Brasier, M. D., and D. McIlroy. 1998. Neonereites uniserialis from c. 600 Ma year old rocks in western Scotland and the emergence of animals. Journal of the Geological Society of London, 155:5-12.

Bromham, L., Rambaut, A., Fortey, R., Cooper, A., and D. Penny. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. Proceedings of the National Academy of Sciences, USA, 95:12386-12389.

Brown, J. R., and W. F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proceedings of the National Academy of Sciences, USA, 92:2441-2445.

Cavalier-Smith, T. 1987. The origin of Fungi and pseudofungi, p. 339-353. In Rayner, A., Brasier, C., and D. Moore (eds.), Evolutionary Biology of Fungi. Cambridge University Press, Cambridge.

CoBabe, E.A., and A. Ptak. (in press_a) In situ lipids in invertebrate skeletons: Evidence of diet in modern invertebrates. Paleobiology.

CoBabe, E.A. (in press_b) Chemosynthesis and chemosymbiosis in the fossil record: Detecting unusual communities using isotope geochemistry. Paleontological Society Short Course: Isotopes in Paleobiology.

Crick, F. H. C. 1958. On protein synthesis, p. 138-163. In Symposia of the Society for Experimental Biology, no. 12. The Biological Replication of Macromolecules. Cambridge University Press, Cambridge.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenberg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and B. H. Hahn. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397:436-441.

Gehling, J. G., and J. K. Rigby. 1996. Long expected sponges from the Neoproterozoic Ediacara fauna of South Australia. Journal of Paleontology, 70(2):185-195.

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E., Bowman, M., Manolsen, M. F., Poole, R. J., Date, T., Oshima, T., Konisha, T., Denda, K., and M. Yoshida. 1989. Evolution of vacuolar H+-ATPase: Implications for the origin of eucaryotes. Proceedings of the National Academy of Sciences, USA, 86:6661-6665.

Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J., Gunnell, G. and C. P. Groves. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Molecular Phylogenetics and Evolution, 9:585-598.

Gribaldo, S. and P. Cammarano. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. Journal of Molecular Evolution, 47:508-516.

Gu, X. 1998. Early metazoan divergence was about 830 million years ago. Journal of Molecular Evolution, 47:369-371.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R., and I. J. Molineux. 1992. Experimental phylogenetics: Generation of a known phylogeny. Science, 255:589-592.

Hillis, D. M., Huelsenbeck, J. P., and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science, 264:671-677.

Hillis, D. M., Mable, B. K. and C. Moritz. 1996a. Applications of molecular systematics, p. 515-543. In Hillis, D., Moritz, C., and B. K. Mable (eds.), Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.

Hillis, D. M., Moritz, C., and B. K. Mable. 1996b. Molecular Systematics, Second Edition. Sinauer Associates, Inc., Sunderland, Massachusetts, 655 p.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and T. Myata. 1989. Evolutionary relationship of the archaebacteria, eubacteria, and eukaryotes

inferred from phylogenetic trees of duplicated genes. Proceedings of the National Academy of Sciences, USA, 86:9335-9359.
Lawson, F. S., Charlebois, R. L., and J. A. R. Dillon. 1996. Phylogenetic analysis of carbomoyl-phosphate genes: Evolution involving multiple gene duplications, gene fusions, and insertions and deletions of surrounding sequences. Molecular Biology and Evolution, 13:970-977.
Li, C.-W., Chen, J.-Y., and T.-E. Hua. 1998. Precambrian sponges with cellular structures. Science, 279:879-882.
McCaffrey, M., Moldowan, J., Lipton, P., Summons, R., Peters, K., Jeganathan, A., and D. Watt. 1994. Paleoenvironmental implications of novel C30 steranes in Precambrian to Cenozoic age petroleum and bitumen. Geochimica et Cosmochimica Acta, 58:529-532.
Meyer, C. P. 1998. Phylogenetic Systematics, Biogeography and Diversification Patterns in Cowries. Unpublished Ph.D. Dissertation. University of California, Berkeley, 223 p.
Nikoh, N., Iwabe, N., Kuma, K.-i., Ohno, M., Sugiyama, T., Watanabe, Y., Yasui, K., Zhang, S.-c., Hori, K., Shimura, Y., and T. Miyata. An estimate of divergence time of Parazoa and Eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. Journal of Molecular Evolution, 45:97-106.
Noro, M., Masuda, R., Dubrovo, I. A., Yoshida, M. C., and M. Kato. 1998. Molecular phylogenetic inference of the Woolly Mammoth Mammuthus primigenius, based on complete sequences of mitochondrial cytochrome b and 12S ribosomal RNA genes. Journal of Molecular Evolution, 46:314-326.
Ozawa, T., Hayashi, S., and V. M. Mikhelson. 1997. Phylogenetic position of mammoth and stellar's sea cow within Tethytheria demonstrated by mitochondrial DNA sequences. Journal of Molecular Evolution, 44:406-413.
Rosing, M. 1999. 13C-Depleted Carbon Microparticles in >3700-Ma Sea-Floor Sedimentary Rocks from West Greenland. Science, 283:674-676.
Runnegar, B. 1982. A molecular-clock date for the origin of the animal phyla. Lethaia 15:199-205.
Swofford, D. L., Olsen, G. J., Waddell, P. J., and D. M. Hillis. 1996. Phylogenetic Inference, p. 407-514. In Hillis, D., Moritz, C., and B. K. Mable (eds.), Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.
Wainwright, P. O., Hinkle, G., Sogin, M. L., and S. K. Stickel. 1993. Monophyletic origins of the Metazoa: an evolutionary link with fungi. Science, 260:340-342.
Woese, C. R., Kandler, O., and M. L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences, USA, 87:4576-4579.
Wray, G., Levinton, A., and L. H. Shapiro. 1996. Molecular evidence for deep Precambrian divergences among metazoan phyla. Science 274:568-573.
Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution and genic heterogeneity, p. 189-225. In Kasha, M., and B. Pullman (eds.), Horizons in Biochemistry. Academic Press, New York.
Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8:357-366.