The Tree of Life

Joel D. Velasco

COMMON ANCESTRY IS one of the pillars of Darwin's theory of evolution. Today, the tree of life, which represents how all life is genealogically related, is often thought of as an essential component in the foundations of biological systematics and so therefore of evolutionary theory – and perhaps all of biology itself. It is an iconic representation in biology and even penetrates into popular culture.

Massive amounts of time, effort, and money are being put into understanding and reconstructing the tree. Yet there are serious debates as to the usefulness and even the very existence of the tree. Here I will attempt to critically evaluate the merits of some of these worries. In doing so, we will see that questions about the tree and the foundations of systematics can be answered in the light of a wide range not only of empirical considerations but of philosophical considerations as well. A historically informed picture of how and why we got to where we are today is important for understanding these debates; however, here I can give only the briefest of introductions to the history of the tree as it has been used in systematics before turning to contemporary and future considerations.

A POTTED HISTORY

Many authors before Darwin had considered the possibility or even promoted the idea that some species were directly genealogically related to each other (Fig. 41.1). Some, including Jean Baptiste Lamarck, had even proposed treelike structures to capture these relationships (Ragan 2009). But it was Darwin who revolutionized our understanding of the diversity of life with his *On the Origin of Species* (1859). It is in the *Origin* that we first see the importance of genealogy on a grand scale where Darwin convincingly argues that common ancestry explains both the striking similarities between different species and the apparent naturalness of a groups-within-groups hierarchical classification. In the *Origin*, Darwin (1859, 129–30) introduces the metaphor of the tree of life, which connects all life through common descent:

> The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth.... The green and budding twigs may represent existing species; and those produced during

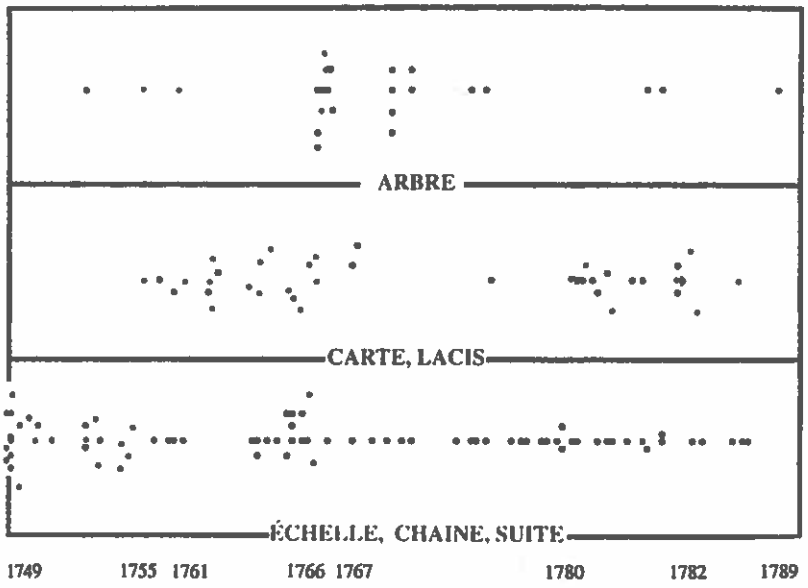

Figure 41.1. Before the idea of evolution took firm root, it was by no means obvious that a tree was the best way of portraying life's history. As can be seen from this chart mapping his different metaphors, Georges Buffon toyed with trees, maps, and chains (the dots record the usages of the respective images; the y-axis shows multiple usages in the same year), and indeed trees became less prominent in his thinking in later years. Drawing, inspired by G. Barsanti, Buffon et l'image de la nature, in *Buffon 8*, ed. J. Gayon (Paris: Vrin, 1992)

> former years may represent the long succession of extinct species.... the great *Tree of Life* ... covers the earth with ever-branching and beautiful ramifications. (emphasis added)

To help us understand descent with modification, which is essential for his theory of natural selection, Darwin gives us a figure – the only figure in the entire *Origin* – to which he then repeatedly refers (116) (Fig. 6.3). This tree represents real genealogical history and is not simply a classification scheme representing subordination of groups within groups, such as the diagrams previously given by Linnaeus, among others.

This idea of a tree that connects all life has been part of the biological literature since Darwin, but it would require twin revolutions in methodology and in the types of data available before serious attempts could be made at building truly universal phylogenies. By the 1950s, despite great advances in the knowledge of the phylogeny of eukaryotes, bacteriologists had generally given up on the idea of that it was possible to build a comprehensive phylogeny for most groups of bacteria. Morphological and physiological data just seemed too sparse and often conflicted (Sapp 2009). But in the early 1960s, Emile Zuckerkandl and Linus Pauling, among others, suggested that molecules such as genes, amino acids, or proteins could be used to track phylogenetic history. Zuckerkandl and Pauling (1965a, 1965b) proposed that some changes might occur at a constant rate forming a "molecular clock," which would aid in phylogenetic reconstruction as well as in determining the timing of evolutionary events.

At the same time, Carl Woese was working on the evolution of the genetic system itself. To examine the early evolution of life, one needs to know the broad-scale phylogenetic history of all life. In 1977, after painstakingly cataloging numerous rRNA sequences (and then searching for further kinds of data to validate their findings), Woese and George Fox announced that they had discovered a third kind of life: what they called the Archaebacteria. Despite being prokaryotic, the Archaebacteria lacked the typical signature found in all bacterial rRNA and, in addition, also shared many deep similarities with eukaryotes, such as the way that they performed transcription and translation. Over the next thirteen years, Woese and colleagues produced the first universal phylogenies (Fox et al. 1980; Pace, Olson, and Woese 1986; Woese 1987) and eventually proposed the three-domain model in which the Archaebacteria were renamed the Archaea, as opposed to the Bacteria and the Eucarya (Woese, Kandler, and Wheelis 1990). Today, the most common representations of the tree are akin to the phylogenetic tree depicted by Woese et al. in Figure 49.5. While some of the details of the tree are no longer accepted, this division of life into three great domains – the Bacteria, the Archaea, and the Eucarya – has been generally (though not universally) accepted as can be seen in Figure 41.2 taken from the back cover of an evolution textbook (Barton et al. 2007).

Woese was not concerned primarily with classification but was trying to answer a particular question: What is the correct evolutionary branching sequence for "major groups" of taxa? More recent reconstructions of the tree may add more taxa or use more or different kinds of data and may come to conclusions different from Woese's, but fundamentally they are working on the same project. While this certainly seems like a perfectly objective task, it depends on the idea that there is a unique, objectively correct tree of life. If there is not, then what purpose is there for us to infer *the* tree?

WHAT IS THE TREE?

A standard way to describe the tree is to propose that it is a universal phylogenetic tree depicting the genealogical relationships of all species through time. Thus, the tree of life is meant to be universal, to be a phylogeny, and to be a tree. Critics have directly or indirectly attacked each of these three apparently essential features.

What does it mean to say that the tree of life is a tree? Modern depictions of the tree of life do not look at all like *biological* trees, such as in the familiar drawings of Ernst Haeckel, but rather are *phylogenetic* trees, which are trees in the mathematical sense of a special kind of object in graph theory. It is conceptually helpful to think of a tree as a set of directed branches connecting nodes where there is a root node with

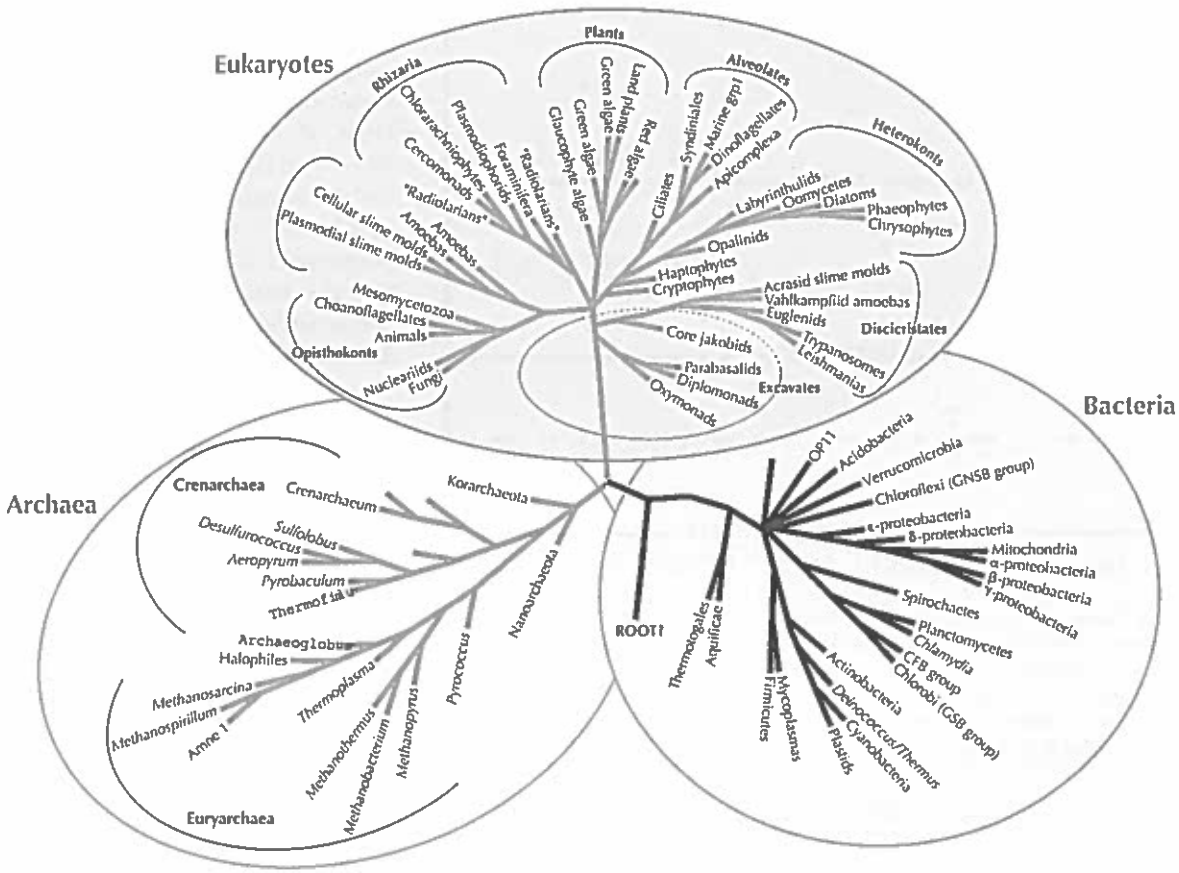

Figure 41.2. A textbook representation of the modern thinking about the tree of life. As can be seen, there are three great Domains – the Bacteria, the Archaea, and the Eucarya. This is based on the work of Carl Woese in the 1970s. Permission: Sandie Baldauf

no parental nodes, interior nodes that have exactly one parent and two (or sometimes more) offspring nodes, and leaf tips that are nodes that have one parent but no offspring nodes. Important features of trees for systematics is that between any two points there is a unique path on the tree and that each node (other than the root) has exactly one parent. To say that the tree of life is a tree is to say that it is a phylogenetic tree in this sense.

To say that the tree is universal implies that the tree should depict the relationships between all living things. Are viruses alive? Canonical representations of the tree typically do not mention viruses, but it is worth mentioning viruses in this context, because they place limitations on those wishing to defend the tree. One must either embrace the idea that the tree is not universal or deny that viruses are alive or have the kind of evolutionary history that the tree is supposed to be tracking. The typical assumption is that the tree must connect all species but that viruses (along with mobile genetic elements like transposons and plasmids) do not form species, at least in the way relevant for inclusion on the tree. But do all *organisms* form species in the relevant way? Many bacteriologists, as well as systematists and philosophers of systematics, deny that prokaryotes form species (Gevers et al. 2005; Ereshefsky 2010b; Lawrence and Retchless 2010). Worse, many of those who do accept that there is a good species concept that applies to prokaryotes will deny that these groups are phylogenetic groups and have branching histories.

It is sometimes thought that a simple change in how we describe the tree can solve this problem. The tree of life shows how *organisms* (or perhaps genomes instead) are genealogically related. But many organisms are not related to each other in a treelike hierarchy of descent. Rather, they form a reticulated network. This is even clearer for genomes where recombination is present. The defender of the tree needs to say something about how, at the appropriate level of description (perhaps when talking about populations or lineages or clades of organisms directly), these entities can form a tree. It is not clear how this can be done, and the burden of proof is surely on the defender of the tree here.

HYBRIDS

If we do manage to muddle through the species problem and say that the tree can connect all species, we then have the empirical question of just how treelike this evolutionary

history is. There is massive reticulation in the form of gene flow through hybridization and introgression between species. While some have attempted to minimize the problem, we now know that even in the best-behaved groups (plants and animals) hybrids regularly form. Mallet (2005) surveys a variety of studies on hybrids and concludes that at least 25 percent of plant and 10 percent of animal species form hybrids with other species in nature. This usually leads to introgression and therefore gene flow between species. This problem is far worse with populations at the tips (now any migration is reticulation) or any kind of lineages.

Of course, like Darwin, we can allow that some hybridization is consistent with the tree. But how much reticulation is it reasonable to allow? This is a difficult question and can be realistically answered only in a context where we know what the purpose of the tree is. If the tree is supposed be allow us to make inferences about genetic history, similarity, biogeography, and other factors, then it is okay if it sometimes leads to errors – any possible model will do that – but it must have a good balance of simplicity, explanatory power, predictive power, and perhaps other less easily describable virtues. If systematists were aided in their research by using the tree, that would count in its favor. If they were positively misled, that would count against its use. Exactly how these have to be balanced against each other is a perennial question in the philosophy of science and one that is unlikely to have a general answer; rather, it needs to be examined carefully in the particular case at hand.

LATERAL GENE TRANSFER

The problem of reticulation might plausibly be thought to be manageable in eukaryotes, but when we generalize to all forms of reticulation, we face what is arguably the most serious problem for the tree: the phenomenon of lateral gene transfer. Lateral gene transfer (LGT), also called horizontal gene transfer, is the name for any instance of a variety of processes where genetic material moves from one organism to another by some process other than reproduction. This includes transformation, transduction, and conjugation.

It is now widely agreed that LGT has been, and still is, a major force in evolutionary history (Gogarten, Doolittle, and Lawrence. 2002; Dagan, Artzy-Randrup, and Martin 2008). The epistemological question of what can be inferred about genetic history is a serious one, given that genes do not in general track the same history and that, as we go deeper in time, any trace of signal may be lost. But the metaphysical question is serious as well – what could the tree be tracking, since clearly the history of *all* genes is not a single tree. It is not clear exactly what this means for the tree because different proposals about what the tree is will be affected differently. (For arguments that widespread LGT undermines the tree concept and possibly traditional phylogenetics as a whole, see Bapteste et al. 2004; Bapteste et al. 2005; and Bapteste and Boucher 2008.) Before looking at different responses to lateral transfer, we first consider more potential problems for the tree.

ENDOSYMBIOSIS

Another source of problems for the tree is endosymbiosis. In endosymbiosis, one organism comes to live inside another, and eventually its descendants become obligate symbiotes. Over evolutionary time, they reach the point where they are so tightly interconnected, often because of extensive LGT between host and symbiote, that it is appropriate to think of the host plus symbiote as one integrated organism. For example, most eukaryotic cells contain many mitochondria in the cytoplasm surrounding the nucleus of the cell. Mitochrondria are clearly functional parts of the cells today and are not organisms in their own right. But mitochondria have their own genomes, and it is now clear that historically they are closely related to various groups of the alpha proteobacteria. Likewise, the chloroplasts that give plants and other organisms such as some algae the ability to photosynthesize were once free-living cyanobacteria. A natural way to depict these genealogical relationships is with a fusion of lineages of very distant branches on the tree as in Figure 41.2. Endosymbiotic events have occurred a number of times in the history of life (Lane and Archibald 2008), but even with a strict understanding of "new lineage," if in some respects these events might be rare, they could hardly be more important. If there is any sense to be made of "key" events in evolutionary history, the origins of mitochondria and of chloroplasts surely count. Any purported universal phylogeny that fails to represent these events is lacking in a very important respect.

THE ROOT OF THE TREE

A major feature of the tree is its root. The root it typically thought to represent LUCA: the last universal common ancestor. Understanding the root is essential for studying the evolution of various ancient biological features, such as the genetic code, protein synthesis, cellular membranes, and, indeed, the cell itself.

As with the tree, different authors have a different conception of what would count as a LUCA, and different conceptions lead to different conclusions about its existence. In phylogenetics with trees, it is assumed that each descendant node gets its traits through common descent with modification. Thus allowing for mutational or other changes, the genes present in organisms today would have to have their ancestors in LUCA. But if LUCA is a single organism with a single genome, this leads to the absurd conclusion that LUCA contained genes for nearly all types of biochemical reactions known in bacteria and archaea today and had a genome larger than any known prokaryotic genome today. This is what Doolittle et al. (2003) termed "the genome of Eden." Such an entity surely never existed.

Different genes have genealogical histories that coalesce in the past at vastly different times. In describing his view of early life, Woese (1998, 6858) says, "The universal ancestor is not an entity, not a thing. It is a process characteristic of a particular evolutionary stage," from which he believes multiple

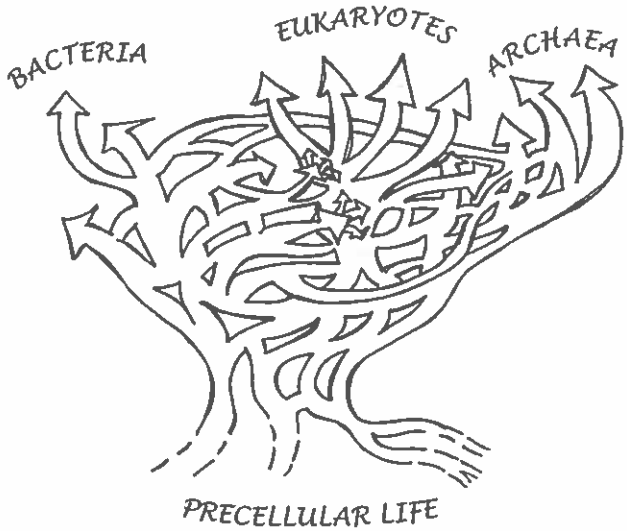

FIGURE 41.3. Does the existence of lateral gene transfer, where genetic information can hop from one branch directly to another, make the traditional tree of life otiose? Was Buffon on the right track after all? W. Ford Doolittle sees more of a net in life's history than a conventional tree. Permission: W. F. Doolittle

communities and independent lineages emerged. Theobald (2010, 220) argues for universal common ancestry and the existence of LUCA but clarifies what he means, saying, "Rather, the last universal common ancestor may have comprised a population of organisms with different genotypes that lived in different places at different times." While Doolittle and others do not consider this "population" dispersed in space and time worthy of being called an "ancestor" of anything, even granting that we should call it LUCA, it is clear that it would be inappropriate to depict this as a single node (the root) on the tree of life. If we attempted to use this tree as we would any other phylogenetic tree (say the tree of primates), we would be led to make mistaken inferences about evolutionary history. As the critics would say, a tree without a root is no tree at all.

If we attempt to represent lateral gene transfer, endosymbiosis, and the base of the tree all on the same diagram, even a very conservative picture will look something like Figure 41.3, which has been drawn by W. Ford Doolittle (see also Doolittle 2000). This is certainly not a phylogenetic tree, but whether that is an essential feature of anything appropriately called the tree of life is not clear.

SAVING THE TREE?

Given the multitude of problems, clearly defenders of the tree must deny that the tree represents the history of species, is universal, and is fully branching with no reticulations at all. But given that is not the case, it is no longer exactly clear what the tree is. A first pass might be as simple as saying that the tree is an idealization that is not perfect but still gets it mostly right and is extremely useful. This view makes the existence of the tree dependent on things like the extent of actual reticulation. I would guess that this is the most common view of practicing biologists – especially those who work on eukaryotes. But the problems are serious, and the extent of reticulation, especially in prokaryotes, far too great to simply idealize away.

But what of those who work explicitly on reconstructing the tree to include prokaryotes? What exactly are they reconstructing? Not a phylogenetic tree that just idealizes out a few reticulations. Here, different, incompatible views of the tree have been proposed.

One idea might be that, as a practical matter, we simply need some kind of reference tree on which to base our classifications and to locate clades so that we can make sense of such things as lateral gene transfer in the first place. The 16S SSU rRNA gene is often used this way. We know that many gene histories disagree. Why not just pick one for a reference tree? For a great many taxa, we have sequenced the 16S gene, and so it has become the default classification tool for prokaryotes. To say that something is an alpha-proteobacteria or a haloarchaen is just to say that its RNA sequence fits in at a certain place in the universal 16S tree. While this may be practical for classification, it provides no defense of anything like the Darwinian hypothesis that there is a unique tree. Deciding which tree is *the* tree cannot depend on conventional choices by us. A plausible tree could be one of two things – a tree that is as reliable for phylogenetic inferences as possible or a tree that represents the actual genealogical history of some kind or other but not the full genetic history.

Galtier and Daubin (2008) explicitly stick to the idea that the tree is a tree of species. On their view, LGT is not obviously a problem metaphysically, though it would be if it meant there were no species (Lawrence 2002). Similarly, if we are building a tree of organisms, it might seem that LGT is only an epistemological issue. Organisms come from other organisms. This history is often referred to as "the tree of cells" because it tracks the cellular history and not necessarily the history of the genes inside the cells. Cicarrelli et al. (2006) claim that the tree is a tree of organisms. Given this, it is easy to see why they simply remove from their data set genes that they have reason to believe have been transferred. If they do not represent the organism's genealogy, why include these misleading data?

But even if there is some genuine tree of species or a tree of cells, this leaves open just how useful it is to reconstruct it and raises a serious question as to what extent the tree could play in the foundational role it is sometimes claimed to play – which may be relevant to whether this object is properly called the tree of life. For example, as Galtier and Daubin (2008) themselves point out, their species tree may not be consistent with *any* single gene's history. Cicarrelli et al. (2006) are criticized by Dagan and Martin (2006) for producing a "tree of one percent" because their tree is based on only thirty-one genes, which are consistent with at most 1 percent of the typical prokaryotic genome of more than three thousand genes.

In order to preserve the idea that the tree is supposed to represent something like a dominant pattern, Koonin, Wolf, and Puigbo (2009) examine whether there is a "statistically

significant trend" in the forest of life that represents all genetic history. They argue that there is and that this could plausibly be called a tree of life. Wu et al. (2009) construct a genome tree based on a concatenation of all the gene data they have and compare this to the known rRNA tree. Although they do not explicitly present it this way, one could reasonably say that the tree of life is this genome tree, which represents something like an average signal that may not be the actual signal of any particular gene.

If we wanted the tree to play the role of representing the history of all genes or genomes, then it is now clear that no such tree can play that role. Some have proposed alternate names and conceptualizations to play the "represent everything" role such as Koonin et al.'s "forest of life." Other suggestions include the "ring of life" (Rivera and Lake 2004), the "net of life" (Kunin et al. 2005), or a "web," "coral," or "potato of life" (Olendzenski and Gogarten 2009). If they are right, then it seems that the Darwin's hypothesis was wrong.

THE FUTURE OF THE TREE

So should we continue to talk of the tree of life and attempt to reconstruct it? There are two kinds of factors relevant to this question. One set of factors relies on the empirical facts. Just how common is the tree of life in LGT, and what kinds of patterns does it produce? What was the early evolution of life like? Was there some single universal common ancestor? What is the extent of hybridization and introgression between sexual species? The answers to each of these questions can tell us something about the utility of talk of the tree and the importance of reconstructing it.

But another equally important set of factors concerns questions about what the tree is supposed to represent, about how we do and ought to use the tree in biological inferences. Does defending the tree entail defending a particular history of life and perhaps even a particular view about what counts as life? Or does it mean defending a particular set of practices? Or perhaps defending the explanatory power and heuristic uses of a particular model? Depending on how these questions are answered, empirical facts like the extent and pattern of LGT may or may not dictate abandoning the tree.

Critics of the tree are certainly right that the tree has been used for many different and sometimes inconsistent purposes. Some practices, such as assuming that one gene will have the same broad-scale genealogy of another, are bad practices. A weak reading of the pluralism defended by pattern pluralists such as Doolittle and Bapteste (2007) is surely right: just as there are a multitude of evolutionary processes besides natural selection, there are a multitude of genealogical patterns besides the single tree pattern. But this is consistent with the tree being one of those patterns – and perhaps a very important one at that. But the critics of the tree want to claim something stronger – that the history and usage of the phrase the "tree of life" dictate that it is the unique pattern or at least a special kind of universal pattern. Further, any specific way of understanding the tree, such as the tree of cells, simply fails to have the power to play the role that the tree was supposed to play.

Is this stronger view correct? Phylogenetic trees really are of central importance in a variety of contexts; trees will continue to be built, and the phrase "tree of life" has a special kind of importance (appearing in both the Bible and Darwin). But it is now clear that different understandings of what the tree is supposed to be and how it can be used come apart, and so the phrase must be used more carefully and only in restricted contexts. What about the prospects for "universal" tree building? Research programs investigating questions about the origin of life, the genetic code, the cell, the eukaryotic cell, and the connections between apparently very disparate forms of life will continue. It is clear that this research will involve phylogenetic trees as well as patterns other than trees and processes other than vertical descent. What this research will uncover is unknown, but we can be certain that it will be a fascinating story of the deep evolutionary connections between all humans, the *Escherichia coli* in our guts, the archaea living in hydrothermal vents deep in the ocean, the roses in our gardens, and the penguins in the Antarctic. At least in this respect, we can surely claim that Darwin was right.