Horizontal Gene Transfer and the Emergence of Darwinian Evolution

Thomas C. Butler

May 8, 2006

Abstract

In this article I discuss the broader view of evolution that is growing out of our increased understanding of the role played by Horizontal Gene Transfer (HGT) in early evolution. Darwinian evolution, where genetic information and variation are passed to an organism by its ancestors, is seen to be only one aspect of evolution, and to have emerged from a world where the primary mode of evolution was the horizontal transfer of genetic information between organisms.

1 Introduction

In recent decades we have seen enormous growth in our understanding of evolution. Much of this has been driven by new experimental methods that have allowed us to begin to uncover a detailed history of life on earth. However, the picture of early life that these methods are uncovering differs in key aspects from the expectations of many evolutionary biologists [1]. The result is that, just as we are uncovering a detailed history of life on earth, that very history is forcing us to reconsider core concepts in the theory of evolution. Many key ideas in traditional Darwinian evolution (to be clarified below), can now be seen as phenomena emerging from the evolutionary process itself. Even the standard explanations for the universality of biochemistry and the genetic code in terms of a last common ancestor, and the whole of life emerging from well defined ancestral organisms are being reinterpreted [1].

As a result of these investigations, evolution in the Darwinian sense is now beginning to be considered simply one aspect of a larger evolutionary framework. In the framework of traditional evolutionary biology, common traits between two organisms are considered to be inherited from a common ancestor [4]. However, tracing the history of life has led to the inevitable conclusion that under certain circumstances, an organism's genetic information is not shared exclusively with offspring (vertical gene transfer, VGT) as expected in classical Darwinian evolution, but can be horizontally transferred (horizontal gene transfer, HGT) to another living organism. This has been inferred from traits arising in organisms that have clearly been borrowed from other organisms that are not their ancestors [2, 6]. Lending further support to these findings are direct observations of such events today. For example, HGT enabled evolution is now understood to be a primary mechanism for the evolution of antibiotic resistant strains of bacteria. The development of antibiotic resistance is often observed to develop much more rapidly than simple vertical inheritance of traits would suggest. This has been shown to be due to antibiotic resistant bacteria sharing the genetic material required for resistance with non-resistant bacteria [7, 5].

Particularly disruptive to traditional views of evolution has been the discovery of pervasive HGT events in early evolution [1, 9]. The picture now forming is that in the early history of life, communities of organisms shared innovations freely, their evolution dominated by HGT. They grew more complex over time and slowly diverged from each other as differences between organisms arose that inhibited HGT [1]. This differs radically from the tra-

ditional view of early evolution with a single Last Common Ancestor (LCA) species dividing into new species. Integrating these findings into evolutionary biology is requiring the introduction of new concepts to understand and classify early life and its evolution[4, 9]. In turn, this new understanding of early evolution is enriching our understanding of all evolutionary processes [1, 3, 5, 9].

2 The Character and Findings of Darwinian Evolution

Traditional Darwinian evolution is based on a small set of key ideas, presented here in the condensed form discussed by Ernst Mayr in *What Evolution Is* [4]:

1. All species evolve gradually through occasional mutations

2. All organisms on earth have common ancestors (the Doctrine of Common Descent)

3. Natural selection determines which mutations will survive

4. The gradual change in the properties of organisms leads to new kinds of organisms (speciation)

Taking these statements as the core of evolutionary theory, much of the task of understanding evolution becomes to discover and interpret the history of life through the lense of this Darwinian synthesis. In this synthesis, we can represent the history of life as a tree (known as a phylogenetic tree), where a line represents a species, and a branch in the tree represents the divergence of one species into two (or more), and a branch of a line coming off another branch indicates a later speciation (in fig. 1 this means that Eucarya and Archaea branched later than Bacteria). Thus how closely related two species are is determined not by how close they are in space on the tree diagram, but by how close the nearest shared branch point is [8]. In this representation, the task of assembling the history of life is represented by the task of drawing and labeling the main branches and roots of the tree of life. Much of this task has been accomplished, in particular the structure of the early history

of life has been greatly clarified by the monumental work of Woese and Fox in the 70's (to be discussed below) [11].

At risk of oversimplification, most of the methods used by Woese, Fox, and others for determining the evolutionary history (phylogeny) of life on earth boil down to one basic idea: common characteristics indicate common ancestors. While there are many subtleties involved in such an analysis, such as similar properties in two organisms evolving independently (convergent evolution), the general method can be applied in many powerful ways when used with care [4, 8]. In essence, we assume that genetic material is passed exclusively from an organism to its offspring. Therefore, a characteristic showing up in two different organisms indicates that in the past, they had a common ancestor [4]. This is in line with traditional Darwinian evolution, as it contains the theory of common descent nas an explicit assumption.

The most reliable way put this program in practice is to compare protein and/or rRNA (a molecule used in the production of proteins) sequence information between organisms [1]. Based on these methods, Carl Woese and collaborators used rRNA information to infer the basic structure of the tree of life. Their analysis indicated that life branched from one root, spreading into three major branches called domains: Bacteria, Archaea, Eucarya (see fig. 1). Bacteria and Archaea are what would traditionally be called prokaryotes because they do not contain a cell nucleus, but the Archaea differ in important ways from Bacteria in that the machinery that takes genetic information and translates it into proteins needed by the cell are much more like Eucarya [11].

Woese and his collaborator's discovery of the structure of the early tree of life is rightly considered one of the great triumphs in evolutionary biology. Taken at face value, it has a number of interesting implications. First, it lends support to the principle of common descent. This explains the existence of a universal genetic code and the universality of biochemistry by showing that these were the biochemistry and genetic code of the Last Common Ancestor (LCA) at the base of the tree [4, 3]. It also clarifies the relationships between different kinds of organisms in explicitly evolutionary terms. The tree of life in fig. 1 is the completion of the search for the tree of life in the framework of classical Darwinian evolution.

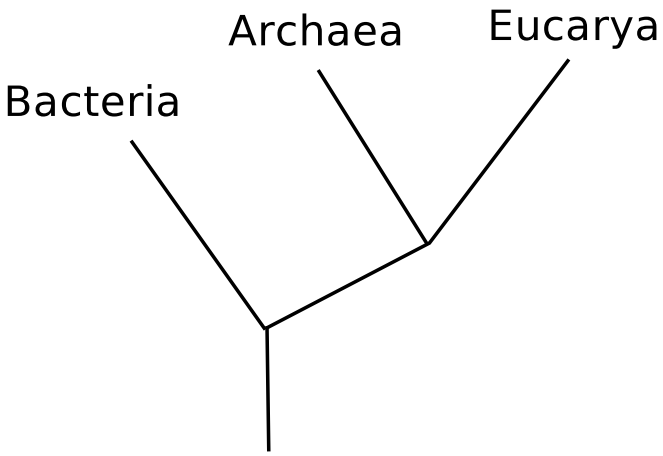

Figure 1: The base of the tree of life under the assumption of the Doctrine of Common Descent, showing the three evolutionarily distinct domains. This graph, an example of a phylogenetic tree, should be read from bottom to top, with the vertical axis indicating time sequencing of branches. Thus Bacteria branched off before Archaea and Eucarya, and Archaea and Eucarya are more closely related to each other than Bacteria.

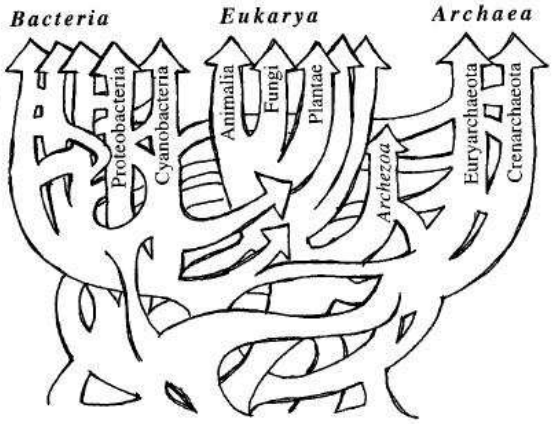

Figure 2: An attempt to include the effects of known HGT events into the base of the tree of life. It is clear from this picture that there needs to be some new thinking about evolution and what a history of life means. Figure taken from [9].

3 Horizontal Gene Transfer and the Tree of Life

The picture presented above is very appealing, but, as was pointed out by Woese himself, is incomplete [3]. Much of the problem is that sequencing work continually uncovers genes where they shouldn't be. This means that in disconnected areas of the tree of life similar genes are found, indicating that the genes have been shared horizontally. In many cases this is so pervasive that entire alternative phylogenies can be constructed [9] for an organism. This means that lots of genetic information has been transferred horizontally across the tree, through HGT, as briefly discussed above [9, 2]. As we go back toward the root of the tree of life, this becomes increasingly true, though the scope is still hotly debated [6, 2]. What is clear is that HGT is crucially important for very early evolution, and that this severely disrupts the pure representation of the history of life as a tree (see fig. 2). Abandoning the tree representation of early life also forces us to reconsider the concept of species and to decouple the evolutionary history of an organism from its genealogy [5].

The emerging picture is that near the base of the tree, HGT was the dominant mode of genetic information sharing [2, 3]. As life emerged, each organism resembled all other organisms closely. The machinery that made up an organism was simple and modular. Such modularity made sharing information quite easy. Innovations spread quickly and easily throughout the system, giving rise to very rapid evolution (remember the rapid development of antibiotic resistance discussed in the introduction). Species, in this context, is meaningless. Evolution happened over the entire community of organisms, and each organism took part in essentially all of the innovations, at least when averaged over time [2, 1].

This picture reinterprets several of the conclusions drawn from the construction of the Tree of Life above. No longer is the universality of biochemistry or the genetic code evidence of a single Last Common Ancestor. Rather, it is evidence of pervasive HGT in early life [1]. The Doctrine of Common Descent collapses together with the simple tree representation of the history of life. We are forced (see fig. 2) to reevaluate how separate the branches of the universal tree are. In what meaningful way are two organisms part of separate species, or even separate domains, when they successfully transfer large amounts of genetic information to each other?

Some possible ways of thinking about these issues are available, but first we need to understand more about HGT. HGT works best under two complimentary circumstances: First, when it addresses simple, modular structures that can be altered without affecting large amounts of sensitive equipment elsewhere in the organism, Second, when the donor and recipient are very similar. Note these two circumstances are really different manifestations of the same idea: HGT works well when the structures affected by the transfer aren't destroyed by the transfer. Horizontal gene transfer works fine, even among very complex organisms if the two organisms are very similar, so that nothing too drastic is done to the sensitive biological machinery of the organism. In fact, Carl Woese has characterized sexual reproduction as an example of modern HGT [1]. To put it another way, if the donor and recipient do not contain fundamental incompatibilities related to the genes being transferred, then HGT is a possibility. If there are fundamental incompatibilities between the organisms tied to the genes being transferred, then the HGT will simply kill the recipient, or the donated genes will be destroyed. Note this also means that organisms can be partially compatible for HGT. That is, some HGT events between the organisms would be fatal, and others would lead to a new type of organism that survives and continues

evolving. One can say that two organisms are highly HGT compatible if most genes can be swapped between them without disrupting the generation of fertile offspring. Of course such a correspondence is an oversimplification. For example: organism A may be very receptive to genes from organism B, but not vice versa, or organism A can receive only a certain subset of genes from organism B [1, 3].

These observations about the prevalence and nature of HGT are the starting point for a reasonable approach to defining a meaningful classification for organisms (an analogue of species), as well as an understanding of the evolutionary dynamics at the roots of modern life. From the above analysis, it is clear that the root of the tree of life doesn't indicate a single common ancestor, but some sort of Last Common Community (LCC) [5]. This LCC had massive HGT events as a primary characteristic, but as it became more complex (i.e. more interconnected machinery within the organism), two substantially HGT-incompatible translation mechanisms began to emerge [3, 11]. This is the origin of Bacteria. This does not mean that the new regions could not engage in all sorts of HGT with each other, but that there were now two pools that couldn't share most innovations regarding the translation apparatus. Thus the two new pools were still deeply interconnected, but in some sense, each of them looked just like the original pool, in that HGT was widely possible within the pool, just as in the last common community. So instead of HGT happening widely throughout the community of all organisms, it was at that time happening primarily within two separate pools. A further, identical division within the non-Bacterial pool led to the origin of the distinction between Archaea and Eucarya, and divisions within these pools led further to species within these great divisions. These divisions are not qualitatively different (there is an interesting sort of scale invariance in this proliferation of bubbles). They only differ in that as the life forms within the pool become more complex, vertical gene transfer, which has been present all along, becomes more and more important. In fact, in this view, species is simply a category defining the boundaries of a community of organisms in which the organisms contained have high levels of HGT (or simply genetic) compatibility with each other. From this perspective, perhaps species as a well defined concept is itself emergent. It seems to emerge with growing complexity as the organisms develop such complex and interconnected internal machinery that only genetic material from nearly identical organisms can be shared, e.g., through sexual reproduction. A species is simply another bubble, like those in fig. 3, where the community has very limited overlap with other

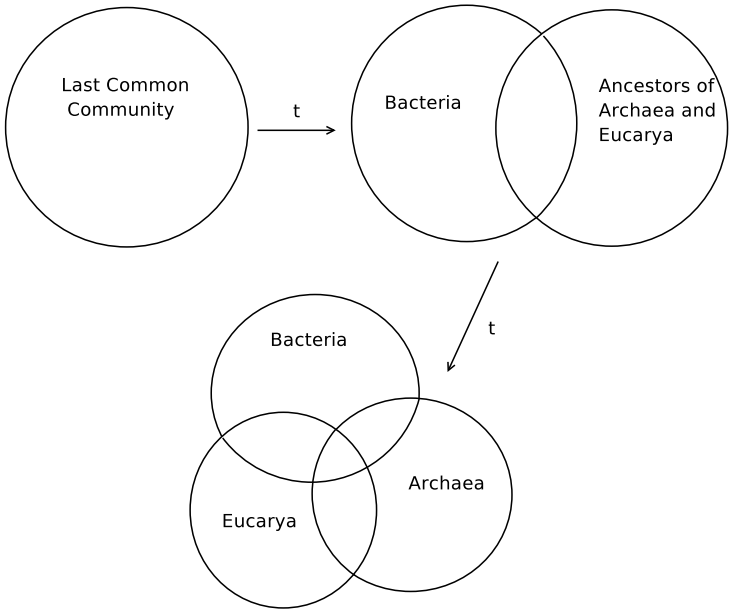

Figure 3: An attempt to represent diagrammatically the roots of the tree of life that were drawn in the conventional manner in fig. 1. The circles represent communities of organisms with high degrees of HGT compatibility. The overlap in the circles is representative of the fact that the communities represented by the area of the circles are not completely isolated from HGT events between different circles.

bubbles, and the bubble itself contains limited variation.

It is important to note that the role of HGT in the sense of direct sharing and acquisition of new genes is most important in studying early life. While the example of antibiotic resistance shows that it remains relevant today, especially in the context of the microbial world, HGT was not necessarily the primary mode of spreading new genetic innovations throughout much of the history of life [2]. Rather, the conventional Darwinian method of inheritance and selection seem to dominate. In particular, it is easy to mistakenly identify HGT events using bioinformatics data bases, because they may not have data on enough organisms to show that there exists a traditional Darwinian path between two organisms that are being examined for possible HGT events between them [6]. The complexity that increased so rapidly in early life due to HGT, caused interconnectivity in intra-organism mechanics

to skyrocket. As noted above, each level of interconnectedness inhibits HGT further, and even though an organism may be able to accept genes from near neighbors, so many of the world's genes are hostile that selection will favor organisms that develop ways of inhibiting HGT. Thus HGT dwindles as complexity increases, unless sexual reproduction, or some other method for tightly controlled HGT is developed. This change, with increasing complexity, from HGT to Darwinian evolution as the primary mode of evolution is often called the Darwinian transition [1, 3, 5].

4 Conclusions and Final Questions

It is clear from the data discussed above that to understand the early history of life, we need to broaden our concept of evolution. In this paper, some of the efforts people have made to do so have been summarized. Tree thinking, while it can be tweaked, does not seem to capture the full breadth of early, and even modern evolution [9, 1]. More useful are concepts that accept the dynamic interplay between organismal groups that were once thought to be distinct. These concepts are beginning to emerge, drawing us into a realization that the framework of evolution is broader than that encapsulated by the traditional Darwinian consensus. What this traditional version of evolution seems to be is a phase of evolution that dominates in the context of multicellular life and complex single celled life. HGT evolution is another phase, with the difference being analogous to a phase transition in statistical mechanics. Evolution is the essence of life [10], but this essence manifests itself in ways that are as different as liquid and solid water.

But the discussion above is painfully imprecise. To really understand the roots of the tree of life, and to understand the deeper properties of evolution, including its phase diagram, a great deal more systematic study is going to be required. Much of the terminology and conceptual discussion above is at best heuristic, and needs to be carefully formulated in terms that make new predictions. More detailed study of the extent of HGT, ancient and modern needs to be done, and its relationship with classical evolution needs to be explored. The transition, with increasing complexity, to Darwinian evolution from HGT suggests computer modeling, and continuing careful analysis of experimental data with an eye toward the phase diagram of evolution will continue to expand and modify our understanding of the structure of evolution [1].

References

[1] Carl R. Woese, 2004, *Microbiology and Molecular Biology Reviews* **68**, 173.

[2] James R. Brown, 2003, *Nature Reviews: Genetics* **4**,121. (Springer, New York, 1999).

[3] Carl R. Woese, 2002, *Proceedings of the National Academy of Science* **99**, 8742.

[4] Ernest Mayr, 2001, *What Evolution Is* (Basic Books, New York).

[5] Carl R. Woese, 2000, *Proceedings of the National Academy of Science* **97**, .

[6] C. G. Kurland, B. Canback, Otto G. Berg, 2003, *Proceedings of the National Academy of Science* **100** 9658.

[7] Maxim D. Frank-Kamenetskii, 1997 *Unraveling DNA: The Most Important Molecule of Life* (Perseus Books, Reading, Massachusetts).

[8] David A. Baum, Stacy Dewitt Smith, Samuel S.S. Donovan, 2005 *Science* **310**, 979.

[9] W. Ford Doolittle, 1999, *Science* **284**, 2124.

[10] Carl R. Woese, 2006, Personal communication.

[11] Carl R Woese, George Fox, 1977, *Proceedings of the National Academy of Science*, **74**, 5088.