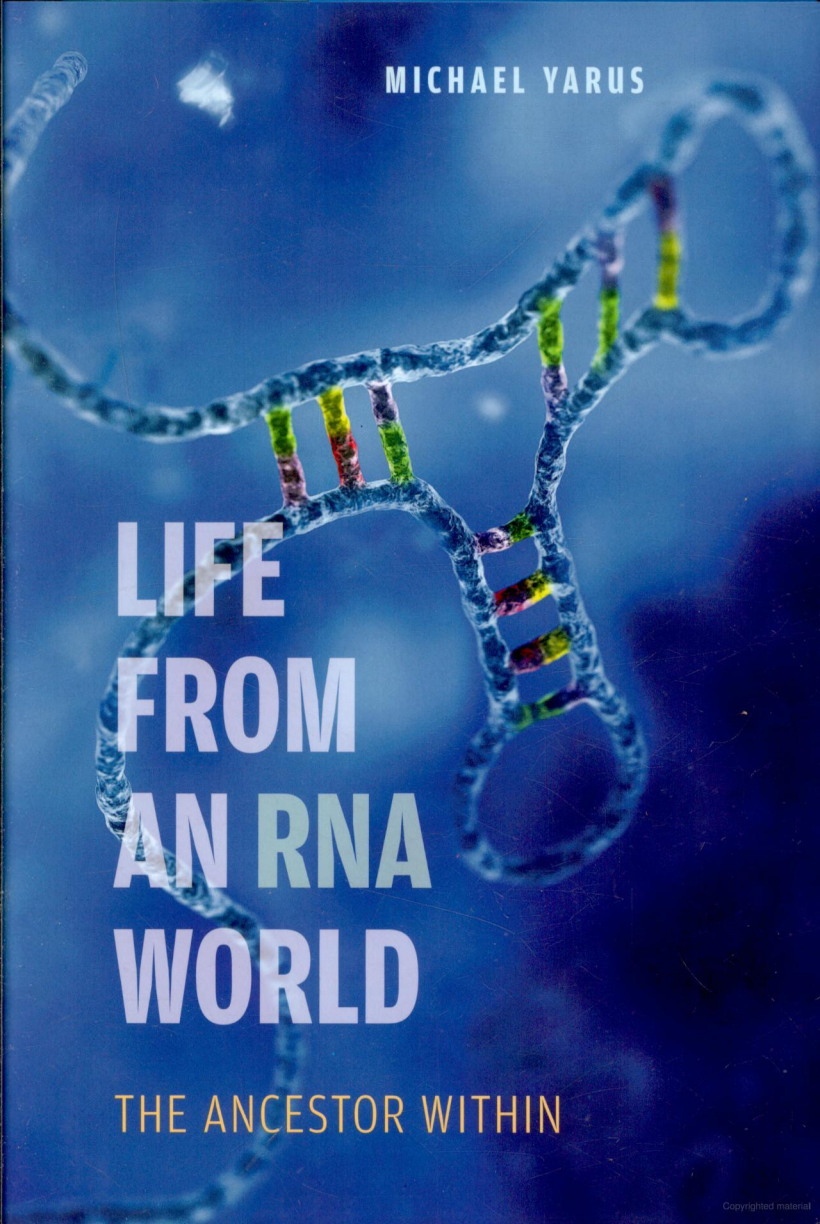


Life from an RNA World

THE ANCESTOR WITHIN

MICHAEL YARUS

HARVARD UNIVERSITY PRESS
Cambridge, Massachusetts
London, England
2010

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Yarus, Michael, 1940–
Life from an RNA world : the ancestor within / Michael Yarus.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-674-05075-4 (alk. paper)
1. RNA. 2. Molecular biology. I. Title.
QP623.Y367 2010
572.8′8—dc22 2009044011

Contents

Introduction to Your Ancestor vii

1 Before We Begin: A Voluntary Chapter 1

2 Framing the Problem: The Buffalo and the Bacterium 5

3 The Big Tree: No Jackalopes Please 13

4 A Dance of Atoms 30

5 *Allegro Agitato:* The Origin of Life 36

6 The Winds That Blow through the Starry Ways 46

7 Tornados in a Junkyard 59

8 Between Genomes and Creatures 79

9 A Thumbnail Molecular Biology 84

10 RNA Structure: A Tape with a Shape 99

11 Intimations of an RNA World 109

12 The Experimentally Impaired Sciences 119

13 Test Tube RNA Evolution: First Light 126

14 Selection Amplification: Interrogating RNA's Possibilities 132

15 RNA Duplication: Replicase Activity in Real RNAs 141

16 RNA Capabilities and the Origins of Translation 149

17 The Quest for the Peptidyl Transferase 155

18 A Language Much Older Than Hieroglyphics: The Genetic Code 161

19 Assume a Spherical Cow: The Ribocyte 176

20 The Future of the RNA World 183

Lexicon 187

Index 195

Introduction to Your Ancestor

Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed.

—Charles Darwin, *On the Origin of Species* (1859)

Remove literary, grammatical and syntactical inhibition.

—Jack Kerouac, "Belief and Technique for Modern Prose"

Greetings to you, who look with good-spirited curiosity at this congenial anachronism. It is unclear how much longer people will write on dried and flattened wood. Trees do so much for humans and for our planet that it hardly seems fair to ask them to carry our thoughts as well. But fair or not, archaic or not, it still appears oddly plausible to make a book in order to collect a substantial number of thoughts in one container. Having made one, I am thankful that it yet seems plausible to you to look at the result. Therefore, a heartfelt welcome. Here's hoping that wooden books survive the ascent of digital text.

This particular book was provoked by an advance in biological thinking about life on Earth. Its subject is a small slice of the action in Darwin's breathtaking summary above. A

majority of evolutionary biologists believe that we now can envision our biological predecessors on this planet, though we have never seen them. *Life from an RNA World* is about these vanished old ones, sketching them at a long-ago time just as their workings began to resemble closely our own. What was the difference between our early relatives and their later offspring—us?

Sketching such a portrait takes some effort, for we will end up in quest of nearly the first among the living beings of our planet. Nearly, but not quite the first. So we must see our quarry across the billions of years between their heyday and ours. The effort that brings them into focus has been widely talked about within the field of biology—but surprisingly little of the story is known by those outside the laboratories who might also be interested. So here is an album to introduce interested non-biologists to our relatives in deep time—slouching along between the origin of the first rudimentary life on Earth and the appearance of more complex beings—who had nearly mastered the intricate informational handicrafts that make modern cells. This era between is called the RNA world.

It is endlessly interesting to inquire into our ancestors and to try to guess in what sense that Ukrainian or Lebanese great grandparent left us his or her gifts. But while this book is concerned with our personal genealogies, it is also about something deeper. If you step back far enough, your genealogy merges into the history of life on Earth—indeed, the only life we know. And within the RNA world lies the solution to a major mystery about the path life has taken on Earth.

To initiate that billion-year view, we need to excavate for some foundations. In particular we need some notions of what evolution is, and of what life itself is, so we can coherently speak of its youth. An image of the small, whirling galaxy of

our genes will help us to appreciate the extent to which we are still the children of our great . . . great grand-ancestors. We will spend a lot of time, then, tracing the shadows of an earlier RNA world within our own cells. Such preoccupation is appropriate because we now flourish by wielding ancient, borrowed genetic recipes. Indeed it seems, surprisingly, that here in the early days of the twenty-first century we are still at the beginning of appreciating our common ancient RNA patrimony.

I am a professional biologist, doing research and teaching in areas that sometimes overlap the subject of this book. This means that later on in the book, when there is not much to go on, you will hear my opinions. I will try to state this sufficiently plainly to make these sections clearly distinguishable from those setting forth concepts that rest on concrete and wide-ranging support and consensus.

Many thanks to those who read drafts of the manuscript and made suggestions—John Abelson, Richard Byyny, Tom Cech, Nataliya Chumachenko, Shelley Copley, James Dahlberg, Matt DeYarus, Larry Gold, Teresa Janas, Leslie Leinwand, Elsebet Lund, Irene Majerfeld, Bill McClain, Peter "ribosaur" Moore, Norm Pace, Alyson Yarus—and to the indefatigable students of MCDB 4100, "The RNA World." Particular thanks to all those who made known their views about good and bad words, and to the artist who helped me picture these thoughts, Greg Kuebler. Many thanks also to the Graduate School of the University of Colorado, whose Council on Creative Work gave me a year free of other responsibilities to work on this book.

I also explicitly acknowledge those in my Boulder laboratory who actually did the work summarized in Chapter 16 on

the reactions of translation. I do this because I wish to credit them, and also to emphasize that experiments not occur without the labors of real people, who often give years to an investigational campaign. They are as follows: *amino acid activation*—Dr. Krishna Kumar; *aminoacyl-RNA synthesis*—Dr. Nataliya Chumachenko, Dr. Mali Illangasekare, Dr. Oleg Kovalchuke, and Rebecca Turk; *peptidyl transferase reaction*—Dr. Mark Welch; *amino acid binding to RNA and genetic coding*—Dr. Shankar Changayil, Dr. Greg Connell, Dr. Mali Illangasekare, Dr. Michal Legiewicz, Dr. Cathy Lozupone, Dr. Irene Majerfeld, and Dr. Shawn Zinnen.

The book you hold is much better for these many good-spirited contributions from others. The text gathers ideas from many sources, and I am certain that I have not explicitly remembered everyone who expressed an essential thought. Accordingly, there will be those whose work changed my mind and illuminated my path, but whose exertions will not be expressly described. To these numerous unnamed thinkers and experimentalists, my profound thanks and sincere apologies, humbly offered. Nevertheless, bookmaking requires that the opinionated, crotchety author reject and accept ideas as he goes. All remaining errors within these pages are mine—sometimes achieved despite the best advice imaginable.

a good general molecular biology textbook will be a help. You might keep one, like *Molecular Biology of the Cell* or *Molecular Biology of the Gene,* close by. Remember also the lexicon at the end of this book, which attempts to outline a basic RNA vocabulary for the reader.

But keep all this in perspective: if you encounter something incomprehensible while reading, pass it by in utter serenity. Bear in mind that experts say many indecipherable things. It is their job. Making sense of things in your own terms is the only useful goal. Because some references are books, feel free to read only those parts relevant to your current questions and interests. Your reward, should you care to claim it, will be a conversational acquaintance with this kind of science—at least as conversational as the formalities of a written article allow. A scientific journal article will be more spontaneous than any textbook, but less vivid than conversation with the person who wrote it. However, I hope none of this appears to be Required Reading: I have tried to make this book reasonably comprehensible without external expeditions.

Then again, if you would prefer to encounter the RNA world in surroundings replete with a professional outlook toward scientific literature and all its trappings, you can and should read *The RNA World.* This is the leading professional anthology in the field—and I have no financial interest in it.

Readings

Molecular Biology of the Cell, Fourth Edition. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Garland, New York (2002).
This useful book is freely accessible online at http://www.ncbi.nlm.nih.gov/books/.

The RNA World, Third Edition. Raymond Gesteland, Thomas Cech, and John Atkins, eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (2006).
A collection of authoritative essays by workers in the areas tangent to the RNA world, written for professionals who want a glance into the topic. The first (1993) and second (1999) editions are also available, and they lack only the recent updates for most articles. A fourth edition is on the way in 2010.

Molecular Biology of the Gene, Fifth Edition. James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine, and Richard Losick. Benjamin Cummings, Upper Saddle River, N.J. (2004).
The latest edition of a classic text; earlier editions may serve almost as well as a first resort on basic questions of molecular biology.

Framing the Problem: The Buffalo and the Bacterium

Go on, I tell you. You have the stomach for it!

—Franz Liszt, to Edvard Grieg

Life on Earth immediately presents us with the striking constancy of individual descent. Wombats, with few exceptions, give birth to exemplary wombats. Nothing could be more obvious. All the same, this humdrum reflection presents a vast impasse to thoughtful examination: how is the plentiful detail of every creature recorded and accurately replayed in its offspring?

Alongside this constancy of tiny details is the contradictory reality of pervasive genetic change across vast time. Where did all those varied creatures come from? A protobovine ancestor becomes both water buffalo and miniature Holsteins in the long run. Why a dingo *and* a fox? Wolves have become both Chihuahua and Shar-Pei, and this last divergence has happened almost within living human memory. How can each dizzyingly complex being be successfully altered by its residence in the world? When a hesitant Charles Darwin (at his ease in England) and a fever-wracked Alfred Russel Wallace

(during an illness in the Indonesian Moluccan Islands, but thinking on a peculiarly parallel track) put forward their ideas about the destiny of biological types in 1858–1859, what were they talking about? We need to agree about this now, because later on we will often talk about these ideas, and even more often assume them.

Recording and propagation first, then. We are now quite clear that each creature describes itself in an essentially linear digital recipe, which is broken into groups called chromosomes and stored inside the membrane-bound, microscopic, subcellular compartment called a cell nucleus. The text of the recipe is written in four characters, called nucleotides, and these are strung together to make long linear texts in the molecule DNA, which is coiled tightly within the chromosomes. The chromosomes consist of many linked individual recipes called genes, each describing one or a few slightly varied products. Closely linked groups of genes may be related entries, like the books shelved near each other in a library. Or, more often, neighboring genes in a chromosome may have no immediately obvious relation to each other, like books at a garage sale. These chromosomes taken altogether are called the genome, the collective name for the genes. There may be other forms of inheritance not embodied directly in DNA sequences, but those contributions to the genome are probably tiny in comparison to that written in chromosomes.

As a result, genetic texts written as chains of nucleotides—symbolized by the letters C, A, G, and T—are more monotonous than hexadecimal computer text, which can use 16 different characters for each position. And genes are much more monotonous than English text with its 26 letters. But with three billion genetic characters in the genome of mammals like

us, a lot of instructions can be encoded as strings of this four-character text (compare Chapter 8).

Genomic texts are a recipe for the creation of a creature. Genetic instructions for creature assembly can be roughly translated into words: "Make *this much* of *that* gene product and put it *there* at *this time*." Crucially, the idea of a digital creature recipe includes the idea of a change in one or more characters of the string—a mutation. That is, writing now in nucleotide text, "TAG A CAT" could mutate to become "GAG A CAT." Mutations like these change what gene products are made, how much of each is made, and where and when they are formed. They potentially yield a more or less functional organism. If, for example, you change a protein that binds oxygen in the blood, you may have a deleterious mutation, recognizable as the genetic disease called anemia. If you lose body hair, the change might be innocuous or even helpful—if you are an Olympic swimmer. Whether mutations are destructive, helpful, or somewhere in between is determined by the process called natural selection.

Darwin's answer to the second question, dealing with divergence into new creatures, dominates biology. It is still shocking —yes, shocking—to encounter his exceptionally simple explanation for the way inevitable mutations in the genetic text are sorted out, and therefore, ultimately, his explanation for the stunning variety that is life on the Earth. And in writing "exceptionally simple," I am not minimizing Darwin's accomplishment. Philosopher Daniel Dennett was not polishing his hyperbole when he wrote that Darwinian evolution is the greatest idea anyone has ever had—way ahead of relativity and the germ theory of disease, for example. There is no question that agreement about Darwin's meaning will be worth our time.

replication and repair. Not all such mutations have effects on the organism. Many genetic changes do not change either the time, the type, or the amount of the output of the genes. These are called neutral mutations—a major source of change in genomes, but not usually contributors to the variation that is acted on by natural selection. Neutral mutations accumulate (the changes are collectively called genetic drift), and they are important in any census of genome change (such as in studies of evolutionary trees or descent, as described in Chapter 3), but they are mere bystanders at the Darwinian concert, their presence as arbitrary as the hiss from a detuned radio. That arbitrary chemical and physical alteration in genomes is one of the significant differences between us and our ancestors may seem odd at first, but it is true.

Two other types of mutational sequence changes in the genetic text are selected. One type of selected mutation is the deleterious change. Deleterious changes are probably more numerous than potentially favorable changes, those that the adaptive Darwinian mill uses to make an organism better fit its world. Most likely an undirected change in a complex system will mean that the system will not work as well as the original, which had many parts selected precisely because they functioned well in their existing forms. If you hit your Stradivarius with a hammer, you are unlikely to improve its tone. Deleterious mutations are selected against and fade from a population (are less abundant in descendants) because they impair reproduction, only to potentially recur in the future because the mutational process is blind and enduring.

A second, smaller class of selected mutations is the favorable ones, rare but crucial in their impact. Thus, it is wrong to think of mutation as completely random in the Darwinian

world view; many mutations do not play (the neutrals), many are eliminated (the deleterious), and only a minority finally end up persisting and changing the evolutionary fate of an organism. This minority status has the curious effect of making the adaptive mutations a small, special class not necessarily representative of total change in the genome. Thus, a significant storm of change surrounds genomes (discussed further in Chapter 4), but only neutral changes, and a selected, adaptive trickle descend into deep time.

Evolution thus has a cost, which must be visible. Because the majority must usually be less fit than the selected few whenever there is evolutionary advance, evolutionary advance implies the waning of the creatures who do not succeed most brilliantly. For a gene indeed, evolutionary success is defined only in the context of the lesser reproductive success of many other genes (and their bearers) who do less well and decline into history. Darwinian deficiencies are among the less appreciated marks of evolution—but here lies a large fraction of human art. Romeo and Juliet; Frankie and Johnnie; sex, drugs, and rock and roll—a Darwinian torrent flows onward, and it will not end until humans end.

Because mutations are usually small alterations of a huge, divided, linear digital message, progenitors and descendants usually resemble each other pretty closely. Evolution therefore advances almost continuously, rather than by broad or general change between successive generations. This is the principle of continuity, which I first heard formalized by Leslie Orgel of the Scripps Institute, and it will be useful later in this book.

And thus were formed the orchid and the spider, the buffalo and the bacterium. From arbitrary changes in the genetic text, within constraints set by the chemistry and physics of

genomes, mindless but purposeful selection from this myriad, and then—forms most varied, sublime, and pertinent to the worlds that bear them.

Readings

A Farewell to Alms: A Brief Economic History of the World. Gregory Clark. Princeton University Press, Princeton, N.J. (2007).
A consideration of human Darwinian genetic change as a possible route upward from a long-stable stone age culture, providing a changed worker who could participate in and profit from the Industrial Revolution.

Darwin's Dangerous Idea: Evolution and the Meanings of Life. Daniel C. Dennett. Simon and Schuster, New York (1996).
An exceedingly rare combination of vigor and rigor, in service of the idea and explanatory power of Darwinian evolution.

The Moral Animal: Why We Are the Way We Are: The New Science of Evolutionary Psychology. Robert Wright. Vintage, New York (1995).
A Darwinian take on the possible evolutionary origins of human behavior and psychology—human nature, in short.

The Big Tree: No Jackalopes Please

The fruit of the righteous is a tree of life.

—Proverbs 11:30

The framework of bones being the same in the hand of a man, wing of a bat, fin of the porpoise, and leg of the horse,—the same number of vertebrae forming the neck of the giraffe and of the elephant,—and innumerable other such facts, at once explain themselves on the theory of descent with slow and slight successive modifications.

—Charles Darwin, *On the Origin of Species* (1859)

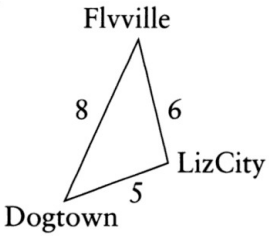
As I sit writing these words, birds glide in to drink from a pool nearby, insects buzz, and a short distance away evergreen treetops sway as the sun stirs the atmosphere. Look anywhere on the temperate or tropical Earth—you cannot help but be struck by the riotous success of life. And this despite its invisibility; life's successes are mostly unseen because virtually everywhere microbial cells outnumber the cells of visible beings. Notably, this is true even within the bodies of the visible beings—our own cells are outnumbered by the cells of microbes on and in us. A huge, usually underappreciated world of small, diverse creatures vibrates with activity below the resolution of

unaided human vision. Yet there is still a simple way to picture all life on Earth together, in impartial and proper array, and it is our subject here: the Big Tree.

The Big Tree appears at this point because, in order to locate the RNA world (the era of RNA creatures) in relation to life today, it is vital to visualize the complete course of life on Earth. So in this chapter I describe how we can draw a rational, objective picture which lays out life's history, using information that is subtle, quantitative, and now within easy reach.

The key idea is this: each creature preserves a record of its history, and particularly of its relation to its nearest relatives in deep time, in its genome. That is, each living thing is most similar to the now-separate type of creature that most recently shared a common ancestor. Humans are similar to chimps, but less similar to dogs. Each is less similar to the lemur from which it diverged next most long ago, and so on. Based on this concept, and given a simple measure of genomic similarity (or conversely, evolutionary distance), we can draw a map.

The implied map is no more complex than that of the Earth's surface. Suppose that LizCity is 5 miles from Dogtown and 6 miles away from Flyville. If you know that it is 8 miles from Dogtown to Flyville, you can draw a simple map that relates the cities, like this one:



domly and particularly, frequently suffering minor invasions by others. In fact, 50% of an admirable genome (our own) is the result of this kind of recent incursion!

To stray back to the topic at hand: gene transfer between different creatures confuses the evolutionary distances we assign by counting nucleotide sequence changes. Given such jumping, the history of a creature is not necessarily the same as the history of all its genes. We want the genes used for our Big Tree to have stayed completely put over deep time, so that a gene's sequence records only the history of its present organism.

The ribosomal RNA (rRNA) gene is a frequent choice for such evolutionary studies. rRNA is wound through the functional heart of the cell's protein assembly robots, the ribosomes (see the lexicon and Chapter 17). rRNA changes slowly and is present in all cells, so that organisms can be compared across deep time. Moreover, the protein synthesis machine—of which the ribosome, with its RNA, is a complicated major part—is even larger and more complex and therefore difficult to transfer successfully. It is unlikely that grafting the front end of a Toyota Yaris onto the back end of a Rolls-Royce will yield a means of transportation as good as either a complete Yaris or a complete Rolls-Royce. Because the mechanisms in front must work with those in the rear, the inexpensively engineered Yaris is unlikely to be improved by an indiscriminate graft from the haughty Rolls. In the same way, because you can't transfer part of the huge assembly that is the protein synthesis machine, you must either take it as a whole or, more likely, not transfer it or its parts at all.

Carl Woese of the University of Illinois began making the measurements for such maps by counting nucleotide differences between rRNA genes (even before large-scale sequencing was possible) and interpreting them as evolutionary dis-

tances. The cumulative number of sequence changes increased with greater separation in time. The idea that such molecular changes roughly measure the time since two genes were one (inside an ancestor) is even older, having first been proposed by the chemist Linus Pauling and the biologist Emile Zuckerkandl, who were in fact thinking about other biomolecules, proteins. While it is tempting to count sequence changes as direct indices of time, the rate of change varies in response to factors we cannot always know. So nucleotide changes are only roughly interpretable as ticks of an evolutionary clock. The clock keeps time best over intermediate separations, because these tend to keep the history between two organisms simple.

Time's approximate, rather than exact, influence is evident in the drawing of the tree, where the present is at the periphery for all lineages. The distance to the old center is usually somewhat different along any two paths, even though the underlying times are necessarily always equal. This is because nucleotide changes are accepted into the genomes of separated creatures as arbitrary, independent events. The total numbers are somewhat different as a result of this variation. So use of the mutational distance (which is what you can actually measure) as a stand-in for time (which is a deduction and not directly observable) must be undertaken with caution. Usually, the smaller the length of the branches (depth of time and complexity of descent), the better. However, our earlier example of a distance two mutations (the distance up one branch and down the other branch to the second sequence) errs too far on the side of simplicity—it is so small that it might be inaccurate because of the variability that always accompanies small numbers of events.

But, with distances between all organisms measured by counting the sequence changes, we can ask for the diagram of evolutionary descent (the tree) that best represents the set of measured distances. The best trees must take mutational chance into account. Actually, the tree in Figure 3.1 is even a bit more sophisticated: it is the result of asking which tree would be most likely to give rise to the observed distances, given that certain kinds of mutations are more probable than others. (For example, the nucleotide bases vary in size: interchange of the small bases C and T is more likely than exchange of a small one for a large one, such as C becoming A.)

What then do we see when all this computation is finished? The answer is amazing and encouraging. Genes (or at least rRNA genes) behave just as evolution by descent from common ancestors suggests they should. These simple ideas about sequence resemblance and descent successfully order all life on Earth into a simple treelike diagram. Just by counting sequence changes, we have reproduced most of what we would have concluded by using all other information we can bring to bear about the macroscopic and microscopic look of organisms. The Big Tree is one of the (lesser known) triumphs of biological science; it summarizes and orders Earth's creatures in somewhat the way that the periodic table organizes the chemical elements.

The tree shows none of the myriad of groupings that would have seemed immediately crazy, for example, humans grouped with (having RNA sequences more similar to) butterflies to the exclusion of other insects. Among organisms big enough to see (the "crown group" at the middle right of Figure 3.1), cats group with other felines, dogs with wolves, apes with humans, and plants with other plants. Distance on the molecular tree

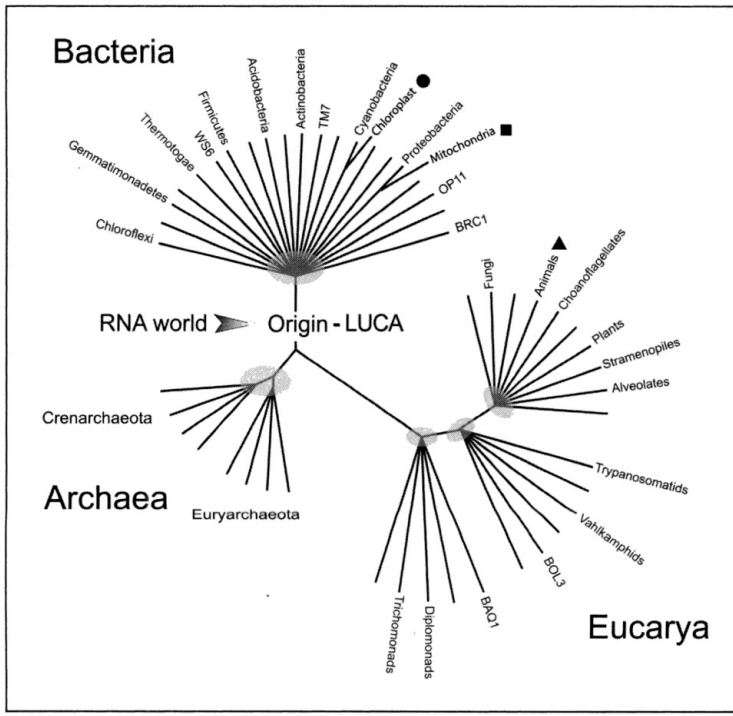


Figure 3.1. The Big Tree. This simplified tree shows the three great domains of life, defined by ribosomal RNA sequences. Grey ovoids mark areas within which evolutionary divergence is not yet confidently resolved. A square marks the mitochondrion and a circle, the chloroplast, each closely related to an existing bacterium. Humans are not resolved from other animals at this low resolution, but are within the branch indicated by a triangle. Alphanumeric names refer to creatures known only from their nucleic acid sequences. The Last Universal Common Ancestor or origin is the root of the tree, behind or before which lies the RNA world.

usually agrees with other indications of relatedness. However, molecular distances have a crucial advantage: they are available even in difficult cases, when appearance is not useful. And, finally, rRNA molecular data link all known creatures, spanning all cases in which new species appeared and still have descendants on Earth.

That descent with modification makes sense of the cousinship of most of the world's biota is a strong argument for the Darwinian process, and thereby for the descent of all of Earth's biota via genetic change from a single origin. Conversely, the tree shows that all present creatures are surprisingly related, much more so than their outward appearances suggest. The snake and the platypus and the whale have different, but related, DNA sequences that belie their outward differences.

The tree both links the various and distinguishes the similar. With all due respect, are you related to bread fungus? The Big Tree shows that you are (along with the rest of us), and tells you just how close that relationship is. The tree's distinctions among apparent similarities are notably useful for microbes. Even under a microscope, there is not much to go on in judging relationships between similar bacteria, which look like similar spheres or oblongs. However, molecular sequence data resolves them easily. Two bacteria, for example, a Bacillus from the soil and a human pathogenic Clostridium, are about as different as a human and bread fungus. In fact, most of the length of the tree's branches is in lines leading to microbes. Therefore, most of Earth's diversity resides in microbes. That is, most of evolution has led to microorganisms, not to the more familiar world of animals and plants and insects. It's a microbial world—now as at life's beginning.

Furthermore, the reach and bushiness of the microbial branches of the Big Tree reflect the fact that the microscopics

Yarus captivates with skilled character development — but here, the "characters" are the prebiotic molecules that gave rise to everything that has ever lived or is alive today on our planet.

— THOMAS CECH, Distinguished Professor, University of Colorado-Boulder, Nobel Laureate in Chemistry, 1989

ISBN 978-0-674-05075-4