Chapter 3

Immortal Genes: Running in Place for Eons

To be sure, everything in nature is change but behind the change there is something eternal.
—Johann Wolfgang von Goethe

HE WASN'T LOOKING FOR A NEW KINGDOM.

Microbiologist Tom Brock and his student Hudson Freeze were prowling around the geysers and hot springs of Yellowstone National Park one day late in the summer of 1966. They were interested in finding out what kinds of microbes lived around the pools and were drawn to the orange mats that colored the outflows of several springs.

They collected samples of microbes from Mushroom Spring, a large pool in the Lower Geyser Basin whose source was exactly 163 degrees F, thought at the time to be the upper temperature limit for life. They were able to isolate a new bacterium from this site, a species that thrived in hot water. In fact, its optimal growth temperature was right around that of the hot spring. They dubbed this “thermophilic” creature *Thermus aquaticus*. Brock also noticed
some pink filaments around some even hotter springs, which raised his suspicion that life might occur at even higher temperatures.

The next year, Brock tried a new approach to “fishing” for microbes in the hot springs of Yellowstone. His fishing tackle was simple: he tied one or two microscope slides to a piece of string, dropped it in the pool, and tied the other end to a log or a rock (don’t try this on your own—you will be arrested and quite likely scalded or worse). Days later, upon retrieving the slides, he could see heavy growth, sometimes so much that the slides had a visible film. Brock was right that organisms were living at higher temperatures than had previously been thought, but he did not imagine that they were living in boiling water. And they weren’t just tolerating 200 degrees F or more—these organisms were thriving in smoky, acidic, boiling pots such as Sulphur Cauldron, in the Mud Volcano area of the park. Brock’s Yellowstone explorations opened eyes and minds to the extraordinary range of life’s adaptability, identified bizarre but important new species such as Sulfolobus and Thermoplasma, and launched the scientific study of what he called “hyperthermophiles,” lovers of superheat.

Brock’s discovery of hyperthermophiles would, in time, lead to three more discoveries with profound impacts on biology. Brock lumped all of his new species into the classification “bacteria.” Under the microscope, they did appear a lot like ordinary bacteria (figure 3.1). But, a decade later, Carl Woese and George Fox at the University of Illinois discovered that various sulfur-, methane-, and salt-loving species actually formed an entire kingdom unto themselves. They were as different from bacteria as bacteria are from eukaryotes (the division of life to which we animals belong, as well as plants, fungi, and protists). This new third domain, or division, of life is now referred to as the Archaea.

The second discovery from Brock’s world was a practical one. A heat-stable enzyme that could copy DNA at high temperatures was isolated from Thermus aquaticus. This enzyme led to the invention of a new, efficient, and very fast technique for the study of genes in any species. This technique catalyzed a vast expansion in the amount and diversity of DNA information that could be obtained from nature, as
well as the creation of a multi-hundred-million-dollar market in DNA diagnostics and forensics.

The third and most recent discovery has emerged from the study of archaean genomes. Scrutiny of archaean genes has revealed critical clues about the making of our own eukaryotic ancestors nearly 2 billion years ago. Still preserved in the DNA of these primitive organisms are many pieces of DNA code that also exist in humans and all other eukaryotes. This shared text forms the remaining traces of an early event that gave rise to the first eukaryote, and is crucial evidence that an archaean was one of our original genetic parents.

In this chapter, we are going to examine some of the oldest DNA text on Earth. The fact that such ancient text has endured over eons of time, against the steady bombardment of mutations that could have

**FIG. 3.1. A sample of microbes from a hot spring.** This scanning electron micrograph reveals the growth of a variety of microbes on a slide immersed in the Obsidian Pool of Yellowstone National Park. Figure from P. Hogenholtz et al. (1998), Journal of Bacteriology 180:366.
erased it many times over, is itself remarkable. But these “immortal”
genres are also powerful evidence of two key elements of the evolution­
ary process—the power of natural selection to preserve the DNA
record and the descent of life from common ancestors.

The immortal genes vividly reveal the evidence for one very impor­
tant but somewhat underappreciated face of natural selection. More
thought and attention has been directed to the “creative” dimension of
natural selection and how new traits evolve, but this is only one aspect
of the evolutionary process. Natural selection also acts to remove, in
Darwin’s words, “injurious change.” I will explain how the effect of
the removal of harmful mutations by natural selection is manifest in
the DNA records of species, in the form of hundreds of genes that
have been preserved across kingdoms of life for more than two billion
years. In these immortal genes, the steps of evolution we see are just a
matter of “running in place” as the gene’s text changes only within
narrow limits set by natural selection.

The survival of individual genes over vast geological periods pro­
vides more than unimpeachable evidence of the preservative force of
natural selection. They are clues to the history of life’s evolution from
ancient ancestors, a new kind of evidence that Darwin could never
have imagined. I will show how these immortal genes are powerful
genealogical records that reflect the degree of relatedness among king­
doms and help us retrieve and reconstruct events in the history of life
that are not visible in the fossil record.

Looking at DNA and Reading the Code

The skyscrapers of DNA sequences that we now own contain a lot of
text, about forty thousand volumes, at a million characters each. The
records of some species, such as humans, require a whole encyclopedia
of about three hundred volumes while others, such as a bacterium,
just a three- or four-volume set. No matter which volume one looks in,
the text at first looks pretty much the same, like this:
ACGGCTATGGGCTAATCGACAT
AATGGCTATAACACATAGCATGCACCTATGGTGAGCTCTAGCCTGGGG
TCCGGCTAAACGACCTTGTAATCGGCTACCAAAGTGCACATACCTTCTACT
CAGACCTTGGCTACCAAAGCTACACATACCCAAATCTTCTATGGACGGCT
TAATGGAGACAAACCTTTACTATGGCTACCAAAATTACTATGGACAT

etc., for hundreds of pages.

How can such a monotonous text composed of just four different characters encode the instructions for a complex creature? Moreover, how the heck do we read this stuff?

To make sense of the language of DNA, we need to learn how to look at genomes and genes and how to read DNA code. We can then make comparisons between species at many different scales, from very close relatives to vastly different life-forms whose lines split off from one another early in life’s history. The clues to evolution emerge from understanding the meaning of the similarities and differences we find.

In order to decipher the natural history that resides in the DNA record, we have to have a firm grasp of the language of DNA, and of how DNA information is decoded in making the working parts of living organisms. Don’t be intimidated—you can learn the language of DNA. It has a very small alphabet and a very limited vocabulary, and its rules of grammar are simple. The payoff for learning about the DNA code is being able to see, and therefore to much better understand, the process of evolution at its most fundamental level. I understand that new terms can get confusing, so you might want to bookmark this short section for future reference.

Here we go.

Proteins are the molecules that do all of the work in every organism—from carrying oxygen, to building tissue, to copying DNA for the next generation. The DNA of each species carries the specific instructions (in code) necessary for the building of these proteins.

DNA is made of two strands of four distinct bases. These chemical building blocks are represented by the single letters A, C, G, and T. The strands of DNA are held together by strong chemical bonds
between pairs of bases that lie on opposite strands—A always pairs with T, C always pairs with G—as shown here:

```
-AGTCAGTC-
-----
-TCAGTCAG-  
```

so, if we know the sequence of one strand of DNA, we automatically know the sequence of the other strand. It is the unique order of bases in a sequence of DNA (ACGTTGGATAA, etc.) that forms the unique instructions for building each protein. The most amazing fact about DNA is that all of life's diversity is generated through the permutations of just these four bases. So, if we want to understand diversity, we have to crack the code.

How are proteins built and how do proteins know what their job is? Proteins themselves are made up of building blocks called amino acids. Each amino acid is encoded as a combination of three bases or a triplet (ACT, GAA, etc.) in the DNA molecule. The chemical properties of these amino acids, when assembled into chains averaging about 400 amino acids in length, determine the unique activity of each protein. The length of DNA that codes for an individual protein is called a gene.

The relationship between the DNA code and the unique sequence of each protein is well understood because biologists cracked the genetic code forty years ago. The decoding of DNA in the making of proteins occurs in two steps, which I will now describe. In the first step of decoding DNA, the sequence of bases on one of the strands of the DNA molecule is transcribed into a single strand of what is called messenger RNA (mRNA). Then, in the second step, the mRNA is translated into the amino acids that build the protein. In the cell, the genetic code is read (from the mRNA transcript) three bases at a time, with one amino acid determined by each triplet of bases (a short example is shown in the right half of figure 3.2).

There are sixty-four different triplet combinations of A, C, G, and T in DNA, but just twenty amino acids. Multiple triplets code for particu-
Fig. 3.2. The expression and decoding of DNA information. An overview of the major steps in decoding DNA into a functional protein. Left, long DNA molecules contain many genes. The decoding of a portion of one gene is shown in two steps. First, the complement of one DNA strand is transcribed into mRNA. Then, the mRNA is translated into protein, with three bases of the mRNA encoding each amino acid of the proteins (shown as the letters L, N, P, and Q here). In mRNA, the base U is used in place of the T in DNA. Figure by Leanne Olds.

lar amino acids (and three triplets code for nothing, and mark the stopping point in the translation of mRNA and the making of a protein—as periods mark the end of a sentence). Much to our convenience, but also of profound evolutionary significance, this code is, with few minor exceptions, the same in every species (this is why bacteria can be used to produce human proteins for pharmaceutical use, such as insulin).

Thus, given a specific DNA sequence, it is easy to decipher the protein sequence that that DNA encodes. However, not every base in DNA is part of a message for a protein. In fact, a large portion of DNA is "noncoding." The first challenge biologists have when given long reams
of DNA text is to figure out where the “coded” messages begin and end. With whole genome sequence data this, thankfully, is now carried out largely on computers using algorithms that are really good at searching for, and finding, the needles in the haystacks of DNA sequence.

The coding sequence of an average gene is about 1200 base pairs in length. In some species—particularly, microbes such as bacteria or yeast—genes are very closely packed with relatively small spaces of noncoding DNA between the thousands of genes in the entire genome. In humans, and many other complex species, genes occupy only a small fraction of all of the DNA, and are separated by long intervals of noncoding DNA. Some of this noncoding DNA functions in the control of how genes are used, but a lot of it is what is called “junk.” This junk accumulates by various mechanisms and often contains long repetitive tracts with no informational content; it is not purged unless it has adverse effects. I will generally ignore this junk, but it is worth mentioning in order to have a picture of the structure of our genomes as archipelagoes of islands (genes) separated by vast areas of open sea (junk DNA).

The Fates of Genes: The Immortal Core

When scientists look at entire genomes, their first aim is to locate all of the genes within the entire DNA sequence. This allows them to take an inventory of a species’ genes that includes the total number of genes and a list of every individual gene. Because biologists have been studying the genes and proteins from species for a while now, we can sort genes and the proteins they encode into categories based upon their function and resemblance to existing genes and proteins.

The most interesting fact from the comparison of genomes is that while the number and kinds of genes differ considerably both between and within the three major divisions of life, great increases in complexity do not require proportionate changes in gene number. As shown in table 3.1, most bacteria possess on average around 3000
Table 3.1. The number of genes in genomes

<table>
<thead>
<tr>
<th></th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Aquifex aeolicus</strong></td>
<td>1560</td>
<td><strong>Sulfolobus solfataricus</strong></td>
<td>2977</td>
</tr>
<tr>
<td><strong>Neisseria meningitidis</strong></td>
<td>2079</td>
<td><strong>Methanocaldococcus jannaschi</strong></td>
<td>1758</td>
</tr>
<tr>
<td><strong>Vibrio cholerae</strong></td>
<td>3463</td>
<td><strong>Halobacterium sp.</strong></td>
<td>2622</td>
</tr>
<tr>
<td><strong>Staphylococcus aureus</strong></td>
<td>2625</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Escherichia coli K12</strong></td>
<td>4279</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Salmonella typhi</strong></td>
<td>4553</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Archaea</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Saccharomyces cerevisiae (yeast)</strong></td>
<td>6338</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Drosophila melanogaster (fruit fly)</strong></td>
<td>13,468</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Caenorhabditis elegans (nematode worm)</strong></td>
<td>20,275</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tetraodon nigroviridis (fish)</strong></td>
<td>20–25,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mus musculus (mouse)</strong></td>
<td>20–25,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Homo sapiens (human)</strong></td>
<td>20–25,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arabidopsis thaliana (plant)</strong></td>
<td>25,749</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
genes, with the smallest genome of a free-living species containing about 1600 genes. Any two bacterial species may differ in size, however, by as many as 3000 genes. Animals possess roughly 13,000 to 25,000 genes, with some animals differing by many thousands of genes. Note that complex creatures, such as a fruit fly, have only roughly twice as many genes as a single-celled brewer’s yeast, and that humans have almost twice as many genes as a fruit fly. But we humans have just about the same number of genes as a mouse.

However, gene number is just a raw figure. More detailed clues about evolution emerge from the direct comparison of the fates of individual genes. The differences in gene number tell us that certain genes must be present in some species and absent from others. Before I discuss some specific comparisons, it is important to think about what we might find when we compare the genes of species that belong to different groups. How similar or different should we expect the genes of different species to be?

Before DNA sequencing was possible, some of the great minds of evolutionary biology in the mid-twentieth century contemplated this question. They knew a bit about mutation and concluded that, over geologic time, mutation would eventually change just about every base pair in a genome. For example, with a mutation rate of about 1 mutation per 100 million base pairs per generation, in 100 million generations, most sites in a gene would be mutated at least once, on average. Given the very short generation times of microbes (on the order of hours), and the modest generation times of plants and small animals (a year or less), then one might expect little trace of similarity between the gene of any two species whose lineages diverged 100 million years ago. Indeed, in his 1963 book Animal Species and Evolution, the great biologist Ernst Mayr remarked, “Much that has been learned about gene physiology makes it evident that the search for homologous genes [the same gene in different species] is quite futile except in very close relatives.”

But, when we compare different kinds of bacteria with one another, or different animals (whose ancestors diverged well over 100 million
years ago) with each other, we find extensive similarities in their genes. For example, when the genome of the infamous delicacy the puffer fish is compared with the genome of the gourmand stupid enough to eat this deadly creature (the human), at least 7350 genes are found that are clearly shared between the two species. Furthermore, the proteins encoded by these genes are on average 61 percent identical. Since the evolutionary lines of fish and other vertebrates (including humans) separated about 450 million years ago, this is a much more extensive similarity than would be expected if mutations were simply allowed to accumulate over time.

More stunning, when we compare the genomes of Archaea, bacteria, fungi, plants, and animals, we find about 500 genes that exist in all domains of life. We know from the fossil record that eukaryotes are at least 1.8 billion years old and the Archaea and bacteria well over 2 billion years old. The genes these organisms all share have withstood more than 2 billion years of the steady bombardment of mutation and stand out as threads of text whose sequence and meaning have not changed significantly despite the vast differences among the species that carry them. These genes are immortal.

The functions of immortal genes are central to fundamental, universal processes in the cell, such as the decoding of DNA and RNA and the making of proteins. All forms of life have depended upon these genes since the origin of complex DNA-encoded life early in Earth’s history. These genes have survived through an immense arc of time, and life will continue to depend upon this core set of genes as it evolves in the future.

Immortal genes have survived not because they avoid mutation—they are as vulnerable to mutation as all other genes. The genes are immortal in the sense that the gene as a unit endures; however, not every letter of the gene’s code endures. This fact can be seen upon more detailed inspection of their DNA sequences and of the sequences of the proteins they encode, and it is a key demonstration of one aspect of the process of natural selection.
Running in Place:
The Conservative Face of Natural Selection

Upon closer examination, we find that while the immortal genes of different species encode very similar proteins, the sequences of bases making up the same genes are less similar than the resulting protein sequences. This discrepancy can exist because of the "redundancy" of the genetic code, which allows different triplets of bases to encode the same amino acid. This feature of the genetic code insulates DNA against injurious changes in that mutations that change a base in DNA do not necessarily change the sequence of the encoded protein. Such changes that do not alter the "meaning" of a triplet are called synonymous changes because the original and mutant triplet are synonyms that encode the same amino acid. Mutations that change the meaning of a triplet and cause one amino acid in the protein to be replaced with another are called nonsynonymous mutations.

It is straightforward to calculate the odds of a mutation being either synonymous or nonsynonymous. There are 64 possible triplets of bases. For any individual triplet there are 3 possible changes at each base, 9 possible single mutations in all. Multiply the 64 triplets by 9 possible mutations and this equals 576 possible different random base mutations. Consultation of the genetic code reveals that 135 of the 576 possible mutations (about 23 percent) are synonymous, while the remaining 77 percent are nonsynonymous. The key prediction from these calculations is that, without the intervention of natural selection, the expected ratio of nonsynonymous to synonymous changes in gene sequences is about 3:1 (77:23).

However, in nature the ratio we typically find is about 1:3 in favor of synonymous changes. This ratio is ten-fold lower than what we expect from random base mutations. Clearly, only a small fraction of the nonsynonymous mutations that occur are retained over time. What accounts for the far fewer than expected nonsynonymous mutations?
Fig. 3.3. An immortal gene. A short portion of the sequence of a protein found in all domains of life (called elongation factor 1-α). Several amino acids, indicated by shading, have not changed over a span of 3 billion years. Figure by Jamie Carroll.

Natural selection. There is no other explanation. This skew in the ratio is very clear evidence of a type of selection, called purifying selection, that maintains the “purity” of the amino acid sequences of proteins by ridding them of changes that would compromise their function.

We can see the signature of purifying natural selection in the sequences of the majority of genes, but it is most striking in the immortal genes that have been preserved across all domains of life. For example, many of the proteins involved in the making of key pieces of the machinery for decoding mRNA are shared among all species. Looking at a specific part of just one of these proteins (for simplicity’s sake each amino acid is designated by a letter), in representative Archaea, bacteria, plants, fungi, and animals great similarity has been maintained over this part of the protein for over 2 billion years (figure 3.3). Note that fourteen individual amino acids have been
absolutely maintained throughout evolution. These fourteen letters (amino acids) are effectively immortal.

If we compare the sequences of the DNA encoding this piece of protein in each species, however, we find that the DNA sequences are less similar than the protein sequence. For example, inspection of the human and tomato versions of this gene reveals that they are identical at just 65 out of 78 positions (83 percent), while the encoded proteins are identical at 25 out of 26 positions (96 percent). The reason for the greater similarity of the protein sequences than of the DNA sequences is the occurrence of 12 synonymous changes in the DNA sequence—these are mutations that are allowed to accumulate.

The pattern of evolution of genes under purifying selection is one of "running in place." That is, the bases may be changing but their translated meaning is not. Consider, for example, the triplet TTA, which encodes the amino acid leucine, in the DNA sequence of a gene. This triplet can change in two different ways and still encode leucine, and these mutated triplets can change further and still encode leucine:

| Original triplet | TTA → leucine |
| Mutated triplets | { TTG → leucine, GTA → leucine, GTI → leucine, GTG → leucine } |
| Double mutated triplets | { GTC → leucine, CTG → leucine } |

Most amino acids are encoded by at least two different triplets, and several amino acids are encoded by three or more (or, in the case of leucine, six). So triplets in DNA sequences can "run" (change from sequence to sequence) but selection usually sees to it that they do not run so far as to change the protein's sequence and function.

Selection prevents protein sequences from changing by favoring one particular sequence over variants in which one or more letters of the
sequence is altered. If a variant works less well than other forms of proteins, even if that is just 0.001 percent less well, selection over time will, by the algebra we saw in the last chapter, purge the variant from a large population. This purging is so efficient that individual letters in a protein's text can remain unchanged among virtually all species. **Realize that an immortal letter in a protein sequence has experienced mutation again and again, in uncountable numbers of individuals, in millions of species, over billions of years, but that all of these mutations have been purged by selection over and over again.**

From the alignment of protein sequences in figure 3.3, we can see that constraints exist on many amino acids in this protein in that only a few positions are allowed to differ between species. Many more synonymous mutations are allowed than are nonsynonymous mutations. I have shown this relationship for just one gene, but I could have selected thousands more genes, either from the 500 or so immortal genes or the majority of other genes in any group of species. This pattern of the strong preservation of the protein sequence at most sites, with the synonymous evolution of the corresponding DNA sequence, and diversity limited to a few sites in the protein, is the predominant pattern of evolution in the DNA record.

DNA sequences that encode the same protein sequence but that are substantially different are unmistakable evidence of natural selection allowing mutations that do not change protein function, while acting to eliminate mutations that would. The preservation of genes among different species over vast periods of time is thus definitive proof of one face of natural selection—its power, in Darwin's words, to "rigidly destroy injurious variations."

But the evolving genomes of species provide more than signatures of natural selection. In the DNA record, there is more information than just the history of a particular gene—there is information about the species that carries it, and about all of the preceding species that also carried it, right back through eons of life's history. Because of the power of natural selection to preserve information that would otherwise be erased in time, genomes contain a record of the history of life.
The new wealth of data from genomes offers unique insights into the deep past that could not be deciphered from any other source. I will close this chapter with the story of the evolution of the domain we belong to (eukaryotes), and the unique contributions that archaea and bacteria appear to have made to our ancestry.

The Making of Eukaryotes:
A Marriage of Two Very Different Parents?

The time will come, I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great Kingdom of Nature.

—Charles Darwin, letter to T. H. Huxley, September 26, 1857

Our understanding of the structure of nature has come a long way since Darwin's time. In his day, the living world was divided into just the plant and animal kingdoms. This dual system had been recognized since Aristotle and was formalized by Carl von Linné in 1735. Ernst Haeckel, in 1866, with his remarkable studies of protists, added a third kingdom to life's tree. The bacteria and fungi were not added as full-fledged kingdoms until the twentieth century.

Within this five-kingdom scheme, a higher primary division was also recognized, based upon fundamental differences in the types of cells found in different kingdoms, in 1938, French biologist Edouard Chatton proposed the names "prokaryote" and "eukaryote" based solely on the absence and presence, respectively, of nuclei. These two "superkingdoms" encompassed all of the known living world, until Carl Woese started studying the genes of the kinds of species Tom Brock found in Yellowstone.

Woese believed that bacterial taxonomy was a mess and needed more objective means of determining the evolutionary relationships
among species than their appearance or physiological characteristics. He turned to molecules. The possibilities for building species trees based upon DNA, RNA, and protein sequences were quickly recognized by scientists (such as Francis Crick, Emile Zuckerkandl, and Linus Pauling), as soon as protein sequencing began to reveal the similarities and differences in proteins shared among groups of species. The basic idea is quite straightforward. Sites in the text of DNA, RNA, or protein sequences that differ among a group of species, but are shared among subsets of these species, reflect their degree of kinship. Just as we build family trees based upon degrees of genetic relationships, we build species trees based upon their genetic kinship. But, as I will explain, sometimes there is a marriage that really confuses the family tree.

Woese used an abundant type of RNA molecule to make trees of bacteria. But when he included the thermophilic, methane-producing species with conventional bacteria, he found that “these ‘bacteria’ appear to be no more related to typical bacteria than they are to eukaryotic cytoplasms.” He proposed there existed a third superkingdom which, because of the adaptation of these species to the sorts of extreme environments that were presumed to exist early in Earth’s history, might be the original or ur-kingdom, so he suggested calling this new superkingdom “archaeabacteria.” This name was modified later to Archaea, in part to avoid confusion with the bacteria, and the superkingdom category was renamed “domain.”

While the division of life into three domains—Eukarya, Archaea, and Bacteria—has held up, the relationships between these three groups has been challenging to sort out. Darwin described the genealogy of species as trees, with speciation producing ramifying branches. But in the world of microbes, unknown to Darwin, some events happen that violate the pattern of treelike evolution. Microbes exchange genes, and some microbes live within other host species in a process called endosymbiosis. These processes enable the transfer of genes between very distant relatives, and thus confuse the family tree. In
order to figure out the relationships between eukaryotes, archaea, and bacteria, biologists have to sort out the history of lots of genes, not all of which may have the same family resemblances.

For example, some of the first studies of archaean molecules turned up some striking resemblances between some archaea and eukaryotes. Proteins that archaea use to package their DNA in chromosomes, to transcribe DNA, and to decode messages bear such similarity to those in eukaryotes that it suggested to many that eukaryotes evolved from some archaean. Some of these provocative similarities are in short “signature” sequences in proteins that are shared among some archaeans and eukaryotes, and no other group. For example, there is a short insertion of eleven amino acids in one of the immortal proteins involved in decoding messages. Table 3.2 shows the sequence of this insertion in different eukaryotes and archaea:

Table 3.2. Insert sequences

<table>
<thead>
<tr>
<th>Eukaryotes</th>
<th>Insert sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>GEFEAGISKNG</td>
</tr>
<tr>
<td>Yeast</td>
<td>GEFEABISKDG</td>
</tr>
<tr>
<td>Tomato</td>
<td>GEFEAGISKDG</td>
</tr>
<tr>
<td>Archaea</td>
<td></td>
</tr>
<tr>
<td>Sulfolobus</td>
<td>GEYEAGMSAEG</td>
</tr>
<tr>
<td>Pyrodictum</td>
<td>GEFEAGMSAEG</td>
</tr>
<tr>
<td>Acidianus</td>
<td>GEFEAGMSEEG</td>
</tr>
<tr>
<td>Bacteria</td>
<td>(Absent)</td>
</tr>
</tbody>
</table>

The existence of this sequence in two domains but not in the third would be most logically explained by the archaea and eukaryotes being more closely related to each other on life’s tree than to bacteria. The resulting picture of the tree would posit that there was a common ancestor of all three domains (the last “universal” common ancestor, or LUCA) that then split into two domains, the Bacteria and Archaea,
and the eukaryotes arose later from a branch of the Archaea. The tree of life would then be as shown in figure 3.4.

However, sequencing of whole archaean and bacterial genomes revealed, somewhat unexpectedly, that the majority of archaean genes show the greatest similarity to bacterial counterparts. Then, as more eukaryotic genomes were sequenced, their analysis suggested that many eukaryotic genes were more related to those of bacteria than of archaea. The story was taking on the nature of one of those riddles like “If your sister is also your aunt, then who is your father?” In short, the answer to the question of which groups are most closely related was muddled.

The resolution of the riddle stems from the pursuit of a key observation. It was noted that most of the similarities between archaea and eukaryotes were in so-called informational genes whose products dealt with the copying and decoding of DNA. Furthermore, most of the similarities between eukaryotes and bacteria were in operational genes involved in the metabolism of various nutrients and basic cellular materials. From the viewpoint of eukaryotes, it appeared as though they got their “brains” (informational genes) from one parent, and their “looks” (operational genes) from another.

This raised the specter that eukaryotes were the product of a mixed
marriage—a genetic fusion of archaean and bacterial parents. The notion of a fusion between vastly different species is not new. In 1970, Lynn Margulis proposed that mitochondria and chloroplasts, two key energy-producing organelles in eukaryote cells, arose from bacteria living within eukaryotes (this fusion process is endosymbiosis). This view is now widely accepted.

But what about the making of a eukaryote from an archaean and bacterial ancestor? Maria Rivera and James Lake of UCLA have concluded that, indeed, eukaryotes are of dual origin, from parents belonging to different branches of life. Rivera and Lake analyzed bacterial, archaean, and eukaryotic genomes for the sets of genes shared in all, all but one, all but two, all but three, etc., of the major divisions within these three domains. Comprehensive analysis of these patterns of shared genes indicate that the eukaryote genome is the product of a fusion between a relative of a type of archaean and a type of bacterium. Because symbiotic relationships are common among organisms living together (for instance, Yellowstone’s *Thermus aquaticus* gets its energy from photosynthetic cyanobacteria that color those bacterial mats), and this occasionally leads to endosymbiosis, the likely explanation for eukaryote origins was as a product of the fusion of genomes between an endosymbiont and its host. The resulting base of the tree of life is then not a trunk, but a ring from which our tree ascends and branches (figure 3.5).

So, if you happen to visit magnificent Yellowstone, don’t turn away from the stinking, boiling soup of those hot cauldrons or be revolted by the colorful strings and mats of slime that ooze around their edges. That’s no way to respect one’s relatives, no matter how distant. Ponder the amazing fact that you share hundreds of genes with members of this community. And, in this sort of community, somewhere an unfathomably long time ago, perhaps along a deep sea vent, in a belch of methane there emerged the ancestor of all of the familiar and visible kingdoms on Earth.

Of course, if all natural selection did thereafter was to maintain the status quo within very strict limits, life would be uniform and
unchanging, and not the riot of diversity we see in the world today and in the past 3 billion years of the fossil record. The figures on gene number in table 3.1 tell us that vast differences exist in gene content between life-forms. Beyond the core of 500 or so immortal genes, species vary widely in gene number. The differences in gene number tell us that, in the course of evolutionary time, new genes must be born. They are indeed, and that creative dimension of evolution will be the focus of the next chapter. It is also a hint that genes might also die. They do die, and I will take up that twist and what it can tell us about evolution in chapter 5.