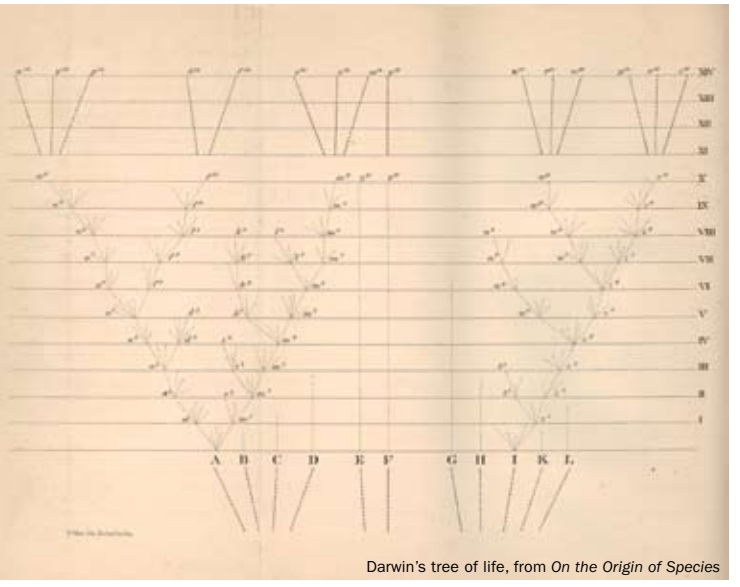
Computing Evolution

Scientists sift through genetic data sets to better map twisting branches in the tree of life

By Patrick Barry



Darwin's tree of life, from *On the Origin of Species*

Among its many prose-filled pages, Charles Darwin's *On the Origin of Species* includes only one illustration. It's a diagram of short lines leading upward from the base — a few lines at the bottom branch out repeatedly as they extend up. Darwin meant for the image to depict what he dubbed the "tree of life." This figure embodied Darwin's vision for how the tremendous diversity of life on Earth arose. A few species — the base of the tree — mutate and evolve over time, sometimes branching to form new species. An ancient species of bird might colonize a chain of islands and slowly evolve narrower beaks or other features specialized for the birds' new habitats. Eventually, groups in different habitats become separate species, and each species continues to evolve and adapt, perhaps branching again. In this way, the first fishlike land animals gave rise to the great diversity of amphibians, lizards, insects, rodents, marsupials, primates and birds.

It was a sweeping vision of life, revealing it to be a giant family with a vast genealogy. Branches of the tree show the kinship among creatures and the history of change and adaptation. Darwin toiled for much of his life to understand the relationships among species, the branches of this immense tree, by gathering countless specimens and scrutinizing their similarities and differences — a longer neck, a brighter-colored shell. Expanding this tree has been the painstaking work of generations of naturalists, biologists, taxonomists and paleontologists during the 150 years since Darwin published his seminal book.

Now that slow slog has quickened to an all-out sprint. Rather than divining clues to an organism's evolutionary history from observed traits, scientists are going straight to the genetic ledger sheet. Modern tools for rapidly reading species' DNA are laying bare those species' genetic inheritances, the patterns of genetic code shaped by eons of mutation and natural selection. And ever more powerful computers are churning through gigabytes and gigabytes of this genetic data to decipher which species are like sisters and which are only distant cousins.

"We've really learned more about relationships [among species] in the last 10 years than we did in the previous 200 years," says Doug Soltis, an evolutionary biologist at the University of Florida in Gainesville. "This is definitely going to be viewed as a golden era in our study of biodiversity. And it's just now taking off."

Already, large branches of the tree are being redrawn as scientists compare the DNA of dozens or hundreds of distantly related species. Within years, rather than decades, this computational excavation of life's past will achieve an important milestone in the history of science: a highly accurate map of the major branches in Darwin's tree of life.

"It's Darwin come full circle," Soltis says. "Starting from his tree figure [in the *Origin*], we're now putting together a basic tree of life for a large portion of known species. It's just incredibly exciting."

Such genetic comparisons have already overturned long-held ideas about the evolution of birds and have shed new light on the origins of animals. Scientists are also getting close to mapping the rapid diversification of the first flowering plants — which happened so quickly and recently, on a geologic timescale, that Darwin called it an "abominable mystery." And studies are refining ideas about the roots of all life, the initial emergence of the three superkingdoms: bacteria, archaea and

1983
Homeobox genes are discovered. The homeobox proteins turn on other genes in precise patterns at certain times during development to determine an animal's body plan.

1996
Dolly the Sheep is the first mammal ever cloned from an adult cell.

1990
The U.S. federally funded Human Genome Project begins.

1998
Celera Genomics, a private company headed by J. Craig Venter, announces it will also sequence the human genome.

eukaryotes, the group that includes all plants and animals.

For the past five years, the National Science Foundation has allocated $12 million each year for these genome comparisons in a program called Assembling the Tree of Life, or AToL. Its goal is producing a tree that maps the evolutionary relationships among all the roughly 1.7 million known species. The Human Genome Project pales in comparison.

As with mapping the human genome, which led to the enormous task of understanding *how* the genome works in health and disease, completing a basic tree will mark the fulfillment of one challenge and the beginning of larger ones. Filling in all the twigs and leaves—every genus and species—will probably take decades. And in the near term, having the major branches of the tree and many of its leaves in hand will point biologists toward another set of questions to answer: Once scientists know what evolution did, they can ask better questions about how it did it.

"We can sit down and say, how did these species evolve? Why did they evolve this way instead of that way?" says Rebecca Kimball, an evolutionary biologist at the University of Florida. "When we're not certain if a chicken is related to a duck, that limits us from looking at this bigger picture. As we begin to get definitive trees of life for many groups, then maybe we can understand better how evolution works."

Ducks in a row

The concept behind these data-intensive comparisons is simple: The genomes of two closely related species should be very similar to each other, while genomes of species that have evolved separately for a longer time will have accumulated more differences.

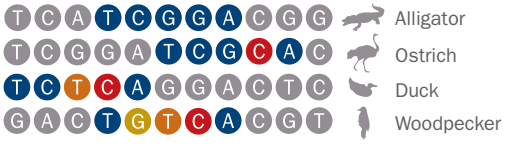
It seems easy enough. First put chromosomes from each species through a DNA sequencer to read the genetic catalog: the long sequences of A's, T's, C's and G's that represent how information-carrying chemicals in DNA are strung together. Then line up the matching parts of those data strings and note all the spots along the strings where the letters differ. Organisms that share large segments of genetic coding or a given mutation in their DNA are more closely related than organisms that don't.

But actual comparisons are a lot more complex. Just lining up the matching fragments of many genomes can be a tremendous challenge. Random mutations to DNA that drive evolutionary change sometimes come in the form of wholesale photocopying of large sections of DNA, or by the loss of a segment containing an entire gene. Species often have different numbers or types of chromosomes. And available genome sequences for infrequently studied

Avian Branches: Building a tree of life

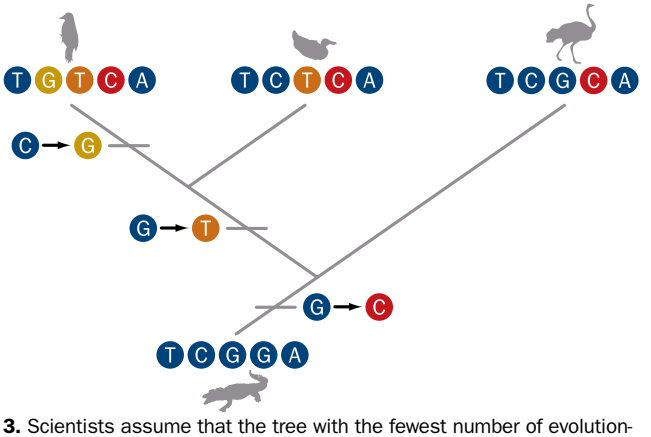
Instead of relying on observable traits to guess evolutionary relationships among species, scientists can now go straight to the source: The DNA that is marked by evolution. Here's one simple way to reconstruct the history of bird species using genetic sequences. (Although the evolutionary relationships shown here are real, the specific genetic codes and mutations shown are representative and are not based on actual DNA data.)

To construct a tree of related species, scientists need to compare these species to one that is not in the group of interest, but is closely related (the "outgroup"). Here, the outgroup is an alligator species that roots the tree by serving as a reference point for all of the species.



1. Researchers pick a region of DNA that has counterparts in all of the species including the outgroup. Usually, many different regions of DNA are compared to make the most accurate tree (just one such segment of DNA is shown here).

2. The strings of DNA from different bird species and the outgroup are lined up and compared. Differences in the letters of DNA code, shown here in color, are identified. Then the species are ordered from fewest to most differences.



3. Scientists assume that the tree with the fewest number of evolutionary changes best represents the species relationships.

Constructing evolutionary trees is complicated. Missing DNA sequences, repeated stretches of DNA and a daunting amount of information all confound the task of generating accurate species histories.

1999
Human Genome Project completes first sequence of a human chromosome.

2001
Working drafts of the human genome sequence are published in *Nature* (mainly reported by the Human Genome Project) and *Science* (mainly reported by Celera Genomics).

2004
Human Genome Project reports the near-complete sequence of the human genome. Later, private companies announce full sequences for individual genomes.

SOURCE: LAURA SANDERS, GRAPHIC: AVIK NANDY; TIMELINE: DRA_SCHWARTZ/ISTOCKPHOTO

species are usually fragmented and incomplete.

"There aren't that many genomes available," says Jonathan Eisen, an evolutionary biologist at the University of California, Davis. Public databases contain partial genomes for more than 140 plants, 250 mammals, 390 invertebrates and 1,600 microbes — a sliver of life's astonishing diversity.

Even in a well-studied group like mammals, scientists have found only about 2,000 genes having counterparts across the whole group that can be lined up for comparison. And limited budgets for fast computers and DNA sampling mean that in these kinds of comparison studies, dubbed by Eisen as "phylogenomics," scientists typically compare only a few hundred or a few dozen genes.

Then there's the question of how to translate the differences among those fragments into maps of the branches in the tree of life. Some genes accumulate changes faster than others, so comparing one gene might tell a different story about the species' histories than comparing another gene would. And genes can sometimes jump from one organism to a distantly related one, mixing up the genetic clues. This "noise" in the data poses a challenge to scientists trying to draw the correct evolutionary tree for a certain set of organisms from a dizzying number of possibilities.

"When you're trying to build trees, once you get over a few hundred organisms there are more possible trees than there are atoms in the universe," Soltis explains. "So it's a huge problem."

Even on fast computers, crunching the numbers for this problem can take months of continuous calculation.

"We managed to crash a few computers with the size of our data set," Kimball says. "We had an analysis running for two months on a computer one time and then a power outage hit. Although we joked about it, it was frustrating at times."

Kimball and her colleagues were analyzing about 32,000 letters of genetic coding from each of 169 bird species to decipher the early branches in the evolution of birds. The results, reported last June in *Science*, confirmed some long-held ideas about bird evolution, but upended others. Surprisingly, perching birds such as the house sparrow are actually closely related to parrots. Flamingos are indeed closely related to water-loving grebes — a relationship that had been disputed — though neither is part of the main branch of waterbirds.

Beyond an aquatic lifestyle, several other traits that might outwardly suggest kinship also evolved more than once in separate groups, according to the team's analysis. An order of daytime birds that includes hummingbirds actually evolved from nocturnal ancestors, which shows that being active during the day must have re-evolved in this lineage. And as Kimball's team reported last September in *Proceedings of the National Academy of Sciences*, flightlessness among birds such as ostriches, emus and kiwis evolved not just once, as scientists had thought, but at least three times.

Similar studies have begun to unravel the "abominable mystery" of flowering plants' rapid emergence. Comparisons of plant genomes show that these diverse plants arose between 140 million and 180 million years ago — earlier than suggested by the oldest known flowering plant fossil, which is only 132 million years old, Soltis and his colleagues noted last year in the *Annals of the New York Academy of Sciences*. Also, water lilies appear to be one of the first lineages to diverge. Although more evidence is needed, this research could settle a long-standing debate about whether flowering plants began as forest shrubs or aquatic herbs.

Soltis says unpublished research by his team takes this work further, outlining many of the major branches of flowering plants' evolutionary history. "We've now got most of [these branches], and we're getting the last papers out now on most of those deep-level relationships," he says.

Within four or five years, Soltis says, scientists are likely to have a complete, basic tree for the roughly 15,000 genera spanning 300,000 to 400,000 species in this diverse family of plants.

Such in-depth studies can flesh out the tree's details piecemeal. That's part of the design of large projects such as AToL: All the work need not be done in a single, giant effort. Individual teams can riddle out parts of the tree and then snap those parts into the master tree like pieces of a jigsaw puzzle.

To reveal the largest branches that form the overall framework of this master tree, scientists use a broader set of DNA samples that includes a wider range of species. Lineages for species as different as slime molds and squirrels diverged hundreds of millions of years ago, so with enough genomic data from diverse species such as these, researchers can map those ancient branches.

For example, illuminating the oldest and largest branches of the animal kingdom required crunching the data for nearly 40 million letters of genetic code from 29 animals representing 21 major groups. The results shook scientists' ideas about how the first animals evolved. Biologists have long believed that the ancestors of sea sponges, which have very simple bodies, were the first to branch off from the rest

2004
Geologists ratify naming the Ediacaran period, a time just before the Cambrian period that hosts fossils suggesting significant diversity before the Cambrian explosion.

Ediacaran fossil

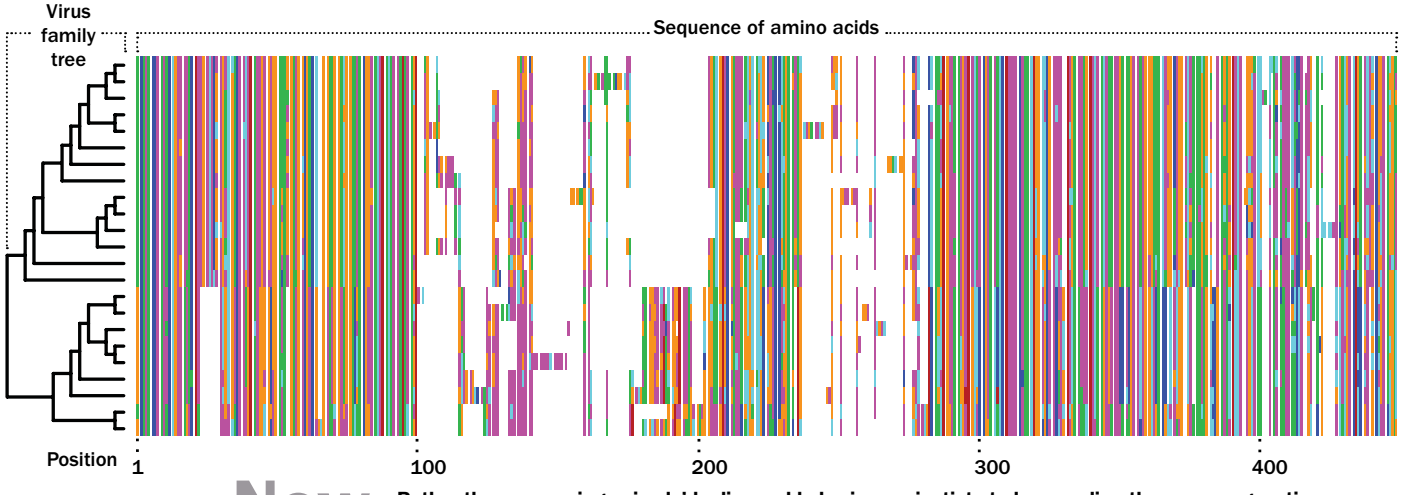
2006
Benjamin Voight and colleagues publish data showing that, within human history, a large portion of the human genome has changed in response to "selective pressures."

2008
Scientists report the first complete sequence of a Neandertal mitochondrial genome, showing no evidence of Neandertal interbreeding with humans.

G. NARBONNE; PINOPIC/ISTOCKPHOTO; PHILIPPE PLAILLY & ATELIER DAYNES/PHOTO RESEARCHERS

Then

Darwin and others collected specimens and scrutinized the similarities and differences among the species' bodies and behaviors. From these comparisons, scientists inferred the evolutionary histories of species. The insects shown here are from Darwin's personal collection.



Now

Rather than comparing animals' bodies and behaviors, scientists today can directly compare genetic codes. Computer-aided analysis of reams of genetic data reveals which species share segments of code or certain mutations, allowing scientists to infer the evolutionary history of life with high accuracy. The branched diagram (far left) represents a family tree of viruses. A section of each virus's genome, depicted here in terms of the string of amino acids encoded by the genetic sequences, reads from left to right. Matching amino acids are the same color. At position 1, for example, viruses in the top section of the tree share code for a particular amino acid (green). Viruses at the bottom section have code for a different amino acid (yellow) except for one strain (green).

FROM TOP: ENGLISH HERITAGE PHOTO LIBRARY; A. LÖYTYNOJA, *ET AL.*, *SCIENCE*, P. 1632-1635, 20 JUNE 2008

of the animal tree and start evolving independently. But the new work, reported by evolutionary biologist Casey Dunn of Brown University in Providence, R.I., and his colleagues last year in *Nature*, suggested that comb jellies, which have more complex bodies, branched off first instead (*SN: 4/5/08, p. 214*). If so, this discovery would imply that the last shared ancestor of sponges and comb jellies either had evolved a complex body already — in which case sponges' bodies must have become simpler over time — or that the common ancestor had a simple body, implying that complex body plans evolved separately in the comb jelly lineage and in the branch containing the rest of the animal kingdom.

Finding the last common ancestor of plants, animals, fungi and protozoa — all of which are called eukaryotes and all of which have much larger and more complex cells than bacteria — is more difficult. Eukaryotes, bacteria and single-celled organisms called archaea constitute the three largest, most fundamental branches in the tree of life, diverging billions of years ago. No consensus yet exists on when and how eukaryotes branched off from the other two superkingdoms, but studies are beginning to illuminate even this deep history.

A team led by Takao Shinozawa, a visiting professor at Waseda University's campus in Saitama, Japan, compared the genomes of 46 species: 36 bacteria, eight archaea and two eukaryotes. The analysis, reported in the August *Genes & Genetic Systems*, suggests that the main DNA in the nucleus of all eukaryotes descends from an archaea. But the DNA in mitochondria, energy-producing organelles in eukaryotic cells, has a different origin. Mitochondria were once free-living cells that became incorporated into eukaryotic cells long ago, most biologists believe. Shinozawa's comparison suggests that the free-living forebears of mitochondria belonged to a group of ancient bacteria called alphaproteobacteria.

"The bacterial taxonomy has been totally changed in recent years," says Bernard Labedan, an evolutionary biologist at the University of Paris-Sud 11 in Orsay, France. "Before it was a mess, and now it's clearer and clearer."

However, Labedan adds, "there are still a lot of things to make more precise." Pinning down these earliest branches with confidence will take more gigabytes of genomic data, more computer horsepower and more time.

Braiding the branches

Microbes, in particular, will be hard to deal with, in part because they swap genes like 13-year-old boys once traded baseball cards. Though direct exchange of genes among distantly related species is fairly rare in large, multicellular organisms such as plants and animals, single-celled microbes are masters of the gene trade. Snippets of DNA can float out of one cell, let's call it Alice, and get picked up by a cell of another species that we'll call Bill. Scientists who base their comparisons on this snippet of DNA will get the false impression that Bill is close kin to Alice and her relatives.

Such gene swapping braids the evolutionary branches, so that the collection of genes in a microbe's DNA may descend from many far-flung species. Some scientists argue that, for this reason, the evolutionary history of microbes is better imagined as a heavily crisscrossed web, rather than a branching tree. This braided genetic past doomed earlier studies that attempted to find a tree-shaped history based on a single gene shared by many species.

> "Could there still be some fuzziness? You bet — that's how science works. Fuzziness is not always bad. Sometimes the areas of fuzziness are telling you that something else is going on here, something that you might want to look at in more detail."
>
> Doug Soltis

But recent work that compares larger swaths of DNA can partially overcome this problem. Some regions of a microbe's genome — parts involved in cell division and other essential functions — are resistant to this lateral swapping of genes, and so follow more predictable rules of inheritance.

"The cores of these microbial genomes do have a tree," Eisen says. "There didn't seem to be any hanky-panky going on." However, this stable core represents only 5 to 10 percent of the microbes' genes, forcing scientists to study these microbes' evolutionary histories as if looking through a keyhole.

For ancient microbes, these lingering genetic data are usually the only clues available. Larger creatures occasionally leave behind fossils that scientists use as a reality check, showing when certain adaptations arose and calibrating the timeline suggested by the DNA. Microbes aren't so helpful.

More genomic data from more species will eventually bring the picture into clearer focus, even if some details of the tree will never be known with 100 percent certainty.

"Could there still be some fuzziness? You bet — that's how science works," Soltis says. "Fuzziness is not always bad. Sometimes the areas of fuzziness are telling you that something else is going on here, something that you might want to look at in more detail."

Despite a few lingering spots of uncertainty, having a highly accurate map of the historical tree of life that is freely available will be a boon for biology, and perhaps for education too. "This is my vision," Soltis says, "where schoolkids are going to be able to navigate the tree of life by clicking on one branch, and they can go down that branch and navigate and explore life's history." ■