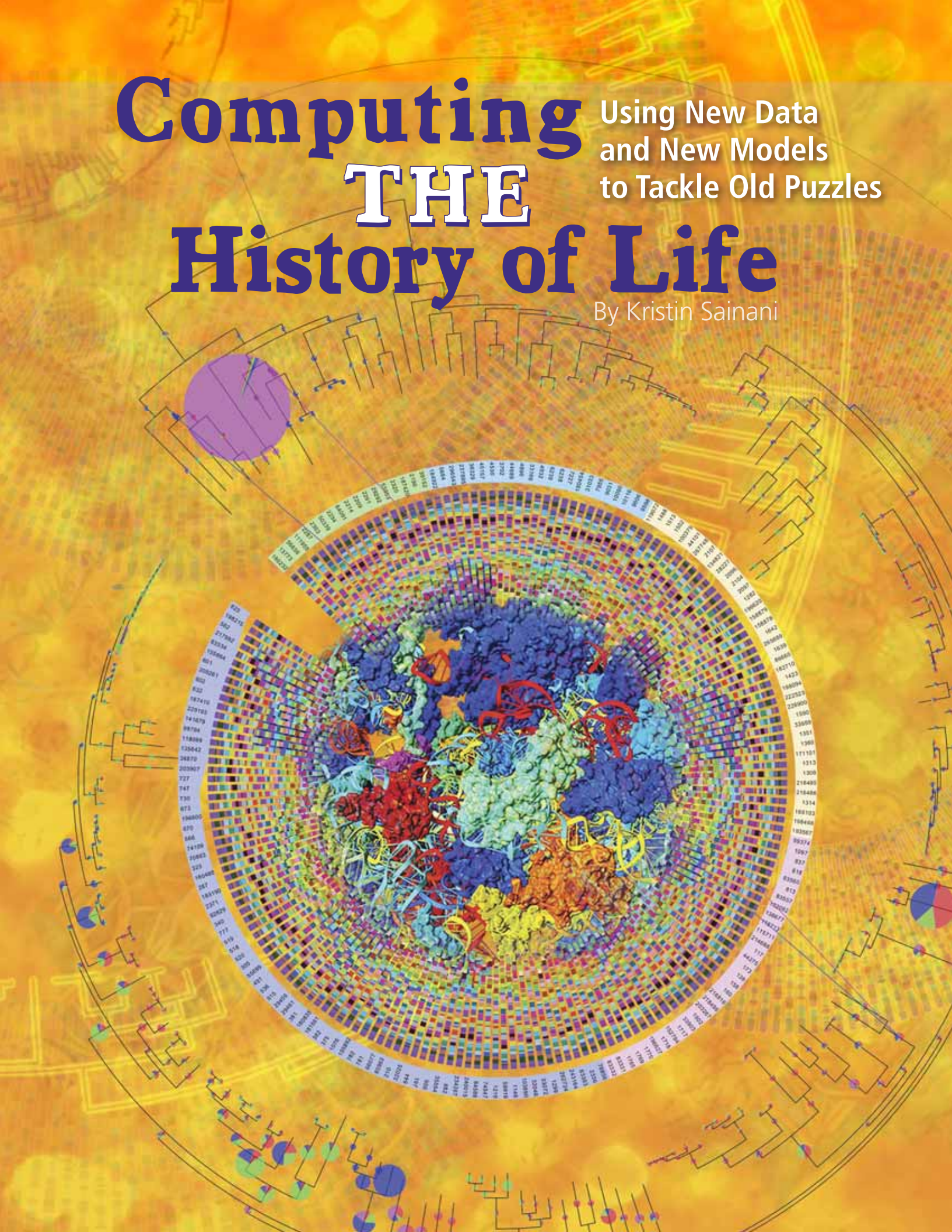
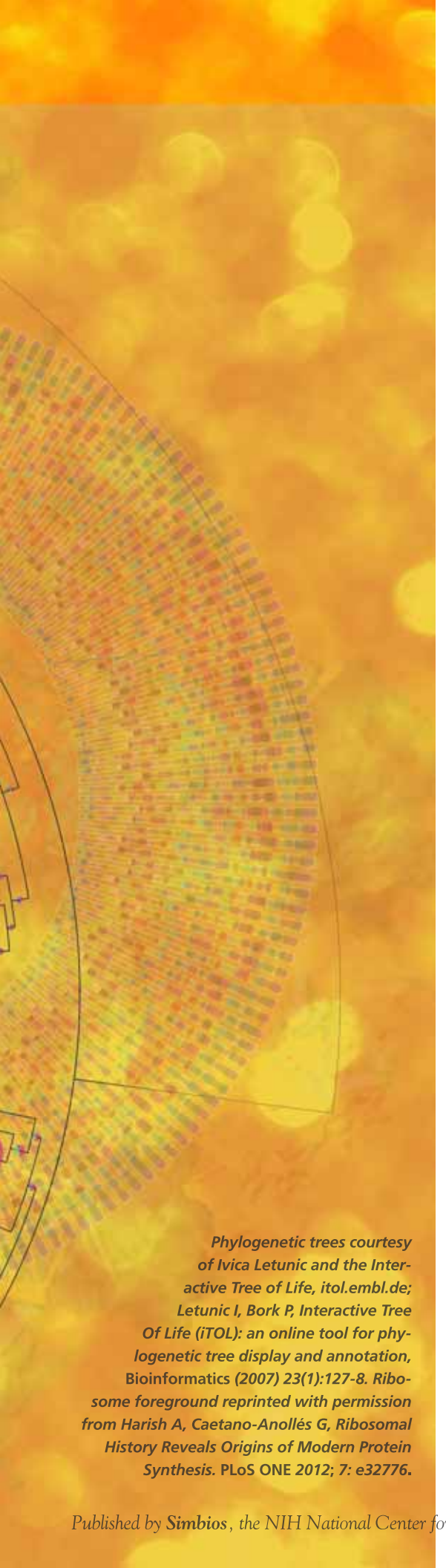
# Computing THE History of Life

## Using New Data and New Models to Tackle Old Puzzles

By Kristin Sainani





*Phylogenetic trees courtesy of Ivica Letunic and the Interactive Tree of Life, itol.embl.de; Letunic I, Bork P, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation,* Bioinformatics *(2007) 23(1):127-8. Ribosome foreground reprinted with permission from Harish A, Caetano-Anollés G, Ribosomal History Reveals Origins of Modern Protein Synthesis.* PLoS ONE *2012; 7: e32776.*

# In 1977,

the late **Carl Woese, PhD**, shook up biology when he published the first tree of life based on genetic sequence data. His team showed that, contrary to popular belief, eukaryotes did not evolve from prokaryotes; instead, three distinct domains of life (bacteria, archaea, and eukaryotes) all arose from a common ancestor.

Woese's revelation is now considered one of the greatest biological discoveries of all time. But it was initially met with vehement skepticism from fellow scientists. He challenged a beloved paradigm in biology, and he initially paid the price. It took a decade for his findings to become widely accepted.

The question "where did we come from?" is one that philosophers, theologians, and scientists alike have been trying to answer for millennia. But reconstructing events that took place millions to billions of years ago is fraught with difficulties. The further back one goes, the less information there is; and the more people resort to filling in the gaps with ideas and stories. These ideas are often so neat and elegant and pleasing that it's hard to give them up, even when new data clearly contradict them.

Today we are in a data-rich era in evolutionary biology. For decades, computational biologists who work in phylogenetics have built evolutionary trees by inferring evolutionary distance from the similarities of DNA sequences for one gene. Now they can build trees using whole-genome sequences (currently available for numerous species). Armed with such data, as well as increasing computational power and sophisticated new computational models and tools, it finally might be possible to answer some of the toughest and oldest puzzles in evolution.

"It used to be that data were the limiting thing. But of course now, keeping up with the data is the problem. I've been around a long time and watched it all. It's been exciting," says Russell Doolittle.

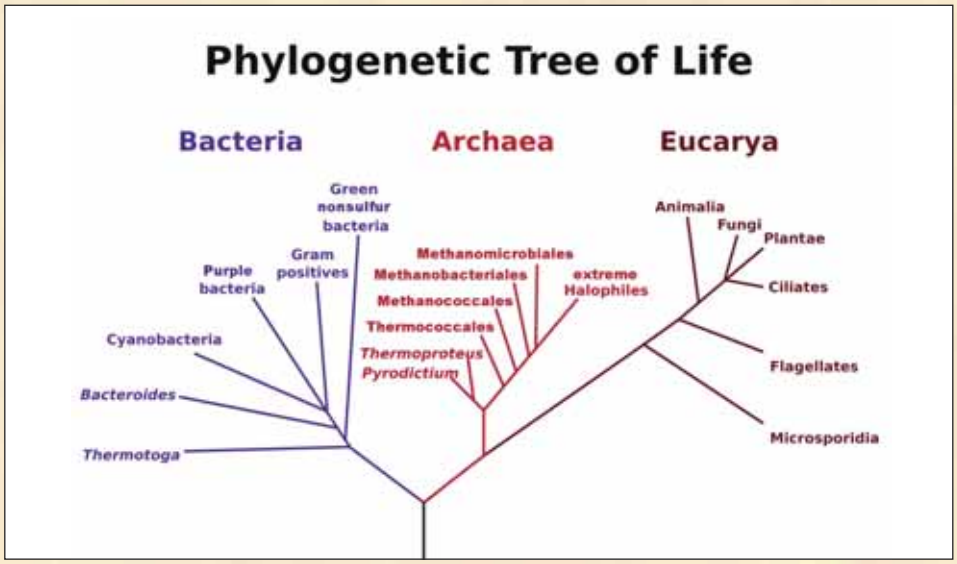
"It used to be that data were the limiting thing. But of course now, keeping up with the data is the problem. I've been around a long time and watched it all. It's been exciting," says **Russell Doolittle, PhD**, emeritus professor of molecular biology at the

University of California, San Diego. A pioneer like Woese, he reconstructed animal evolution using protein sequence data in the 1960s.

This article reviews seven history-of-life puzzles on which computational biologists and bioinformaticians are making headway: How did life begin? Which came first: RNA or proteins? Or did metabolism come first? Is there a fourth domain of life? How have proteins evolved since life began? Why did introns evolve? And what drives the evolution of form?

To answer these questions, many computational biologists are venturing beyond phylogenetics and simple Darwinian tenets by incorporating chemistry, physics, protein structure, epigenetics, morphology, ecology, and development into their algorithms.

The answers to these puzzles may surprise you, and some remain hotly contended. "People are still arguing many of the same arguments that they had before all of the data were there," Doolittle says. But if there is one thing evolutionary biology needs, it's a few renegades who aren't afraid to challenge the status quo. Not all their revolutionary ideas will hold up to scrutiny, but those that do could forever change our understanding of life itself.




***Tripartite Life.*** *In 1977, Carl Woese first proposed the radical idea that three domains of life arose from a common ancestor. He inferred evolutionary relationships by comparing sequence similarities in ribosomal RNAs across multiple organisms. His three-branch tree of life is now widely accepted. Created by Maulucioni from figure 1 in Woese CR, Kandler O, Wheelis M (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya,"* **Proc Natl Acad Sci USA** *87, and made available through the Wikipedia Commons at http://en.wikipedia.org/wiki/File:Phylogenetic-Tree,_Woese_1990.png.*

## How did life begin?

Life on Earth began about 3.8 billion years ago. But exactly how this happened—how non-living chemicals transformed into organic building blocks and then living cells—remains a mystery.

Phylogenetics can only answer questions about what happened more recently than about 3.5 billion years ago, when the last universal common ancestor (the primitive cells that gave rise to bacteria, archaea, and eukaryotes) first appeared on Earth. "The earliest thing you're ever going to see by direct sequence analysis is already an incredibly complicated organism. It had a lot more than DNA; it had RNA, proteins, RNA machinery, transport, homeostasis, and bioenergetics," says **Eric Smith, PhD**, an external professor at the Santa Fe Institute. "So you have to dig back way further than that in time."

Experimental scientists have established several different scenarios for how organic molecules might have first appeared on Earth. For example, in 1953, **Stanley Miller** and **Harold Urey** famously created a "primordial soup" of amino acids by passing electricity (simulating lightning) through an airtight flask of water plus methane, ammonia, and hydrogen gases (which they believed, at the time, to be present on early Earth). After organic molecules first appeared, however, it is unclear how they joined together to build the basic machinery of life.

Some scientists have proposed that "autocatalytic sets"—groups of molecules capable of producing each other through mutual catalysis—were necessary to get things going. But others have argued that autocatalytic sets could not have arisen spontaneously. "Some say it's equivalent to a tornado blowing through a junkyard and randomly assembling pieces of metal and plastic into a Boeing 747," says **Wim Hordijk, PhD**, a computer scientist and owner of SmartAnalytix.com.

So, Hordijk and his collaborator **Mike Steel, PhD**, professor of mathematics and statistics at the University of Canterbury in New Zealand, decided to actually calculate the probability. "Nobody had ever looked at this in a concrete mathematical way," Hordijk says. "So that is what we've done. We proved mathematical theorems about it and ran computer simulations." The mathematical framework integrates probability theory and graph theory—with molecules as nodes and interactions between them as edges in the graph.

In a 2012 paper in *Acta Biotheoretica*, they showed that autocatalytic sets do appear spontaneously with high probability. "In this simple model of a chemical reaction system where you have polymers floating around that could be glued together or broken apart and can do catalysis, it's actually very likely that you will get these autocatalytic sets," Hordijk says. Plus, smaller autocatalytic sets can team up together. "The smaller ones can grow into bigger ones. That's necessary to get some sort of evolutionary process going," Hordijk says.

It's difficult to create and study autocatalytic sets experimentally. But, in an October 2012 paper in *Nature*, experimentalists reported that small RNA molecules can spontaneously form a cooperative self-replicating network. Next, Hordijk and Steel simulated that system using their model, virtually replicating the experimental results, as published in a 2013 paper in the *Journal of Systems Chemistry*. They also made new predictions about the behavior of the system that the experimentalists are now testing. "The hope is that by doing these computer simulations, we can actually guide the experimentalists," Hordijk says. This particular experiment started with already assembled RNA, so it doesn't answer the question of how RNA formed in the first place, he notes.

Also, Hordijk and Steel's model makes no assumptions about the type of molecule involved; their mathematical framework could just as easily be applied to proteins or metal complexes. So, it doesn't answer the question of which type of molecule got life going.

## Which came first: RNA or proteins?

Nucleic acids store the information that is needed to make proteins, but proteins are the workhorses that allow nucleic acids to replicate. So, scientists have long puzzled over which came first. In the 1980s, with the discovery that RNA can both store information and catalyze reactions, many scientists believed they had the answer: RNA came first (note that DNA is a more stable molecule believed to have evolved from RNA). The "RNA world" hypothesis—which purports that RNA got things going and was gradually replaced by proteinaceous enzymes and DNA—still prevails today. "I still accept the idea of an RNA world as real," Doolittle says. "There are RNA surrogates for many proteins. RNA could have easily been the intermediate that was gradually replaced by proteins."

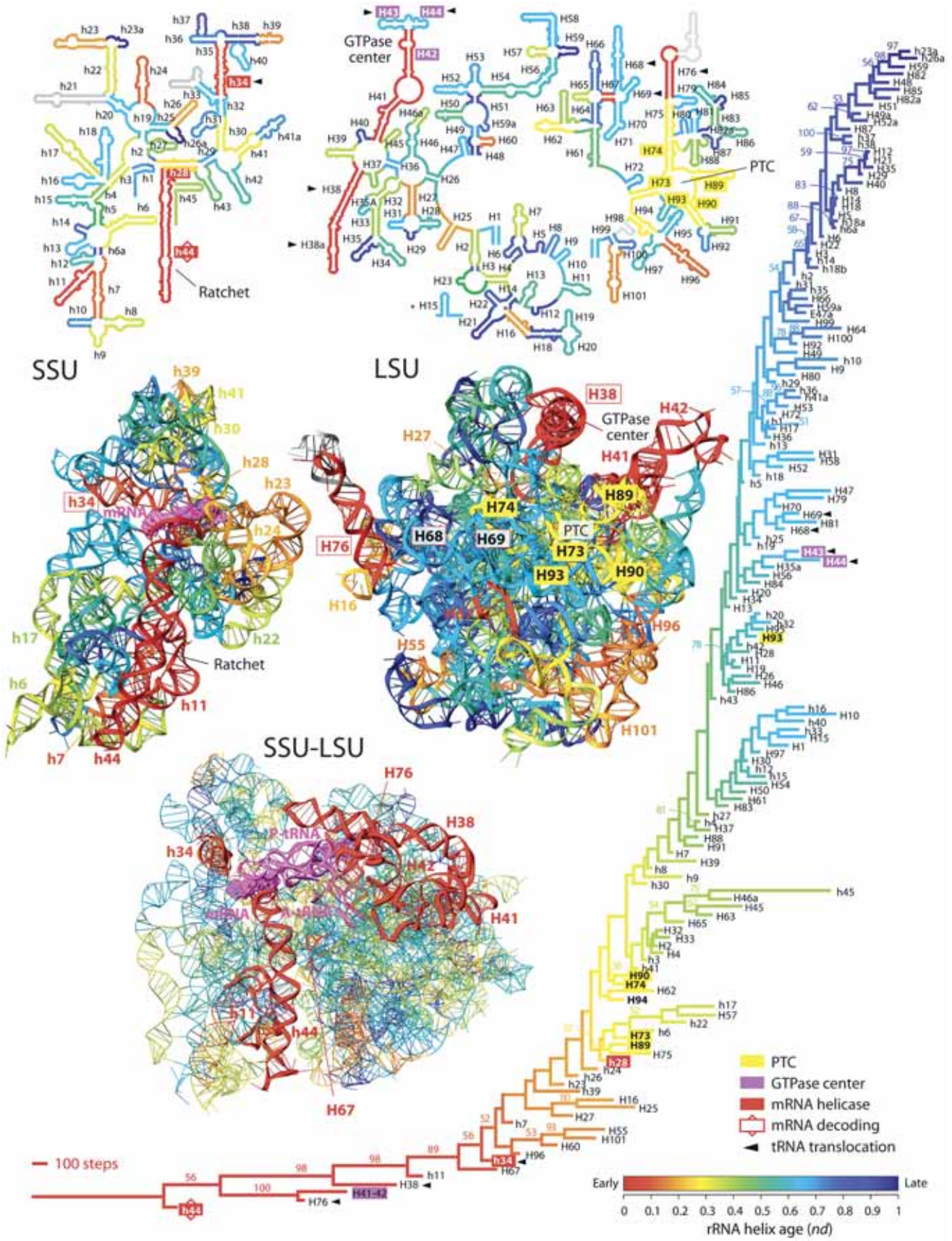
But not everyone is convinced. For one thing, no one has ever synthesized ribose, the sugar backbone of RNA, in abiotic conditions, says **Jean-Michel Claverie, PhD**, professor of medical genomics and bioinformatics at the University of the Mediterranean in France. "I'm not in that field, but I had to review a book about the RNA world. And this is when I realized how weak the evidence is," he says. "The existence of an RNA world, although it would make a lot of sense and would elegantly explain the central role of the ribozyme in protein synthesis, is still not founded on anything solid."

In a 2012 paper in *PLoS ONE*, **Gustavo Caetano-Anollés, PhD**, professor of crop sciences at the University of Illinois at Urbana-Champaign (where Woese once worked), and his colleagues challenged the RNA world hypothesis. Caetano-Anollés builds evolutionary timelines by looking for similarities across organisms in 3-D structures—RNA secondary structures and protein folds—rather than in genetic sequences. "I have always been suspect of explorations that come from sequence and target very deep evolutionary divergence," Caetano-Anollés says. "How can people make judgments about what happened so far back in time with something that is changing so incredibly fast?" Structures change at a much slower pace than sequence; structure comparisons are also much less sensitive to messy evolutionary phenomena such as convergent evolution (independent evolution of similar features) and horizontal gene transfer (the exchange of genes between unrelated organisms), he says.

"I have always been suspect of explorations that come from sequence and target very deep evolutionary divergence," Caetano-Anollés says. "How can people make judgments about what happened so far back in time with something that is changing so incredibly fast?"

His team traced the evolutionary history of the ribosome using data from the SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology, Homology) protein structure classification databases (which group proteins into fold groups). They computationally compared ribosomal protein and ribosomal RNA structural domains across nearly 1000 organisms, including bacteria, archaea, and eukaryotes. The idea: structural domains that are the most universal are the oldest, whereas domains that appear in only a few organisms are the youngest. RNA-world proponents believe that the first ribosomes were composed solely of RNA. But Caetano-Anollés' team found that ribosomal proteins are just as old as ribosomal RNA; and that the two evolved together. Thus, early Earth was in fact a ribonucleoprotein world, he says. "My stance may not be popular with those who focus on sequence. However, structural genomic data have been analyzed and the interpretation is against the RNA world."

"Caetano-Anollés has certainly done some provocative stuff," Doolittle comments. "I think he's been mistaken about some of it, but his approach is so refreshing that I read all of his work, even though I'm skeptical of some of his conclusions." Doolittle wonders how the proteins could have been propagated without a memory system. "You can't have all the information for a particular kind of fold passed on from




***A Structural History of Life.*** *Gustavo Caetano-Anollés builds evolutionary trees based on the 3-D structures of RNAs and proteins. This tree reconstructs the evolutionary history of ribosomal RNA helices. The oldest structures are in red and the youngest structures are in blue. Similar analyses of ribosomal proteins (not pictured here) suggest that ribosomal proteins and ribosomal RNAs coevolved, refuting the idea that RNA appeared on Earth before proteins (the so-called "RNA world hypothesis"). Reproduced from figure 2 of: Harish A, Caetano-Anollés G. Ribosomal History Reveals Origins of Modern Protein Synthesis.* **PLoS ONE** *2012; 7: e32776.*

one generation to another until you can explain how this information is stored. At the moment, that's still a fatal flaw," Doolittle says.

But, in a 2013 study in *PLoS ONE*, Caetano-Anollés' team provides evidence that protein synthesis occurred before there was a memory system (before there was a genetic code or ribosomes). "The ancestors of synthetases [the enzymes that load amino acids onto transfer RNA], are responsible for the specificity of the genetic code," Caetano-Anollés says. During transcription in the ribosome, tRNA molecules bring the correct amino acid into the growing protein sequence by matching their three-letter anticodons to codons in the messenger RNA. Synthetases contain two types of domains: those that perform the loading and those that read the anticodon to determine exactly which amino acid to load. Caetano-Anollés team found that the former are more ancient than the latter; this and other evidence suggest that these enzymes were originally involved in non-ribosomal protein synthesis. The genetic code only arose later, likely as a way to improve protein flexibility and function, Caetano-Anollés says.

## Or did metabolism come first?

Smith also disputes what he terms the "radical RNA-first view." Though life may have gone through a stage in which RNA was the main molecule of both heredity and catalysis, he doesn't believe that RNA was the first mover. Rather, he says, metabolism began as a system of chemical reactions that did not involve RNA. Early metabolic networks could have arisen spontaneously and been catalyzed by minerals or perhaps simple small-molecule/metal complexes. "For early chemistry, we're not looking for something that undergoes Darwinian adaptation, because the early chemistry is universal stuff that's never changed. We're just looking for stuff that will transduce energy, fix carbon, do the same things over and over again, and provide an ordered framework—out of which more molecular complexity comes later," Smith says.

In the metabolism-first view, the chemistry that eventually became life must have included methods for carbon fixation—converting inorganic carbon to organic carbon. Two carbon fixation pathways—the reductive citric acid cycle (also known as the reverse Kreb's cycle) and the Wood-Ljungdahl pathway—are believed to be the most ancient. But scientists have long debated which of these two evolved first.

Smith tackles these types of history-of-life questions by focusing on chemistry. "When you talk about the low-level chemistry, you don't need to refer to the genomic era of modern cells because whatever preceded them was also using low-level chemistry," Smith says. The metabolic networks that organisms use today are highly conserved. For example, modern autotrophs (organisms that can fix carbon and thus make their own food) use one of only six different pathways for carbon fixation. "This suggests that even the long-range evolution of complicated organisms has been strongly constrained by the principles of very low-level chemistry," Smith says.

To reconstruct the evolutionary history of biological carbon-fixation, Smith teamed up with **Rogier Braakman, PhD**, a fellow at the Santa Fe Institute. Braakman developed a novel computational technique called phylometabolic reconstruction, which integrates phylogenetics with flux-balance analysis, a type of metabolic analysis. In flux-balance analysis, researchers derive a series of equations to represent all the inputs and outputs in a metabolic network; then they simulate the flux of metabolites through this network, assuming constraints such as conservation of energy and mass. Braakman and Smith added the further constraint that early life must have been self-sufficient—able to make all its own building blocks. This limit confines the sequence-based phylogenetic reconstruction to a set of allowed configurations. "What we're doing here is saying: One thing that we know about autotrophs is that they made everything that they needed."

"For early chemistry, we're not looking for something that undergoes Darwinian adaptation, because the early chemistry is universal stuff that's never changed. We're just looking for stuff that will transduce energy, fix carbon, do the same things over and over again, and provide an ordered framework—out of which more molecular complexity comes later," Smith says.

Their paper, published in *PLoS Computational Biology* in 2012, surprisingly concluded that neither the reductive citric acid cycle nor the Wood-Ljungdahl pathway evolved first; instead, primordial life contained both pathways. This redundancy may have

been an important failsafe since early life forms were probably fragile, Smith explains. Braakman and Smith also showed that further innovations in carbon-fixation were driven by the invasion of specific chemically novel environments (e.g., alkaline or oxidizing environments) more than by chance innovations in the genome.

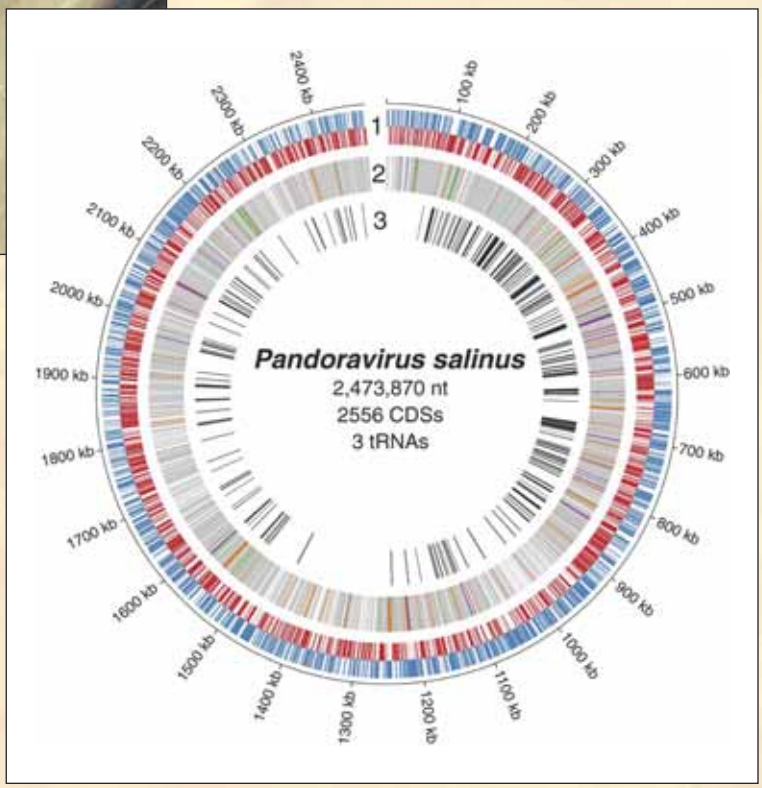
## Is there a fourth domain of life?

When it comes to reconstructing the history of life, viruses have traditionally been ignored. After all, it's not clear that viruses are even alive, given their lack of a cellular structure and dependence on cellular organisms. But with the recent discovery of giant viruses—which are as large and complex as some bacteria—viruses have suddenly taken center stage in evolutionary debates. Some researchers even argue that viruses comprise a fourth domain of life.

In 2003, French scientists identified the first giant virus, which they named Mimivirus, short for "microbe mimicking virus." The Mimivirus, which infects amoeba, can be seen under a light microscope and has more than 1000 genes, including some involved in protein translation and metabolism (hallmarks of cellular organisms). "This was a challenge for the classic paradigm of viruses," says Claverie, who was involved in the discovery. Since then, Claverie's team has uncovered several other giant viruses, including the Megavirus in 2011 and the most perplexing, the Pandoravirus, in 2013. The genome of Pandoravirus is twice as large as that of other giant viruses; and 93 percent of its genes resemble nothing ever sequenced before.

Mimivirus and Megavirus share certain protein translation genes, but are also highly genetically distinct. Claverie's explanation: giant viruses descended from an ancient, cell-like common ancestor (one that has no modern cellular descendents). Over time, they lost genes and became parasitic. "We believe: the bigger the viral genome, the closer you are to the origin," Claverie says. In phylogenetic reconstructions, Mimivirus and Megavirus wind up either at the base of the eukaryotic branch of life or on a completely new branch distinct from eukaryotes, archaea, and bacteria. Pandoravirus is so dissimilar to any known organism on Earth that its existence also challenges Woese's tripartite tree of life. "It is an increasingly complicated story," Claverie says.

Others strongly dispute this view, however. They believe that giant viruses are the ultimate gene robbers, and that their genomes are growing rather than shrinking. Giant viruses could have picked up their large and crazy genomes through horizontal gene transfer with their amoebal hosts (or other amoebal parasites). These looted genes may then

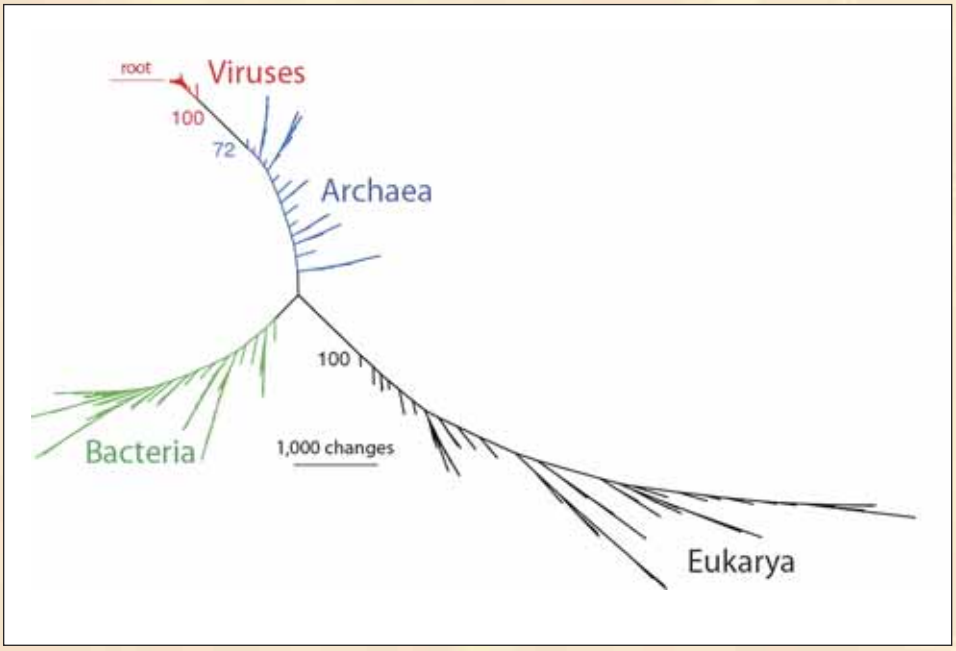



***World's Biggest Virus.*** *Electron microscope image of* **Pandoravirus salinus** *(above) and a diagram of its genome (right). With nearly 2.5 million nucleotides (nt's), the genome of Pandoravirus is as large as some eukaryotic cells and twice as large as any other known virus on Earth. But 93 percent of its genes resemble nothing ever sequenced before—opening up a Pandora's box of questions about the history of life. In the genome picture, CDS=putative protein-coding sequences; CDSs on the direct (blue) and reverse (red) strands of DNA are indicated in the outermost circle. In circle 2, CDSs that match known genes or motifs are indicated in orange, green, purple, and white; CDSs with no match are shown in gray. Photo courtesy of: Chantal Abergel and Jean-Michel Claverie. Genome picture reproduced with permission from: Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes.* **Science** *19 July 2013; 341:281-286.*

have evolved rapidly within the viruses, creating their puzzling genetic diversity. Using alternate models that account for such possibilities, other research groups have published phylogenetic reconstructions that place giant viruses squarely within the three domains of life, next to their ameobal hosts.

But Claverie isn't convinced by these arguments. "The thing is, if those viruses are picking up genes from the environment, where are those cells? Because what has characterized those new viruses that we keep sequencing is that they don't look like anything else," he says. "They appear to steal genes from cells we haven't sequenced yet. And I don't think many people are prepared to believe that there is such a big loophole, such a big [set of] missing data."

Phylogenetic reconstructions are highly sensitive to models and assumptions, especially when dealing with viruses, as this debate reveals. But Caetano-Anollés also performed a structural reconstruction of evolutionary history including giant viruses and other DNA viruses. And, like Claverie, he found that viruses clustered into a separate domain of life that either predated or coexisted with the last universal common ancestor. "Until now, the universal tree is a tree of cellular lineages, not a tree of everything. From my point of view, that's an omission," Caetano-Anollés says.




***A Fourth Domain?*** *The discovery of giant viruses has raised the possibility that viruses comprise a fourth domain of life. Gustavo Caetano-Anollés' team built this evolutionary tree by comparing protein fold structures from the proteomes of archaea, eukarya, bacteria, and viruses/giant viruses (50 organisms each). They conclude that viruses are a distinct form of life that either predated or coexisted with the last universal common ancestor. Reproduced from: Nasir A, Kim KM, and Caetano-Anollés G. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya.* **BMC Evolutionary Biology** ***2012, 12:156.***

"Until now, the universal tree is a tree of cellular lineages, not a tree of everything [including viruses]. From my point of view, that's an omission," Caetano-Anollés says.

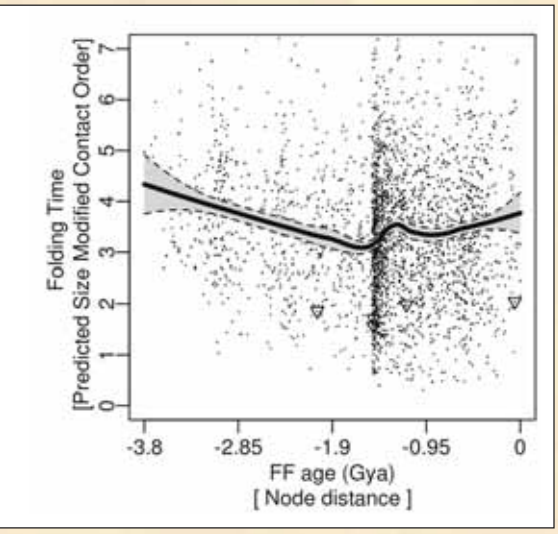
## How have proteins evolved since life began?

The earliest proteins to evolve were likely versatile but not optimized. Many researchers are trying to understand how proteins became optimized over the course of evolution. For example, what drove the evolution of different protein folds and of multidomain complexes?

**Frauke Gräter, PhD**, an expert in protein folding, has long wondered about the evolution of folds. Her team made use of a model for predicting protein folding times for all proteins structurally known to date, based on the distance between contact points—amino acids that touch in the folded molecule—in the unfolded sequence. Contact points that start farther apart take longer to come together. To add an evolutionary perspective to this concept, she teamed up with Caetano-Anollés. "His way of mapping proteins structures on a timeline from four billion years ago to today was exactly what was needed to combine with our proteome-wide prediction of folding times," says Gräter, who is a group leader at the Heidelberg Institute for Theoretical Studies in Germany.

In a 2013 paper in *PLoS Computational Biology*, Gräter and Caetano-Anollés showed that protein folding became progressively faster from 3.8 billion to 1.5 billion years ago. (After this, alpha but not

beta folds continued to fold faster.) "Proteins were apparently folding faster and faster for most of the time during evolution. So there was pressure for efficient folding over time," Gräter says. Faster protein folding likely prevents diseases that are caused by protein misfolding and aggregation, such as Alzheimer's, she explains. "Once proteins are in their native fold, they are not prone to aggregation anymore." Her team is now exploring evolutionary trends in other protein properties, such as floppiness and mechanical stability.




***Folding Faster and Faster.*** ***By coupling a computational model that predicts protein folding times with a structural reconstruction of the history of different folds, Frauke Gräter's team was able to trace how protein folding times have changed since the beginning of life. They found that protein folding became progressively faster from 3.8 billion to 1.5 billion years ago, at which time there was an explosion in protein fold diversity. After this, alpha folds continued to fold faster, but beta folds did not. Reproduced from: Debès C, Wang M, Caetano-Anollés G, Gräter F. Evolutionary Optimization of Protein Folding.*** **PLoS Comput Biol** ***2013; 9(1): e1002861.***

To achieve complex functions, proteins have evolved to work in multi-domain complexes that assemble after protein translation. **Sarah Teichmann, PhD**, program leader in genome evolution at the EMBL-European Bioinformatics Institute and Wellcome Trust Sanger Institute in the United Kingdom, wondered if the order of assembly is under selective pressure.

To test this theory, her team first developed a mathematical model that predicts the order in which protein complexes assemble based on 3-D structures and the surface area at the interfaces of different subunits. Then they looked for gene fusion events between genes encoding different subunits of the same protein complex. A gene fusion occurs when separate genes are shuffled into the same open reading frame, and thus become translated together in the order in which they appear in the genome. Teichmann reasoned that if the order of assembly of protein complexes is under selection pressure, then only certain gene fusions—those that preserve this order—would be favored in evolution. "The neat computational trick here is that we are combining the structural bioinformatics with genomics. We go from the 3-D protein level to the 2-D genomic arrangement," she says.

Indeed, she showed that fusion events that preserve the mathematically predicted order of assembly appeared statistically more frequently in the genome than those that did not. The results were published in *Cell* in 2013. "It's intuitive in the sense that you want to have the subunits of a protein complex find each other quickly; you don't want to have them floating around the cell in an unbound state for a long time," she says. Unbound proteins could aggregate and cause disease.

## Why did introns evolve?

One of evolution's biggest puzzles is the intron. These extra pieces of DNA interrupt genes and have to be spliced out before protein translation. When, why, and how did they evolve in the history of genes?

The question of "when?" has largely been solved, says **Scott Roy, PhD**, assistant professor of cell and molecular biology at the University of California, San Francisco. Though a few reputable naysayers argue that introns are as old as the genetic code itself (and helped make genes possible), "the consensus perspective is that a large number of introns arose for the first time in the last common ancestor of all eukaryotes," Roy says. This would be about 1.5 billion years ago.

More perplexing is the why question. In higher eukaryotes such as humans, introns help create protein diversity through alternative splicing to produce more than one protein from a gene sequence. But until recently, scientists believed that alternative splicing was rare in lower eukaryotes and thus couldn't be their *raison d'être*. "That turns out to be at best a gross simplification and in some cases just completely wrong," Roy says. For example, recent microarray analyses showed that almost all of yeast's 200 intron-containing genes are alternatively spliced, Roy says.

His team is hunting for examples of functional, evolutionarily conserved alternative splicing in fungi. Functionally important variants may represent only a fraction of transcripts, "so you have to sequence the heck out of the transcriptome," Roy says. Analyzing the data is a major computational challenge because the transcripts have already had the introns removed, and the algorithm has to guess where these splicing events happened. "You get these short reads—about 100 nucleotides. And then you have this huge genome and you need to figure out where does this 100 nucleotide read come from in the genome," Roy says. "There are a lot of programs out there that do it, but they're not very consistent." His team uses multiple programs as well as in-house software to arrive at a consensus.

They have found some alternative splicing events that appear to be conserved over long timescales and in different species; but "it remains to be seen whether it's true conservation or just coincidence," he says. "I don't even know where my money is at this point. Which is exciting, actually," he says.

The purpose of introns may also be related to the 3-D genomic architecture of eukaryotes, says **Liya Wang, PhD**, a research scientist at Cold Spring Harbor Laboratory. In eukaryotes, DNA is organized into nucleosomes: 140-base-pair stretches of DNA are coiled around proteins called histones. The DNA coiled around a histone is more likely to be an exon than an intron, suggesting that this 3-D structure helps to prevent introns from interrupting a functional stretch of DNA, Wang explains.

To study the mechanisms of intron gain and loss, he and **Lincoln D. Stein, PhD**, program director of informatics and bio-computing at the Ontario Institute for Cancer Research and a professor at Cold Spring Harbor, came up with a computational model that could recreate the distribution of exon sizes for the genomes of 14 different species. Surprisingly, their model predicted that the probability that an

exon will gain an intron is proportional to its size to the third power, suggesting a 3-D volumetric relationship rather than one based just on sequence. "One hypothesis is when the introns try to attack, they are attacking a ball that the exon occupies by its dynamic motion; the larger the ball, the higher the chance," he says. The results were published in *BMC Evolutionary Biology* in 2013.

Wang and Stein are now modeling whether CG content (the frequency of cytosine/guanine nucleotide pairs, which is related to methylation), also affects intron insertion. Their work reflects a growing recognition of the importance of higher-order features, such as epigenetics and morphology, in shaping evolution.

## What drives the evolution of form?

The first multicellular organisms appeared about 565 million years ago, followed by an abrupt explosion of body plans from about 550 to 530 million years ago (visible in the fossil record). Nearly all modern shapes appeared then; and there have been few innovations since. This observation has long puzzled scientists; how could gradual, Darwinian evolution result in such rapid changes in form?

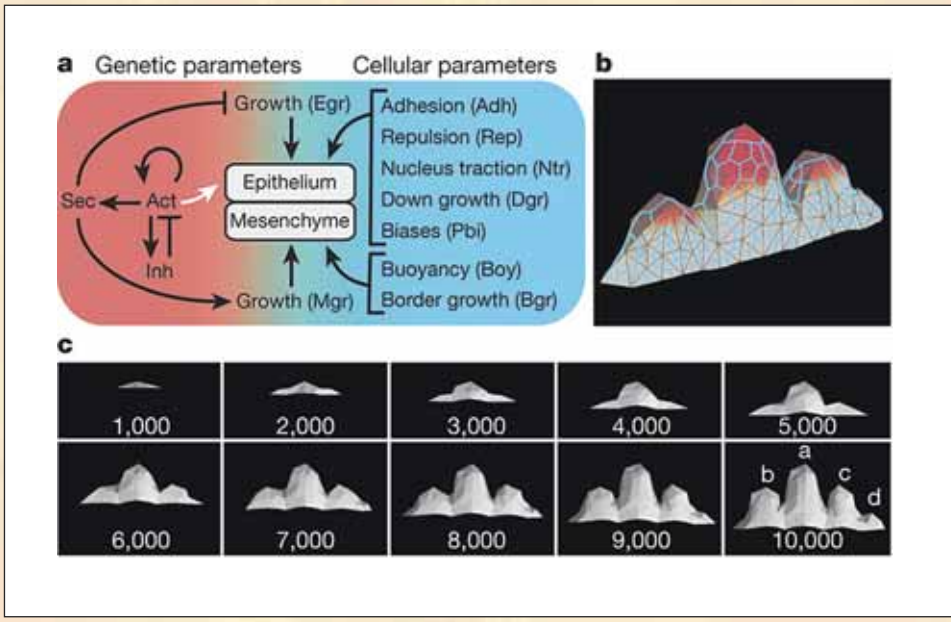
**Stuart Newman, PhD**, professor of cell biology and anatomy at New York Medical College, believes that the answer lies in physics. In a 2012 paper in *Science*, Newman argues that genes that evolved for other purposes in unicellular organisms (such as those for adhesion), suddenly found new roles in the physical landscape of multicellular organisms. "You have a way through physics of generating radically new forms by very small genetic changes," he says.

"If you look at the logic of the Darwinian perspective, it says you can't have abrupt change. But this is a 19th century view. We now know with 20th century advances in the physics of materials that things like tissue masses can change abruptly and discontinuously," he says. Physical laws also limit what morphological motifs are possible, which explains why there's been little diversification in form in the past half billion years.

Newman's team simulates limb development using a finite-element model. When they virtually evolve limbs, they end up with a variety of shapes that never existed in any animals, but that still resemble natural limbs. "So there's both a great plasticity but then there's also a constraint in that. With the Darwinian paradigm you can in principle get from anywhere to anywhere by adaptation, but this kind of mathematical modeling approach shows that there are really deep constraints in the kinds of forms you can come up with. You can't get just anything."

**Isaac Salazar-Ciudad, PhD**, a senior researcher at the University of Helsinki and the Autonomous University of Barcelona in Spain, also looks beyond Darwin to study the evolution of form. His team has developed a computational model of tooth development. "We have a set of cells and those cells have genes inside; those genes affect each other in gene networks," Salazar-Ciudad says. "Then at the same time, those cells are actually moving and interacting mechanically with each other." This is one of the first models to combine these two components, he says.

In a 2013 paper in *Nature*, Salazar-Ciudad used his model to explore the relationship between genotype and phenotype in the evolution of morphology. He virtually evolved teeth by gradually mutating them, and then explored the resulting 3-D phenotypes. "We found that the mapping between genotype and phenotype is so complex that natural selection cannot fine tune every aspect of morphology," he says. "We say that natural selection is indeed acting all the time and it is very important, but there is a restriction on what kinds of things it can do."




***A Model with Teeth.*** *Isaac Salazar-Ciudad's team created a morphological model of seal tooth development and evolution. Panel (a) shows the cellular and genetic parameters included in the model. Panel (b) shows how tissue morphology is modeled in three dimensions; cells are allowed to move and interact with each other, creating shape. Panel (c) shows how the tooth shape evolves from the initial conditions until 10,000 time points. Salazar-Ciudad's uses the model to study the evolution of teeth as well as their development. Reproduced with permission from: Salazar-Ciudad I, Jernvall J. A computational model of teeth and the developmental origins of morphological variation.* **Nature** *2010; 464: 583-586.*

## And more puzzles remain...

History-of-life puzzles spur passionate debate precisely because the scientific questions are so tied to existential ones—who we are, where we came from, why we're here. But answering these questions isn't just about satisfying deep-seated human curiosity; it's also about practical ends. "Obviously there's just a big curiosity behind it. People want to know where did we come from, where did it all start?" Hordijk says. "But, besides that, I think great medical things will come out of this. If we understand how life started, that automatically gives us a better understanding of how life works. That will certainly have a lot of important medical implications." □