Reconstructing Ancestral Proteins

Chase Nelson

Lucas Wheeler and Michael Harms, "Were Ancestral Proteins Less Specific?" *Molecular Biology and Evolution* 38, no. 6 (2021): 2,227–39, doi:10.1093/molbev/msab019.

Members of a family resemble one another. Such similarities are most often noticed in contemporaries, as when siblings have the same smile. But they also provide insight about individuals long departed. Absent a historical record, the best bet for inferring the traits possessed by a forebear is to examine all their descendants in the present. It matters neither how long ago the ancestor lived nor how copious their progeny; if 90% of the offspring are tall, it seems reasonable to suppose the primogenitor was too. As the aphorism suggests, the present is the key to the past.

What is true of people is also true of proteins, and for the same reasons. Proteins are molecular machines built from strings of amino acids, typically a few hundred residues in length, chained together with covalent bonds like beads on a necklace. Most organisms employ an alphabet of twenty amino acids in their proteins, with an overall function defined by their precise order—leucine at position 5, lysine at position 26. If two proteins contain identical amino acids at a sufficient number of positions, common ancestry can be inferred and they can be grouped into a family. If every currently known family member contains, say, lysine at position 5, chances are high that the same was true of the ancestral protein.

But proteins aren't always what they used to be. This is true in the sense that one amino acid occasionally gets swapped for another over the course of evolution. But it is also true in terms of general properties. Each amino acid has its own size, hydrophobicity, and charge. As a result, neighboring residues along the chain undergo biophysical interactions, giving rise to a series of twists and turns. The chain progressively folds into a particular three-dimensional shape and thereby assumes a particular function. Swapping an amino acid can disturb this shape and thereby alter the function, most commonly by changing the suite of molecules with which the protein has the ability to interact.

Inferring the exact amino acids in ancient proteins has long been a topic of interest among theoreticians. Writing in 1963, Linus Pauling and Emile Zuckerkandl introduced the concept of chemical paleogenetics. They reasoned that sequencing enough proteins in the present could allow sufficient numbers to be grouped into families, thus allowing their extinct protein ancestors to be inferred and synthesized in the lab.[1] These proteins could then be used to test a wide range of evolutionary hypotheses about ancient organisms, even those that left no trace in the fossil record. The field blossomed in the 1980s with leaders such as Steven Benner, followed later by Gina Cannarozzi, Betül Kaçar, Joseph Thornton, and many others.

An influential example of ancestral protein reconstruction, conducted by Eric Gaucher and colleagues, involved bacterial elongation factors of the Tu family.[2] Focusing on bacteria that are currently mesophilic—organisms that grow optimally at temperatures between 20 and 40°C—these researchers found the ancestral protein to function optimally between 55 and 65°C, suggesting that the ancestor was instead thermophilic. This example highlights the necessity of protein reconstruction: it may not be enough to examine the average properties of a protein's modern descendants. The ancestral molecules themselves must be constructed.

Other applications of ancestral protein reconstruction have involved testing general hypotheses about genome evolution and evaluating specific gene functions to gain a better understanding of the connection between genotype and phenotype.[3] Proteins can even be compared directly to their ancestors over an evolutionary tree, circumventing the historical, and therefore statistical, dependencies that have long plagued comparisons among modern forms alone.[4]

Given that proteins are associated with repertoires of target molecules, a particular protein can be categorized as relatively specific, having few targets, or general, having many targets. From a theoretical standpoint, the existence of an overall historical trend in the direction of, say, general-to-specific could allow a

principle of protein evolution to be established. Indeed, Roy Jensen proposed just such a principle in 1976, reasoning that early cells must have possessed a small collection of genes encoding proteins with relatively general activities.[5]

Protein specificity may seem a mere theoretical curiosity, but it is perhaps the most far-reaching line of research to spring from ancestral protein reconstruction and for one lucrative reason: protein engineering. A protein binding its target is not the end but the beginning of its functional story. The binding of a single molecule might set off a change in the protein's conformation that ultimately stretches the target until it rips in two. Or, the binding of multiple molecules might stimulate the protein to smash them together until a new product is forged. In either case, the binding of targets by proteins speeds up perhaps every chemical reaction carried out by the cell.

If ancestral proteins tended to be more general than their modern forms, they would have been capable of binding a wider range of targets and driving a greater diversity of potential reactions. They might also have exhibited other valuable features, including increased stability at high temperatures and the ability to fold without the aid of modern so-called chaperone proteins.[6] In such cases, efforts to engineer new protein applications would be much better spent investigating variations on ancestral proteins instead of their more limited descendants. There have indeed been glimmers that ancestral proteins were more general.[7]

In a 2018 study, a team including evolutionary biochemists Lucas Wheeler and Michael Harms addressed the question of protein generality by focusing on the S100 protein family, so named due to their solubility in 100% saturated ammonium sulfate.[8] This family comprises just over twenty members in the human genome, most of them encoded by a cluster of genes located on the long arm of chromosome 1. For the vast majority of S100 proteins, the binding of Ca^{2+} ions transiently exposes a hydrophobic binding pocket which, if occupied by a well-fitting target molecule, forms a stable structure that holds both the Ca^{2+} and the target in place.9 Members of the S100 protein family function as calcium sensors, maintaining homeostasis by responding to intracellular Ca^{2+} fluctuations in real time. Many play key roles in cell replication and immune detection.[10]

Wheeler and colleagues focused on just two S100 family members: S100A5 and S100A6. This pair is thought to have resulted from the duplication of a single ancestral gene, dubbed ancA5/A6, around 318 million years ago in the common ancestor of mammals, birds, and reptiles.[11] The initial result was two identical and therefore redundant copies, but such tandem gene duplication events eventually allow each descendant gene to specialize in just a subset of the ancestor's tasks.[12] Both S100A5 and S100A6 are known to be overexpressed in certain cancers and to bind some of the same targets, but have slightly different, if poorly defined, functions.[13]

S100 is a particularly well-chosen family for investigating the evolution of protein specificity: this family binds a set of targets so diverse that no one pattern or motif can be used to summarize them. It can at least be said that the targets are typically short protein fragments, or peptides, twelve or more amino acids in length and occurring within larger proteins. Binding probably occurs through a combination of shape complementarity and hydrophobic interactions.[14] Known peptide targets include regions of the proteins sodium/calcium exchanger 1 (NCX1), Siah-interacting protein (SIP), and two commercially available peptides. Focusing on just these four targets, the team first determined which were bound by each human S100 copy: S100A5 bound the two commercial peptides and NCX1, but not SIP, while S100A6 bound only one commercial peptide and SIP, but not NCX1 or the other commercial peptide. The team then used maximum likelihood methods to reconstruct the ancestral ancA5/A6 protein and assess its own binding: remarkably, all four targets were bound by the ancestor.15 It seemed that in the time since their duplication from ancA5/A6, S100A5 and S100A6 had partitioned binding partners as one might split belongings after a divorce. What was one had become two, and neither quite amounted to what it once was. These results hinted at a trend toward increasingly specific proteins in evolution.

Despite these results, a sample size of four targets was a limited basis on which to draw any conclusions. As the authors themselves cautioned,

> [I]t could be that the proteins both acquired more peptides that we did not sample ... while becoming more specific for the chosen set of targets. ... Particularly given the large number of targets for these proteins, distinguishing these possibilities will require an unbiased, high-through[p]ut approach to measuring specificity.[16]

In a new study, Wheeler and Harms increase the sample size of targets from four to approximately 100,000.[17] To do this, they first note that S100 proteins might conceivably bind peptide targets not produced by a present-day organism. Thus, rather than limiting to known interaction partners, they instead examine a random sample drawn from all 20^{12}—that is, 4.096×10^{15}—possible peptides that are twelve amino acids in length. Their final sample is a commercial library of some 10^9 unique peptides.

Not each of the 10^9 potential targets is actually bound by S100A5 or S100A6. To home in on those that are, the authors turn to quantitative phage display, in which gene fragments encoding random peptide targets are inserted

into the surface protein gene of bacteriophages. Each phage expresses one target. The whole pool of phages is then mixed with S100 to allow binding, after which the S100 is isolated. Those phages expressing a target bound by an S100 will remain stuck, while the others are washed away. Finally, the bound phages undergo sequencing of their genes en masse to identify which 12-amino acid peptides were bound. Another experiment is run in parallel with a competitor peptide already known to occupy S100's binding pocket, allowing the results to be normalized in the event that some of the targets bind indiscriminately to other parts of the S100s.

The results are something of a surprise: instead of increasing specificity with fewer targets, the authors observe a pattern of shifting specificity. S100A5 and S100A6 still bind just a subset of the ancestor's targets; but, when assaying this large sample, it becomes clear that the loss of old targets is offset by a gain of new targets. The specificities of the S100s have not, in fact, changed in comparison to their ancestor. Using alternative, and less likely but still plausible, ancestors yields generally consistent results.[18]

Limiting studies to a handful of known interaction partners had given the wrong picture. To more accurately characterize a protein's intrinsic specificity, unbiased sampling of targets—including those not known in modern organisms—is necessary.

The study of molecular evolution seeks to reveal how the building blocks of organisms have changed over time, both in pattern and in process. Wheeler and Harms have charted new territory in describing the pattern of evolution at the protein level, albeit for a single pair of closely related proteins. Can any implications be drawn from their work about what drives protein evolution?

One of the most impressive answers to this question comes from a 2008 study by Shozo Yokoyama and colleagues.[19] These researchers resurrected the protein ancestors and intermediates of thirty-eight modern-day rhodopsins, visual pigments used by vertebrates to see dim light. Such light is most prevalent at dusk and has a wavelength between 400 and 500 nanometers (nm), with lower wavelengths penetrating to deep sea environments. Because the wavelength of maximal absorption, λ_{max}, can be easily assayed in the lab, its biological function is straightforward to assess quantitatively. As might be expected, deep-sea fish tend to have rhodopsins with a lower λ_{max} than shallow water fish and terrestrial animals, ~480 nm versus ~500 nm.[20] The researchers therefore sought to characterize each amino acid change leading to these different λ_{max} values.

Across the rhodopsin's length of 354 amino acids, 203 positions underwent an amino acid change but only twelve of these affected λ_{max}. The team's more recent work on visual pigments yields similar results: "[T]he effects of the small proportion of adaptive sites on the evolutionary rates are buried among those of the neutral changes."[21] Most amino acid changes that are accepted during visual pigment evolution have little or no discernible effect on function. Statistical methods of sequence analysis generally fail to predict those that do.[22] With notable exceptions, the same is true of most proteins—neutral changes dominate.[23]

The researchers made another stunning observation: similar shifts in λ_{max} and specific amino acid changes both recurred multiple times in evolution. The independent recurrence of the same evolutionary substitutions could result from at least two mechanisms. In the first, all mutations occur at similar rates, but those causing functional differences are promoted by natural selection. In the second, specific mutations recur at disproportionately high rates. Although selection is sure to have played an important role in protein evolution, the fact that specific mutations recur, sometimes multiple times, points to the contribution of the second option: there is a bias in the mutational input itself.

Take the most important example in mammals: the CpG dinucleotide, that is, a C followed by a G on the same DNA strand. In most of the genome, C → T mutation rates are ten times higher at CpG sites than at Cs in other contexts.[24] Because they are more likely to occur, they are also more likely to recur, substantially biasing the pool of variation upon which selection can act. Although the C → T mutation rate is not quite so high in protein-coding genes, it is still elevated. Given that new amino acids arise due to changes in the underlying DNA, it is conceivable that the known CpG bias could give rise to a bias in protein evolution. The DNA triplet ACG, for example, encodes the amino acid threonine. But because ACG contains a CpG dinucleotide, it is subject to an elevated rate of C → T transitions resulting in ATG, which encodes methionine. The extent to which protein evolution is driven by directionality in the mutational input itself remains to be systematically evaluated, but work by Jay Storz and colleagues suggests it is substantial.[25]

Ancestral proteins can only be inferred for modern proteins similar enough to be grouped into families, of which S100s are one example. This leaves out deeper protein relationships between families. As a rule of thumb, a pair of proteins matching at fewer than ~30% of their positions cannot be confidently aligned. This is because such levels of similarity are likely due to chance alone.[26] As a result, it is only possible to scratch the surface of evolutionary history—only those proteins which diverged relatively recently remain similar enough to compare with confidence. The deepest questions about the origins of novel gene families remain shrouded in mystery.

This includes protein specificity. The ancestor of S100A5 and S100A6 may not have been more general than its descendants—but is the same true of other proteins and protein families? If so, how far back can this trend, or lack thereof, be extrapolated? Which targets were actually present in their environments? How did the primordial archetypes which gave rise to the modern protein families evolve, and had their own ancestors been more general? And, are more general proteins easier to chance upon in sequence space, that is, could they be reasonably expected to have arisen as evolutionary starting points?

To answer these questions, a lot more work just like that of Wheeler and Harms will need to be done.[27]

Chase Nelson is a Postdoctoral Research Fellow at Academia Sinica in Taipei and Visiting Scientist at the American Museum of Natural History in New York City.

1. Linus Pauling and Emile Zuckerkandl, "Chemical Paleogenetics: Molecular 'Restoration Studies' of Extinct Forms of Life," *Acta Chemica Scandinavica* 17 (1963): S9–16.
2. Eric Gaucher et al., "Inferring the Palaeoenvironment of Ancient Bacteria on the Basis of Resurrected Proteins," *Nature* 425, no. 6,955 (2003): 285–88, doi:10.1038/nature01977.
3. David Liberles, ed., *Ancestral Sequence Reconstruction* (New York: Oxford University Press, 2007).
4. Joseph Felsenstein, "Phylogenies and the Comparative Method," *The American Naturalist* 125, no. 1 (1985): 1–15, doi:10.1086/284325.
5. Roy Jensen, "Enzyme Recruitment in Evolution of New Function," *Annual Review of Microbiology* 30, no. 1 (1976): 409–25, doi:10.1146/annurev.mi.30.100176.002205.
6. Valeria Risso, Jose Sanchez-Ruiz, and S. Banu Ozkan, "Biotechnological and Protein-Engineering Implications of Ancestral Protein Resurrection," *Current Opinion in Structural Biology* 51 (2018): 106–15, doi:10.1016/j.sbi.2018.02.007.
7. Olga Khersonsky and Dan Tawfik, "Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective," *Annual Review of Biochemistry* 79, no. 1 (2010): 471–505, doi:10.1146/annurev-biochem-030409-143718. For a detailed survey of the work undertaken by Dan Tawfik and laboratory, see Tyler Hampton, "Dan S. Tawfik Group: The New View of Proteins," *Inference: International Review of Science* 1, no. 1 (2014), doi:10.37282/991819.14.8.
8. Lucas Wheeler et al., "Conservation of Specificity in Two Low-Specificity Proteins," *Biochemistry* 57, no. 5 (2018): 684–95, doi:10.1021/acs.biochem.7b01086; Anne Bresnick, David Weber, and Danna Zimmer, "S100 Proteins in Cancer," *Nature Reviews Cancer* 15, no. 2 (2015): 96–109, doi:10.1038/nrc3893.
9. Rosario Donato et al., "Functions of S100 Proteins," *Current Molecular Medicine* 13, no. 1 (2013): 24–57, doi:10.2174/156652413804486214.
10. Bresnick, Weber, and Zimmer, "S100 Proteins in Cancer."
11. Danna Zimmer et al., "Evolution of the S100 Family of Calcium Sensor Proteins," *Cell Calcium* 53, no. 3 (2013): 170–79, doi:10.1016/j.ceca.2012.11.006; Sudhir Kumar et al., "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times," *Molecular Biology and Evolution* 34, no. 7 (2017): 1,812–19, doi:10.1093/molbev/msx116.
12. Susumu Ohno, *Evolution by Gene Duplication* (New York: Springer-Verlag, 1970); Austin Hughes, "The Evolution of Functionally Novel Proteins after Gene Duplication," *Proceedings of the Royal Society B: Biological Sciences* 256, no. 1,346 (1994): 119–24, doi:10.1098/rspb.1994.0058.
13. Bresnick, Weber, and Zimmer, "S100 Proteins in Cancer."
14. Lucas Wheeler et al., "Learning Peptide Recognition Rules for a Low-Specificity Protein," *Protein Science* 29, no. 11 (2020): 2,259–73, doi:10.1002/pro.3958.
15. Ziheng Yang, "PAML 4: Phylogenetic Analysis by Maximum Likelihood," *Molecular Biology and Evolution* 24, no. 8 (2007): 1,586–91, doi:10.1093/molbev/msm088.
16. Wheeler et al., "Conservation of Specificity in Two Low-Specificity Proteins," 693.
17. Lucas Wheeler and Michael Harms, "Were Ancestral Proteins Less Specific?" *Molecular Biology and Evolution* 38, no. 6 (2021): 2,227–39, doi:10.1093/molbev/msab019.
18. Wheeler and Harms, "Were Ancestral Proteins Less Specific?"; Geeta Eick et al., "Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty," *Molecular Biology and Evolution* 34, no. 2 (2017): 247–61, doi:10.1093/molbev/msw223.
19. Shozo Yokoyama et al., "Elucidation of Phenotypic Adaptations: Molecular Analyses of Dim-Light Vision Proteins in Vertebrates," *Proceedings of the National Academy of Sciences* 105, no. 36 (2008): 13,480–85, doi:10.1073/pnas.0802426105.
20. Austin Hughes, "The Origin of Adaptive Phenotypes," *Proceedings of the National Academy of Sciences* 105, no. 36 (2008): 13,193–94, doi:10.1073/pnas.0807440105.
21. Shozo Yokoyama et al., "A Simple Method for Studying the Molecular Mechanisms of Ultraviolet and Violet Reception in Vertebrates," *BMC Evolutionary Biology* 16, no. 1 (2016): 64, doi:10.1186/s12862-016-0637-9.
22. Hughes, "Origin of Adaptive Phenotypes."
23. Masatoshi Nei, *Mutation-Driven Evolution* (Oxford: Oxford University Press, 2013).
24. Alan Hodgkinson and Adam Eyre-Walker, "Variation in the Mutation Rate across Mammalian Genomes," *Nature Reviews Genetics* 12, no. 11 (2011): 756–66, doi:10.1038/nrg3098.
25. Jay Storz et al., "The Role of Mutation Bias in Adaptive Molecular Evolution: Insights from Convergent Changes in Protein Function," *Philosophical Transactions of the Royal*

Society B: Biological Sciences 374, no. 1,777 (2019): 20180238, doi:10.1098/rstb.2018.0238.
26. Burkhard Rost, "Twilight Zone of Protein Sequence Alignments," *Protein Engineering, Design and Selection* 12, no. 2 (1999): 85–94, doi:10.1093/protein/12.2.85.
27. I am indebted to Michael Harms for discussion and Zachary Ardern for feedback.

DOI: 10.37282/991819.21.36

Published on August 11, 2021

https://inference-review.com/article/reconstructing-ancestral-proteins