The Internal Language of Proteins

Zachary Ardern

Lijia Yu, Deepak Kumar Tanwar, Emanuel Diego Penha, Yuri Wolf, Eugene Koonin, and Malay Kumar Basu, "Grammar of Protein Domain Architectures," *Proceedings of the National Academy of Sciences of the United States of America* 116, no. 9 (2019) : 3,635–45, doi:10.1073/pnas.1814684116.

The question of whether proteins share a common language across the major branches of life is an intriguing one. In their paper, Lijia Yu et al. investigated the arrangements of protein domains, considering them as a kind of grammar that orders the internal structure of proteins. They found that only a small subset of all the possible orderings occur, and that the properties of such a grammar are common across most of the major taxonomic groupings.

Proteins in plants and animals are typically composed of multiple domains in the form of functional subunits on the order of 100–200 amino acids in length. These are often highly conserved and widely distributed across the tree of life. A protein domain architecture is the arrangement of these domains within a particular gene. Recognized protein domains number in the tens of thousands, and there are a vast number of potential domain architectures.[1] Despite these possibilities, less than 5% of the possible sequential pairings are realized in any genome studied to date.[2] Given that genomes are continuously shuffling, it seems that there must be constraints acting on the orderings that are realized. In the same way that linguistic grammars limit languages, these constraints could be considered as a grammar limiting possible functional arrangements.

The concept of a language for protein assembly has been widely discussed in different contexts. This should come as no surprise, because biology is increasingly being viewed as an information science.[3] John Maynard Smith compared protein evolution to a word game, noting that the word "WORD" can be converted to the word "GENE" by a series of single-letter mutations in which each step is a functional intermediate word.[4] Moving up a level in the hierarchy, the composition of proteins can be considered as small motifs, or seqlets, that reoccur across proteins either through descent or convergent evolution.[5] Much as languages are composed of words, seqlet combinations can, depending on the thresholds used, account for the majority, or even the whole, of the elements in protein space. Such words can be categorized into different types, including elements that describe, modify, or connect—akin to nouns, adjectives, and conjunctions. A number of related ideas have been summarized by Mario Gimona in a paper that attempts to consider not just syntax, but also semantics within proteins.[6] He concludes that the functional characteristics of sequences within proteins are likely to be determined by their context. Gimona cautions against approaches that merely consider lists of so-called words—that is, sequences or domains—within proteins. Instead he suggests considering word order as a path toward ascertaining the context.

Rather than attempting to articulate the full language of proteins, whatever that might have meant in practice, the objectives of Yu et al. were far more modest and precise. Seeking to quantify the informational difference between random and actual protein domain architectures, they focused on the simplest dimension of word order within proteins: the probability of observing particular consecutive pairs of domains. Using this approach, much can be learned about the complexity of different genomes. A motivating observation for the researchers was that the major transitions in life are associated with increases in complexity.[7]

As part of their study, Yu et al. examined the distribution of consecutive pairs of domains, known as domain bigrams, across genomes. In this context, a shift away from randomness toward a smaller subset can be considered a gain in information relative to a default or background condition—namely, the magnitude of the change in entropy from a shuffled set.

The researchers found that there was a reasonably consistent difference between random sets and actual genomes, which they interpreted as the minimal information gain required to maintain a functioning living cell. Although not immediately obvious from first principles, this finding makes sense in light of the data. It is conceivable that the information encoded in genomes might not be crucially dependent on domain architecture, so that information could be randomized without issue. Although some caveats need to be taken into account, the finding that genomes are biased against randomized domain structures is a substantive result. Nevertheless, it can still be interpreted as just one aspect of the minimal information required for a modern cell, and a small aspect at that. There are large informational requirements both above and below domain grammar in the reductionist hierarchy, notably in the sequences of the domains themselves, and in the interaction networks between proteins.

Two exceptions to a near-universal pattern are worthy of further exploration. First, there exist a few genomes that are relatively simple in comparison to the vast majority. These are limited to a specific subset of the single-celled microorganisms known as archaea.[8] Whether the lower genome complexity in some branches of archaea that is apparent from protein domain bigrams is an artifact of features such as genome size or duplication histories is a question that needs further attention. Second, Yu et al. find that animal genomes "show the highest information gain among the analyzed groups."[9] They claim that this accords "with the notion that domain architectures in animals are more elaborate and evolve under stronger constraints than those in other organisms." At first glance, this appears to make sense. Animals are indeed the most complex organisms on earth. But, upon closer inspection, the reasoning behind this claim appears problematic. As a result of population sizes that are typically between two and four orders of magnitude lower than bacteria, animal genomes experience much weaker selective pressure. The findings of Yu et al. could also be explained by selection acting on quite different features of the two types of genomes, effectively limiting domain combinations in animals to a much higher degree. It might also be due to constraints from factors other than selection.

As Yu et al. acknowledge in their paper, bigram analysis does not capture higher-level informational structures in protein architecture. If the interpretations described by the authors are even roughly correct, it seems likely that their conclusions would be amplified rather than diminished by an analysis that included higher-level structures, such as trigrams. Many proteins and structures are, in fact, inadequately described by such an analysis. Broadly speaking, this is due to exceptions to the assumed domain structure and difficulties in annotation. These factors are at play with disordered proteins, alternative splicing, unannotated proteins, and unrecognized domain structures. Intrinsically disordered proteins are particularly prevalent in eukaryotic proteomes, where they apparently play an important role. Alternative splicing of RNA molecules prior to translation produces alternative forms of proteins and greatly expands the proteome in higher eukaryotes. The alternatively spliced transcripts can have differently ordered domains, but for the purposes of a tractable study, only the canonical version was used by Yu et al. It should also be noted that many proteins are not yet annotated, even in well-studied and small bacterial genomes.[10] Any number of these unexamined proteins likely contain interesting domain structures that have not yet been catalogued in databases.

Comparisons between domain orderings across genomes are complicated by gene duplications and common ancestry. Due to both common descent and rampant horizontal gene transfer, there are probably no two genomes that can be considered fully independent. Genome and gene duplications are particularly important in plants and animals, which both exhibit genome structures that have been significantly shaped by major duplication events. Such correlations are mitigated by the frequency of domain shuffling that occurs during short-term evolutionary processes. The degree to which duplications in animals contribute to their apparent exceptionality in bigram statistics is a topic that requires further investigation.

A study centered on the grammar of protein domains may touch on any number of fundamental issues in biology. These could include topics such as contingency and biological laws, convergence and stochasticity in evolution, the nature of biological information, and associated questions concerning complexity, gene origins, and the causes of evolutionary change.

The role played by historical contingency and its implications for lawlike consistency in the processes that shape living forms is perhaps the biggest question in biology. In the case of protein domains, if contingency rather than law is dominant, different taxa should exhibit different increases in information content compared with shuffled genomes. In the reverse scenario, there should be complete consistency across taxa, or a clear relationship between information content and some other factor, such as morphological complexity. The findings of Yu et al. appear to be more closely aligned with the latter scenario, but whether this is a result of selective forces, structural factors, or something else entirely remains unclear.

A few decades ago it was thought that almost all genes were ancient and that modern genes were derived through processes such as duplication and divergence.[11] It has since become clear that many genes emerged from noncoding sequences over the course of evolution.[12] Such genes would presumably begin with a minimal number of functional units and accumulate additional domains over time. The resulting protein domain architectures reflect the relative contributions of de novo gene processes and the number of domain shuffling events that are retained. Although researchers are currently limited by a number of methodological issues, especially in relation to determining the relevant rates, the study of gene origins remains one of the fastest developing fields within evolutionary biology.

Two conclusions can be drawn from the work of Yu et al. First, there is apparently an information increase within proteins at the level of domain arrangement, which is associated with functional cells. Second, complexity can be objectively measured. This paper adds to the growing body of evidence that there have been genuine increases in complexity over the course of evolutionary history, and that this is particularly evident in animals. Biologists have become so accustomed to considering notions of human uniqueness as thoroughly debunked that any hint of so-called progress within evolution is treated with great skepticism. Whatever one makes of such a loaded term, increased complexity in some lineages is observable across multiple biological features, including protein domain architecture.

Zachary Ardern is a Junior Group Leader in Microbial Evolutionary Genetics at the Technical University of Munich. He has a PhD from the University of Auckland in New Zealand.

[1] Sara El-Gebali et al., "The Pfam Protein Families Database in 2019," *Nucleic Acids Research* 47, no. D1 (2019): D427–32, doi:10.1093/nar/gky995.

[2] Lijia Yu et al., "Grammar of Protein Domain Architectures," *Proceedings of the National Academy of Sciences of the United States of America* 116, no. 9 (2019): 3,635–45, doi:10.1073/pnas.1814684116.

[3] John Maynard Smith, "The Concept of Information in Biology," *Philosophy of Science* 67, no. 2 (2000): 177–94, doi:10.1086/392768.

[4] John Maynard Smith, "Natural Selection and the Concept of a Protein Space," *Nature* 225, no. 5,232 (1970): 563–64, doi:10.1038/225563a0.

[5] Isidore Rigoutsos et al., "Dictionary Building via Unsupervised Hierarchical Motif Discovery in the Sequence Space of Natural Proteins," *Proteins* 37, no. 2 (1999): 264–77, doi:10.1002/(SICI)1097-0134(19991101)37:2<264:: AID-PROT11>3.0.CO;2-C.

[6] Mario Gimona, "Protein Linguistics—A Grammar for Modular Protein Assembly?" *Nature Reviews Molecular Cell Biology* 7, no. 1 (2006): 68–73, doi:10.1038/nrm1785.

[7] Eörs Szathmáry and John Maynard Smith, "The Major Evolutionary Transitions," *Nature* 374, no. 6,519 (1995): 227–32, doi:10.1038/374227a0.

[8] The main general differences between archaebacteria ("ancient bacteria") and aubacteria ("true bacteria") lie in their membranes and/or cell walls, and archaea are generally considered simpler than eubacteria—though this is not true in every aspect, as for instance the RNA polymerase is more complex in archaea. Yu et al. report that among archaea,

> Euryarchaeota and Nanoarchaeota show an information gain value close to that in Bacteria, ~1.2 bits … whereas the rest of the archaea have a lower value of ~1.04 bits. Thus, these archaea are characterized by anomalously low proteomic complexity.

Yu et al., "*Grammar*," 3,638.

[9] Yu et al., "*Grammar*," 3,638.

[10] Sarah Hücker et al., "Discovery of Numerous Novel Small Genes in the Intergenic Regions of the *Escherichia coli* O157:H7 Sakai Genome," *PLOS One* 12, no. 9 (2017): e0184119, doi:10.1371/journal.pone.0184119.

[11] Robert Dorit, Lloyd Schoenbach, and Walter Gilbert, "How Big Is the Universe of Exons?" *Science* 250, no. 4,986 (1990): 1,377–82, doi:10.1126/science.2255907.

[12] Stephen Branden Van Oss and Anne-Ruxandra Carvunis, "De Novo Gene Birth," *PLOS Genetics* 15, no. 5 (2019): e1008160, doi:10.1371/journal.pgen.1008160.

Published on September 28, 2020

https://inference-review.com/article/the-internal-language-of-proteins