Ninety-nine Point Nine Percent of the Time, Nature Uses the Same Group of Amino Acids, and We Don't Know Exactly Why

Abstract

For over 2 billion years, all life on earth has built itself on a collection of 20 specific amino acids. In light of the far greater number of amino acids that have been found to exist naturally, the universality of this group of 20 stands out as a curious feature. Explanations as to why have ranged from the simplistic to the fantastical . . . and none of them has yet proved completely satisfactory. However, progress has been made in understanding how amino acids function in relation to the genetic code. An important feature, it seems, of this nearly universal arrangement is its resilience against genetic error. The details and subtle complexities of this feature, many of which have yet to be understood, may provide a context for understanding the group of 20 amino acids as an optimal arrangement selected by evolution. At the same time, the engineering of modified organisms that incorporate a 21st non-natural amino acid may ultimately prove the viability of alternative arrangements to the standard group of 20. In any course, much will be revealed as exploration of amino acid modification continues.

There are a handful of numbers that are foundational to the most basic biological processes common to all living organisms. *Four* nucleotides carry all of the genetic information particular to each organism. *Two* strands of nucleic acid form the double helix—DNA—that stores this genetic information. *Three* nucleotides in sequence—called a codon—signify a particular amino acid during the production of proteins, and a group of *20* particular amino acids are the sole building blocks of those various proteins, the catalogue of which runs into the tens of thousands. Most of these numbers, though not *a priori* necessary, make a degree of sense. The quaternary code created by four nucleotides is simple yet compact. The two strands of DNA create a highly stable structure that effectively preserves genetic information, and, additionally, store the genetic information in duplicate, a redundancy that helps to eliminate errors during replication. The three-unit codon, especially, seems to be just the right size. A two unit codon, with only 16 possible variations, would be to small to represent a functionally diverse library of amino acids, and a four unit codon, with 256 possible variations, might be unnecessarily large and cumbersome. The number of biological amino acids, however, remains apparently inexplicable. In a ballpark estimation, a catalogue of 20 amino acids satisfies the intuitive criterion, "not too small, not too large;" it is large enough to be functional in the creation of a wide variety of proteins, but not so large (it could, theoretically, number in the hundreds) that it creates unnecessary complexities. This, however, is only a generally satisfying explanation. 19 fits this criterion. So does 21. On the face of it, these variations would be just as functional. Yet for at least 2 billion years

of evolutionary history, with few exceptions, all of the millions of species that have existed have built themselves out of the same particular group of 20 amino acids. Why?

Attempts to answer this question have often produced creative and daring hypotheses, from the high level mathematical analysis of the symmetry between amino acids and codons (Yang 2003) to a theoretical speculation of the quantum mechanical efficiency of a group of 20 during molecular bonding (Buchanan 2000). The simplest explanation, however, is that the genetic code and its correspondence to a particular set of amino acids are part of what the famed biologist Francis Crick termed a "frozen accident." According to this explanation, there is nothing particularly special about the natural amino acids, their number or their representation in the language of the genetic code; the arrangement was merely the first to be stumbled across by evolution. Once it had been solidified, "any further changes would have been catastrophic (Freeland 2004, 87)." This assumption dominated conventional wisdom for several decades. Problematically, this assumption also predated—by several decades—any detailed and expansive body of knowledge that could even begin to verify or discredit such a theory. When in the 90's the volume of genetic and molecular biological data began to expand exponentially, it was discovered that the use of amino acids was not so frozen after all. Several organisms, including representatives from all three domains of life, have been found to synthesize and use a 21st amino acid, selenocysteine, in addition to their standard repertoire of 20; a 22nd amino acid, pyrrolysine, has been discovered in some bacterial and archaeal organisms (Freeland 2004). Although these organisms are the exception rather than the rule, their existence demonstrates that life's amino acids are not bound into a ridged cannon. It is possible, and apparently advantageous in some instances, for

species to incorporate new amino acids. This revelation, however, displaced a satisfyingly simple yet erroneous theory with a more slippery question: if is possible for life to incorporate additional amino acids, why, in all but a handful of cases, doesn't it?

Another possible avenue of explanation lies in the relation between amino acids and the language of the genetic code. There are three features of the relationship between codons and amino acids that are significant in their tendency to eliminate or reduce the potentially destructive consequences of genetic errors. First, because DNA's representation of amino acids is written in a sequence of three nucleotides—termed a codon—there are 64 unique codons. With the exception of three codons that signal the end of a protein sequence, each codon represents a particular amino acid. This arrangement, however, leaves a conspicuous imbalance between the number of representational codons—61—and the number of amino acids they represent—20. The consequence is a fortuitous redundancy in the genetic code. Most amino acids can be represented by more than one codon, and more than half are represented by three or four. The immediate advantage of this redundancy is that it softens the consequences of errors in the replication, transcription or translation of genes. Second, a further advantage is added by the fact that redundant codons differ from each other by a single letter, and most often, the variable letter lies in the third position of the codon—which also happens to be the most likely location of an error during the reading of the codon (Freeland 2004). By structuring redundancy around the most probable errors, the likelihood that errors will have consequences is even further decreased.

A third feature of the codon language that contributes to a softening of the consequences of errors is a pattern of codon similarity between codons that represent functionally similar amino acids. It was proposed as early as the 1960's that "nearly all transitions between functionally closely related amino acids can be brought about by one single mutational step.'' (Xia 1998) This increases the chances that in a non-synonymous mutation the functionality of the amino acid chain will be preserved, for although no two amino acids are identical, some seem to be interchangeable in certain situations. Though this phenomenon is still not completely understood, research has shown that of over a hundred different properties that have been measured in amino acids, there are a mere few that play a dominant role in governing functionality. The uncanny effectiveness of the codon language in exploiting similarity among these select properties has recently been demonstrated by researchers Stephen Freeland and Lawrence Hurst, who have shown that compared with large samplings of randomly generated alternative languages, "nature's code outperforms all but one in a million of the alternatives (Freeland 2004)."

These features, taken together, indicate a highly developed resiliency in the codon language and its relation to amino acids. This does not cohere with the "frozen accident" view, which supposes that the genetic code and the selection of the 20 amino acids evolved relatively quickly and randomly, solidifying into a stable, yet fragile and inflexible system. Instead, it appears that the evolution was dynamic and prolonged, leading to highly refined and flexible system of stability. Incidentally, this view fits well with newly emerging theories of early evolution. It has for some time been widely held that life began with what is now the intermediary between DNA and proteins, RNA, and that at some later date DNA and protein appeared. But which appeared first, DNA or

protein, has long been a topic of speculation. New evidence, however, suggests that protein emerged first and that RNA and protein may have exclusively formed the basis of life for a period in the early history of evolution. Because RNA is far less stable than DNA and is more libel to randomly mutate, it makes sense that during this period the genetic code and the selection of amino acids would have evolved to maximize resiliency against random mutation. The result would have been a system that was not so much "frozen" in a coincidental arrangement, as it was optimally adapted.

Optimal adaptation may sufficiently explain the particulars of the genetic code, and may provide a general context for understanding the universality of the 20 natural amino acids, but it nevertheless falls far short of being able to explain specifically why and how exactly 20 particular amino acids are optimal. And now, not only do the naturally occurring exceptions to this universality call into question the potential viability of 21 (or more) amino acid organisms, artificially designed organisms do as well. In 2001, the Schultz Lab of the Scripps Research Institute successfully modified a bacterial species to incorporate a 21st amino acid (Chin 2002). In 2003, advancements lead to modified eukaryotic organisms (yeast organisms) which are able to autonomously synthesis non-naturally occurring amino acids and incorporate them into proteins. This modification has been repeated successfully with five different non-natural amino acids. The resulting organisms are fully viable, and both replicate themselves and synthesize non-natural protein with an accuracy and efficiency that is on par with natural organisms (Chin 2003). Once again, this led us again to question why nature has strongly favored 20 rather than 21.

More clues may be revealed in the near future. The development of modified organisms that produce non-natural amino acids has an incredible potential usefulness for both science and industry. For instance, enzymes tainted with heavy metal ions can be used to trace metabolic pathways in little understood cellular processes. Also, pharmaceutical drugs containing non-natural amino acids could be produced cheaply and efficiently by populations of modified single cell organisms (Schultz Lab Website) This diverse potential usefulness almost guarantees that both many more 21 amino acid organisms will be created and that attempts will be made to expand the number of incorporated amino acids to its functional limit, whatever that may be. The Schultz lab, optimistically, speculates that when the limitation of a 3 unit codon with its 64 possible combinations becomes a limiting factor, that it will be possible to engineer a new genetic language based on a 4 unit codon, expanding the total number of codons to 256 (Schultz Lab Website). As more organisms are artificially modified in these ways, successes and failures may provide clues about the history of natural evolution and lead to a more conclusive answer to the "why 20?" question.

Works Cited

Buchanan, Mark. "Life Force." New Scientist 15 April 2000: 21-24.

Chin, Jason W, Martin, Andrew B, King, David S, and Schultz, Peter G. "Addition of a Photocrosslinking Amino Acid to the Genetic Code of *Escherichia coli*." PNAS 99 (20 Aug. 2002): 11020-11024.

Chin, Jason W, Cropp, Ashton, Anderson, J Christopher, Mukherji, Mridul, Zhang, Zhiwen and Schultz, Peter G. "An Expanded Eukaryotic Genetic Code." Science 301 (15 August 2003): 964-966.

Freeland, Stephen J, Hurst, Laurence D. "Evolution Encoded." Scientific American Apr. 2004: 85-91.

Freeland, Stephen J, Knight, Robin D, and Landweber, Laura F. "Do Proteins Predate DNA?" Science 286 (22 Oct. 1999): 690-692.

Shultz Lab website. The Scripps Research Institute. 23 Mar. 2004.

Yang, Chi Ming. The Naturally Designed Spherical Symmetry in the Genetic Code. Tian Jin: Nankai University, 2003.

Xia, Xuhua, and Li, Wen-Hsiung. "What Amino Acid Properties Affect Protein Evolution?" Journal of Molecular Evolution 47 (1998): 557-564.