Evolution and multilevel optimization of the genetic code

Tobias Bollenbach, Kalin Vetsigian and Roy Kishony

Genome Res. 2007 17: 401-404; originally published online Mar 9, 2007;
Access the most recent version at doi:10.1101/gr.6144007

Perspective

Evolution and multilevel optimization of the genetic code

Tobias Bollenbach,[1] Kalin Vetsigian,[1] and Roy Kishony[1,2,3]

[1]*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA;* [2]*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*

The discovery of the genetic code was one of the most important advances of modern biology. But there is more to a DNA code than protein sequence; DNA carries signals for splicing, localization, folding, and regulation that are often embedded within the protein-coding sequence. In this issue, Itzkovitz and Alon show that the specific 64-to-20 mapping found in the genetic code may have been optimized for permitting protein-coding regions to carry this extra information and suggest that this property may have evolved as a side benefit of selection to minimize the negative effects of frameshift errors.

The Wild West of early code theories and the comma-less code

The first glimmer of light in the story of the code came when Dounce (1952) proposed the extraordinary for its time idea that the order of nucleotides determines the order of amino acids in polypeptide chains. After the discovery of the double-helix structure of DNA (Watson and Crick 1953), which accounted for replication, the race was on to crack the secrets of genetic expression. Gamow (1954) suggested a "key-and-lock" mechanism in which the amino acids specifically bound to "holes" in the DNA formed from four nucleotides. The shape of the hole determined which amino acid could bind to it, allowing the sequence of the DNA to encode a specific amino acid sequence. Two of the four bases forming each hole were complementary, which implied that each amino acid was effectively defined by a base triplet. Careful analysis revealed that this "diamond code" could encode a maximum of 20 different amino acids, making the theory particularly appealing. But the codons (as we would call them now) were envisaged to overlap: the three nucleotides that encoded the amino acid at position 1 in the protein would include two of the nucleotides that encoded amino acid 2. This would restrict the amino acid sequences that are possible (Gamow 1954).

But analyzing Gamow's results (Gamow 1954; Gamow et al. 1956), Crick and coworkers highlighted that no such restrictions could be found in nature (Crick et al. 1957). In the same year, Brenner (1957) showed that an overlapping triplet code could be ruled out. But the alternative, having a code with nonoverlapping triplets, poses the problem of selecting and enforcing the correct reading frame. Acknowledging that the problem could be solved by translating the sequence sequentially, but not knowing about start codons, Crick and coworkers proposed a "code without commas" (Crick et al. 1957). The comma-less code allowed coding for arbitrary amino acid sequences and could only be read in one reading frame—any attempt to read in the wrong frame would be immediately recognized as nonsense. This was achieved by using a very large number of nonsense codons, such that any out-of-frame sequence consisted exclusively of them. As in the case of the diamond code, the maximum number of amino acids that could be encoded by a comma-less code was exactly 20—a seductive coincidence with the number of amino acids used in nature.

[3]Corresponding author.
E-mail roy_kishony@hms.harvard.edu; fax (617) 432-5012.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6144007.

The guessing game continued. Sinsheimer (1959) suggested a code with only two letters in which A and C were equivalent as were G and U. To encode the 20 different amino acids in this scheme would take at least five bases. The next year, Yčas (1960) proposed a hypothesis based on empirical observations from several viruses that showed different amino acid abundances in proteins and base abundances in RNA. In this idea, single nucleotides (not nucleotide triplets) encoded the amino acids, but the nucleotide sequence contained only part of the information required. Other codes in which more than three nucleotides encode one amino acid were considered: Golomb (1962) suggested a sextuplet code that was comma-less and had additional properties that would make translation very reliable. These early theoretical approaches to the genetic code have been reviewed by Hayes (1998).

Optimal features of a redundant code

The discovery of the actual genetic code by Nirenberg and coworkers (Nirenberg and Matthaei 1961; Nirenberg 2004) put an end to the theoretical speculations and led to the quick rejection of the comma-less code and the other earlier theories. The redundancy of the code was something of a surprise and was the focus of much early interest. For example, family boxes and wobble rules (Crick 1966), which describe the system by which the same amino acid is assigned to several similar codons, were identified. Other striking properties of the code that seem far from random were revealed. Woese observed that similar codons are assigned to amino acids with similar chemical properties, most notably, similar polar requirement (Woese 1965b; Woese et al. 1966a). He proposed that the code is optimized for minimizing the impact of mistranslation errors. These errors occur when a codon is translated via a tRNA with a near cognate anticodon. The finding that the genetic code is optimized with respect to minimizing the impact of translational misread errors was statistically quantified by Haig and Hurst (1991) and further strengthened by taking into account biased mistranslation and mutation (Freeland and Hurst 1998).

There has been much speculation about how the code evolved (Osawa et al. 1992; Knight et al. 1999, 2001a; Di Giulio 2004). Is it a "frozen accident" (Crick 1968)? If so, why does it have such useful features? Several factors influencing the early evolution of the code were suggested: (1) the code has evolved under selection pressure to optimize certain functions such as minimization of the impact of mutations (Sonneborn 1965) or translation errors (Woese 1965a); (2) the number of amino acids

in the code has increased over evolutionary time according to evolution of the pathways for amino acid biosynthesis (Wong 1975); and (3) direct chemical interactions between amino acids and short nucleic acid sequences originally led to corresponding assignments in the genetic code (Woese et al. 1966b).

These hypotheses are not mutually exclusive, and there is some support for all of them ruling out Crick's frozen accident hypothesis. Evidence for optimization of the code for certain functions exists, as discussed above, and there are indications that the usage frequencies of some amino acids in proteins are decreasing, while those of others are increasing (Jordan et al. 2005, but see also Hurst et al. 2006; Wong 2005), suggesting that some amino acids were added to the genetic code relatively recently. Random RNAs that bind arginine are enriched in arginine codons (Knight and Landweber 2000), and the simplest RNA molecules that bind the amino acid isoleucine have sequence motifs that are very similar to its associated codons and anticodons (Lozupone et al. 2003).

The discovery of variant codes (Barrell et al. 1979; Fox 1987; Knight et al. 2001a) made the connection between evolvability and universality even more puzzling. On one hand, they prove that the genetic codes can evolve; on the other hand, if they could easily evolve, why are all variations minor? It was recently proposed that extensive horizontal gene transfer during early evolution can account for both evolution toward optimality and the near universality of the genetic code (Vetsigian et al. 2006).

Shifting the frame of optimality

But if the code was optimized for some functions, are there other, less obvious, functions for which it is also optimal? Frameshift mutations might be important because they result in nonfunctional proteins, which waste resources and could also be toxic. A way to minimize the resource waste is to terminate elongation as quickly as possible after the error. There are some bioinformatics clues that the impact of frameshift errors was minimized in evolution. It has been observed that in many (albeit not all) organisms, codon usage frequencies are biased toward codons that can contribute to stop codons if read off-frame (Seligmann and Pollock 2004).

If optimization of fast termination after a frameshift error is built into the genetic code itself, what would an optimal code look like? Crick's comma-less code, interpreted such that all nonsense codons correspond to "stop codons" in today's terminology, is the perfect code in this respect: it stops translation immediately after a translational frameshift. However, such extreme optimization comes at a high price. Since there are no synonymous codons for any amino acid in the comma-less code, the majority of point mutations result in nonsense codons, essentially equivalent to null mutations. This would highly increase the mutational load. In the actual genetic code, only about one of 20 point mutations results in a new stop codon (Osawa et al. 1992), and many of the other 19 give functional proteins (possibly with altered properties). The question then is whether the genetic code could be optimized for fast termination after a frameshift error, while maintaining its optimization for other functions.

Optimality of the genetic code with known properties as constraints

In this issue of *Genome Research*, Itzkovitz and Alon report on the intriguing discovery of two new properties for which the genetic code seems to be optimized. They compared the actual genetic code with an ensemble of all other codes that are equally optimized with respect to mistranslation or mutation (for more on this statistical approach, see also Alff-Steinberger 1969; Haig and Hurst 1991; Freeland and Hurst 1998). Assuming that the usage frequencies of the different amino acids are fixed, while their codon assignments vary in the ensemble, they find that the actual code is far better than other possible codes in minimizing the number of amino acids incorporated until translation is interrupted after a frameshift error occurred. This new observation by Itzkovitz and Alon could therefore be seen as reviving the basis for Crick's theory of a comma-less code, modified by the constraints imposed on the code by the need to be robust to other kinds of translation errors and mutations. Another possible interpretation of their result is that the amino acid usage has adjusted to reduce the effects of frameshift errors; alternative genetic codes would have had a different amino acid usage coadapted to them. It has been shown previously that amino acid usage is rather malleable, and, for example, influenced by GC content (Knight et al. 2001b).

Itzkovitz and Alon suggest another, quite unanticipated, type of optimality: the code is highly optimal for encoding arbitrary additional information, i.e., information other than the amino acid sequence in protein-coding sequences. Optimality for encoding additional information is particularly important and relevant given the known signals contained in the nucleotide sequence of coding regions. These include RNA splicing signals, which are encoded in the nucleotide sequence together with the amino acid sequence of the prospective protein (Cartegni et al. 2002), as well as signals recognized by the translation apparatus. For example, a few codons that are usually read as stop signals are translated as the rarely used amino acid selenocysteine if they appear in a special context on the mRNA strand (Fox 1987). Information about where nucleosomes should be positioned on the DNA is also contained in the base sequence (Yuan et al. 2005; Segal et al. 2006). Sequences for RNA secondary structure are another source of information that has recently been found to be over-represented in protein-coding sequences (Zuker and Stiegler 1981; Shpaer 1985; Konecny et al. 2000; Katz and Burge 2003).

Interestingly, the optimal structure of the code for both information encoding and translation interruption after frameshift appear to derive from the same root cause, namely, the fact that stop codons can easily be concealed within a sequence. For example, the UGA stop codon is only one frameshift away from NNU|GAN; the GAN codons encode Asp and Glu, which are very common in protein sequences. Similarly, UAA and UAG can be frameshifted to give NNU|AAN and NNU|AGN (the AAN codons encode Asn or Lys and AGN gives Ser or Arg). Glu, Lys, Asp, Ser, and Arg are relatively common amino acids in the genome, so the probability of a stop codon arising from a misread of a codon from one of these three amino acids is very high. The fact that a stop codon can be "hidden" in this way using a frameshift means that even a signal sequence that happens to include a stop codon (a problem that is bound to arise sooner or later) can be encoded within the protein sequence by using one of the two reading frames in which the stop codon encodes for a frequently used amino acid.

The ability to encode hidden messages is a direct result of the redundancy of the code. Like the universal genetic code, language, such as English, has considerable redundancy, i.e., it takes more letters and words to convey a certain message than necessary from an information theoretical point of view. In other

words, the information content of an English sentence is less than what could be encoded in a sequence of Latin letters and punctuation marks of equal length. This redundancy allows for communicating several messages in parallel—a property occasionally used in human history for sending secret messages that are "camouflaged" in unsuspicious looking communications (steganography). An illustrating example can be found in the following sentence from the "Sherlock Holmes" story, *The Adventure of the Gloria Scott* (Conan Doyle 1893):

> "The supply of game for London is going steadily up. Headkeeper Hudson, we believe, has been now told to receive all orders for fly-paper and for preservation of your hen pheasant's life."

Reading every third word starting with the first (and adding a few punctuation marks), the hidden message emerges: "The game is up. Hudson has told all. Fly for your life." It becomes increasingly difficult to convey such additional messages in a communication with decreasing redundancy of the language or code that is used.

This concept of simultaneously communicating two messages, one of which is more obvious and detailed than the other, is similar to that of providing a template for an amino acid sequence together with noncoding information in a nucleotide sequence. However, unlike in human communication, where the main message is used as a camouflage, secrecy is certainly not the reason for the use of this approach in nature. Rather, selection pressure for using resources efficiently may be the reason that the genetic code adapted this property. But, was it really a clear advantage in the early evolution of the code to be able to encode additional noncoding information? The correlation between the ability to encode additional information and the property of optimality of translational termination following frameshift errors offers a possible evolutionary scenario, in which selection for resource waste minimization favored codes that efficiently terminate translation, and the ability of the code to carry additional information was a byproduct. This second property may have become important only later on, when additional complex regulatory programs and regulatory motifs started to develop. A possible exception is the ability to include sequences for stabilizing RNA secondary structure. RNA molecules that possessed this ability in parallel to their protein-coding function might have had an advantage over RNAs that were less effective in this ability.

As we learn more about the functions of the genetic code, it becomes ever clearer that the degeneracy in the genetic code is not exploited in such a way as to optimize one function, but rather to optimize a combination of several different functions simultaneously. Looking deeper into the structure of the code, we wonder what other remarkable properties it may bear. While our understanding of the genetic code has increased substantially over the last decades, it seems that exciting discoveries are waiting to be made.

Acknowledgments

We thank R. Ward for inspiring suggestions and P. Yeh for useful comments on the manuscript.

References

Alff-Steinberger, C. 1969. The genetic code and error transmission. *Proc. Natl. Acad. Sci.* **64:** 584–591.

Barrell, B.G., Bankier, A.T., and Drouin, J. 1979. A different genetic code in human mitochondria. *Nature* **282:** 189–194.

Brenner, S. 1957. On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proc. Natl. Acad. Sci.* **43:** 687–694.

Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **3:** 285–298.

Conan Doyle, A. 1893. *The Memoirs of Sherlock Holmes*. Murray, London.

Crick, F.H. 1966. Codon–anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **19:** 548–555.

Crick, F.H. 1968. The origin of the genetic code. *J. Mol. Biol.* **38:** 367–379.

Crick, F.H., Griffith, J.S., and Orgel, L.E. 1957. Codes without commas. *Proc. Natl. Acad. Sci.* **43:** 416–421.

Di Giulio, M. 2004. The origin of the genetic code: Theories and their relationships, a review. *Biosystems* **80:** 175–184.

Dounce, A.L. 1952. Duplicating mechanism for peptide chain and nucleic acid synthesis. *Enzymologia* **15:** 251–258.

Fox, T.D. 1987. Natural variation in the genetic code. *Annu. Rev. Genet.* **21:** 67–91.

Freeland, S.J. and Hurst, L.D. 1998. The genetic code is one in a million. *J. Mol. Evol.* **47:** 238–248.

Gamow, G. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173:** 318.

Gamow, G., Rich, A., and Yčas, M. 1956. The problem of information transfer from nucleic acids to proteins. In *Advances in Biological and Medical Physics,* Vol. 4, pp. 23–68. Academic Press, New York

Golomb, S.W. 1962. Efficient coding for the desoxyribonucleic channel. In *Proceedings of Symposia in Applied Mathematics,* pp. 87–100. American Mathematical Society, Providence, RI.

Haig, D. and Hurst, L.D. 1991. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33:** 412–417.

Hayes, B. 1998. The invention of the genetic code. *Am. Sci.* **86:** 8–14.

Hurst, L.D., Feil, E.J., and Rocha, E.P.C. 2006. Causes of trends in amino-acid gain and loss. *Nature* **442:** E11–E12.

Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S., and Sunyaev, S. 2005. A universal trend of amino acid gain and loss in protein evolution. *Nature* **433:** 633–638.

Katz, L. and Burge, C.B. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **13:** 2042–2051.

Knight, R.D. and Landweber, L.F. 2000. Guilt by association: The arginine case revisited. *RNA* **6:** 499–510.

Knight, R.D., Freeland, S.J., and Landweber, L.F. 1999. Selection, history and chemistry: The three faces of the genetic code. *Trends Biochem. Sci.* **24:** 241–247.

Knight, R.D., Freeland, S.J., and Landweber, L.F. 2001a. Rewiring the keyboard: Evolvability of the genetic code. *Nat. Rev. Genet.* **2:** 49–58.

Knight, R.D., Freeland, S.J., and Landweber, L.F. 2001b. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2:** RESEARCH0010.

Konecny, J., Schoniger, M., Hofacker, I., Weitze, M.D., and Hofacker, G.L. 2000. Concurrent neutral evolution of mRNA secondary structures and encoded proteins. *J. Mol. Evol.* **50:** 238–242.

Lozupone, C., Changayil, S., Majerfeld, I., and Yarus, M. 2003. Selection of the simplest RNA that binds isoleucine. *RNA* **9:** 1315–1322.

Nirenberg, M. 2004. Historical review: Deciphering the genetic code—A personal account. *Trends Biochem. Sci.* **29:** 46–54.

Nirenberg, M.W. and Matthaei, J.H. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.* **47:** 1588–1602.

Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A. 1992. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56:** 229–264.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772–778.

Seligmann, H. and Pollock, D.D. 2004. The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* **23:** 701–705.

Shpaer, E.G. 1985. The secondary structure of mRNAs from *Escherichia coli*: Its possible role in increasing the accuracy of translation. *Nucleic Acids Res.* **13:** 275–288.

Sinsheimer, R.L. 1959. Is the nucleic acid message in a 2-symbol code? *J. Mol. Biol.* **1:** 218–220.

Sonneborn, T. 1965. Degeneracy of the genetic code: Extent, nature, and genetic implications. In *Evolving genes and proteins* (eds. V. Bryson and H. Vogel), pp. 377–397. Academic Press, New York.

Vetsigian, K., Woese, C., and Goldenfeld, N. 2006. Collective evolution and the genetic code. *Proc. Natl. Acad. Sci.* **103:** 10696–10701.

Watson, J.D. and Crick, F.H. 1953. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* **171:** 737–738.
Woese, C.R. 1965a. On the evolution of the genetic code. *Proc. Natl. Acad. Sci.* **54:** 1546–1552.
Woese, C.R. 1965b. Order in the genetic code. *Proc. Natl. Acad. Sci.* **54:** 71–75.
Woese, C.R., Dugre, D.H., Dugre, S.A., Kondo, M., and Saxinger, W.C. 1966a. On the fundamental nature and evolution of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **31:** 723–736.
Woese, C.R., Dugre, D.H., Saxinger, W.C., and Dugre, S.A. 1966b. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci.* **55:** 966–974.
Wong, J.T. 1975. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci.* **72:** 1909–1912.
Wong, J.T. 2005. Coevolution theory of the genetic code at age thirty. *Bioessays* **27:** 416–425.
Yčas, M. 1960. Correlation of viral ribonucleic acid and protein composition. *Nature* **188:** 209–212.
Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309:** 626–630.
Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.