EVOLUTION

The genetic code sees off rivals

DOI:
10.1038/nrg2076

URLs

There are many possible three-letter genetic codes that could adequately encode protein sequences, but what about the need to encode higher-order information on binding and splicing sites? New research shows that the actual genetic code is better than potential alternatives at encoding such information at the same time as encoding protein.

The main function of coding regions is to specify the amino-acid sequence. However, these regions also need to include other elements — sequences for splicing out introns and for mRNA secondary structures, sequences that allow the regular binding of histones, and even some regulatory-protein binding sequences lie within coding regions. These higher-order requirements can often conflict with the main task of protein coding, so Itzkovitz and Alon looked at whether this conflict was minimized in the actual genetic code in comparison with the alternatives.

It is already known that there are some constraints on the code. The effect of translational misreads is minimized by having similar codons encoding similar amino acids, and smaller amino acids have more codons as they are required more often in protein assembly. The authors therefore compared only the possible three-letter codes that conformed to these constraints. They then considered higher-level sequence requirements for motifs of different lengths. For example, if a binding protein requires a particular five-base sequence, how likely is it that that sequence can be included in an average gene without compromising the structure of the protein that it encodes? The motif can appear in any of the three reading frames, but a reading frame will be excluded if the motif creates a stop codon in it. Less extremely, the probability that a particular motif can be included in a particular reading frame will depend on the usage frequencies of the codons that it contains.

The authors added together the probabilities of all motifs between 4 and 25 bases in length appearing in each of their potential genetic codes. They found that the real genetic code could accommodate more arbitrary motifs in coding sequence than almost any of the other possibilities — it has a higher information content. One reason for the real genetic code's superiority is the fact that its stop codons, when frame-shifted, tend to form common codons, whereas in other codes frame-shifted stop codons form rarer codons or even other stop codons.

The reverse of this — that common codons can be frame-shifted into stop codons — might explain the initial adoption of the real genetic code. When a ribosome skips a reading frame energy is expended in creating a useless or even harmful protein. In the real genetic code the probability of encountering a stop codon after such a mistake, and therefore saving energy, is maximized. The authors suggest that selection on this, rather than for the ability to encode higher-level information, might be the explanation for the actual code, as the code was probably fixed before such higher-level complexity arose. The mammoth task now is to uncover all the higher-order codes that are contained in the genome.

Patrick Goymer

ORIGINAL RESEARCH PAPER Itzkovitz, S. & Alon, U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* 9 February 2007 (doi:10.1101/gr.5987307)
FURTHER READING Condon, A. Designed DNA molecules: principles and applications of molecular nanotechnology. *Nature Rev. Genet.* **7**, 565–575 (2006)