ANTISENSE ORFS, CODON BIAS AND THE EVOLUTION OF THE GENETIC CODE*

W. L. Duax[1,3], A. Addlagatta[1], V. Pletnev[2], R. Huether and P. Yu[1]

[1]*Hauptman-Woodward Institute, Buffalo, NY 14203*
[2]*Shemyakin Institute, Moscow, Russia*
[3]*SUNY at Buffalo, Buffalo, NY*

The short chain oxidoreductase (SCOR) family of enzymes includes over 5000 members, extending from bacteria and Archaea to humans, for which 36 have known crystal structures and 4500 have unknown function. The superfamily has at most one fully conserved residue. The signatures of subgroups of the superfamily are composed of 30-40 residues conserved at approximately 80 % identity distributed throughout the 250 aa proteins. Nucleic acid sequence analysis reveals that 21% of the SCOR genes (342/1612) have an antisense open reading frame (ORF) overlapping the entire sense gene. Furthermore 5 % have a third frame shift ORF. In all cases a double ORF consists of a pair of in-frame sense/antisense totally overlapping ORF's (SASORFs). The three ORF's of a triple ORF (TORF) are always composed of the two SASORFs and a third ORF in the "sense" frame related by a double frame shift. SCOR open reading frames are never found in the frame in which the wobble base of the sense codon is in the center of the nucleic acid triple. Analysis of 260 SCOR genes having double open reading frames (DORFs) and 82 having TORFs reveals that over 85 % of the 250 amino acids in the proteins encoded by these genes are coded for by the GC-rich codons. When nucleic acid triple frequency is analyzed in the two alternate frames the same codon bias is observed. [In the 260,600 nucleic acid triples in the SCOR family genes having DORFs and TORFs the frequency of appearance of GC-only triples is at least ten times that of the AT-only triples.] Examination of the population of the 32 complementary pairs of nucleic acid triples rather than 64 individual triples results in the most pronounced separation of the four classes of codons (GC-only, GC-rich, AT-rich and AT-only). In addition to the complete separation of the GC-rich and AT-rich halves of the 64 codons, there is a pattern of clustering of sets of permuted nucleic acid triples. Analysis of triple frequencies in a random collection of genes reveals no comparable bias or pattern.

The recent discovery of a 22nd amino acid (Srinivasan, G., *et al*., Science **296**, 1459-1461, 2002) draws renewed attention to the fact that at least fourteen codons have different definitions in different species (Maeshiro, T. *et al*., PNAS **95**, 5088-5093, 1998). The majority of the codons having variable definition are AT-only or AT-rich and have as many as four known definitions. Half of the reported variations are in the mitochondria of Eukaryotes and the other half are in bacteria and Archaea. Since we only know the genetic codes of a small percentage of all species it is highly probable that additional variations exist. The patterns thus far observed predict that further codon variation will involve primarily AT-only and AT-rich codons. Thus far none of the GC-only codons have been found to have variable definitions. Blast searches for proteins having sequence homology with the hypothetical protein product of "nonsense" SCOR ORFs reveal no fragments of more than a dozen residues with 50 % conserved identity.

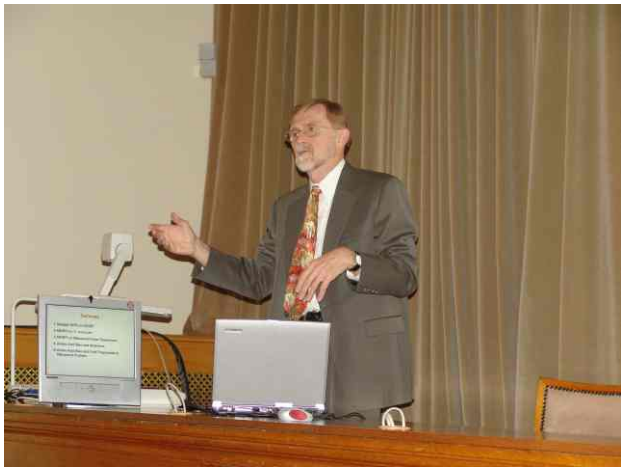
In 1994 LeJohn described the isolation and characterization of two proteins that were products of the sense and antisense strand of the same gene, a 1063 amino acid eukaryotic glutamine dehydrogenase and a 600 amino acid heat shock protein (J. Biol. Chem. **269**, 4513-4531, 1994). Co-isolation of the two proteins using a monoclonal antibody suggested that the two protein products of the same gene formed a complex that was recognized by the antibody. Seventeen codons were entirely unused in the sense/antisense overlapping open reading frames (SAS ORFs) and half of the 64 codons are responsible for the coding of 89% of the 1282 amino acids in the protein products of the overlapping strands. Our subsequent analysis of the frequency of occurrence of SASORFs in the HSP70 family revealed the presence of DORF's in 10% (36/365) of the HSP-70 genes. Two thirds of these DORFs have the same triple bias found in the SCOR SASORFs. Over 80% of the nucleic acid triples in the genes are GC-rich. However the HSP-70 codon-use pattern differs from that of the SCORs in several important ways. Members of the HSP-70 protein family share 70-80% sequence identities. No TORFs were found in the 365 HSP-70 genes examined. A BLAST search for characterized proteins having sequence homology with the hypothetical proteins in the antisense frames consistently detected homologies with fragments of glutamine dehydrogenases, calcium binding proteins, and calcium receptors. The HSP-70 DORFs that do not exhibit the GC-rich codon bias of the SCOR family have an alternate bias best characterized as the GC-rich bias with an AT-drift. This AT-drift pattern occurs in HSP-70 proteins of AT-rich species and illustrates the role of the wobble base in the evolution of genes for specific proteins from a GC-rich to an AT-rich species. Together with Carter (Cell 10, 705-808, 2002) we have suggested that (a) the existence of SASORFs in the SCOR and

* The lecture of the president of the International Union of Crystallography, Professor Duax, at the seminar of the Czech and Slovak Crystallographic Association, April 20, 2005.

HSP-70 families of proteins, (b) the *bonafide* example of a eukaryotic gene that encodes an HSP-70 and a Rossmann fold containing enzyme reported by LeJohn, and (c) the structural similarities between tRNA synthetases I and II and the Rossmann fold dehydrogenase and HSP-70s respectively, support the postulate of Rodin and Ohno (Orign. Life Evo. Biosph **25**, 569-589, 1995) that tRNA synthetase I and II may have evolved from SASORFs of the same primordial gene.

One possible explanation for the conservation of 260 DORFS and 82 TORFS in the family of Rossmann folds is that the primordial member of this family was encoded so early in evolution that most productive mutational development had occurred without introduction of any codons having two or more adenine or threonine residues. It has been repeatedly demonstrated that AT-rich DNA melts at a lower temperature than GC-rich DNA. This is consistent with computational data concerning the relative stability of AT- and GC-rich DNA. It would be reasonable to assume that when the genetic code was evolving the more stable GC-only and GC-rich codons would gain and retain consistent definitions and code specific amino acids more consistently and efficiently than less stable AT-only and AT-rich codons. The fact that the majority of the codons that are least used in coding in the 82 SCOR TORFs and 260 SCOR DORFs are AT-rich and include those that have multiple definitions in different species, is consistent with such a hypothesis. The fact that the partitioning of nucleic acid of triples into subgroups that are GC-only, GC-rich, AT-rich and AT-only is much more pronounced when examined as a function of complementary pairs of triples rather than codons suggests that patterns of relative population are directly related to the relative energies of complementary pairs of triples. The patterns of relative energies of triples are less pronounced in the coding frame because the coding frame is influenced by biological factors. These and other data suggest that the SCOR families of enzymes diverged from a common ancestor that evolved before the AT-rich half of the genetic code was defined.

This work is supported in part by NIH Grant No DK26546.



W. L. Duax