Evolution of modern proteins

SIR—In a recent communication Brenner[1] has outlined an approach to understanding the evolution of modern proteins. In particular he has drawn attention to the active serine residue in a number of proteins and suggested that its progenitor was an active cysteine residue (the serine codon AGY being derived from the cysteine codon TGY by a single base change). Brenner also makes the point that the limited catalytic function of primitive peptides might have been enhanced by the binding of metal ions.

It is of interest that the isolation and sequencing of genes for the selenoenzymes mammalian glutathione peroxidase (GSH-Px)[2-7], and bacterial formate dehydrogenase (FDH)[8] has shown that the active-site selenocysteine residue is specified by the 'stop' codon TGA. Also selenocysteine transfer RNA has been shown to be derived by modification of a specific serine-charged tRNA (possibly via a phospho-seryl tRNA intermediate)[9,10], suggesting that TGA was originally a sense codon in a primitive genetic code but that this general function was lost on the introduction of oxygen to the biosphere[10].

Thus it is possible that serine was specified by the codon TGN in a more primitive genetic code but that the seryl tRNA could be modified to contain other more catalytically active amino acids such as phosphoserine, cysteine, selenocysteine or tryptophan (TGA is read as tryptophan in mitochondria). Insertion of the appropriate modified serine into the peptide sequence would have been dependent on the surrounding context sequences in the messengerRNA. Thus a diversity of active peptides could have been generated using a restricted amount of genetic information, with further diversification dependent on the evolution of greater exclusivity in tRNA charging, and the development of pathways for the *in vivo* production of complex amino acids.

The present-day retention of the selenocysteine coding function of TGA in GSH-Px and FDH is presumed to be dependent on the surrounding nucleotide sequences in the mRNA. It is of note that in GSH-Px the strongly conserved amino-acid sequence around the active site consists entirely of mono-basic, mono-acidic amino acids, suggesting a primitive origin for this region of the enzyme. In fact the whole GSH-Px polypeptide is deficient in aromatic and heterocyclic amino acids, again suggesting an ancient origin for this two exon polypeptide of relative molecular mass 21,000 (21K). FDH on the other hand, appears to be of more recent origin being larger (70K) and containing a higher proportion of more complex amino acids.

Comparison of the region around the selenocysteine in GSH-Px and FDH does not indicate any conservation of amino-acid or nucleotide sequence which might in itself be a signal for translation of the TGA codon. However, it is of note that in both cases the adjacent nucleotide sequences form inverted repeats which are themselves preceded by inverted repeats. Whether such structural motifs play a direct role in translation of the TGA codon as selenocysteine, and are indeed of ancient origin, remains to be seen.

Finally, it is of note that in neither GSH-Px nor FDH is the selenocysteine residue associated with the conserved GXSXG or GXCXG motifs[1] found in many enzymes containing active serine or cysteine residues. This would support the hypothesis that in a primitive genetic code TGN coded mainly for modified serine, the particular modified tRNAs being specified by context sequences in the mRNA. Subsequent evolution led peptides containing an active cysteine residue, and its context sequences, down a separate phylogenetic pathway leading, as proposed by Brenner[1], to the active serine enzymes.

PETER S. GOLDFARB

Molecular Toxicology Group, Department of Biochemistry, University of Surrey, Guildford, Surrey GU2 5XH, UK

1. Brenner, S. *Nature* **334**, 528–530 (1988).
2. Goldfarb, P. *et al. Nucleic Acids Res.* **4**, 3517–3530 (1983).
3. Chambers, I *et al. EMBO J.* **5**, 1221–1227 (1986).
4. Mullenbach, G. *et al. Nucleic Acids Res.* **15**, 5484 (1987).
5. Sukenaga, Y. *et al. Nucleic Acids Res.* **15**, 7178 (1987).
6. Ishida, K. *et al. Nucleic Acids Res.* **15**, 10051 (1987).
7. Ho, Y. *et al. Nucleic Acids Res* **16**, 5027 (1988).
8. Zinoni, F *et al. Proc natn. Acad. Sci. U.S.A.* **83**, 4650 (1986).
9. Sunde, r & Evenson, J. *J. biol. Chem.* **262**, 933–937 (1987).
10. Leinfelder, W. *et al. Nature* **331**, 723–725 (1988).

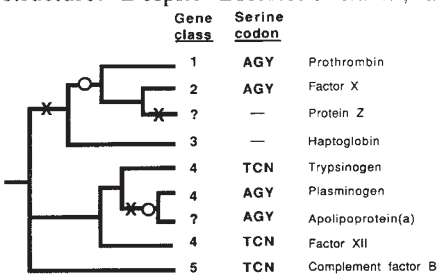
Evolution of an active-site codon in serine proteases

SIR—Brenner[1] proposes that the last common ancestor to vertebrate serine proteases was a cysteine protease that existed billions of years ago, based on the observation that the active-site serine of some serine proteases is encoded by the codon TCN, whereas in others it is AGY. Serine is the only amino acid where it is impossible to change from any one codon to all other possible codons by single nucleotide substitutions without having to pass through an intermediate codon that does not encode serine. I have a simpler explanation for the evolutionary history of this gene family.

The TCN codon for the active-site serine is found in serine protease genes of eubacteria and invertebrates and in some vertebrate genes. In contrast, the AGY codon is found only in serine protease genes of vertebrates, suggesting a late origin from an ancestral serine protease that was of the TCN type. Moreover, the vertebrate AGY proteases are involved in physiological processes found only in vertebrates[2]; thus there is no evidence for an ancient lineage associated with these genes.

Molecular phylogenies of the vertebrate serine proteases have been produced[3]. None of these phylogenies supports Brenner's[1] suggestion that the TCN and AGY proteases represent separate lines of descent from an ancestral protease. Rather, the phylogenies of the genes suggest that the TCN codon evolved on two occasions to an AGY codon, once on the lineage leading to the vitamin K-dependent coagulation factors, and once along the lineage leading to plasminogen and to apolipoprotein(a) (see figure).

Evidence of multiple origins of the AGY serine codon is also found in gene structure. Despite Brenner's claim[1], a strict association of gene structure and serine codon is not observed (see figure). Vertebrate serine protease-like genes have been grouped into five classes[2]. The AGY proteases fall into two clusters within the phylogeny of vertebrate serine proteases, with representatives in three classes. The larger group represents the vitamin K-dependent coagulation factors; the second group includes plasminogen and apolipoprotein(a), which are like the trypsin-like proteases. The incompletely characterized gene for plasminogen[4] seems to have one intron shared with the trypsin-like genes. Thus there seems to be less correlation between active-site codon and gene structure than between amino-acid sequence and gene structure[2].



Tree relating selected vertebrate serine proteases and related proteins together with codon used and exon–intron structure. The phylogeny is based on previous alignments[3], with the introduction of protein Z (ref. 5) and apolipoprotein(a) (ref. 6) adjacent to the most similar sequences. Gene classes, identified by numbers[2], indicate genes which share similar intron–exon structure; uncharacterized genes are indicated by question marks. Protein Z and haptoglobin do not have serine at the active site. X denotes the loss of a serine codon and an open circle the gain of a new serine codon. Brenner[1] proposed that haptoglobin and protein Z are descendants of TCN proteases; molecular phylogenies indicate that both are more closely related to AGY proteases.

My model is based not only on the sequence of the active-site serine codon, but also on amino-acid sequence and gene structure. It seems that the TCN codon has twice been converted into an AGY

© 1988 Nature Publishing Group