J Mol Evol (1983) 19:214–218

Journal of Molecular Evolution

Codon Usage in Muscle Genes and Liver Genes

Kenneth E.M. Hastings and Charles P. Emerson, Jr.

Department of Biology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22901, USA

Summary. Synonymous codon usage frequencies, derived from cDNA clone sequences, were compared for several sets of vertebrate genes. Gene sets as diverse as those expressed in avian skeletal muscle and in mammalian liver showed similar patterns of synonymous codon usage. There were no significant differences suggesting tissue-specific co-adaptation of codon usage patterns and tRNA anticodon profiles. The results indicate a consensus codon usage pattern for vertebrate genes which is largely independent of taxonomic class, tissue of expression, and the cellular fate and rate of evolution of the encoded proteins. Certain elements of the consensus codon usage pattern indicate that it is the product of natural selection and not simply a mutational equilibrium among phenotypically equivalent synonyms.

Key words: Codon: anticodon adaptation – Mutation – Selection

Although the various codon synonyms for any given animo acid are phenotypically equivalent in terms of protein structure, these synonyms are not necessarily used with equal frequency in genes. The observation of non random synonymous codon usage is firmly established in a wide variety of organisms (Grantham et al. 1980; Wain-Hobson et al. 1981) but its biological significance is not entirely clear. In one-celled organisms (bacteria and yeast) codon usage patterns can be correlated with levels of gene activity and with tRNA anticodon profiles. That is, actively expressed genes preferentially use codons that can be read by the most abundant anticodons in the tRNA population (Bennetzen and Hall 1982). This may reflect an optimization of the rate of protein synthesis per mRNA molecule Bennetzen and Hall 1982). A similar relationship, but with an added developmental aspect, has also been considered for multicellular organisms. Tissue-specific features of tRNA populations (Garel 1974; Sprague et al. 1977) and amino acyl-tRNA synthetases (Streheler et al. 1967) have supported the idea that genes encoding tissue-specific proteins might utilize a variety of different codon usage patterns especially adapted for efficient translation in their different cytoplasmic environments. This hypothesis raises the more general question of whether gene primary structure is subject to tissue-specific influences during evolution in multicellular organisms. In order to address these issues, we have compared synonymous codon usage in genes expressed in vertebrate skeletal muscle and genes expressed in vertebrate liver.

We have chosen to compare skeletal muscle and liver in this analysis because both tissues actively express a variety of apparently unrelated genes. By pooling (evolutionarily) unrelated genes we are less likely to be misled by unique gene-specific codon usage features than if we examined individual genes or evolutionarily related gene families. Muscle and liver are the only tissues for which this kind of information is presently available. The data base, summarized in Table 1, consists of sequenced regions of cDNA clones of mRNAs expressed in either embryonic quail skeletal muscle cultures or in adult mammalian liver. Most of the liver data (90% of the codons) relate to the blood proteins serum albumin (Sargent et al. 1981), β fibrinogen (Chung et al. 1981), and prothrombin (MacGillivray et al. 1980) which are synthesized predominantly in the liver. Although other tissues may produce β_2 microglobulin (Parnes et al. 1981) and metallothionein (Durnam et al. 1980), these proteins were included in the analysis because the sequence data concern those particular genes that are

Offprint requests to: K.E.M. Hastings

expressed in liver. The muscle data concern six skeletal muscle-specific components of the contractile apparatus (Hastings and Emerson 1982). Among contractile proteins, only myosin light chain 2 and troponin C are thought to be related by descent from a common ancestral gene (Weeds and McLachlan 1974).

The raw codon usage data were obtained by summing up the number of occurrences of each codon in the appropriate reading frame in each gene segment (see Table 2). ATG and TGG are excluded from all subsequent analyses because these codons have no synonyms. To compare preferences directly among synonymous codons, we calculated the relative use (R) of each of the 59 degenerate codons in each gene set as follows:

$$R = \frac{N_{codon}}{N_{amino\ acid}} \times D$$

where N_{codon} is the total number of times a given codon was used in the gene set, $N_{amino\ acid}$ is the total number of times the amino acid specified by that codon (and its synonyms) is encoded in the gene set, and D is the degeneracy of that amino acid; *i.e.*, the number of synonymous codons for that amino acid. (Note that R = 1 would be expected for each codon if all synonyms are used equally.)

R values determined for liver and muscle gene sets are presented graphically in Fig. 1. In general, codons which are preferred synonyms (R > 1) in the muscle gene set are also preferred in the liver gene set, codons which are disfavored synonyms (R < 1) in the muscle gene set are also disfavored in the liver gene set, and codons which are indifferent synonyms (R ~ 1) in the muscle gene set are also indifferent in the liver gene set. Thus, there is a high degree of similarity of synonymous codon usage patterns in muscle genes and liver genes.

In order to evaluate differences between the two gene sets, codon usage patterns for each amino acid were compared by χ^2 analysis. Only in the cases of alanine, arginine, and leucine did muscle gene synonym usage differ significantly (at the $P < 0.05$ level) from that expected on the basis of liver gene R values. (When Yates' correction was included, the difference in leucine

Table 1. The gene sets compared. All sequence data pertain to protein coding regions and were obtained by cDNA cloning and DNA sequencing methods. The liver data were collected from the literature (references indicated). The muscle data result from our work on contractile protein mRNAs in differentiated muscle cultures derived from embryonic quail (*Coturnix coturnix*) (Hastings and Emerson 1982)

Organism	Protein encoded	Fate of protein	Unit evolutionary period*	Number of codons sequenced
		Muscle genes		
Quail	α actin	Contractile apparatus	>260	42
Quail	α tropomyosin	Contractile apparatus	~150	59
Quail	Myosin light chain 2 (fast)	Contractile apparatus	~ 26	79
Quail	Myosin heavy chain	Contractile apparatus	~ 16	61
Quail	Troponin C (slow)	Contractile apparatus	~130	106
Quail	Troponin I (fast)	Contractile apparatus	~ 15	155
				502 Total
		Liver genes		
Rat	Serum albumin	Secreted	3	608 (Sargent et al. 1981)
Cow	β fibrinogen	Secreted	~ 5	425 (Chung et al. 1981)
Cow	Prothrombin	Secreted	?	160 (MacGillivray et al. 1980)
Mouse	β_2 microglobulin	Cell surface	~ 3	61 (Parnes et al. 1981)
Mouse	Metallothionein	Intracellular	~ 4	61 (Durnam et al. 1980)
				1315 Total

*Unit evolutionary period (U.E.P.) is the average time, in Myear, required for a 1% difference in amino acid sequence to arise between two lineages (Wilson et al. 1977). With the exception of serum albumin (Wilson et al. 1977) the U.E.P.'s shown are our own rough estimates based on a single comparison, in some cases involving less than the entire protein molecule. Divergence times of 260 Myear, and 85 Myear were assumed for avian and mammalian lineages, and different mammalian orders, respectively (Wilson et al. 1977). The calculations were based on the following information. Avian and mammalian α-actins are apparently identical (Vandekerckhove and Weber 1978). The chicken and rabbit homologues of myosin light chain 2 (fast) and troponin I (fast) differ by 15%, and 18% respectively (Matsuda et al. 1977; Wilkinson and Grand 1978). There is one difference between quail and rabbit α-tropomyosins in a stretch of 59 amino acids, and two differences between quail and rabbit troponin C (slow) in a stretch of 106 amino acids (Hastings and Emerson 1982). There are 10 differences between quail and rabbit myosin heavy chains in the C-terminal 61 amino acids (Hastings and Emerson 1982). Metallothionein from mouse liver and horse kidney differ by 21% (Huang et al. 1977). These last two could be paralogous comparisons (see Wilson et al. 1977). Human and bovine β fibrinogen differ by 18% (Chung et al. 1981) and human and rabbit β_2 microglobulin differ by 29% (Parnes et al. 1981). Only the bovine prothrombin sequence is known (Magnusson et al. 1975), so no U.E.P. estimate was possible for this protein

Table 2. Codon usage in muscle genes and liver genes

		Muscle	Liver			Muscle	Liver			Muscle	Liver			Muscle	Liver
phe	TTT	3	17	ser	TCT	2	14	tyr	TAT	4	25	cys	TGT	3	32
	TTC	15	27		TCC	10	24		TAC	4	26		TGC	4	41
					TCA	4	9		TAA	terminate			TGA	terminate	
					TCG	4	2		TAG						
	TTA	1	8									trp	TGG	2	21
	TTG	0	12												
leu	CTT	8	15		CCT	3	17	his	CAT	3	10		CGT	7	9
	CTC	8	19	pro	CCC	9	25		CAC	3	22	arg	CGC	6	6
	CTA	4	3		CCA	4	18	gln	CAA	3	15		CGA	1	6
	CTG	21	41		CCG	0	7		CAG	14	38		CGG	5	10
	ATT	12	13		ACT	8	20	asn	AAT	4	22	ser	AGT	4	7
ile	ATC	10	22	thr	ACC	8	31		AAC	12	35		AGC	6	18
	ATA	4	9		ACA	1	24	lys	AAA	18	45	arg	AGA	0	22
met	ATG	20	26		AGC	1	9		AAG	33	66		AGG	11	13
	GTT	1	12		GCT	21	28	asp	GAT	21	31		GGT	4	16
val	GTC	5	17	ala	GCC	10	40		GAC	19	36	gly	GGC	13	29
	GTA	0	12		GCA	5	21	glu	GAA	34	45		GGA	5	26
	GTG	14	39		GCG	0	4		GAG	43	51		GGG	5	8

The numbers shown are the total number of occurrences of each codon (in the correct reading frame) in each of the gene sets introduced in Table 1. Initiation codons were included in those cases where they were present in the cDNA sequences. Termination codons were not counted

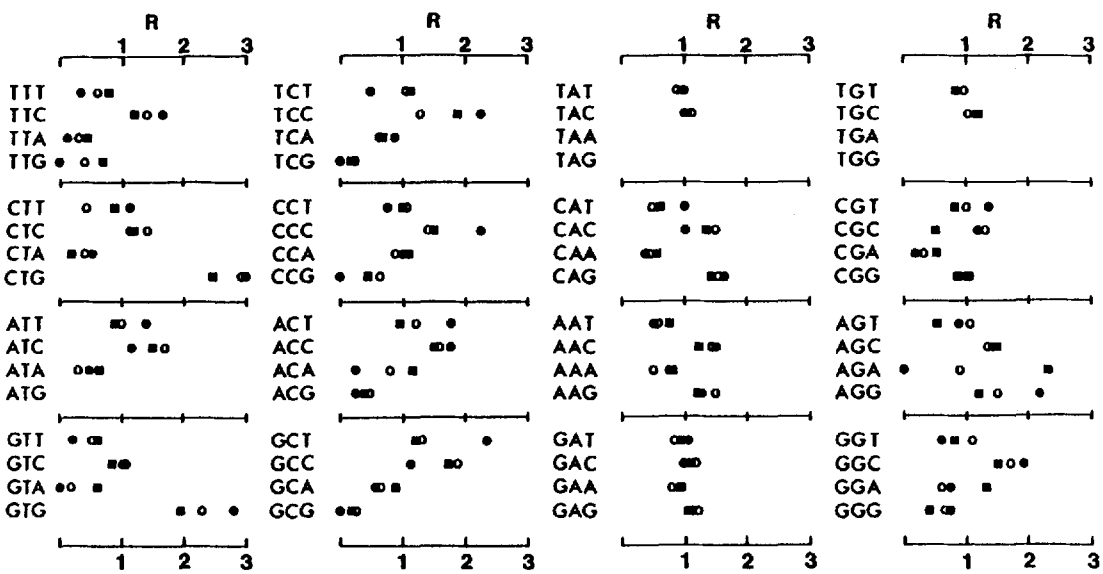

Fig. 1. Relative use (R) of the 59 degenerate codons in three gene sets: Muscle genes (•), liver genes (▪), immunoglobulin, globin, and peptide hormone genes (○)

codon usage was only marginally significant ($0.05 < P < 0.10$)). The muscle and liver synonym usage patterns for the remaining 15 amino acids could not be distinguished at this level of confidence in the χ^2 test, which is consistent with the overall similarity suggested by Fig. 1.

It seems unlikely that the small number of statistically significant differences between the muscle and liver gene sets could represent tissue-specific "matching" of synonymous codon usage patterns in mRNA with the anticodon profiles of the corresponding tRNA populations. The major differences concern the comparative use of GCT vs GCC for alanine, and AGA vs AGG for arginine (Table 2 and Fig. 1). Current evidence indicates that in eukaryotes GCU and GCC are decoded with the same anticodon, and the same may also be true for AGA and AGG (see Nishimura 1979). Thus there is no reason to think that muscle and liver tRNA populations could differ in their relative

abilities to decode GCU vs GCC, or AGA vs AGG. This suggests that the observed differences in alanine and arginine codon usage are probably not related to any differences between muscle and liver tRNA populations.

The best evidence for tissue-specific coadaptation of codon usage patterns and tRNA anticodon profiles comes from studies of the silk gland of the silk worm (Garel et al. 1974; Sprague et al. 1977). Our analysis of synonymous codon usage in muscle and liver genes, and in a third set of genes encoding tissue-specific proteins (see below) indicates that this phenomenon is not of general importance in vertebrates. Several interpretations are possible at this point. First, there may be no relationship at all between codon usage patterns and tRNA populations in vertebrate tissues. Second, there may be a functionally important relationship, but one which does not vary greatly in different tissues. For example, the various abundant mRNAs could all be adapted to the same population of tRNAs in the manner described by Bennetzen and Hall (1982) for bacteria and yeast. Third, tissue-specific codon: anticodon adaptations may be restricted to tissues that, like the silk gland, but unlike muscle and liver, devote much of their translational activity to proteins of unusually simple amino acid composition.

We have also calculated codon R values for a pooled set of 18 genes (genes 73–90 in the compilation of Grantham et al. (1980)) encoding mammalian immunoglobulins, globins, and peptide hormones. These are presented in Fig. 1, along with R values for the muscle and liver gene sets. The obvious tendency for the R values of the three gene sets to cluster for most codons indicates that all three gene sets display very similar patterns of synonymous codon usage. This result argues further against the idea that the use of distinct tissue-specific codon usage patterns might be a generally important facet of tissue-specific gene expression at the translational, or any other level. On the other hand, the codon usage pattern homogeneity indicated in Fig. 1 is entirely consistent with the "genome" hypothesis, or rule, of Grantham et al. (1980), that all genes in a given genome (or type of genome) tend to conform to a common codon usage pattern. The small number of codon usage differences between the muscle and liver gene sets discussed above could reflect incomplete averaging out of gene-specific peculiarities by the gene pooling process. If this is the case, we should expect that these differences will not be maintained when the analysis is extended to include a greater number of muscle and liver genes than is now available.

In light of their similar patterns of synonymous codon usage it is instructive to consider some of the differences between muscle and liver and the particular gene sets compared. Skeletal muscle is derived from embryonic mesoderm, whereas liver is derived from the endoderm (Balinsky 1970). The muscle gene set is entirely avian, the liver gene set entirely mammalian. Most of the liver gene products are secreted proteins (see Table 1), but none of the muscle gene products (contractile proteins) is. Contractile proteins are highly conserved evolutionarily whereas most of the liver proteins are less highly conserved (see Table 1). Thus, apparently none of these factors contributes greatly to synonymous codon usage patterns of vertebrate genes.

The consensus codon usage pattern that emerges in Fig. 1 is apparently a general feature of vertebrate gene primary structure. How this pattern is established and maintained, and what, if any, is its functional significance, are questions we cannot answer directly. But we can consider a related question - can the observed codon usage pattern be explained by mutation alone (among synonyms which are in every sense phenotypically equivalent), or are we obliged to invoke natural selection (with some codons making a greater contribution to the fitness of the organism than other, formally synonymous, codons)?

Consider any codon in genomic DNA. The third position is occupied by either an A · T, or a G · C, base pair. If the distribution were an equilibrium determined simply by the rates of the various single base mutations one would expect that both orientations of, say, a G · C pair would occur equally often. That is, there should be no preference for having the C in the coding DNA strand and the G in the anticoding strand, or vice versa. The same would also be true for A · T base pairs. And, although any ratio of A · T to G · C base pairs could be accomodated by a simple mutation model, one would expect that this ratio would be the same for all synonym groups free of natural selection in any one organism.

If we look at the whole group of two-codon families in Fig. 1 (codons for amino acids having only two codons), we see that the above expectations are a fairly good description of the observed codon usage pattern. In this group (TTPy[1], TAPy, CAPy, CAPu, AAPy, AAPu, GAPy, GAPu, and TGPy) there is no preferred net orientation of A · T or G · C base pairs in the third codon position (i.e., $R_{NNA} \sim R_{NNT}$, and $R_{NNG} \sim R_{NNC}$), and there is a rather homogeneous A · T to G · C ratio, with the latter slightly in excess (i.e., $R_{NNA}, R_{NNT} \leq 1$; $R_{NNG}, R_{NNC} > 1$). Thus a simple model based only on single base mutation rates without any influence of natural selection could explain the general features[2] of synonymous codon usage in two-codon families in Fig. 1. According to this model the slight excess of G · C would be interpreted to indicate that A · T pairs to mutate to G · C pairs more frequently than vice versa.

---

1 Py = T or C, Pu = A or G, N = A, C, G or T, XY = any specified dinucleotide

2 To carry the analysis further, the greater R value inequality for TTPy, CAN, and AAN, as compared with TAPy, TGPy, and GAN, may indicate the operation of additional factors

A different picture emerges in the case of four-codon families (XYN, where all four XYN codons are synonyms). In most four-codon families in Fig. 1 there is a decided preference for one or the other orientation of G · C base pairs in the third codon position. Where XY= TC, CC, AC, GC, or GG, the preference is for C in the coding DNA strand and G in the anticoding strand, i.e., $R_{XYC} > R_{XYG}$. The reverse preference is observed where XY = CT, or GT. This distribution is difficult to explain by models based on mutation alone[3] and therefore seems to indicate that, in four-codon families in vertebrate genes, some synonyms contribute a greater degree of fitness than others and are favored by natural selection. In whatever sense this fitness is reflected phenotypically, it does not appear to vary in different tissues, for the foregoing observations apply without exception to all three gene sets compared in Fig. 1.

The question of the exact functional significance of synonymous codon usage patterns is probably best approached by experiment. It should be possible to interconvert synonymous codons by in vitro mutagenesis in a cloned gene, introduce the altered gene into a host cell, and determine whether any of its functions, e.g., replication, transcription, translation, have been impaired. Our analysis suggests that in designing experiments of this nature concerning vertebrate genes the taxonomic class and tissue of origin of the host cell should be less important considerations than its overall suitability for the experiment. Also, alterations within four-codon families, particularly XYC ⇔ XYG interconversions are more likely to have an effect on function than alterations within two-codon families.

Acknowledgements. We thank O. Colin Stine for discussions that greatly simplified our statistical approach and Benjamin D. Hall for information in advance of publication. This work was supported by research grants from the National Institutes of Health and the Muscular Dystrophy Association (to C.P.E.), and a Muscular Dystrophy Association Post-Doctoral Fellowship (to K.E.M.H.).

References

Balinsky BI (1970) An Introduction to Embryology, Third Edition (W.B. Saunders Company, Philadelphia)

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Chung DW, Rixon MW, MacGillivary RTA, Davie EW (1981) Characterization of a cDNA clone coding for the β Chain of bovine fibrinogen. Proc Natl Acad Sci USA 78:1466–1470

Durnam DM, Perrin F, Gannon F, Palmiter RD (1980) Isolation and characterization of the mouse metallothionein-I gene. Proc Natl Acad Sci USA 77:6511–6515

Garel JP (1974) Functional adaptation of tRNA population. J Theor Biol 43:211–225

Garel JP, Hentzen D, Daillie J (1974) Codon responses of $tRNA^{Ala}$, $tRNA^{Gly}$, and $tRNA^{Ser}$ from the posterior part of the silkgland of *Bombyx mori* L. FEBS Letters 39:359–363

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

Hastings KEM, Emerson CP Jr (1982) cDNA clone analysis of six co-regulated mRNAs encoding skeletal muscle contractile proteins. Proc Natl Acad Sci USA 79:1553–1557

Huang I, Yoshida A, Tsunoo H, Nakajima H (1977) Mouse liver metallothioneins: Complete amino acid sequence of metallothionein-I. J Biol Chem 252:8217–8221

MacGillivary RTA, Degan SJF, Chandra T, Woo SL, Davie EW (1980) Cloning and analysis of cDNA coding for bovine prothrombin. Proc Natl Acad Sci USA 77:5133–5157

Magnusson S, Peterson TE, Sottrup-Jensen L, Claeys H (1975) In: Reich E, Rifkin DB, Shaw E (eds) Complete primary structure of prothrombin: Isolation, structure and reactivity of ten carboxylated glutamic acid residues and regulation of prothrombin activation by thrombin in Proteases in Biological Control. Cold Spring Harbor Laboratory, NY, pp 123–149

Matsuda G, Suzuyama Y, Maita T, Umegane T (1977) The L-2 light chain of chicken skeletal muscle myosin. FEBS Letters 84:53–56

Nishimura S (1979) In: Schimmel PR, Soll D, Abelson JN (eds) Modified nucleosides in tRNA in Transfer RNA: Structure, Properties, and Recognition. Cold Spring Laboratory, NY, pp 59–79

Parnes JR, Velan B, Felsenfeld A, Ramanathan L, Ferrini U, Appella E, Seidman JG (1981) Mouse β2-microglobulin cDNA clones: A screening procedure for cDNA clones corresponding to rare mRNAs. Proc Natl Acad Sci USA 78: 2253–2257

Sargent TD, Yang M, Bonner (1981) Nucleotide sequence of cloned rat serum albumin messenger RNA. Proc Natl Acad Sci USA 78:243–246

Sprague KU, Hagenbüchle O, Zuniga MC (1977) The nucleotide sequence of two silk gland alanine tRNAs: Implications for fibroin synthesis and for initiator tRNA structure. Cell 11: 561–570

Streheler BL, Hendley DD, Hirsch GP (1967) Evidence on a codon restriction hypothesis of cellular differentiation: Multiplicity of mammalian leucyl-sRNA-specific synthetases and tissue-specific deficiency in an alanyl-sRNA synthetase. Proc Natl Acad Sci USA 57:1751–1758

Vandekerckhove J, Weber K (1978) Mammalian cytoplasmic actins are the products of at least two genes and differ in primary structure in at least 25 identified positions from skeletal muscle actins. Proc Natl Acad Sci USA 75:1106–1110

Wain-Hobson S, Nussinov R, Brown RJ, Sussman JL (1981) Preferential codon usage in genes. Gene 13:335–364

Weeds AG, McLachlan AD (1974) Structural homology of myosin alkali light chains, troponin C, and carp calcium binding protein. Nature 252:646–649

Wilkinson JM, Grand RJA (1978) Comparison of amino acid sequence of troponin I from different striated muscles. Nature 271:31–35

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Annu Rev Biochem 46:573–639

3 Such an explanation could be contrived by making the additional assumption that single base mutation rates are determined not only by the identity of the base actually undergoing mutation, but also by the identities of the next one or two adjacent bases. By arbitrary choice of such neighbor effects, any codon usage pattern could be explained without recourse to selection

Received March 30, 1982/Revised September 15, 1982