J Mol Evol (1996) 42:377–381

JOURNAL OF MOLECULAR EVOLUTION

On the Prevalence of Certain Codons (''RNY'') in Genes for Proteins

Thomas H. Jukes

Space Sciences Laboratory, University of California, Berkeley, Berkeley, CA 94720, USA

Received: 23 June 1995 / Accepted: 11 November 1995

Abstract. J.C. Shepherd notes that codons of the type RNY (R = purine, N = any nucleotide base, Y = pyrimidine) predominate over RNR in the genes for proteins. He has hypothesized that RNY codons are the relics of ''a primitive code'' composed of repeating RNY triplets. He found that RNY codons predominated in fourfold RNN codon sets (family boxes). These family boxes code for valine, threonine, alanine, and glycine. We argue that the proposed ''comma-less'' code composed of RNY never existed, and that, in any case, survival of such a code would have long since been erased by mutations. The excess of RNY codons in family boxes is probably attributable to preference for the corresponding tRNAs.

Key words: Genetic code — RNY codons — Comma-less code — Genes for proteins — Amino acid distribution

Introduction

J.C. Shepherd has proposed in various publications (Shepherd 1981a,b, 1983, 1984, 1986, 1990) a hypothesis to the effect that primitive messages were formed by repetition of coding triplets having the form RNY (R = A or G; Y = C or T(U); N = R or Y) and that relics of such messages are found almost universally in present genomes. The hypothesis was criticized by Wong and Cedergren (1986), who found no ''significant support'' in various proteins ''for a primitive repeating-RNY or repeating RRY gene structure.'' They pointed out that an enrichment of RNY codons could be the outcome of natural selection rather than a primitive remnant. Nevertheless, Shepherd's claim has been supported in some publications, including *The Molecular Biology of the Gene* by Watson et al. (1987), which reproduced diagrams by Shepherd (1983) to uphold Shepherd's hypothesis. We have therefore reexamined the ''vestigial'' aspects of Shepherd's proposal, which were challenged in 1988 by Wong and Cedergren.

Shepherd (1981a,b) says that the first code was composed of repeating sequences of coding triplets having the form RNY, that these have been replaced by the present universal code, but that vestiges of the repeating sequences are still detectable. His hypothesis concerning the nature of the first code is founded on publications by Crick et al. (1957) dealing with a comma-less code and with an ancient coding system based on the comma-less code but composed of RRY triplets (Crick et al. 1976).

Discussion

Crick et al. (1957) described a ''theoretical'' code that was readable in only one frame and required no start signal. This code was termed ''comma-less,'' referring to the lack of functional demarcation between coding triplets in its sequence. For example, a sequence could be written as:

. . . U C A, C G G, A U A, U G C . . . , or
. . . U, C A C, G G A, U A U, G C . . .
. . . U C, A C G, G A U, A U G, C . . .

Present address: Space Sciences Lab, University of California, 6701 San Pablo Avenue, Oakland, CA 94608, USA

in which the commas divide the letters into groups of three, each representing one amino acid. ''The problem is how to read the code if the commas are [removed], i.e., a comma-less code'' (Crick et al. 1957). In all justice, it must be noted that Crick and co-workers presented their comma-less code as a ''tentative hypothesis.''

The term ''comma-less'' is a confusing one because commas do not exist in any code; for that matter, they do not exist in nature. The genetic code is comma-less. The translation process is now well known to depend on a ribosome-mRNA binding complex that binds a triplet of nucleotides and goes through a cycle in which the triplet is translated into an amino acid. The process is repeated until a chain-termination codon is reached. There are no ''commas.'' There is an initiation codon, usually AUG, that starts the three-at-a-time translation process.

Crick and co-workers listed two difficulties in the ''coding problem'': First, why are there not 64 kinds of amino acids? Second, how does one know how to choose the groups of three nucleotides?

The first ''difficulty'' is a question rather than a difficulty. It is unanswerable, but the authors get rid of it by devising a code that included only 20 amino acids (the ''magic twenty'') that participate in the biological synthesis of proteins. Crick (1958) provided a penetrating explanation of the listing of the ''magic twenty'' in a classical article.

Crick and co-workers offered a solution to the second difficulty by proposing that ''certain triplets (codons) make sense and some make nonsense'' and that a code could be devised in which the maximum number of amino acids coded was not greater than 20. A solution for 20 was given by first excluding the four triplets AAA, CCC, GGG, and UUU, because if AAA is placed next to AAA, the sequence can be misinterpreted by reading it in the wrong frame. The remaining 60 triplets were grouped into 20 sets of three, each set being cyclic permutations of one another. One such set could be the actual code, and the other two would be excluded. This code would contain no ambiguity (''wobble'') in the third positions of codons.

For some reason, Crick and colleagues thought there should be only one codon per amino acid, and their objective was to reduce the number of codons to 20. An evolutionary difficulty with such a code is that most point mutations would give rise to untranslatable codons. Their comma-less code would stop evolution in its tracks. Earlier, Dounce (1952) had proposed, more logically and correctly, that the code could contain 40 or more codons divided among the amino acids that occur in proteins.

Another difficulty with Crick's comma-less code is in explaining how any messenger sequence requiring it could have originated. There would be only one chance in three (actually, in 3.1) that a nucleotide triplet could be a translatable codon. The odds would be very great against obtaining sequences of as few as 20 consecutive translatable codons by chance. A typical gene for a protein contains about 900 nucleotides. This sequence would have to be composed of 300 translatable codons, and only 1 triplet in 3 is translatable according to Crick's comma-less code. The odds against such a sequence, or even a shorter sequence, occurring or being maintained against the accumulation of mutations are very great. Indeed, such a code would seem to preclude DNA-based life from ever originating. Furthermore, untranslatable codons are not stop codons unless they interact with a release factor for terminating an amino acid sequence (Caskey 1980).

In contrast, the present code has only 3 stop codons in 64 codons; about 1 in 21, on average.

Shepherd's reliance on the comma-less code as the primeval design of $(RNY)_n$ sequences is therefore misplaced and should be rejected. Some other reason must be sought for any prevalence of RNY in coding sequences of genes.

In 1961, Nirenberg and Matthaei discovered that polyuridylic acid could function as a ''synthetic messenger RNA'' for the biological synthesis of polyphenylalanine. Ironically, the triplet UUU which coded for phenylalanine was excluded from comma-free codes, and thus died the 1957 proposal by Crick, Griffith, and Orgel. Discovery of the translation mechanism showed later how the codons were chosen by being placed in a reading frame. However, Crick said in 1988, of the 1957 proposal: ''the correct genetic code . . . has proved decisively that the whole idea is quite erroneous. However, it is just conceivable that *it may have played a role near the origin of life, when the code first began to evolve*, but this is speculation.'' (Emphasis added.) Surely the comma-less code would have been just as useless ''when the code first began to evolve'' as it is today!

In 1976, Crick and colleagues proposed that the primitive message was a ''repeating family of sequences . . . RRY, RRY, RRY . . . where the commas are written to show the correct phase of reading, and for the anticodon [loop] the family 3′ UGYYRUU with the triplet . . . in italics. . . . this restricted base sequence is comma-free in the sense of Crick, Griffith and Orgel (1957)'' and ''the codons allowed are GGY(Gly), GAY(Asp), AGY (Ser), and AAY(Asn).'' This is indeed a meager quota of amino acids for protein synthesis.

So, despite Crick's awareness of the danger—he said in 1988, ''Theorists almost always become too fond of their own ideas. It is difficult to believe that one's cherished theory, which really works rather nicely in some respects, may be completely false''—he has shown fondness for his idea by reviving the comma-less code as a possibility for an era in which its existence cannot be tested. The revised idea (Crick et al. 1976) is that the primitive genetic message was carried in a comma-less code with the sequence $(RRY)_n$. Eigen and Schuster

Table 1. Distribution of RNY, RNR, YNY, and YNR codons in family boxes and nonfamily boxes, stop codons excluded[a]

		No./thousand per codon						
Number in code	Type of codons	HU	YE	EC	HU	YE	EC	Average
8 RNY	Family box	147	157	168	18.4	19.6	21.0	19.7
8 RNR	Family box	114	78	126	14.2	9.8	15.8	13.3
8 RNY	Nonfamily box	157	187	170	19.6	23.4	21.3	21.4
8 RNR	Nonfamily box	175	207	145	21.9	25.8	18.2	22.0
32 RNN	Total	593	629	609				
	Average				18.5 ± 3.3	19.7 ± 7.1	19.0 ± 2.2	19.1 ± 4.0
8 YNY	Family box	115	82	97	14.4	10.2	12.1	12.2
8 YNR	Family box	100	73	112	12.5	9.1	14.0	11.9
8 YNY	Nonfamily box	116	108	99	14.5	13.5	12.4	13.5
5 YNR	Nonfamily box	75	106	78	15.0	21.2	15.7	17.3
29 YNN	Total	406	369	386				
	Average				14.0 ± 0.9	12.7 ± 5.5	13.8 ± 2.2	13.4 ± 2.3

[a] Data from 2,681 human (HU) sequences, 891 yeast (YE) sequences, and 1,562 *E. coli* sequences; 2.8 stop per thousand *found;* 47 stop per thousand *expected*

(1978), building on the model proposed by Crick and co-workers (1976), suggested $(RNY)_n$ instead of $(RRY)_n$, thus bringing the proposal back full circle. They pointed out that it would be advantageous if RNY coded for eight amino acids rather than four as in the case of RRY, and if RNY, but not RRY, is symmetric with respect to both plus and minus strands of DNA, so RNY can be complemented by another RNY. This would enable both strands ''to become equivalent targets for specific recognition by enzymes.'' RNY codons are AAY-(Asn), AGY(Ser), ACY(Thr), ATY(Ile), GAY(Asp), GGY(Gly), GCY(Ala), and GTY(Val). No basic amino acids have RNY codons. This might be a problem in the functioning of proteins.

The case for RNY was examined by Wong and Cedergren (1986). They found no evidence for present existence of a primitive RNY repeating structure. Their conclusion was based on the rates of silent substitutions, on the frequency of base doublets, and on synonymous codon ratios for 52 proteins of *E. coli,* 56 of yeast, 1 of *Drosophila,* and 1 of *Xenopus.* They noted that both yeast and *E. coli* preferred UCY over AGY or UCR codes for serine. Thus, for serine, YNY and YNR were preferred over RNY. They considered the four RNN four-codon sets: ACN, GUN, GCN, and GGN. The rate of synonymous mutations is about 5.1×10^{-9} for yeast. This showed that the excess of RNY cannot be a historical relic traceable to primitive repeating-RNY genes. They discussed the effect of tRNA, of greater abundance leading to higher levels of RNY codons.

Shepherd (1990), 4 years later, made no response to Wong and Cedergren (1986). Shepherd thus allowed their objections to his RNY theory to stand unchallenged.

In an $(RNY)_n$ coding sequence, mutations can take place between A and G in position 1, to any base in position 2, and between C and T in position 3, without removing the RNY format. Thus, AGC can mutate to GTT, which is another RNY, but not to AGA, which is an RNR, while preserving the RNY motif. C and T(U) are always synonymous in the third codon position, and are recognized by the same anticodon, e.g., GAA pairs with UUU and UUC.

There is, indeed, on average, a majority of codons with the form RNY as compared with RNR in coding sequences (see below), but this is quite independent of the discredited comma-less code.

Shepherd (1986) found that there is a predominance of Y over R in third positions of fourfold-degenerate sites (family boxes) but not when RNY and RNR code for different amino acids (nonfamily boxes). He concludes that in primeval times, when the coded protein was improving by many changes of amino acids, those mutated phenotypes having much more efficient proteins would dominate in the population and be accepted; however, ''changes from RNY to RNR giving no change of amino acid could have little chance of surviving.'' But, said Shepherd, when RNY and RNR code for different amino acids, the ratio RNY:RNR may be less than one, because the requirement for the RNR-coded amino acid may be higher. This second conclusion by Shepherd is correct (see below, Table 2).

A reexamination of the question concerning any RNY excess was made with the codon usage in three different types of organisms—human, yeast, and *E. coli* (Wada et al. 1992). This sample had 2,681 sequences (GenBank records) from humans, 891 from yeast, and 1,562 from *E. coli* (Wada et al. 1992). This sample should be adequate, since it contains sequences from three types of organisms, even though there may be some redundancy. The results are given in Table 1.

Table 1 shows that R greatly predominates over Y, 1,834:1,157 = 1.59:1 in first positions of the total of all 61 amino acid codons, i.e., RNN is greater than YNN. Next, Y must be greater than R in third codon positions of RNN for RNY to predominate. Y is greater than R, 1.48:1 in these positions in RNN family boxes, but not in

Table 2. RNR and RNY coding in family boxes and two-codon sets

Amino acid	RNY	RNR	Amino acid	RNY	Amino acid	RNR
Val	99	94	Ile(AUN)	142	Ile(AUA) + Met	95
Thr	105	63	Asn	138	Lys	180
Ala	131	96	Asp	164	Glu	197
Gly	136	67	Ser(AGY)	71	Arg(AGA)	56
Totals	471	320		515		528
Average	118	80		129		132

nonfamily boxes (Table 1). The explanation for the global excess of RNY must be sought in the distribution of synonymous codons in family boxes—a point which Shepherd (1986) emphasized (see above). This is set forth in Table 2.

If all 61 codons in the genetic code are used equally, any 8 codons would be used at a rate of 131 per thousand as compared with an average value of 157 for RNY and 106 for RNR in RNN family boxes (Table 1).

Most of the preference for RNY is attributable to three amino acids—Thr, Ala, and Gly—and this preference for RNY over RNR occurs in family boxes rather than in two-codon sets (nonfamily boxes, Table 1). From Table 2, it appears at first sight that Ile(AUN) contributes to RNY predominance. However, this is outweighed by the RNR data for Lys and Glu. Therefore, RNY preference over RNR is narrowed to Thr, Ala, and Gly, because Val does not participate in RNY preference over RNR (Tables 2 and 3) except in yeast. In contrast to yeast, the values for GUY and GUR are, respectively, 35 and 34; 34 and 36; and 35 and 34 in maize, *Pseudomonas,* and *Bacillus subtilis.*

Preference for RNY is channeled as follows:

1. RNN over YNN 1.59:1. Nine percent of this effect is because of avoidance of stop codons.
2. Remaining RNN preference is 1.45:1. This is divided between RNY and RNR in the proportions of 1.17:1.
3. The RNY:RNR preference is all located in family boxes. RNY:RNR in family boxes = 1.48:1.
4. The ratio of RNY to RNR in family boxes is as follows: Thr 1.67:1; Ala 1.36:1, Gly 2.03:1; Val 1.05:1.

RNY preference is confined to Thr, Ala, and Gly. These 3 amino acids have 19.7% (61/12) of the 61 codons in the code for amino acids, and they constitute 20.0% of amino acids in the sample. However, in the 20.0%, 12.3% is RNY and 7.6% is RNR (instead of 10.0:10.0).

Shepherd (1990) found that RNY codons predominated in protein genes and in exons but did not predominate in noncoding sequences or in introns. In consequence, his program shows that these two categories—protein genes + exons, and noncoding regions + introns—are separable in his Figs. 1–4 (Shepherd 1990).

Why is there a preference for Y-termination codons over R-terminated codons in the case of Thr, Ala, and Gly? A possible explanation lies in a greater use arising from an overabundance of tRNAs containing anticodons GNN, which pair with codons NNY.

Table 3. Codons ending in Y and R compiled for threonine, alanine, glycine, and valine distribution shown in Table 1

	T	A	G	V
Human				
Y	35	49	37	26
R	21	22	35	37
Yeast				
Y	36	41	41	39
R	23	20	15	20
E. coli				
Y	34	41	58	34
R	19	54	17	37
Totals				
Y	105	131	136	99
R	63	96	67	94

Shepherd (1981a) used the following sequences in his calculations: ''the complete genomes of the DNA viruses X174, G4 and fd and of the generally weaker correlations with the same characteristic features found in a DNA virus (simian virus 40), a plasmid (pBR322), an RNA virus (MS2) and various prokaryotic and eukaryotic genes—e.g., a ribosomal protein gene cluster of *E. coli* and the sea urchin histone genes.'' He cites references to the virus genes, but to only two of the ''various prokaryotic and eukaryotic genes,'' so it was not possible to reuse the sequences employed by Shepherd. As an alternative, sequences from human, yeast, and *E. coli* were used by us, the codons for which, tabulated from the GenBank genetic sequence data, were listed by Wada et al. (1992). Their compilation had 1,145,022 codons from 2,681 human sequences, 459,247 codons from 891 yeast sequences, and 524,410 codons from 1,562 *E. coli* sequences. These were tabulated as codons per 1,000 of codon use. This should be a substantial base for examining use of RNY codons.

Increased use of RNY codons could be the result of evolution in the presence of transfer RNA populations that contain GNN anticodons, and of selection for efficient translation.

Shepherd has repeatedly (Shepherd 1983, 1984, 1986) estimated that the time of last use of the comma-less code

appears to be on the order of 3,000 million years ago (+500 or even +1,000 million years). We needed to determine the lifetime of the RNY codon during evolution. If we assume that the rates of formation and loss of RNY are equal, we can calculate the rate of loss of RNY by comparing human and rodent sequences and using a value of 90 MYA for the time of the human/rodent divergence (Collins and Jukes 1994). The comparison was made with 121 gene sequences taken from the compilation by Collins and Jukes (1994) totaling 54,552 codons, 17,178 (31.5%) of which were RNY. (This 31.5% is almost identical with the 30.8% found for humans in Table 1.) For every 100 RNY codons in the human sequence, 88.5% corresponded to RNY in the rodent sequence. Therefore the rate of observed mutation in the 121 comparisons was 11.5 per 90 MY (uncorrected for multiple hits, which would increase the mutation rate). If we assume that the rates of formation and loss of RNY codons are equal, the sequences lose RNY at a rate of 5.75% (11.5 ± 2) per 90 MY or 192% per 3 BY (or 128% per 2 BY, Shepherd's lower figure). Therefore, the amount of time needed to lose any initial excess has been far exceeded, and there can be no original RNY codons surviving in existing gene sequences.

Undoubtedly, the attractive and appealing idea of a ''fossilized'' remnant of primeval codons in existing genes has drawn attention to Shepherd's thesis, e.g., by Watson et al. (1987). The probable reason for the excess of RNY in family box codons of Thr, Ala, and Gly—that this results from a prevalence of the cognate tRNAs—is humdrum by comparison.

Summary

Preference of RNY over RNR was found in family boxes for Thr, Ala, and Gly in protein-coding sequences in a sample containing 2,681 human sequences, 891 yeast sequences, and 1,562 *E. coli* sequences. Valine codons GUY were preferred over GUR in yeast, but not in human, *E. coli*, maize (129 sequences), *Pseudomonas* (259 sequences), or *B. subtilis* (636 sequences). Other vertebrates (mouse, chicks) resemble humans in RNY vs RNR content. The results with Thr, Ala, and Gly confirmed Shepherd's findings (e.g., Shepherd 1990).

The excess of RNY over RNR in Thr, Ala, and Gly codons cannot be attributed to vestiges of an ancient coding system. The possibility is suggested that it may arise from an overabundance of tRNAs for Thr, Ala, and Gly with GNN anticodons.

Acknowledgment. The author is grateful to David Collins, Noboru Sueoka, and an anonymous reviewer for advice and suggestions, and to Carol Fegté for preparing the manuscript.

References

Caskey CT (1980) Peptide chain termination. Trends Biochem Sci 5:234–237

Collins DW, Jukes TH (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20:386–396

Crick FHC (1958) On protein synthesis. Symposium society for experimental biology. Academic Press, New York

Crick FHC (1988) What mad pursuit: a personal view of scientific discovery. Basic Books, New York

Crick FHC, Brenner S, Klug A, Pieczenik C (1976) A speculation on the origin of protein synthesis. Orig Life 7:389–397

Crick FHC, Griffith JS, Orgel LE (1957) Codes without commas. Proc Natl Acad Sci USA 43:416–421

Dounce AL (1952) Duplicating mechanism for peptide chain and nucleic acid synthesis. Enzymologia 15:251–258

Eigen M, Schuster P (1978) The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. Die Naturwissenschaften 65:341–369

Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA 47:1588–1602

Shepherd JCW (1981a) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Shepherd JCW (1981b) Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J Mol Evol 17:94–102

Shepherd JCW (1983) From primeval message to present-day gene. Cold Spring Harb Symp Quant Biol 47:1099–1108

Shepherd JCW (1984) Fossil remnants of a primeval genetic code in all forms of life? Trends Biochem Sci 9:8–10

Shepherd JCW (1986) Origins of life and molecular evolution of present-day genes. Chemica Scripta 26B:75–83

Shepherd JCW (1990) Ancient patterns in nucleic acid sequences. In: Doolittle RF (ed) Methods in enzymology 183: molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, San Diego, pp 180–193

Wada et al. (1992) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 20: Suppl pp 2112, 2114

Watson JD et al. (1987) Molecular biology of the gene. Benjamin Cummings, Palo Alto

Wong JTF, Cedergren R (1986) Natural selection versus primitive gene structure as determinant of codon usage. Eur J Biochem 159:175–180