J Mol Evol (1983) 19:226-236

Journal of Molecular Evolution
© Springer-Verlag 1983

Choice of Base at Silent Codon Site 3 is not Selectively Neutral in Eucaryotic Structural Genes: It Maintains Excess Short Runs of Weak and Strong Hydrogen Bonding Bases

B. Edwin Blaisdell

Linus Pauling Institute of Science and Medicine, Palo Alto, California 94306, USA

Summary. On the average in the coding sequences of 30 eucaryotic structural genes the weak hydrogen bonding, W, (A or T) or strong hydrogen bonding, S, (C or G) base in codon site 3 was chosen to be unlike its neighbors on both sides up to two sites away. This preference produced the nonrandom excess of runs W and S of length one and two and the defict of long runs observed earlier (Blaisdell 1982). The neighbors in the different codon, 3' to codon site 3, were as important in determining the choice as were the neighbors 5' in the same codon. Every amino acid except methionine and tryptophan, of least frequent occurrence, permits choice of W or S. The persistence of this preference could explain the observation that the rate of substitution of codon site 3 in fuctional genes is considerably less than in synonymous pseudo genes.

Key words: Eucaryotic DNA – Structural genes – Primary DNA sequence – Nonrandomness – Run lengths – Weak hydrogen bonding bases – Coding sequences – Codon site 3 – Rate of substitution

Introduction

Recent observation (Blaisdell 1982) has found in a collection of about 30 eucaryotic nuclear DNA sequences that the coding and noncoding subsets are both nonrandom, but in different ways. Coding sequences exhibit an excess of runs or length 1 and 2 and a deficit of long runs of the base classes weak hydrogen bonding, W, (A or T) and strong hydrogen bonding, S, (C or G). Noncoding sequences exhibit a deficit of runs of length 1 and 2 and an excess of long runs of the base classes purine, R, (A or G) and pyrimidine, Y, (C or T). These respective kinds of nonrandomness were found in DNA sequences whose coding part coded for proteins of widely different function, in widely different eucaryotic species for the same protein and, in the same species, for related sequences that diverged a long time ago and that now show large differences in base and, if coding, amino acid sequence, Table 1. This paper studies the mechanisms by which the W,S nonrandomness is produced in coding sequences. The earlier paper noted that every amino acid (except methionine and tryptophan which are specified by a single codon and are of least frequent occurrence in proteins (Dayhoff 1978) may be specified by a codon having either W or S in site 3. Thus suitable choice of the base at site 3 in the codons might produce the excess runs of length 1 or 2 of W and S and the deficit of long runs of them. In turn the deficit of long runs of S might prevent the formation of excessively stable hairpin loops in the secondary structure of mRNA and the deficit of long runs of W might insure that the secondary structure be sufficiently stable. The advantage of a mRNA secondary structure of moderate stability, neither too high nor too low, was postulated in the earlier paper (Blaisdell 1982). Too high stability would impede the dissolution of the secondary structure during decoding on the ribosome, too low would produce a mRNA structure open and straggly which might impede diffusion through or along the nuclear membrane and through the cytoplasm to the ribosome. The survival value of a moderation of stability of the secondary structure of mRNA and its insurance by suitable choice of the base in codon site 3 might explain the observation that evolutionary base substitution in codon site 3 is more rapid in pseudogenes than in similar functional genes (Li et al. 1981). Also suitable choice of the amino acids coded for, either singly or by their succession could contribute to the nonrandomness. The collection of DNA sequences studied here is the same as that

Table 1. Identification of Genes

1	Human	Alpha 2 globin	Proudfoot and Maniatis 1980
2	Human	Beta globin	Lawn et al. 1980
3	Human	A gamma globin	Slightom et al. 1980
4	Human	Delta globin	Spritz et al. 1980
5	Human	Epsilon globin	Baralle et al. 1980
6	Mouse	Alpha globin	Nishioka and Leder 1979
7	Mouse	Beta globin major	Konkel et al. 1979
8	Mouse	Beta globin minor	Konkel et al. 1979
9	Rabbit	Beta globin	van Ooyen et al. 1979
10	Chick	Beta globin	Richards et al. 1979
11	Human	Immunoglobulin Kappa constant	Hieter et al. 1980
12	Mouse	Immunoglobulin Kappa constant	Altenburger et al. 1981
13	Mouse	Immunoglobulin Gamma I constant	Takahashi et al. 1980
14	Mouse	Immunoglobulin kappa Variable	Nishioka and Leder 1980
15	Mouse	Immunoglobulin Gamma 2 variable	Sakano et al. 1980
16	Human	Preproinsulin	Ullrich et al. 1980, Bell et al. 1980
17	Rat	Preproinsulin	Lomedico et al. 1979
18	Chick	Preproinsulin	Perler et al. 1980
19	Human	Cortico-lipotropin	Chang et al. 1980
20	Rat	Prolactin	Gubbins et al. 1980
21	Human	Interferon alpha 2	Goeddel et al. 1980
22	Human	Interferon beta	Law et al. 1981
23	Yeast	Glyceraldehyde-3-Phosphate Dehydrogenase	Holland and Holland 1979
24	Yeast	N-(5'phosphoribosyl) -Anthranilate Isomerase	Tschumper and Carbon 1980
25	Mouse	Alpha amylase	Young et al. 1981
26	Yeast	Actin	Ng and Abelson 1980
27	Sea urchin	Histone H2B	Sures et al. 1978
28	Sea urchin	Histone H3	Sures et al. 1978
29	Chick	Ovalbumin (Fragment)	Robertson et al. 1979
30	French bean	Phaseolin	Sun et al. 1981

studied in the earlier paper, Table 1, except that the short immunoglobulin joining sequences are omitted and chick beta globin has been added to provide a wider range of species for the beta globins.

Results and Discussion

Table 2 presents the results of analysis of variance of observed counts minus expected counts for the 23 = 8 possible contiguous W, S triplets in each of the three possible phases, codon sites (1,2,3), (3,1,2) and (2,3,1). Phase (1,2,3) means the triplet evaluated occurs in succession in sites 1, 2, 3 of the same codon, phase (3,1,2) means the triplet evaluated occurs in succession in site 3 of a codon and in sites 1 and 2 of the codon 3' to it, and phase (2,3,1) means the triplet evaluated occurs in succession in sites 2 and 3 of a codon and in site 1 of the codon 3' to it. The expected counts are calculated using the observed fractions, for each gene, of W or S in the separate codon sites, 1, 2 and 3 of the complete coding sequence. For sites 1, 2, 3 respectively the fractions of W range from 31–60, 49–70, 10–65 and medians are 44.5, 59, 35. For each phase separately and for each subsetting of the triplets the average values (observed counts – expected counts) for the 30 genes in the collection are sorted in increasing order. Negative values show a dificit of observed counts with respect to expected, positive values an excess. In coding the triplets W represents a weak hydrogen bonding base (A or T), S a strong hydrogen bonding base, X represents W or S, Y represents W or S but not the same as X, O represents both W and S. The F statistic (Miller 1981) is a test of the hypothesis that there are some differences between the averages of the subsets which are large compared with the differences between the members within the several subsets; it increases with increasing differences between the averages. For example, for phase (1,2,3) column 2, the value -log p = 13 means that the probability, if there were no differences between the averages of the 8 possible W, S triplets that F be 12.61 or larger is about 10(-13) where the number in parenthesis is the exponent of the preceding number, a highly significant result. The mean square error (ms error in the table) is a measure of how well the individual values for each of the 30 sequences center about the average value of each triplet in the analysis. The layouts of Tables 2, 3 and 5 are similar.

The following features can be observed in Table 2. First consider phase (1,2,3). In columns 2 and 3 it is obvious that the 8 triplets have an understandable ordering by complementary pairs. The first pair (SSS, WWW) = XXX where all three codon sites are occupied by a base of the same W, S class has the greatest deficit of observed counts. The second pair (SWW, WSS) = YXX where site 3 is the same as site 2 but different from site 1 has a lesser deficit. The third pair (WSW, SWS) = XYX where site 3 is different from site 2 but the same as site 1 has an excess. And the fourth pair (WWS, SSW) = YYX where site 3 is different from both sites 1 and 2, which are of course the same, has the greatest excess. Application of the multiple range test (Duncan 1955) gives at

Table 2. Analysis of variance of W,S triples: Ordered average values of counts (observed – expected)[a]

Codon sites (1,2,3)								
Triple, value	SSS	-6.66	XXX	-5.82	OXX	-3.95		
	WWW	-4.98						
	SWW	-2.91	YXX	-2.07				
	WSS	-1.24						
	WSW	2.22	XYX	3.06	OYX	3.95		
	SWS	3.89						
	WWS	4.00	YYX	4.84				
	SSW	5.67						
deg. freedom		7,232		3,236		1,238		
F		12.61		28.39		72.30		
-log p		13		16		15		
ms error		50.09		49.95		51.70		
Codon sites (3,1,2)								
Triple, value	SSS	-4.88	XXX	-4.61	XXO	-3.21		
	WWW	-4.35						
	WWS	-2.07	XXY	-1.80				
	SSW	-1.53						
	WSW	2.69	XYX	2.95	XYO	3.21		
	SWS	3.21						
	SWW	3.20	XYY	3.46				
	WSS	3.72						
deg. freedom		7,232		3,236		1,238		
F		28.71		67.36		172.09		
-log p		27		33		27		
ms error		13.57		13.41		14.33		
Codon sites (2,3,1)								
Triple, value	SSS	-7.85	XXX	-7.19	XXO	-3.93	OXX	-3.20
	WWW	-6.54						
	WWS	-1.31	XXY	-0.66			OXY	3.20
	SSW	-0.01						
	SWW	0.13	YXX	0.79	XYO	3.93		
	WSS	1.44						
	WSW	6.41	YXY	7.06				
	SWS	7.72						
deg. freedom		7,232		3,236		1,238		1,238
F		33.73		77.46		101.02		58.79
-log p		31		36		19		12
ms error		26.52		26.50		36.62		41.83

[a] In phase (1,2,3) the W,S occupant of codon site 3 is chosen preferentially to be unlike the base next preceding it in the same codon and to a lesser degree to be unlike the base next but one preceding it in the same codon. Phase (3,1,2) shows the same preference with respect to the bases succeeding site 3 in the succeeding codon and phase (2,3,1) shows an even stronger preference with respect to the immediate neighbors on both sides simultaneously

the 5% level the multiple comparison graph (Lehmann 1975, p. 239)

```
1 2 3 4 5 6 7 8
---     -------
  ---
    ---
```

for the 8 triplets columns 2 and 3. Here the 8 digits 1 through 8 designate by their ranks the 8 triplets in column 2, that is, 1 means SSS, 2 means WWW, 3 means SWW etc. A line is drawn under those (ordered) sets of triplets whose range does not exceed the maximum range expected for a random sample of the triplet averages (the Duncan test statistic). In this case the interpretation of the comparison graph is: triplet 1, SSS, is not less than triplet 2, WWW, but is less than the other 6 triplets, 3 through 8. In turn, triplet 2, WWW, is not less than triplet 3, SWW, but is less than the other 5 triplets, 4 through 8. Further, triplet 3, SWW, is not less than triplet 4, WSS, but is less than the other 4 triplets, 5 through 8. The common line under 5 through 8 means that no one of these is less than any of the others. The absence of a line under both 4 and 5 means that each of 1 through 4 is less than each of 5 through 8; that is there is a clean separation between those triplets 1 through 4 each having the W, S occupant of site 3 the same as the occupant of the preceding site 2 and those triplets 5 through 8 having the occupant of site 3 different from the occupant of the preceding site 2.

Analysis of variance for the above 4 complementary pairs, columns 4 and 5 shows a thousand fold increased significance over columns 2 and 3, -log p 16 > 13. Also there is a small improvement in the mean square error showing that it is quite acceptable to pool the counts for the complementary pairs. Columns 6 and 7 show the results of further pooling into subsets where site 3 is the same as or different from site 2 but the values for both W and S in site 1 are combined. The result is still highly significant but less so than for columns 4 and 5 and the mean square error is considerably larger. These comparisons show that the pooling has now been too comprehensive and that site 1 has an influence on the choice of occupant for site 3, as is also apparent in columns 2 to 5. The multiple comparison of the multiple range test for the 4 complementary pairs in columns 4 and 5 is

```
1 2 3 4 .
    ---
```

In columns 2 and 3 SSS is less than WWW and WWS is less than SSW and both of these differences are significant at the 1% level by the rank sum analog of the Duncan multiple comparison test (Miller 1981, p. 157). These significant differences suggest, in codons, that the continuation of the string of SS is more avoided than the continuation of the string WW and that the cessation of the string SS is more sought than the cessation of the string WW, and all of these effects are stronger than the choice with respect to a single preceding base class in site 2. This bias toward the occupant of site 3 being different from site 2 and even site 1 obviously promotes the previously observed excess of short runs W, S nonrandomness (Blaisdell 1982).

In Table 2 phase (3,1,2) the triplets are formed from site 3 of each codon and sites 1 and 2 of the succeeding codon instead of sites 1 and 2 of the same codon. The tabulated results are similar to those for phase (1,2,3) for each classification. The ordering of the 8 W, S triplets is the same bearing in mind that codon site 3 is now first in the triplet instead of last. The following differences can be noted: for each classification, for the present phase (3,1,2) compared with phase (1,2,3), the significance levels are higher (eg -log p 33 > 16) and the mean square errors are smaller (13.41 < 49.95), both in column 5. Both of these differences indicate that the occupant of codon site 3 is more influenced by the two bases which succeed it in the next codon than by the two bases that precede it in the same codon. The decrease in mean square error shows that the variation within the given sets is more influenced by neglecting the effect of the succeeding bases than by neglecting the effect of the preceding bases. The untabulated results are also similar. The multiple comparison of the multiple range test for the 8 triplets in columns 2 and 3 is

```
1 2 3 4 5 6 7 8
---  ---  -------
```

like that for phase (1,2,3) except that there is also a clear separation between triplets 1 and 2 each having site 3 the same as both succeeding sites 1 and 2 and triplets 3 and 4 each having site 3 the same as only the immediately succeeding site 1. The multiple comparison of the multiple range test for the 4 complementary pairs of triplets in columns 5 and 6 is

```
1 2 3 4 .
    ---
```

And again in columns 2 and 3 SSS is less than WWW and SWW is less than WSS but neither difference is significant at the 5% level by the same test as before. This bias toward site 3 being different from following site 1 and even site 2 also promotes the excess of short runs W, S nonrandomness. This is a broadening to the W, S context of the earlier finding (Bullock and Elton 1972) that the frequency of codons with C in site 3 preceding codons with G in site 1 is less than expected.

In Table 2 phase (2,3,1) the triplets are formed from site 3 of each codon, the preceding site 2 of the same codon and the succeeding site 1 of the succeeding codon. The results are now somewhat different from those for phases (1,2,3) and (3,1,2). For the 8 individual triplets in columns 2 and 3 and for the 4 sets of complementary pairs in columns 4 and 5 the significance levels are in both cases higher than for either of phases (1,2,3) or (3,1,2), -log p 36 > 33 > 16, showing that the selection of the occupant of site 3 is more determined by both its nearest neighbors than by either one alone. The mean square error in these columns is less than for phase (1,2,3) but greater than for phase (3,1,2) indicating that the neglect of the influence of second nearest neighbors has substantial effects. In columns 6 through 9 the significance levels are much lower than in columns 2 through 5 and the mean square errors are much larger. Both these differences indicate that combining the values for both W and S in either immediate neighboring site has large effects, again evidence that the neighbors on both sides of site 3 determine the choice of its occupant. The untabulated results are also different. The 8 triplets again have an understandable ordering but one whose interpretation is different. The multiple comparison of the multiple range test is now

$$\underline{1\quad 2}\quad \underline{3\quad \underline{4\quad 5\quad 6}}\quad \underline{7\quad 8}$$

showing a clean separation between triplets 1 and 2 each having site 3 the same as the immediate neighbors on both sides and triplets 3 through 6 having site 3 the same as the immediate neighbor on only one side (and of course also having site 3 different from the immediate neighbor on only one side) and there is also a clean separation between the latter set of 4 triplets and triplets 7 and 8 having site 3 different from the immediate neighbor on both sides. The multiple comparison of the multiple range test for complementary pairs in columns 4 and 5 is now

$$1\quad \underline{2\quad 3}\quad 4\ .$$

And again in columns 2 and 3 SSS is less than WWW in the most avoided complementary pair 1 and WSW is less than SWS in the most sought complementary pair 4 but neither difference is quite significant at the 5% level. This bias toward site 3 being simultaneously different from site 2 preceding and site 1 succeeding also promotes the excess of short runs W, S nonrandomness.

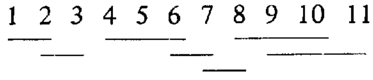
Table 3 presents the results of analyses of variance of the pooled phases (1,2,3), (3,1,2), (2,3,1) coded 1, 3, 2, respectively appended to the W, S codes. In columns 2 and 3 it is obvious that the 24 triplets have an understandable ordering by complementary pairs as listed in column 4. Since comparison is now being made between distinct phases the definition of complementary is extended from complementary bases alone to comple-

Table 3. Analysis of variance of pooled W,S triples, codon sites (1,2,3), (3,1,2), and (2,1,3): Ordered average values of counts (observed – expected)[a]

			Triple and			
Row	Phase	Value	Rows	Value	Rows	Value
1	SSS2	-7.85	1,3	-7.19	1-6	-5.87
2	SSS1	-6.66	2,4	-5.82	7-14	-0.94
3	WWW2	-6.54	5,6	-4.61	15-22	3.58
4	WWW1	-4.98	7,8	-2.49	23-24	7.06
5	SSS3	-4.88	9,11	-1.38		
6	WWW3	-4.35	10,12	-0.66		
7	SWW1	-2.91	13,14	0.79		
8	WWS3	-2.07	15,16	2.45		
9	SSW3	-1.53	17,19	3.46		
10	WWS2	-1.31	18,20	3.55		
11	WSS1	-1.24	21,22	4.84		
12	SSW2	-0.01	23,24	7.06		
13	SWW2	0.13				
14	WSS2	1.44				
15	WSW1	2.22				
16	WSW3	2.69				
17	SWW3	3.20				
18	SWS3	3.21				
19	WSS3	3.72				
20	SWS1	3.89				
21	WWS1	4.00				
22	SSW1	5.67				
23	WSW2	6.41				
24	SWS2	7.72				
deg. freedom		23,696		11,708		3,716
F		19.40		40.07		136.34
-log p		60		67		~75
ms error		30.06		29.89		30.52

[a] The pattern of preference for the occupant of side 3 found for each of the phases separately in Table 2 has a simple understandable relation to the pattern found when they are analyzed jointly in Table 3

mentary bases in complementary phases. For example, SWW1 and WWS3, rows 7 and 8 are complementary since both show W in site 3, preceded or succeeded, respectively by W in the nearest neighbor site and S in the second nearest neighbor site. Analysis of variance for these 12 complementary pairs shows a large increase in significance, -log p 67 > 60 for columns 4 and 5 vs columns 2 and 3, respectively. As before, the slight decrease in the mean square error shows that it is quite acceptable to pool the counts for these complementary pairs. Application of an hierarchical clustering algorithm using the average measure of distance between sets (Becker and Chambers 1981) to the 24 values in column 3 produced the well-separated clusters, rows 1 through 6, 7 through 13, 14 through 21, and 22 through 24. Adjustments of these clusters gives the understandable subsets {1 through 6}, {7 through 14}, {15 through 22}, and {23, 24} corresponding respectively to (a) site 3 the same as both other sites in all three phases, (b) site 3 the same as only one of its adjacent neighbors in phase (2,3, 1) or the same as only its nearest neighbor (and not its second nearest neighbor in phases (1,2,3) and (3,2,1)), (c) site 3 different from its nearest neighbor in phases (1,2,3) and (3,1,2) and (d) site 3 different from both of its adjacent neighbors in phase (2,3,1). Analysis of variance for these 4 subsets shows a further large increase in significance -log p 75 > 67 > 60 at the expense of a very small increase in mean square error. The multiple comparison of the multiple range test for the 24 triplets in columns 2 and 3 shows no readily understandable structure (data not shown). Even the multiple comparison of the multiple range test for the 12 complementary pairs does not have a simple structure



through there is a clean separation between subsets {1 through 3} and {4 through 12}, between {4 through 11} and {12} and a suggestion of a separation between {4 through 7} and {8 through 11}, all in agreement with columns 6 and 7. However, the multiple comparison of rank sum tests for neighboring pairs (Lehmann 1975, pp. 241 and 242) gives the more comprehensible structure

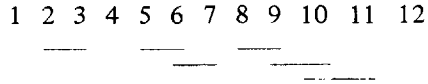


Table 4. Natural logs of binomial probability (too many minus too few counts) for individual genes[a]

tbl 3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
tbl 1																								
1	0	-2	-1	-2	-2	-1	-2	0	2	-3	0	-2	0	0	0	-4	-1	0	3	1	0	2	1	2
2	-6	-3	-5	0	-5	-5	-5	-3	-1	-1	-1	2	-3	1	2	1	2	2	3	3	-1	2	2	6
3	-6	-4	-5	-3	-2	-3	-1	-1	0	0	0	2	1	2	2	2	0	2	1	2	0	2	0	3
4	-8	-3	-4	-1	-2	-4	-4	-3	-2	-2	-2	2	-3	1	2	1	2	1	3	2	1	4	2	7
5	-6	-2	-3	-1	-3	-1	-4	-3	0	-1	-3	1	0	2	2	-1	1	1	3	2	1	2	1	3
6	-2	-2	-2	-2	-5	-5	0	-2	1	0	1	1	-5	-1	-1	-2	2	1	6	0	1	2	2	4
7	-4	-2	-4	-1	-3	-4	-3	-2	-1	-1	-1	1	-2	0	0	0	2	0	4	1	1	3	2	5
8	-5	-2	-3	0	-3	-3	-3	-2	0	-2	-2	1	-2	2	1	-1	2	0	4	2	1	3	1	5
9	-7	-3	-5	0	-5	-3	-4	-7	0	-1	-1	2	-2	1	2	1	1	4	4	3	-1	2	2	6
10	-1	-1	-3	-1	-3	0	0	-2	0	2	0	1	0	-2	0	0	2	1	5	0	0	0	2	3
11	-1	-5	-4	-2	-3	-4	-2	-2	-1	0	1	-2	-2	-2	2	5	-1	4	-1	4	-2	1	4	4
12	-9	-4	-7	-2	-8	-12	-1	-1	-3	2	0	2	-3	-3	2	6	3	2	3	2	0	1	4	7
13	-8	-7	-6	-2	-7	-9	-4	-2	-1	-2	0	1	-3	0	3	5	2	4	2	6	-1	3	5	12
14	-8	-4	-1	-1	-5	-3	-5	1	1	-4	-2	-1	-2	2	4	1	0	2	1	7	-2	1	2	6
15	-2	-7	-2	-2	-2	-1	-3	-1	1	-3	0	-3	0	0	2	1	-2	3	0	4	0	3	4	4
16	-3	-2	-4	-2	-2	0	-2	-2	0	-1	0	1	-3	1	0	0	1	1	3	1	0	2	1	4
17	-3	-2	-5	-3	0	-3	-2	-2	-2	-1	-1	0	-1	0	1	2	1	1	0	1	1	2	2	4
18	-6	-4	-3	-2	-2	-4	-3	-1	0	-2	0	1	-2	2	0	1	2	0	2	2	1	4	1	6
19	0	1	-4	-2	2	2	-2	0	-3	-2	-3	-2	1	-1	0	0	2	-2	1	-2	4	3	2	2
20	-1	-2	-3	-1	-3	-4	-1	0	-1	2	1	-1	-2	-3	0	2	2	1	2	1	-1	0	4	2
21	-4	-6	-5	-3	-4	-3	-2	1	1	-1	-1	-2	2	0	6	3	-1	3	-1	4	-1	0	3	4
22	-7	-6	-4	-2	-5	-2	-1	-2	0	1	1	1	0	1	3	3	-1	4	-1	3	-1	0	1	3
23	-12	-20	-15	-28	1	0	1	-1	-2	-3	-4	-7	11	5	1	0	2	-1	0	-2	21	16	6	4
24	-1	-2	0	0	-2	0	-1	-2	0	-1	0	0	-2	-1	-1	0	0	1	2	1	0	2	1	2
25	-1	-4	-1	0	-3	0	-1	-1	2	0	2	-1	1	0	-2	-2	-1	1	2	1	0	2	1	0
26	-7	-12	-5	-23	0	1	3	-2	-1	-3	-4	-7	5	4	1	-1	0	0	1	-6	23	9	4	4
27	-2	-2	-5	-1	0	-3	-3	-2	-4	0	-2	-2	0	-2	-2	1	2	-2	2	-1	1	4	2	2
28	-5	-1	-8	-3	0	-2	-3	-4	-4	-1	-3	0	0	1	0	1	3	0	2	0	4	3	2	4
29	-11	-2	-3	-3	-4	-3	1	-3	-2	1	-1	3	-2	0	1	2	2	2	3	-1	3	1	2	4
30	-3	-3	-3	-1	0	-2	1	0	-2	2	2	2	1	0	-2	2	1	1	-2	-1	0	1	0	-2

[a] The patterns observed for the averages of 30 genes are generally possessed by the 30 genes severally

in full agreement with columns 6 and 7. Neither of these tests can be used for columns 6 and 7 because interpretive tables are available only for subsets of equal cardinality. However, a multiple F test (Duncan 1955; Miller 1981, p. 97) shows that each of the 4 subsets in columns 6 and 7 differs from each of the others at a highly significant level, $p < 0.0005$ (data not shown).

Table 4 shows that the patterns found for the averages of 30 DNA sequences in Tables 2 and 3 are exhibited by each severally. The columns of Table 4 are designated by the row numbers in Table 3 of the respective triplets and phases, the rows by the row numbers in Table 1 of the respective DNA sequences. The entries in the table are $-\ln(1-B) + \ln B$ where B is the binomial distribution B(k;n,p) = sum m = 0 to k (b(m;n,p)), the binomial probability of observing k or fewer occurrences of a given triplet in a given phase in n codons when the probability of occurrences of the triplet is p = product j = 1 to 3 (f(i,j)) where f(i,j) is the observed fraction of base i, i in {W,S}, in codon site j, j in {1,2,3}. Then 1-B is the probability of observing more than k occurrences of the given triplet in the given phase. Positive entries show more occurrences than expected, negative entries fewer. Columns 1–6 corresponding to all three W, S bases the same in the triplet are generally negative and columns 23 and 24 corresponding to the W, S base in site 3 being different from both the nearest neighbors on its two sides are generally positive. An absolute value of the rounded natural logarithm equal to 2 corresponds to a significance level between 0.08 and 0.22. Note that the outlying values in rows 23 and 26 (both for yeast), columns 1, 2, 3, 4, 21 and 22 probably contribute in large part to the less significant values for -log p and mean square error for phase (1,2,3) compared with phase (3,1,2) in Table 2, and to the contrary ranking of the corresponding triplets in columns 2 and 3, rows 2, 4, 5, 6, 17, 19, 21, 22 in Table 3.

A verification that the results in Tables 2 and 3 are not produced by the large number of globins in the sample was obtained by repeating the calculations for phase (2,3,1) on the same subsample as used before (Blaisdell 1982). This sample of 13 coding sequences consists of 1 of 10 globins (rabbit beta), 1 of 5 immunoglobulins (mouse kappa constant), 1 of 3 insulins (human), 1 of 2 interferons (human beta), 1 of 2 histones (sea urchin H3), and all of the other coding DNA sequences. Pooling the triplets XXY and YXX which do not differ significantly the averages are -8.10, 0.65, 6.79, F = 34.00, -log p = 11. these values may be compared with the values from the total sample, -7.19, 0.13, 7.06, F = 114.34, -log p = 35.

Since Tables 2 and 3 indicated that second nearest neighbors influenced the choice of W, S occupant of site 3, table 5 presents a more direct examination of this influence by analysis of variance of 5-tuples, phase (1,2,3, 1,2). The F test for the 32 5-tuples is highly significant, -log p = 30, the mean square error is only 7.51, considerably less than any of the values found in Tables 2 and 3 for triplets, a clear demonstration that second nearest neighbors have an influence. It is easily observable that the 32 5-triplets have an understandable ordering by complementary pairs, Table 5, columns 4 and 5. Analysis of variance of these 16 complementary pairs shows significance considerably increased from columns 2 and 3 -log p 37 > 30 even though the members of the complementary pairs (7,10) through (27,32) have considerable differences in their ranks in the set of 32. Nonetheless, the small decrease in the mean square error shows that it is quite acceptable to pool the counts for these complementary pairs nonadjacent in the ordering. The nonadjacency of the complementary pairs is apparently a consequence of the conflicting ranks of the phases of the component triplets. For example, in the pairs (8,13) and (9,12) the rank in phase (1,2,3) of the component WSW corresponding to 8 is 5 and of SSS corresponding to 9 is 1, but the rank in phase (3,1,2) of component WWW corresponding to 8 is 2 and of SWS corresponding to 9 is 5. It is not surprising that the values for all 4 rows 8, 9, 12, 13 are about the same and that the pair (9,12) can occur between the elements of the pairs (8,13). In fact, in rows 8, 9, 11–24 where site 3 has different neighbors on its 2 sides, there is a strong negative rank correlation (Spearman statistic = 1184, p = 0.0016) (Lehmann 1975, p. 301) between the ranks of phases (1,2,3) and (3,1,2) as found in Table 2. This observation shows that the extreme ranks for one of these phases are associated with the opposite extreme rank of the other phase which results in relatively small changes in the values in Columns 3 over these 16 rows. The multiple comparisons of the multiple range test of the 32 5-tuples or of the 16 complementary pairs show no readily understandable structure (data not shown).

Application of the hierarchical clustering algorithm with the complete measure of distance between sets to the 32 values in column 3 produces the well separated clusters rows 1–3, 4–6, 7–8, 9–19, 20–25, 26–30, and 31–32. Adjustments to these clusters gives the 7 understandable subsets shown in column 6 corresponding respectively to (a) all 5 sites being the same, (b) 4 sites the same and one different (including sites 2, 3, 4 of course), (c) sites 2, 3, 4 the same but different from sites 1 and 5, (d) 3 contiguous sites the same (including site 3, of course) and exactly one of sites 2 and 4 different from 3, (e) no 3 contiguous sites the same and exactly one of site 2 and 4 different from 3, (f) both of sites 2 and 4 different from 3 with 3 or 4 of the sites occupied by W and (g) both of sites 2 and 4 different from 3 with 3 or 4 of the sites occupied by S. Analysis of variance of these 7 subsets, columns 6 and 7 shows a further increase in significance compared with columns 4 and 5, 40 > 37, and no change in mean square error showing acceptable pooling. The multiple comparison of the multiple F test for these 7 subsets is at the 5% level

Table 5. Analysis of variance of W, S 5-tuples, codon sites (1, 2, 3, 1, 2): Ordered average values of counts (observed – expected)[a]

Row	5-tuple	Value	Rows	Value	Rows	Value	Rows	Value
1	WWWWW	-2.61	1,2	-2.60	1,2	-2.60	1–7 10	-1.79
2	SSSSS	-2.58	3,5	-2.04	3–6	-1.80	8,9, 11–13, 15, 16 22	-0.39
3	SSSSW	-2.46	4,6	-1.57	7,10	-0.96	14, 17–21, 23,24	0.43
4	SWWWW	-1.62	7,10	-0.96	8,9, 11–13, 15,16 22	-0.39	25–32	1.75
5	WWWWS	-1.61	8,13	-0.81	14, 17–21, 23,24,	0.43		
6	WSSSS	-1.52	9,12	-0.72	25–28	1.46		
7	WSSSW	-1.13	11,15	-0.43	29–32	2.03		
8	WSWWW	-1.08	14,20	0.08				
9	SSSWS	-0.82	16,22	0.39				
10	SWWWS	-0.79	17,21	0.42				
11	SSSWW	-0.73	18,23	0.50				
12	WWWSW	-0.62	19,24	0.73				
13	SWSSS	-0.53	25,31	1.64				
14	SWWSW	-0.46	26,30	1.70				
15	WWWSS	-0.12	27,32	1.73				
16	WWSSS	-0.06	28,29	1.92				
17	SWWSS	-0.01						
18	WSWWS	0.04						
19	SSWWS	0.29						
20	WSSWS	0.63						
21	WSSWW	0.84						
22	SSWWW	0.85						
23	SWSSW	0.96						
24	WWSSW	1.17						
25	WWSWW	1.22						
26	WSWSW	1.48						
27	WWSWS	1.51						
28	SWSWW	1.64						
29	WSWSS	1.76						
30	SWSWS	1.80						
31	SSWSS	2.23						
32	SSWSW	2.33						
deg. freedom		31,928		15,944		6,953		3,956
F		7.97		15.45		37.43		69.67
-log p		30		37		40		43
ms error		7.51		7.50		7.49		7.57

[a] The pattern of preference for the occupant of codon site 3 found for the phases (1, 2, 3), (3, 1, 2) and (2, 3, 1) separately in Tables 2 and 3 are maintained when these phases are part of the more comprehensive phase (1, 2, 3, 1, 2)

1 2 3 4 5 6 7

which suggests the coarser equal cardinality subsetting shown in columns 8 and 9. These subsets are (a) 3 or more contiguous sites the same including sites (2,3,4), (b) 3 contiguous sites the same but not including sites (2,3,4) so that site 3 is the same as one of sites 2 or 4 and different from the other, (c) no 3 contiguous sites the same and site 3 is the same as one of sites 2 or 4 and different from the other, and (d) site 3 different from both sites 2 and 4. Analysis of variance of these 4 subsets, columns 8 and 9 shows a further increase in significance compared with columns 6 and 7, -log p 43 > 40, but a small increase in the mean square error indicating that pooling has now been too comprehensive. The multiple comparison of the multiple range test shows that each of these subsets is different from each of the others at the 1% level. The ranking of the 7 subsets in columns

6 and 7 and the very low mean square error compared with Table 3 give very clear evidence that second nearest neighbors on both sides have a substantial influence on the choice of the occupant of site 3 of a codon. The ranks in rows 8, 9, 11–24 of the triplets in sites 1, 3, 5 as found in Table 5, column 2 are highly correlated with their ranks in Table 2 column 2 phase (2,3,1) (Spearman statistic 224, $p = 0.0049$). The bias observed in Table 5 toward the occupant of site 3 being different from its neighbors twice removed also promotes the excess of short runs W, S nonrandomness.

As mentioned at the beginning, the earlier paper (Blaisdell 1982) suggested that the choice of the amino acids coded for, either singly or by their succession could contribute to the observed short W and S run nonrandomness. Analysis of variance of the pairs XX and XY only in codon sites 1 and 2 showed a weak significance of the difference for the values of counts observed minus counts expected, $F(3,116) = 2.30$, $-\log p = 2$. This result shows some preference for amino acids with different W, S bases in codon sites 1 and 2 and obviously contributes to the excess short run W and S nonrandomness. The effect is not nearly as strong as those influencing the choice of occupant of codon site 3, $-\log p$ $2 < 16, 33$ or 36 from columns 4 and 5 of Table 2. Analysis of variance of the pairs XNX and XNY (where N means that the occupant of the site is ignored) in phase (2,3,1) found the respective average values of observed minus expected to be -0.71 and -0.42, showing a slight preference for a W, S difference, but not at a significant level, $p = 0.68$. Similarly analysis of variance for the 16 pairs of W, S pairs in sites 1 and 2 in successive codons found no significant differences, $p = 0.79$. In this analysis expected values were calculated from the observed fraction of W, S pairs in codon sites 1 and 2 and not from the fractions of the separate bases W and S in the respective codon sites. Thus it appears that in the whole set of 30 DNA coding sequences there is no evidence that the choice of amino acid sequence contributes to the excess W, S short run nonrandomness. However, inspection of the raw data seemed to show similar and significant differences for the 10 globins and analysis of variance for the 16 pairs confirmed this impression, $F(15,146) = 3.78$, $-\log p = 5$. Similarly, for the 5 immunoglobulins $F(15,16) = 3.21$, $-\log p = 3$. But in neither of these cases was a readily discernable W, S pair pattern observed. Furthermore, correlation of the values of observed minus expected for these two sets was -0.54, $p = 0.016$ showing that successive W, S pairs giving high values in one set tended to give low values for the other set. The observed significant differences probably reflect the constraints of producing a functioning protein and considerable sequence homology within the sets rather than the constraints of maintaining excess W, S short run nonrandomness. It is concluded that choice of the succession of amino acids does not contribute materially to the observed excess of short runs W, S nonrandomness.

In conclusion, it has been found that on the average in the coding sequences of 30 eucaryotic structural genes the weak hydrogen bonding, W, or strong hydrogen bonding, S, base in codon site 3 is chosen to be unlike its neighbors on both sides up to two sites away. This preference obviously promotes the nonrandom excess of runs of W and S of length one and two and the deficit of long runs observed earlier in coding sequences (Blaisdell 1982). The neighbors in the different codon 3' to codon site 3 are as effective or more so in determining the choice than are the neighbors 5' in the same codon. This observation supports the suggestion (Blaisdell 1982) that the conservation of the cited properties of coding sequences throughout evolution is due to the desirability of having the numerous hairpin loops (Heindel et al. 1978) in mRNA bound with moderate strength, neither too weak nor too strong, or to the desirability of avoiding long W (A or T) runs capable of being confused with the A, T rich signal (Pribnow box) for the initiation of transcription of the A, T rich signal for the attachment of the poly A tail to the transcript, or to the desirability of avoiding long S (C or G) runs capable of forming left-handed Z conformation DNA double helices. The results in this paper support these suggestions rather than the suggestion of the desirability of having the codon-anticodon binding of moderate strength (Grantham et al. 1981). Whatever its function the observed bias in the choice of the occupant of site 3 in codons makes unlikely the supposition that mutations in site 3 are selectively neutral (King and Jukes 1969; Perler et al. 1980) and offers an explanation of the observation (Li et al. 1981) that the rate of mutation of nonfunctional (pseudo) genes at all sites is about twice that of site 3 (generally silent) in the codons of functional genes.

Acknowledgements. I am grateful to E. Zuckerkandl for asking me to examine the statistical properties of DNA sequences and to R.F. Doolittle for suggestions for improving the exposition. I thank C. Gorham for preparation of the manuscript.

References

Altenburger W, Neumaier PS, Steinmetz M, Zachau HG (1981) DNA sequence of the constant region of the mouse immunoglobulin kappa chain. Nucleic Acids Res 9:971–981

Baralle FE, Shoulders CG, Proudfoot NJ (1980a) The primary structure of the human epsilon-globin gene. Cell 21:621–626

Baralle FE, Shoulders CC, Goodbourn S, Jeffreys A, Proudfoot NJ (1980b) The 5' flanking region of human epsilon globin gene. Nucleic Acids Res 8:4393–4404

Becker RA, Chambers JM (1981) S, a language and system for data analysis. Bell Laboratories, Murray Hill

Bell GI, Pictet RL, Rutter WJ, Cordell B, Tischer E, Goodman HM (1980a) Sequence of the human insulin gene. Nature 284:26–32

Bell GI, Pictet R, Rutter WJ (1980b) Analysis of the regions flanking the human insulin gene and sequence of an Alu family member. Nucleic Acids Res 8:4091–4109

Blaisdell BE (1982) A prevalent persistent global nonrandomness that distinguishes coding and noncoding eucaryotic nuclear DNA sequences. J Mol Evol (in press)

Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P (1978) Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc Natl Acad Sci USA 75:4853–4857

Bullock E, Elton RA (1972) Dipeptide frequencies in proteins and the CpG deficiency in vertebrate DNA. J Mol Evol 1: 315–325

Chang ACY, Cochet M, Cohen SN (1980) Structural organization of human genomic DNA encoding the proopiomelanocortin peptide. Proc Natl Acad Sci USA 77:4890–4894

Duncan DB (1955) Multiple range and multiple F tests. Biometrics 11:1–42

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, De Riel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) The structure and evolution of the human beta-globin gene family. Cell 21:653–668

Feller W (1967) An introduction to probability theory and its applications, 3rd edition, John Wiley & Sons, New York

Goeddel DV, Yelverlon E, Ullrich A, Heyneker HL, Miozzari G, Holmes W, Seeburg PH, Dull T, May L, Stebbins N, Crea R, Maeda S, McCandliss R, Sloma A, Tabor JM, Gross M, Familetti PC, Pestka S (1980) Human leukocyte interferon produced by E. coli is biologically active. Nature 287:411–416

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

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage in a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Gubbins EJ, Maurer RA, Lagrimini M, Erwin CR, Donelson JE (1980) Structure of the rat prolactin gene. J Biol Chem 225: 8655–8662

Hardison RC, Butler ET, Lacy E, Maniatis T, Rosenthal N, Efstratiadis A (1979) The structure and transcription of four linked rabbit beta-like globin genes. Cell 18:1285–1297

Heindell HC, Liu A, Paddock GV, Studnicka GM, Salser WA (1978) The primary sequence of rabbit alpha globin in mRNA. Cell 15:43–54

Hieter PA, Max EE, Seidman JG, Meizel JV, Leder P (1980) Cloned human and mouse kappa immunoglobulin constant and J region genes conserve homology in functional segments. Cell 22:197–207

Holland JP, Holland MJ (1979) The primary structure of a glyceraldehyde-3-phosphate dehydrogenase gene from Saccharomyces cerevisiae. J Biol Chem 254:9839–9845

Kafatos FC, Efstratiadis A, Forget BG, Weissman SM (1977) Molecular evolution of human and rabbit beta globin mRNAs. Proc Natl Acad Sci USA 74:5618–5622

Kataoka T, Kawakami T, Takahashi N, Honjo T (1980) Rearrangement of immunoglobulin gamma-1 chain gene and mechanism for heavy-chain class switch. Proc Natl Acad Sci USA 77:919–923

King FL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Konkel DA, Maizel JV, Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosome beta-globin genes. Cell 18:865–873

Lawn RM, Efstratiadis A, O'Connell C, Maniatis T (1980) The nucleotide sequence of the human beta-globin gene. Cell 21: 647,651

Lawn RM, Adelman J, Franke AE, Houck M, Cross M, Najarian R, Coeddel OV (1981) Human fibroblast interferon gene lacks introns. Nucleic Acids Res 9:1045–1052

Lehmann EL (1975) Nonparametrics. Holden-Day, San Fransisco, p 239

Li W, Gojobori T, Nei M (1981) Pseudo genes as a paradigm of neutral evolution. Nature 292:237–239

Lomedico P, Rosenthal N, Efstratiadis A, Gilbert W, Kolodner R, Tizard R (1979) The structure and evolution of the two nonallelic rat preproinsulin genes. Cell 18:545–558

Miller RG (1981) Simultaneous Statistical Inference. 2nd Edition, Springer, New York, p. 157

Newell N, Richards JE, Tucker PW, Blattner FR (1980) J genes for heavy chain immunoglobulins of mouse. Science 209: 1128–1132

Ng R, Abelson J (1980) Isolation and sequence of the gene for actin in Saccharomyces cerevisiae. Proc Natl Acad Sci USA 77:3912–3916

Nishioka Y, Leder P (1979) The complete sequence of a chromosomal mouse alpha-globin gene reveals elements conserved throughout vertebrate evolution. Cell 18:875–882

Nishioka Y, Leder PJ (1980) Organization and complete sequence of identical embryonic and plasmacytoma kappa V-region genes. Biol Chem 255:3691–3694

Pan J, Elder JT, Duncan CH, Weissman SM (1981) Structural analysis of interspersed repetitive polymerase III transcription units in human DNA. Nucleic Acids Res 9:1151–1170

Peck LF, Wang JC (1981) Sequence dependence of the helical repeat of DNA in solution. Nature 292:375–378

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J (1980) The evolution of genes: the chicken preproinsulin gene. Cell 20:555–566

Proudfoot NJ, Brownlee CG (1976) Noncodong region sequences in eucaryotic messenger RNA. Nature 263:211–214

Proudfoot NJ, Maniatis T (1980) The structure of a human alpha globin pseudogene and its relationship to alpha globin gene duplication. Cell 21:537–544

Rhodes D, Klug A (1981) Sequence dependent helical periodicity of DNA. Nature 292:378–380

Robertson MA, Staden R, Tanaka Y, Catterall JF, O'Malley Brownlee CG (1979) Sequence of three introns of the chick ovalbumin gene. Nature 278:370–372

Sakano H, Huppi K, Heinrich G, Tonegawa S (1979) Sequences at the somatic recombination sites of immunoglobulin light chain genes. Nature 280:288–294

Sakano H, Maki R, Kurosawa Y, Roeder W, Tonegawa S (1980) Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy chain genes. Nature 286:676–683

Slightom JL, Blechl AE, Smithies O (1980) Human fetal G-gamma and A-gamma globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638

Spritz RA, De Riel JK, Forget BG, Weissman SM (1980) Complete nucleotide sequence of the human delta-globin gene. Cell 21:639–646

Sun SM, Slightom JL, Hall TC (1981) Intervening sequences in a plant gene: comparison of the partial sequence of cDNA and genomic DNA of French bean phaseolin. Nature 289:37–41

Sures I, Lowry J, Kedes LH (1978) The DNA sequence of sea urchin (S. purpuratus) H2A, H2B and H3 histone coding and spacer regions. Cell 15:1033–1044

Takahashi N, Kataoka T, Honjo T (1980) Nucleotide sequences of class-switch recombination region of the mouse immunoglobulin gamma2b-chain gene. Gene 11:117–127

Tilghman SM, Tiemeier DC, Seidman JG, Peterlin BM, Sullivan M, Maizel JV, Leder P (1978) Intervening sequence of DNA identified in the structural portion of a mouse beta globin gene. Proc Natl Acad Sci USA 75:725–729

Tschumper G, Carbon J (1980) Sequence of a yeast fragment containing a chromosomal replicator and the TRPI gene. Gene 10:157–166

Tsujimoto Y, Suzuki Y (1979) The DNA sequence of B bombyx mori fibroin gene including the 5' flanking, mRNA coding, entire intervening and fibroin protein coding regions. Cell 18:591–600

Ullrich A, Dull TJ, Gray A, Brosius J, Sures I (1980) Genetic variation in the human insulin gene. Science 209:612–615

van Ooyen A, van den Berg J, Mantei N, Weissmann C (1979) Comparison of total sequence of a cloned rabbit beta-globin gene and its flanking regions with a homologous mouse sequence. Science 206:337–344

Young RA, Hagenbuchle O, Schibler U (1981) A single mouse alpha-amylase gene specifies two different tissue-specific mRNAs. Cell 23:451–458

Received November 5, 1982/Revised February 15, 1983