Volume 11 Number 7 1983 Nucleic Acids Research

Statistical characterization of nucleic acid sequence functional domains

T.F.Smith[1] M.S.Waterman[2] and J.R.Sadler[3]

[1]Dept. Physics, Northern Michigan Univ., Marquette, MI 49855, [2]Dept. Biochemistry and Biophysics, School of Medicine, Univ. California, San Francisco, CA 94121, and [3]Dept. Biochemistry/Biophysics/Genetics, Univ. Colorado Health Sciences Center, Denver, CO 80262, USA

Received 1 September 1982; Revised and Accepted 1 March 1983

ABSTRACT

It has long been recognized that various genome classes were distinguishable on the basis of base composition and nearest neighbor frequencies. In addition Grantham et al. (8) have recently presented evidence that these distinctions are preserved at the level of codon usage. As discussed in this report it is now clear that these and related statistics can uniquely characterize the various functional domains of the genome. In particular peptide coding, intervening segments, structural RNA coding and mitochondrial domains of the vertebrate genome are uniquely characterizable. The statistical measures not only reflect understood functional differences among these domains but suggest others. The ability of these simple statistics of nucleic acid sequences to reflect so much of the encoded complex pattern information and/or effects of selective constraints is somewhat surprising.

Here, we investigated the statistical measures most distinctive of the various domains and then linked them to our current understandings in so far as possible.

INTRODUCTION

There now exists a large body of nucleic acid sequence data. In this study we review the statistical characteristics of the various taxonomic classes and functional domains--protein encoding, structural-RNA encoding, introns, etc.--now accessible within the current data base (7). A comparative survey of 83 vertebrate sequences containing more than one domain each along with some viral and bacterial sequences representing over two hundred thousand bases was carried out. The obvious advantage of this large data set is that both the identification and comparison of patterns particular to both individual and classes of sequences can be carried out in concert.

Some properties earlier identified as characteristic of entire genomes are now confirmed as characteristics of individual sequences; in addition, patterns which are diagnostic of vertebrate coding, intervening and mitochondrial domains have been identified.

The statistical analyses employed began by obtaining the oligonucleotide distributions for all sequence domains under investigation. In addition, dyad

frequency distributions were obtained for various fixed intervening intervals (an interval of zero producing the nearest neighbor frequencies while an interval of two allowed for the investigation of codon-codon positional correlations). While such frequency tables can be compared directly, the simplest comparative statistic is the chi-square measure. The chi-square statistic is a measure of the sum of squares of the deviations between observed values, O_i, and particular model values, M_i, compared to an expected error (normally the square root of the model value),

$$x^2 = \sum_{i=1}^{n} \left\{ \frac{O_i - M_i}{\sqrt{M_i}} \right\}^2 = \sum_{i=1}^{n} \frac{(O_i - M_i)^2}{M_i} \quad .$$

Chi-square values on the order of the number of model degrees of freedom are the expected if the proposed model actually underlies the observations. Large chi-square values obtained, for example, for the observed dyad distributions as compared to those predicted from the base composition alone, indicate as expected, that the dyad distributions are the results of constraints beyond those determining base composition. Chi-square values obtained from such a simple model as base composition alone are indicators only of *relative* nonrandomness.

Finally, a simple heuristic measure of strand pairing asymmetry

$$AS = \frac{|[A] - [T]| + |[G] - [C]|}{[A] + [T] + [G] + [C]} \quad ,$$

was calculated for each sequence region investigated. The AS values range from zero (complete symmetry in pairing composition) to unity. Here the square brackets indicate occurrences of each base along the particular strand. For sequences of three to five hundred bases in length of equal base composition, the expected value of AS is less than 0.10.

RESULTS

A survey of 83 vertebrate sequences encoding at least one complete protein and some associated non coding segments supports the hypothesis that the various domains are distinct even at a simple statistical level. The most trivial is base composition: the 35 kilo base pairs of vertebrate peptide coding sequences, CDS's, display a slight C + G preference and a slight A over T asymmetry, while the associated 25 kilo base pairs of intervening sequences, IVS's, display a strong A + T preference as well as T over A and G over C strand asymmetries. The vertebrate rRNA have a relative high G + C content of over 60% as expected from secondary structure stability requirements, an

observation however, which does not hold for many known nonvertebrate rRNA's. The CDS domains of the three vertebrate (7) mitochondria appear uniquely characterized by their overall composition, particularly the extreme lack of G (~ 12.5%) and an A over T asymmetry. (Interestingly, these Mitochondria compositional characteristics are much more yeast-like than bacteria-like.)

However characteristic these compositional properties may be, they are not diagnostic, since they manifest themselves only when averaged over many typical domains. The nearest neighbor frequencies are more diagnostic. The most striking of these, is the CpG suppression (13), and in the CDS's, the additional TpA suppression along with the TpG elevation (relative to that expected from base composition alone). Displayed in Table 1 are the composite averaged nearest neighbor frequencies, relative to their expected values given composition alone.

The sequence chi-square values (relative to base composition expectations) in Table 2 are generally significantly greater for the CDS domains than for the associated IVS's. More than the suppression of TpA (in CDS's) is involved in this difference since the stronger suppression of CpG in the IVS's generally compensates for the reduced TpA suppression. In fact the observed high chi-square values can be generated by other than the suppression of these two doublets. This is shown by the thymidine kinase gene of Herpes simplex virus and the X. Laevis 18srRNA sequence. Both sequences show expected values of CpG and TpA given composition alone, yet they have chi-square values of 25.4 and 127.3 respectively. In addition, both of these sequences show no strand pairing asymmetry.

The viral entries in Table 2 reveal an additional idiosyncrasy of CpG suppression: two complete viral genomes, SV40 and the Bk papovavirus, along with partials from the human influenza and Herpes simplex virus sequences were examined. The SV40 and Bk papovavirus show some of the strongest CpG suppressions observed while the thymidine kinase gene of Herpes simplex and the influenza virus hemagglutinin and neuraminadase genes show little or no suppression of CpG.

Not all the analyzed sequences displayed high chi-square values. For example, the 970 base pair IVS in the silkworm fibroin gene has a very low nearest neighbor chi-square and thus its dyad frequencies are predicted rather well from base composition alone. There are also two vertebrate coding sequences, the mouse kappa light chain and the human alpha-2 globin, which have rather low dyad Chi squares reflecting among other properties the lack of any strong CpG or TpA suppression.

Table 1

Relative Nearest Neighbor Frequencies

A) Vertebrate CDS's 34723 bp (Invertebrate CDS's)

[A] = .246 (.25)		[C] = .265 (.27)		[G] = .264 (.26)		[T] = .226 (.22)	
AA	1.12 (1.1)	AC	.86 (0.9)	AG	1.10 (0.8)	AT	.87 (1.1)
CA	1.20 (1.2)	CC	1.10 (1.0)	CG*	.41 (0.7)	CT*	1.35 (1.1)
GA	1.11 (1.1)	GC	1.02 (1.0)	GG	1.02 (1.0)	GT	.81 (0.9)
TA*	.47 (0.5)	TC	.99 (1.0)	TG*	1.55 (1.5)	TT	.92 (0.9)

B) Vertebrate IVS's 24729 bp (Alu/B1/4.5s repetitive family)

[A] = .268 (.33)		[C] = .207 (.22)		[G] = .222 (.23)		[T] = .303 (.22)	
AA	1.13 (1.16)	AC	.85 (0.82)	AG	1.20 (1.23)	AT	.83 (0.75)
CA	1.18 (1.25)	CC	1.21 (1.16)	CG*	.23 (0.24)	CT*	1.26 (1.28)
GA	.99 (0.91)	GC	.93 (1.01)	GG*	1.26 (1.30)	GT	.83 (0.86)
TA*	.76 (0.68)	TC	1.03 (1.09)	TG	1.12 (1.11)	TT	1.10 (1.25)

C) Mitochondrial CDS's 22778 bp

[A] = .307		[C] = .285		[G] = .126		[T] = .281	
AA	.94	AC	.95	AG	1.11	AT	1.07
CA	1.01	CC	1.08	CG*	.62	CT	1.08
GA	.94	GC	1.06	GG*	1.74	GT*	.63
TA	1.08	TC	.95	TG	.96	TT	.99

D) Bacterial CDS's 84355 bp

[A] = .256		[C] = .244		[G] = .257		[T] = .244	
AA*	1.25	AC	.89	AG	.83	AT	1.01
CA	1.00	CC	.94	CG	1.10	CT	.95
GA	.96	GC	1.19	GG	.93	GT	.91
TA*	.76	TC	.96	TG	1.14	TT	1.12

Nearest neighbor frequencies are presented as ratios to the expected frequencies given only base composition. The asterisk indicates the most diagnostic neighbors. The vertebrate entries do not include any immunoglobulin variable regions. The invertebrate CDS and vertebrate Alu/B1/4.5s values are both based on less than 3000 bp and thus may be subject to rather large sampling errors; however, there is a high correlation with the vertebrate CDS in the first case and with the vertebrate IVS in the second.

It can also be seen in Table 2 that there is no obvious correlation between strand pairing asymmetry and the chi-square values. Compare the 316-bp mouse gamma 2b (Mouse GAM2B) heavy chain IVS and the human fetal gamma globin-b 858-bp IVS (Human HBGAB). The first of these shows total CpG

suppression with a dyad chi-square value of 54.6 and only an 0.08 asymmetry value. While the second has a nearly equal dyad chi-square value of 53.3 (also reflecting in part CpG suppression), yet this sequence contains the largest strand pairing asymmetry measured, of 0.50.

The triad nearest neighbor chi-square values in Table 2 correlate well with the dyad values. In coding sequences these values measure in part the degree to which the codon usage deviates from that expected from base composition alone. The details of such deviations have been well documented by Grantham et al. (8,9). Although codon usage in known vertebrate sequences is dominated by CpG suppression, and to a lesser extent by TpA suppression and TpG elevation, there are exceptions. For example, the alpha hemoglobins show only slight CpG suppression reflecting, in part, their encoding of Arg by CGX rather than only by AGG as in the beta-globins. Another important differential codon usages between the globin genes is the apparent prohibition in alpha globins of TTT codons for Phe.

That such a simple statistic, as a dyad chi-square, can indicate differential constraints on the various genome domains is made clear by calculating dyad frequencies for neighboring triplet positions, how often base X at i follows base Y at i+3. In non-coding regions one expects little deviation from random, but within a coding sequence one might expect strong correlations, particularly since the middle bases of the codons are known to reflect the chemical nature of the encoded amino acid. Thus, selection for particular amino acid neighbors would be reflected in middle base correlations, and this is in fact seen. For example, while Table 2 gives a value of 26.0 for the overall dyad chi-square of mouse KAPJ CDS, chi-square values of 28.6 ($P \simeq .5\%$), 45.1 ($P << .1\%$) and 18.0 ($P \simeq 5\%$) are given in Table 3 for the first to first, second to second, and third to third position correlations among neighboring codons when compared to that expected from base composition. The same calculations for this gene's four IVS's give an average chi-square of only 10.5 ($P \simeq 50\%$) and a maximum of 19.9 ($P \simeq 5\%$).

Careful examination of these data reveals that the high middle base-middle base correlations often reflect high XTX/XTX, XAX/XTX and XTX/XAX frequencies. There are exceptions such as the silkwork fibroin and the rat amylase sequences, neither of which show strong middle-middle base correlations.

The correlation among third base positions led to an analysis of codon-codon boundary bases following the suggestion of Shepherd (10). Here one is looking at the nearest neighbor frequencies restricted to the last and

Table 2

STATISTICS OF EXEMPLARY SEQUENCES

Vertebrate Sequences	Length in bp	Strand asym	Dyad Chi-sq. per d.f.[1]	Triad Chi-sq. per d.f.[2]	CG Suppression[3]	Comments
Chick OVAL (ovalbumin incl. flanks)						
CAP + Leader	348	.23	3.5	1.6	.07	Note high CDS Chi-square indicative
IVS-1	251	.18	2.6	1.2	Total	of the non-random nature of coding
IVS-2	581	.37	5.0	2.7	.14	regions. This is in contrast with
IVS-3	400	.20	2.6	1.5	.30	the CDS in egg white Lysozyme which
CDS	1161	.12	13.2	4.9	.21	shows little CG suppression.
Human FIBIF (fibroblast interferon)						
CDS	570	.15	6.8	3.1	.08	The high CG suppression accounts for over 30% of the high dyad chi-square.
Human HBA2 (alpha-2-globin)						
IVS-1	117	.45	2.6	0.9	.30	IVS's as expected at random from
IVS-2	140	.42	2.0	1.2	None	composition.TTT missing from all
CDS	429	.16	2.9	2.4	.80	frames.
Human HBAPS (alpha-globin, pseudogene alpha 1)						
IVS-1	127	.51	3.0	1.2	.13	
IVS-2	133	.17	3.2	1.4	.18	
CDS	405	.06	5.9	2.5	.28	
Human HBB (beta-globin)						
IVS-1	130	.20	2.3	1.3	Total	CCGG and GCG missing throughout entire
IVS-2	850	.23	3.9	2.1	.15	gene. The CDS shows a high TG value
Leader & Trailer	503	.02	6.4	2.6	.08	expected from CG methylation induced
CDS	444	.18	8.9	3.8	.15	mutation, while the IVS's do not.
Human HBD (delta-globin)						
IVS-1	128	.32	2.6	1.4	Total	All termination codons slightly
IVS-2	890	.19	4.0	2.2	.13	suppressed.
CDS	444	.30	8.8	3.5	.09	

Vertebrate Sequences	Length in bp	Strand asym	Dyad Chi-sq. per d.f.[1]	Triad Chi-sq. per d.f.[2]	CG Suppression[3]	Comments
Human HBE (embryonic epsilon-globin)						
IVS-1	122	.45	1.9	1.2	Total	Has slightly lower coding region values than the two above.
IVS-2	855	.44	6.5	2.7	.20	
CDS	444	.11	7.7	3.3	.19	
Human INS (preproinsulin)						
IVS-1	179	.21	5.9	2.3	.15	While CG is only moderately suppressed, TA is strongly suppressed throughout the gene, resulting in the highest IVS chi-square observed.
IVS-2	786	.08	8.2	3.4	.41	
CDS	333	.09	7.4	2.9	.29	
Mouse GAM1 (c-region heavy chain gamma-1)						
IVS-1	356	.32	4.8	2.0	Total	While the CG suppression is greatly reduced in the CDS, the dyad chi-square is unusually high.
IVS-2+3	219	.10	4.9	2.8	.08	
Mouse HBB (beta-globin major gene)						
IVS-1	116	.20	2.9	1.4	Total	Nearly identical to the values obtained for the other globins.
IVS-2	646	.35	5.6	2.6	.12	
CDS	444	.13	8.4	3.3	.15	
Mouse KAPJ (j-region, kappa light-chain)						
IVS-1	316	.15	3.3	1.5	.21	With and IVS CG suppression of approx-six fold, the slightly higher than expected Cg in the CDS is unexpected. The CDS also has an atypically low dyad chi-square value.
IVS-2	267	.46	4.9	2.3	Total	
IVS-3	284	.22	2.6	1.2	.18	
IVS-4	299	.39	3.5	1.7	.21	
CDS	195	.50	2.9	2.4	None	
Rabbit HBB1A1 (beta-1 globin)						
IVS-1	126	.27	2.0	1.2	.17	Calculated values identical to those for Rabbit beta-2 globin.
IBS-2	573	.28	2.9	1.3	.29	
CDS	444	.24	9.4	3.8	.13	
Rabbit HBB2PS (beta-2 globin pseudogene)						
IVS-1	100	.25	2.6	1.2	Total	This pseudo gene IVS has the expected IVS properties.

Vertebrate Sequences	Length in bp	Strand asym	Dyad Chi-sq. per d.f.[1]	Triad Chi-sq. per d.f.[2]	CG Suppression[3]	Comments
AfishINS (Lophius americanus insulin)						
mRNA Leader & Trailer	306	.38	3.9	1.8	.71	The weak CG suppression is higher in the coding region.
CDS	351	.13	2.7	1.8	.59	
Frg 18sRNA (X. Laevis 18srRNA)	2948	.03	14.2	3.7	None	High chi-square values reflect suppressed CpA and ApC in conjunction with elevated ApA and TpT.
NON-VERTEBRATE CONTINUATION OF TABLE 2						
Worm FIBRUS (Bombyx mori fibroin)						
IVS	970	.02	1.3	1.9	None	Statistics as expected for random sequence except strand asymmetry.
CDS	504	.35	2.5	1.5	None	
Drosophilia (major heat shock induced protein)						
Leader/Trailer	849	.15	5.9	2.7	1.33	High chi-square values for this non-coding region are the result of a 30% increase above that expected from base composition for TT, AA and CG.
E. coli (Lactose permease)						
CDS	1254	.31	5.6	5.8	None	Slight AG and GA suppression; note the high Codon (triad) chi-square reflecting the high usage of TTT (Phe) and CTG (Leu).
Influenza virus a/udorn/72 (nonstructural gene 8)						
IVS	472	.20	3.5	1.5	.50	Here the statistics do not distinguish between IVS and CDS.
CDS	366	.45	3.5	1.5	.50	

Non-Vertebrate Sequences	Length in bp	Strand asym	Dyad Chi-sq. per d.f.[1]	Triad Chi-sq. per d.f.[2]	CG Suppression[3]	Comments
Influenza virus a/aichi/2/68 (hemagglutinin)						
CDS	1740	.25	8.8	3.3	None	
SV40 (Simian virus 40)						
Entire genome	5226	.05	31.3	15.5	.13	The strong CG suppression accounts for 41% of the very large chi-square particularly in the IVS's.
IVS (4572)	346	.29	5.8	2.2	.10	
BVNN (Papovavirus Bk)						
IVS	82	.67	1.6	1.0	Total	
CDS	2229	.18	14.4	8.1	.07	
Herpes simplex (Thymidine kinase)						
CDS	1806	.05	2.9	1.9	None	
IS101 (plasmid pSC101)						
Insertion seq.	310	.22	1.4	1.1	None	Has the statistics expected for a random sequence.
Human mitochondrial						
CDS's	7826	.39	2.5	3.7	.76	Codon (triad) chi-sqaure more than twice dyad value.

Note: 1. For 9 degrees of freedom a chi-square/d.f. of 1.0 gives P $\simeq$ 50%; 1.9, P $\simeq$ 5%; 2.5, P $\simeq$ .5%.
2. For 27 degrees of freedom a chi-square/d.f. of 1.0 gives P $\simeq$ 50%; 1.5, P $\simeq$ 5%; 1.9, P $\simeq$ .5%. Triad chi-squares for IVS's are totals, while for CDS's they include only proper reading frame codons. The expected values were calculated from base composition alone as in the dyad case.
3. CG suppression was calculated simply as the ratio of observed to expected, "none" indicating values between 0.8 and 1.2 and "total" indicating no 5'-CpG-3' observed.

Table 3

EXAMPLES OF CHI-SQUARE [a] VALUES FOR OBSERVED NEIGHBORING CODON POSITION BASE FREQUENCIES

Sequence	Chi-square values for positions in neighboring-codons 1,1	2,2	3,3	Chi-square values for total nearest neighbor frequencies
Mouse KAPJ				
IVS-1	20.0	10.3	12.9	29.7 [b]
IVS-2	5.3	7.0	7.9	44.1 [b]
IVS-3	7.3	12.1	5.1	22.8
IVS-4	17.4	14.3	6.2	31.5
CDS	26.8	45.1	18.0	26.0
Human MTOCL				
CDS 3307-4263	73.6	123.8 [c]	108.8	21.8
Human HBTHAL				
IVS	14.0	13.5	6.4	16.6
CDS	16.1	80.5	63.0	28.4
Human INS				
IVS-1	9.2	11.0	15.1	53.1 [b]
IVS-2	15.3	11.9	23.5	74.5 [b]
CDS	12.1	61.6	34.7	65.9
A Fish INS				
mRNA Leader & Trailer	15.1	21.6	16.6	24.9
CDS	17.9	47.3	37.0	34.7
Chick OVAL				
CDS	77.8 [d]	45.6	26.2	117.6

a. These values were obtained using base positional composition alone to calculate expected values.
b. Even with the highly nonrandom nature of the nearest neighbor frequencies, these IVS's show no triad to position correlations, as expected for a noncoding sequence.
c. This is the strongest middle base codon to codon correlation seen.
d. This is one of the few eukaryotic cases where the middle base does not show the highest value.

first bases in neighboring codons. The analysis revealed a strong preference for codons to end with a pyrimidine and begin with a purine. The ratio of pyrimidine-purine to purine-pyrimidine boundaries is generally a factor of two or more (Table 4, columns 1 and 4). Two important exceptions were noted (*Fig.*) within the data examined: these were the four vertebrate insulin coding regions, and the mitochondrial coding sequences.

These properties are highly diagnostic of protein coding domains. Thus the step-three correlations can be used to identify unknown CDS's, and when combined with the taxa specific nature of codon usage (8,9), may allow for the

Table 4

EXAMPLES OF CODON-CODON BOUNDARY PREFERENCE

	YR*	RR	YY	RY
Human HBD	48	42	36	21
Chick OVAL	128	125	75	58
Worm FIBROS	79	52	20	16
Mouse heavy chain gamma2b	109	95	80	50
Human HBTHAL	29	19	21	15
BKVMM CDS's	212	220	173	137
ϕX 174				
Largest non-overlapping CDS	161	132	61	73
Yeast cytochrome C	40	34	19	17
Average Percentages	36	32	21	11

*Y = Pyrimidine R = Purine

construction of more efficient DNA probes from known amino acid sequences.

CODING DOMAIN STATISTICAL CONCLUSIONS

The most obvious characteristic of the known vertebrate sequences is the CpG suppression (12-14). An intriguing explanation is that CpG neighbors are restricted in frequency because they form an integral component of a regulatory system using DNA cytidine methylation (15-19). Other possible roles may involve Z-DNA (2) formation in terms of regulation (21), recombination control (22), or nucleosome binding. The latter is suggested by the observation that the two nucleosomed vertebrate viruses (SV40 and BKVMM) show high CpG suppression while the non-nucleosomed one (Herpes simplex) does not. Saragosti et. al (50) has shown that a large nucleosome-free region develops about the replication point on the SV40 genome 44 hrs. post-infection, the same region of about nine percent of the genome which contains two thirds of the CpG's.

Bird (13) has suggested that the CpG suppression is basically a passive result of the high mutability of methylated cytidine to thymidine. As corroboration, Bird has cited the elevated frequencies of TpG and CpA, the products of such mutations, in DNA's where CpG is suppressed. However, these elevations exist even in the absence of CpG suppression as in the silkworm fibroin and the *E. coli* lac permease mRNAs. In other cases, the TpG + CpA elevation exceeds the CpG suppression. For example, the rat amylase CDS has 50 less CpG's than expected while the TpG + CpAs are elevated 120 above that

expected from base composition. Finally, the difference between the alpha and beta hemo-globins, which is apparent (7) in both the CDS's and IVS's refutes a simple passive explanation. Since the globin sequence homologies give credence to a common ancestral sequence, the current differential alpha beta CpG suppression across various vertebrate lines of descent indicates selective maintenance of CpG in the alpha globins, while absence of CpG in β-globin genes is certainly consistent with mutational loss. For example, in the mouse alpha hemoglobin CDS there are fifteen CpG's, eleven of which involve a C in the third codon position which should be least resistant to mutation pressure. The other four CpG's in this gene involve the codons CGT for Arg and GCG for Ala. To remove all of these CpG's, is therefore possible with no change in amino acid composition. This suggests that these nearest neighbors are selectively maintained.

In at least one set of sequences, the 5.8srRNAs, there is a predominance of CpG dyads. This is curious since among the other major structural RNAs--the 18s, 26s and transfer RNAs--CpG frequencies occur at the expected value given their 62 to 65 percent C + G composition in the vertebrates. These RNAs also have in addition a ApA and TpT elevation and a slight CpA and ApC suppression relative to that predicted from base composition alone. One assumes that these dyad characteristics are related to yet unidentified (secondary) structural constraints.

In coding segments the TpA suppression might appear to be the result of the lack of terminator codons (at least in frame). However, like the passive explanation of CpG suppression, this is probably too simple; the TpA neighbors are repressed in IVS's and to a lesser degree in both the structural RNAs and the Alu/B1/4.5s middle repetative sequences. The TpA suppression in the IVS's is especially hard to understand since IVS's are compositionally A + T rich. The suppression of both CpG and TpA and the elevation of TpG in vertebrate CDS's may be the cause of the observed (9) differential codon usage rather than the result of it. This hypothesis is supported by the fact that non-CDS domains, show similar though not identical relative, nearest neighbor-frequencies.

The correlations between neighboring codons no doubt reflect selection for protein structure and perhaps production rates as well (9). The frequently observed T:T, T:A and A:T middle base correlations between neighboring codons reflect high alpha helix content (23) in the globins where these correlations extend over at least three and four codons. It is interesting to note that the lac permease sequence, for which the three

dimensional structure is not yet known, shows the same strong T:T, T:A and A:T middle-middle base correlations, suggesting high alpha helix content.

The codon pyrimidine-purine boundary preference could reflect some molecular "hardware" constraint perhaps involving ribosome translocation, rather than any selection at the protein level. Although there are sequences that display a different codon boundary preference, such as the human mitochondria and insulin sequences, the pyrimidine-purine boundary preference persists over a wide taxonomic range, including vertebrates, their viruses, bacteria and their phages. Shepherd's suggestion (10) that this is a vestige of an archaic code structure seems unlikely to us, particularly since it is so strong among vertebrate viruses, which are probably recent evolutionary products. It is important to note that there are only two amino acids, Gln and Trp, which cannot be encoded in codons beginning with a purine and/or ending with a pyrimidine.

NON-CODING DOMAIN STATISTICAL CONCLUSIONS

The non-CDS domains also have identifiable if not easily interpretable statistical properties.

The vertebrate IVS's share at least two of the major sequence characteristics with their associated CDS's: the general, but not universal, suppression of CpG and the high dyad chi-square values. The latter in combination with strand asymmetry strongly suggests there are selective pressures acting at the monomer compositional level. The particularly strong CpG suppression in the IVS's is even less likely to have been the sole result of passive mutational pressure, since there is only a very slight compensating TpG and CpA elevation. These pressures must be in addition to those responsible for the extreme positional stability (2) of these regions within given gene families.

One possible explanation for both the chi-square values and the CpG's suppression is that IVS's contain vestigal coding information. This appears ruled out by the lack of any step-three correlations in the IVS's equivalent to those displayed by the CDS domains (in Fig. 1 and Table 4). This in turn may rule out the IVS's as arising from *copia* or Drosophila *P* like elements since such transposeables appear to code for some self-regulatory proteins. Gilbert's hypothesis (1) that the IVS's divide proteins into structural/function domain* may explain the observed positional stability (2) of the IVS's but not the size stability (3), the compositional or statistical properties reported here. The possibility that IVS's may represent the

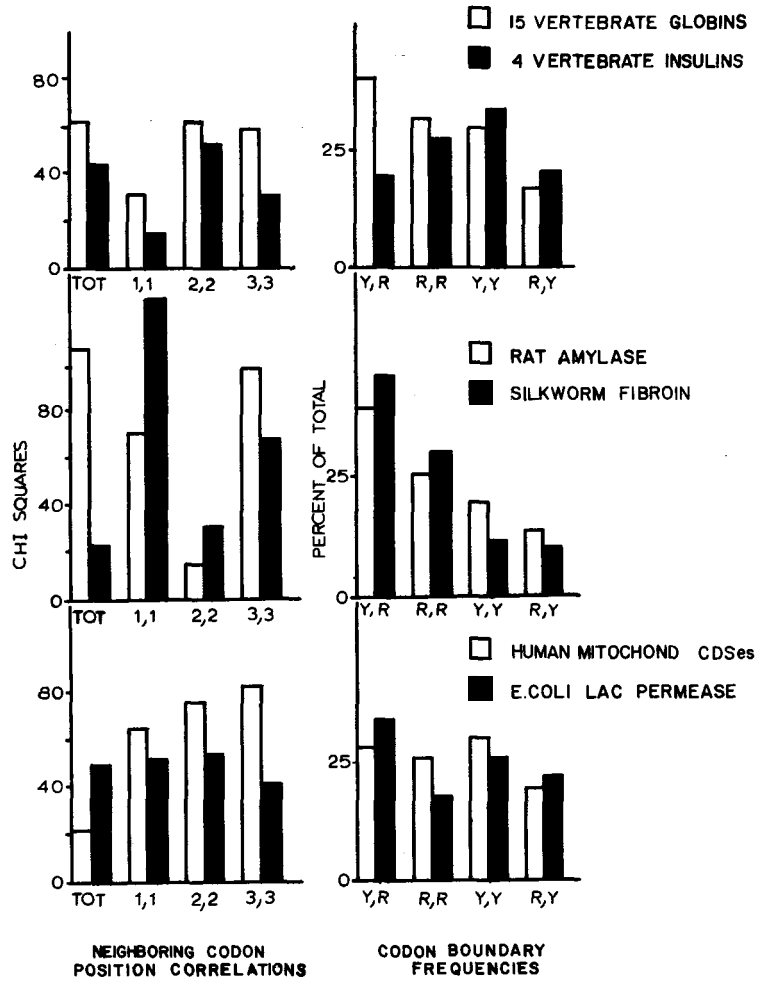

Figure 1. Graphical display of some Codon Boundary Preference and Neighboring Codon Position Correlations as reflected by chi-square values. The values plotted for the 15 vertebrate globins and four insulins are simple averages.

primitive genetic organization also does not clarify the origin of the selective pressures apparently operating, though it does support the idea that deep hierarchical (4), regulatory or molecular constraints are involved. Finally it is worthy of note (Table 1), that the relative dyad frequencies of the vertebrate IVS's and the Alu/B1/4.5s family are remarkable similar.

There are probably many constraints acting in concert on the non-coding

domains to influence their statistics. Since the DNA's local thermal stability (40,41) and/or local twist angles (36,37) are known functions of base composition, it is possible that various protein-DNA interaction-requirements might influence statistical measures. Such tactile characteristics could influence the entire genomic statistics if they were important for DNA replication or non specific histone binding. However, it is not clear how histone interactions or other DNA binding complexes like the DNA replicase could have any strong differential statistical influence on a particular domain, since they interact with nearly all coding and non-coding regions at one time or another.

ACKNOWLEDGEMENT

M.S.W. wishes to thank the Department of Biochemistry and Biophysics of the University of California at San Francisco for providing partial support for this work.

* Since a high proportion of the IVS's interrupt the encoded protein within or at the end of an alpha helix--the thermally most rigid component of globular proteins. The position may be such that new proteins arising from recombination will not have large "peptide loop-outs" exposed to the solvent which could destroy any new globular integrity or function.

REFERENCES

1. Gilbert, W., (1978) Nature 271, 501.
2. Marcker, A., (1980) 6th EMBO Symposium in Heidelberg, Nature 228, 216.
3. Proudfoot, N. J., et al., (1980) Science 209, 1329.
4. Smith, T. F., and H. Morowitz, (1981) J. Mol. Evol. in press.
5. Smith, T. F., and M. Waterman, (1981) J. Mol. Biol. 141, 195.
6. Lerner, M. R., J. A. Boyle, S. L. Wulin, J. A. Steitz, (1980) Nature 283, 220.
7. Goad, W., (1982) Los Alamos National Laboratory, Nucleic Acid Sequence Library.
8. Grantham, R., et al., (1980) Nuc. Acids Res. 8, 1893.
9. Grantham, R., et al., (1981) Nuc. Acids Res. 9, R43.
10. Shepherd, J. C. W., (1981) PNAS (USA) 78, 1596.
11. Marx, J. L., (1981) Science 212, 653.
12. Porter, A. G., et al., (1979) Nature 282, 471.
13. Bird, A. P., (1980) Nuc. Acids Res. 8, 1499.
14. Fitch, W. M., (1980) J. Mol. Evol. 16, 153.
15. Browne, M. J. and R. H. Burdon, (1977) Nuc. Acid Res. 4, 1025.
16. Waalwijk, C. and R. Flavell, (1978) Nuc. Acid Res. 5, 4631.
17. Duncan, B. K. and J. H. Miller, (1980) Nature 287, 560.
18. Razin, A. and A. D. Riggs, (1980) Science 210, 604.
19. Lindahl, T., (1981) Nature 290, 363.
20. Wang, A. H. J., et al., (1981) Science 211, 171.
21. McKay, D. B. and T. A. Steitz, (1981) Nature 290, 744.
22. Slighton, J. L., A. E. Blechl, O. Smithies, (1980) Cell 21, 627.
23. Bourgeois, S., et al., (1979) Biopolymers 18, 2625.

24. Goeddel, D., et al., (1981) Nature 290, 20.
25. Anderson, S., et al., (1981) Nature 290, 457.
26. MacDonald, R., et al., (1980) Nature 287, 117.
27. Hobart, P., et al., (1980) Science 210, 1360.
28. Fields, S., et al., (1981) Nature 290, 213.
29. Sanger, F., et al., (1977) Nature 265, 687.
30. Wagner, M. J., et al., (1981) PNAS (USA) 78, 1441.
31. Fischhoff, D., et al., (1980) J. Mol. Biol. 144, 247.
32. Hartley, J. L. and J. E. Donelson, (1980) Nature 286, 860.
33. Fox, T. D., (1979) PNAS (USA) 76, 6534.
34. Büchel, D., et al., (1980) Nature 283, 541.
35. Garoff, H., et al., (1980) Nature 288, 236.
36. Lomonossoff, G. P., P. J. G. Butler, and A. Klug, (1981) J. Mol. Biol. 149, 745.
37. Dickerson, R. E. and H. R. Drew, (1981) J. Mol. Biol. 149, 761.
38. Benoist, C., et al., (1980) Nucleic Acid Res. 8, 127 .
39. Smith, T. F., et al., (1981) J. Mol. Evol. 18, 38.
40. Gabarro-Arpa and Reiss, (1980) Nature 280, 515.
41. Moreau, J., et al., (1982) Nature 295, 260.