The EMBO Journal vol.4 no.5 pp.1217-1223, 1985

The use of multiple alphabets in kappa-gene immunoglobulin DNA sequence comparisons

Samuel Karlin and Ghassan Ghandour

Department of Mathematics, Stanford University, Stanford, CA 94305, USA

Communicated by L.L.Cavalli-Sforza

Comparisons within and between the human, mouse and rabbit immunoglobulin-kappa gene (J-C region) DNA sequences are carried out in terms of three two-letter nucleotide alphabets: (i) S-W alphabet (W = A or T; S = G or C); (ii) P-Q alphabet which distinguishes purines (P = A or G) from pyrimidines (Q = C or T); and (iii) a 'control' E-F alphabet (E = A or C; F = G or T). All statistically significant direct repeats within each of the three sequences and all significant block identities (a set of consecutive matching letters) shared by two or more sequences are determined for each alphabet. By contrast to the S-W and E-F alphabets, the P-Q alphabet comparisons reveal an abundance of statistically significant block identities not seen at the nucleotide level. Various interpretations of these P-Q structures with respect to control and functional roles are considered.

Key words: immunoglobulin kappa gene/multiple DNA alphabet comparisons/transition and transversion mutations

Introduction

A potentially important concept for the comparisons of nucleic acid or protein sequences is the grouping of letters in one alphabet to form natural new alphabets. In general, there are many ways to partition a given alphabet. For instance, the amino acids can be grouped according to structural, chemical, charge, functional or hydrophobicity criteria to name a few. However, not all possible partitions will have a meaningful biological mapping. The content or interpretation of a sequence may depend on the alphabet used to analyze it. Classifications of amino acids in various alphabets have been used by many (e.g., Fitch, 1966; Haber and Koshland, 1970; McLachlan, 1972; Dayhoff, 1978; Miyata *et al.*, 1979; Jimenez-Montano and Zamora-Cortina, 1981; Karlin *et al.*, 1984).

Extensive DNA sequence comparisons within and between the human, mouse and rabbit immunoglobulin-kappa gene (J-C region) are presented in Karlin *et al.* (1985) and Karlin and Ghandour (1985). (For the organization of the Ig-kappa gene region in these species see legend to Figure 1 below.) In this paper we exploit the concept of multiple alphabet comparisons in the analysis of the human, mouse and rabbit Ig-$\varkappa$ gene DNA sequences in terms of three two-letter nucleotide alphabets. The first termed the S-W nucleotide alphabet identifies A and T (W, weak bonding bases) and joins G with C (S, strong bonding bases) and thus characterizes nucleotides by their bonding strength. A second natural two-letter DNA nucleotide alphabet groups A and G (P = purines) *versus* C and T (Q = pyrimidines) labeled as the P-Q nucleotide alphabet. The P-Q alphabet distinguishes nucleotides by their chemical and structural properties. A third two-letter alphabet constitutes the groups E = [A+C] *versus* F = [T+G] which we refer to as the control (E-F) alphabet since it does not appear to have any natural chemical association.

For a given alphabet we determine all identity blocks (a set of consecutive matching letters, its length is the number of letters in the block) between multiple sequences exceeding a prescribed length. We also ascertain in these alphabets all long repeats within each Ig sequence. In Karlin *et al.* (1985) we presented formulas for determining the expected length and variance of the longest block identity, $K_{r,s}$, in r out of s 'random' sequences ($s = 3$ in the case at hand). We use a 'random' model as a standard in order to ascertain the distributional properties of the variable $K_{r,s}$. The following random model often serves well as a standard by which to assess statistical significance of long block identities. For this model, the s sequences of lengths $N_1,\ldots,N_s$, respectively are generated such that the successive letters of the νth sequence are sampled independently having letter l_i occur with probability $p_i^{(\nu)}$. (The $p_i^{(\nu)}$ are usually specified as the actual letter frequencies of the observed sequences.) We consider as statistically significant those identity blocks sufficiently longer than the theoretical expected length calculated on the basis of the random model. More explicitly, we set the stringent criterion that a block identity be considered statistically significant relative to the corresponding random sequence model provided its length exceeds the expected length of the longest block identity of the random model by at least two standard deviations. Table I lists the relevant significant length levels based on the corresponding random model.

Thus a block identity in any of the three two-letter alphabets common to all three sequences of length exceeding 22 bp and similarly a direct repeat of length ≥ 28 bp would occur by chance with probability ≤ 0.01.

Table II presents all significant block identities between the sequences common to two of three or in all three (3 of 3) sequences. Table III records all the statistically significant repeats in the P-Q, S-W and E-F alphabets within the human, mouse and rabbit Ig-kappa gene sequences.

For locating a specific DNA word (oligonucleotide) we employ the following notation: '130-5′ to J_1' indicates that the start of the specified oligonucleotide is 130 bp upstream from J_1; '77-3′ to J_4' signifies that the given oligonucleotide starts at base position 77 after J_4. The designation '16-5′ in J_3' specifies that the start of the designated oligonucleotide coincides with the 16th base pair of J_3.

Applications of the corresponding analysis to other sequences

Table I. Statistically significant length levels for repeats and identity blocks for the S-W, P-Q and control (E-F) alphabets with respect to the human, mouse and rabbit Ig-kappa sequences

		S-W	P-Q	E-F
Direct repeats	Human	28	28	28
	Mouse	28	28	27
	Rabbit	28	28	28
Identity blocks	2-of-3	31	30	30
	3-of-3	22	22	22

Table II. All significant P-Q-alphabet block identities involving at least two of (human, mouse, rabbit) Ig-kappa gene sequences and flanking J-C regions

No. of bp	Human	Mouse	Rabbit	No. of DNA mismatches	Comments
I. Overlapping the J regions					
28 $QP_5Q_2P_3QPQP_3QP_4Q_3P_3$	26 -5′ in J_1	26 -5′ in J_1	26 -5′ in J_1	Hu-Ms 3 Hu-Rb 4 Ms-Rb 3	Extends a 14-bp DNA common oligonucleotide between Hu and Ms 14 bp in 3′ direction. Also 17-bp DNA block identity between Hu and Rb extends 9 bp in 3′ direction but only 2 bp in 5′ direction.
26 $PQ_2P_3Q_2P_5Q_2P_3QPQP_3QP$	19 -5′ in J_2	19 -5′ in J_1	19 -5′ in J_5	Hu-Ms (1 at position 12 of this identity) Hu-Rb (no mismatches) Ms-Rb (at position 12)	Extends a 3-sequence 11-bp DNA common oligonucleotide (5 occurrences) (Hu:J_2, Ms:$J_1J_2J_5$, Rb:J_5) 15 bp in 3′ direction. Also it is embedded in 28-bp DNA identity between Hu and Rb.
22 $P_2Q_2P_5Q_2P_3QPQP_3QP$	23 -5′ in J_2,J_5	23 -5′ in J_1	23-5′ in J_5	Hu J_2-Ms (1 mismatch) Hu J_5-Ms (6 mismatches) Hu J_2-Rb (0) Hu J_5-Rb (5) Ms-Rb (1)	See block identity of length 26 above.
40 $P_6Q_2P_3Q_2P_5Q_2P_3QPQP_3QPQ_9$	14 -5′ in J_2		14 -5′ in J_5	2	Extends a 28 DNA block identity 4 bp in 5′ direction and 8 bp in 3′ direction.
36 $PQ_2P_4QP_5Q_2P_3QPQP_3QP_4Q_3P_3Q$	19 -5′ in J_1		19 -5′ in J_1	4	See triple 28-bp block identity above.
33 $P_2QP_5Q_2P_3QPQP_3(QP)_2Q_7P_2Q$	24 -5′ in J_4		24 -5′ in J_4	3	Extends a 15 bp DNA identity 14 bp in 5′ direction and 4 bp in 3′ direction.
II. C domains					
33 $PQPQ_3PQP_2Q_4P_2QP_2QPQ_4P_2QPQ_2P_2$	188 -5′ in C		188 -5′ in C	4	Extends a 12-bp DNA block identity 10 bp in 5′ direction and 11 bp in 3′ direction.
31 $Q_3PQ_7PQ_2PQ_3P_2QP_3QP_2Q_2P_2$	22 -5′ in C	22 -5′ in C		6	Extends a triple 9-bp DNA identity 2 bp in 5′ direction and 20 bp in 3′ direction.
III 5′ to J_1					
24 $QPQ_3P_{10}Q_2P_3Q_4$	103 -5′ to J_1	100 -5′ to J_1	93 -5′ to J_1	Hu-Ms (6) Hu-Rb (7) Ms-Rb (7)	Occurs downstream after a (16,16,10) bp gap in (Hu, Ms, Rb) from triple 11-bp DNA block identity.
33 $PQ_4PQ_9P_2QP_8Q_5PQ$	64 -5′ to J_1	64 -5′ to J_1		6	Extends a Hu, Rb 14-bp DNA identity 15 bp in 5′ direction and 4 bp in 3′ direction.
31 $PQPQ_3P_2Q_3PQ_6P_6QP_3QP_2$	279 -5′ to J_1		303 -5′ to J_1	2	Aligns with a Hu, Rb 13-bp DNA block identity (gap of 1 bp).
IV. Between J regions					
22 $P_7QP_{10}QP_2Q$	113 -5′ to J_4	105 -5′ to J_4 1165 -3′ to J_5	100 -5′ to J_4	Hu-Ms (6) Hu-Rb (6) Ms-Rb (8) Non-aligned: Hu-Ms (12) Ms-Rb (14)	The region 1165-3′ to J_5 is in the middle of the second enhancer region (see Karlin and Ghandour, 1985)
43 $P_2Q_5(PQ)_2P_{10}Q_2P_8Q_2PQ_2(PQ)_3P$	41 -5′ to J_2		39 -5′ to J_2	5	Extends triple 10 bp DNA identities 33 bp in 3′ direction (nonamer).
40 $PQP_3Q_3PQP_2QPQ_4P_2Q_2P_3QP_3Q_2PQ_5P_2Q$		56 -3′ to J_4	50 -3′ to J_3	8	Extends a 19-bp Ms, Rb DNA identity 21 bp in 5′ direction.

No. of bp	Human	Mouse	Rabbit	No. of DNA mismatches	Comments
30 $Q_3P_2QP_3Q_4PQ_5P_3Q_2P_3QPQ$	91 -3′ to J_2	87 -3′ to J_2		9	The longest DNA match is 5 bp long. Intersects a Hu-Rb 16 bp DNA block identity.
V. J_5-C intron					
36 $Q_3PQ_3PQ_5P_2(QP)_2Q_4PQP_2Q_2PQ_3PQP$	101 -5′ to C		99 -5′ to C	8	Overlaps a 12-bp DNA identity. Extends mainly in 3′ direction.
VI 3′ to C					
23 $P_2QP_4QP_3Q_5PQPQ_3P$	177 -3′ to C	186 -3′ to C	159 -3′ to C	Hu-Ms (2) Hu-Rb (2) Ms-Rb (2)	Extends a triple 10 bp DNA identity 13 bp in 3′ direction.
38 $P_4QP_3QP_3QP_4QP_3Q_5PQPQ_3PQP_2Q_2$	167-3′ to C		149 -3′ to C	9	Extends a 20-bp DNA identity 10 bp in 5′ direction and 8 bp in 3′ direction.
30 $P_3Q_3PQ_5PQ_2PQP_2Q_6P_2Q_3$	81 -3′ to C		67 -3′ to C	9	In alignment with the 38-bp identity.

P-Q alphabet: P = (A,G), Q = (C,T); S-W alphabet: S = (G,C), W = (A,T); control alphabet: E = (A,C), F = (G,T). The notation P_6Q_2 = PPPPPPQQ; $(PQ)_3$ = PQPQPQ; same notation applies to all three alphabets. Number of mismatches indicates the count of differences in the DNA content of the specified block identities in the various two-letter alphabets.

including various mammalian viral genomes, the globin family and mitochondrial sets for different species will be presented elsewhere. The programs and analyses used to identify the significant block identities are described in Karlin *et al.* (1983).

Results

Significant block identities (Table II, Figures 1 and 2)

S-W alphabet. The only statistically significant block identity in the S-W alphabet is of length 34 bp between human and rabbit appearing at 10 bp 5′ in J_3 in both sequences of composition $S_5WS_3WS_2W_3SWS_2W_4SW_3S_2W_3SW$ (a subscript 5 signifies five consecutive occurrences of a base of that type). This is curious since J_3 in rabbit does not share any significant DNA similarity with either of the human or mouse sequence. At the DNA level this block identity has four mismatches (nucleotide substitutions).

Control [E-F] alphabet. Only one significant block identity of length 30 bp is observed for this alphabet between human and mouse located 653 − 5′ to C and 617 − 5′ to C, respectively, in the vicinity of the established enhancer element (Queen and Baltimore, 1983). This block identity is composed of a 13 bp followed by a 16-bp DNA identity with a single mismatch at the 14th position (cf. Karlin and Ghandour, 1985). Thus, in practical terms, the control alphabet does not provide any new significant block identities not revealed by the DNA comparisons.

P-Q alphabet. In contrast to the S-W and control alphabets, the P-Q alphabet comparisons reveal a striking abundance of statistically significant block identities. We distinguish three kinds of identity blocks: (i) those that extend significant DNA block identities involving relatively few (one or two) DNA mismatches; (ii) those that extend a substantial DNA block identity but include several transition (C ⇌ T or A ⇌ G) mismatches; (iii) those that reveal new regions of correspondence that would not be detected through DNA comparisons.

As an example of (i), human and rabbit share a 28-bp oligonucleotide which extends to a 40-bp identity in the P-Q alphabet (see Table II-I). The 28-bp DNA identity is flanked on both sides by one DNA mismatch (necessarily a transition mutation) and subsequently shares in the 5′direction a 3-bp DNA match and a 7-bp DNA match in the 3′ direction. By contrast, the 24-bp P-Q block identity common to all three sequences 5′ to J_1 (see Table II-III) includes six or seven DNA transition mismatches over the pairwise comparisons for the three sequences.

All two and three sequence long P-Q block identities lying in part in the J-gene segments extend over the 3′ end (splice junctions). The three sequence block identities (Table II-I and Figure 1) relate the human J_1, J_2, J_5 gene segments to mouse J_1 and to rabbit J_1 and J_5. By contrast, J_5 of human showed the least extent of DNA identity. Moreover, it is J_5 of rabbit that shows the strongest DNA similarity to various human and mouse J segments, whereas J_1 of rabbit is the most similar across species in terms of amino acid comparisons (Karlin and Ghandour, in preparation).

It is noteworthy that human and mouse show a 31-bp P-Q block identity in the C domain proximal to the 5′ end associated with a 9-bp common oligonucleotide of the three sequences (Table II-II). This region further embodies strong amino acid similarity between human and mouse.

Significant P-Q block identities in the region 3′ to C also extend DNA identity blocks. The only significant three sequence P-Q block identity (length 23 bp) extends a 10-bp common oligonucleotide 13 bp in the 3′ direction (Table II-VI). In addition, the two sequence block identities (between human and rabbit) extend the corresponding DNA block identities in both directions.

Even though the 2.5-kb J_5-C intron is rich in long DNA block identities across the human, mouse and rabbit Ig-kappa gene regions (see Karlin and Ghandour, 1985, for an extensive analysis), only one statistically significant two sequence block identity occurred with respect to the P-Q classification. This single significant block identity between human and rabbit appears in alignment near the C domain and extends a 12-bp DNA block identity to a 36-bp P-Q block identity involving eight DNA transition mismatches (Table II-V).

A long P-Q identity between mouse and rabbit of 40 bp length (Table II-IV) expands a 19-bp DNA identity 21 bp in the 5′ direction. Generally, long P-Q block identities carrying several transition mismatches that extend DNA identities tend to show the strongest P-Q preservation in the direction of the closest coding region.

The region 5′ to J_1 contains a significant three sequence P-Q

Table III. Statistically significant direct repeats within each Ig region for the three two-letter alphabets

Oligonucleotides	Length	Locations	No. of DNA mismatches
I. Human			
A. P-Q alphabet			
$P_6Q_2P_4QP_5Q_2P_3QPQP_3QP$	31	14 -5′ in J_1;J_4	4
$P_2Q_2P_5Q_2P_3QPQP_3QPQ_6$	28	23 -5′ in J_2;J_5	7
$QP_5Q_2P_3QPQP_3QPQ$ (triple repeat)	20	26 -5′ in J_2,J_4,J_5	(J_2-J_4:1),(J_2-J_5:4),(J_4-J_5:5)
$QP_5Q_2P_3QPQP_3QP$ (four repeat)	19	26 -5′ in J_1,J_2,J_4,J_5	
B. S-W alphabet			
$W_3(SW)_3WSW_3SW_{16}$	31	277 -5′ to J_1,863 -5′ to C	14
$S_3WS_2W_2S_2WS_2WSW_2SW_3S_2W_3SW$	28	16 -5′ in J_2,J_4	1
C. Control alphabet			
No significant direct repeats			
II. Mouse			
A. P-Q alphabet			
$Q_9P_3Q_3P_2Q_3PQ_3P_2Q_2P_3Q_2P$	34	84 -5′ to J_4,86 -5′ to J_5	8
$QP_3QP_3Q_5PQP_9QPQP$	28	47 -5′to J_4,46 -5′ to J_5	5
$QPQP_8QP_{10}QP_2Q_3$	28	109 -5′ to J_4,1161-3′ to J_5	12
$(QP)_4Q_3P_{10}Q_2P_3Q_2$	28	106 -5′ to J_1,2 -5′ to J_2	11
Q_{27} (28 iterated pyrimidines)	27	123 -3′ to C	15
B. S-W alphabet			
No significant direct repeat			
C. Control alphabet			
$E_5LEL_3E_3LE_5L_2E_2L_2EL$	27	19 -5′ in J_1,J_2	1
III. Rabbit			
A. P-Q alphabet			
No significant direct repeats			
B. S-W alphabet			
$W_{22}SW_{22}$	45	1721 -3′ to J_5,1770 -3′ to J_5	18
$W_{29}SW_4$	34	1682 -3′ to J_5,1823 -3′ to J_5	20
W_{34} (run of 35 W-bases)	34	236 -5′ to J_1	28
C. Control alphabet			
No significant direct repeats observed			

See footnotes to Table II.

block identity (24 bp) with no corresponding significant DNA identities. This P-Q block identity starts ~15 bp 3′ to the significant 11-bp DNA block identity found in all three species in Karlin *et al.* (1985) which we hypothesized serves as a 'global recombination' control site. Another region where the P-Q alphabet reveals segments of strong identity with no corresponding significant DNA identity falls in the intervening section between J_3 and J_4 in all three species. A significant 22-bp block identity is located ~100 bp 5′ to J_4 in all three species. This may indicate a common control element specific to J_4.

Two species significant P-Q block identities between the J segments relate the consensus nonamer 40 bp 5′ to J_2 in human and rabbit. Also, a 30-bp human-mouse P-Q block identity pertains to a region ~91 -3′ to J_2 and 87 -3′ to J_2, respectively, with no corresponding substantial DNA matching. This pattern of conserved segments supports the concept of multiple control sites described in Karlin *et al.* (1985).

Direct repeats (Table III).

P-Q alphabet. Examination of Table III reveals that the long direct repeats in the human relate the J_1, J_2, J_4 and J_5 gene segments all in perfect alignment extending DNA oligonucleotide similarities primarily in the 3′ direction over the splice junction. By contrast, all statistically significant P-Q repeats in the mouse sequence fall in non-coding regions, and are localized to intervening regions between the J-segments. Noteworthy are the long P-Q repeat sequences relating a segment between J_3 and J_4 with a segment between J_4 and J_5. These similarities present in the mouse but absent from the human and rabbit sequences strengthen the proposition that, in the mouse, these two intervening sections may be a product of a relatively recent duplication or a result of DNA conversion. No statistically significant direct repeats occur in the rabbit sequence.

S-W alphabet. The degree of similarity is considerably diminished compared with that observed for the P-Q alphabet. In the rabbit, a large number of highly significant S-W repeats are observed. However, all these repeats are almost exclusively W in content and restricted to the A+T-rich stretches found ~200 bp 5′ to J_1 and 1.6−1.9 kb 3′ to J_5. By contrast long weak base stretches are absent in the human and mouse sequences (cf. Karlin *et al.*, 1985, and Karlin and Ghandour, 1985).

Control alphabet. No significant repeats were observed for this alphabet save for a 27-bp repeat relating the J_1 and J_2 segments of mouse. However, this direct repeat is essentially a DNA direct repeat with a single mismatch at the 15th position.

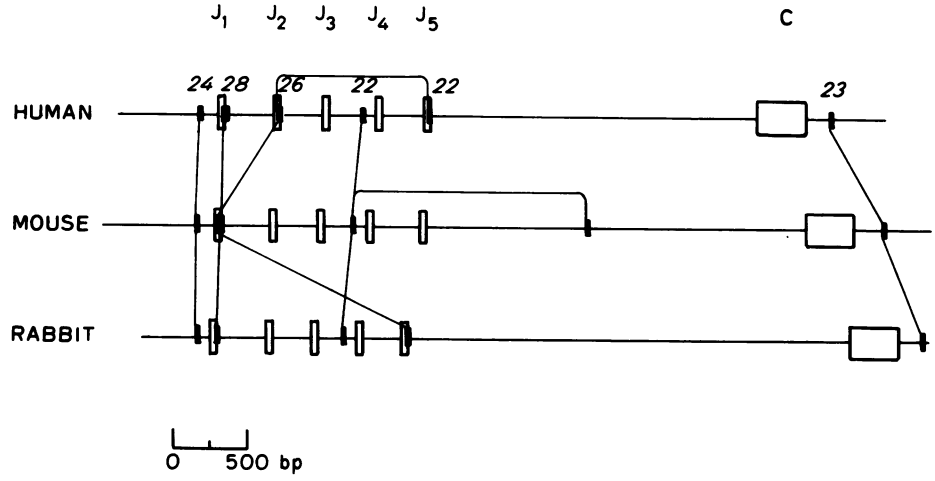

Fig. 1. Statistically significant block identities in the P-Q alphabet involving all three (human, mouse and rabbut) Ig-kappa gene sequences. The three Ig-κ sequences have the following inter-domain distances in bp.

	J_1	J_2	J_3	J_4	J_5	C	
Human	645	323	266	303	276	2180	553
Mouse	740	316	267	285	299	2515	557
Rabbit	482	333	244	251	253	2951	208

Each J is of length 39 bp except for J_2 in rabbit which is 42 bp. The human and mouse C domains are composed of 321 bp each (including the termination codon), while the rabbit C domain is 315 bp long. The numbers on the figure refer to the length of the indicated block identity.

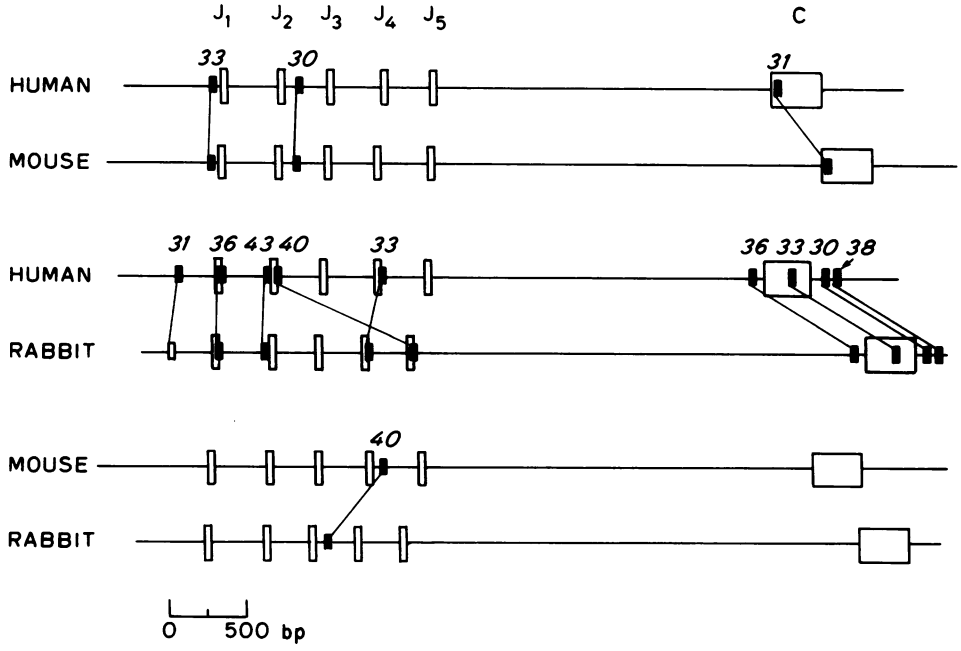

Fig. 2. Statistically significant block identities in the P-Q alphabet involving two of three (human, mouse and rabbit) Ig-kappa sequence.

Discussion

Comparisons of the human, mouse, and rabbit Ig-kappa gene (J-C) DNA sequences in terms of the P-Q, S-W and E-F alphabets contribute in two ways. Firstly, they emphasize certain DNA identities that are significantly extended in one or more of these two-letter nucleotide alphabets. Secondly, they reveal certain significant block identities not detected from DNA comparisons that may reflect on gene control and function. We discuss several of the key results with some comments.

(i) There is an abundance of significant P-Q block identities (Table II, Figures 1 and 2), but only one with respect to the S-W classification and none in the control (E-F) alphabet. Moreover, all significant P-Q block identities are in close parallel alignment with boundaries of J segments or the C domain. Why are there so many P-Q significant block identities, but hardly any in the S-W and none in the E-F control alphabets? The absence of significant block identities in the control [E-F] alphabet is not surprising since this classification of nucleotides entails no apparent chemical or biological rationale. It appears that transition substitutions but not transversions may be acceptable in appropriate control regions. In this perspective it is suggestive that either substantial DNA identity or a sufficiently specific expanded P-Q arrangement may mediate or prevent the required DNA protein interactions. In particular, several long P-Q block identities feature extensions of DNA identity blocks of the J-

segments over the splice junction. Is it possible that a key to J-C splicing depends on two physical characteristics, the existence of a V-J recombined unit, and an appropriate long P-Q configuration covering both sides of the splice junctions plus common P-Q structures further downstream 60−90 bp 3′ to the J-gene segments?

(ii) The significant P-Q identity blocks relate either J segments (usually extending in the 3′ direction across the splice junction), the intervening regions between J segments and 5′ to J_1, or the flanking regions 3′ to C. The P-Q block identities intersecting J segments tend to involve relatively few nucleotide mismatches. However, those corresponding to possible control sites generally contain many more transition mismatches, perhaps underscoring the importance of the P-Q configuration more than the exact DNA content. A substantial P-Q identity in alignment ~100 bp 5′ to all J_4 previously not identified in DNA comparisons may entail some biological control and may be worth experimental investigation. It is noteworthy that with one exception (near the C domain) there are no significant block identities in the J_5-C large intron for any of the two-letter alphabets.

The section between the classical nonamer and heptamer (from 8 bp to ~30 bp 5′ to the J segments) which does not show much DNA identity reveals, however, several cases of pronounced P-Q identity. A number of previous authors (e.g., Hieter *et al.*, 1982) underscored the lack of DNA homology for the intervening section between the consensus nonamer and heptamer signals. However, the P-Q significant identities suggest that selective constraints do operate in some of these regions allowing transition mutations while prohibiting transversions.

(iii) By the nature of the genetic code, synonymous codon replacements are inclined towards transitions as all degeneracy-two amino acids differ by a transition substitution. In particular, coding regions subjected to functional constraints are expected to show a bias toward transitions. This is indeed confirmed in various tabulations for many eucaryotic genes (e.g., see Li *et al.*, 1985; Gojobori *et al.*, 1982). On the other hand, in pseudogene regions and for non-coding regions, under no selection pressures, one might expect on a random basis twice as many transversions compared to transitions as there are eight types of transversions *versus* four types of transitions.

In comparisons of human, mouse and rabbit for both the β-globin and the Ig-kappa genes, synonymous codon changes involve more transitions than transversions by a factor of 3. In fact, in the C domain, human *versus* mouse show 64 synonymous codons out of 106 involving 19 transitions against five transversions. Human and rabbit agree in 48 residues entailing eight transitions and five transversions. Mouse and rabbit also agree in 48 residues embodying 15 transitions and four transversions. For corresponding aligned residues with different amino acids, mouse *versus* rabbit reveal about equal numbers of transversions and transitions. However, for human compared separately with both mouse and rabbit, an excess (factor of 2) of transversions is observed with respect to the unmatched residues of the β-globin gene. By contrast as could be expected by chance in the Ig-kappa gene C domain the number of non-synonymous nucleotide changes in the first and second codon positions show about equal numbers of transitions and transversions.

Current tabulations based on available DNA data suggest that mutation rates among nucleotides are not random (e.g., Li *et al.*, 1985; Gojobori *et al.*, 1982) and in particular cumulate to ~54% transitions. However, this could reflect a statistical artifact of the sequences sampled which tend to encompass mostly genic domains entailing substantial synonymous codon replacements. Along these lines, it is known that the transition of C to T or G to A appear to be abundant in methylated CG doublets (Bird, 1980).

The only significant block identities involving the constant domain among the two-letter DNA alphabets is a long 33-bp P-Q identity between human and rabbit extending the region of highest DNA homology (cf. Karlin *et al.*, 1985) and a 31-bp P-Q block identity between human and mouse expanding the 9-bp triple common oligonucleotide near the 5′ end. These extensions accentuate these regions as attractive segments of potential functional importance.

(iv) Consultation of Table II and Figures 1 and 2 suggests that important non-coding regions appear to tolerate more transitions over transversions. One can speculate that distinguishing and contrasting physical characteristics inherent to purines and pyrimidines more than specific DNA content in their arrangement may be important in mediating or avoiding certain protein binding properties, self regulation needs, or other functional purposes. The absence of significant S-W block identities both in coding and non-coding regions further suggests that DNA hydrogen bonding configurations play a relatively minor role in control or function at any level. Of course, the prevalence of transitions suggests relatively more facile S ⇌ W mutations and concomitantly less likely significant block identities in the S-W alphabet.

The predominance of transition mutations in non-coding regions is highlighted in a number of data sets. For example, in comparing the D-loop of mitochondrial DNA sequences among primates, Brown (1983) (see also Aquadro and Greenberg, 1983) observed 95% of all base substitutions as transitions. Other tabulations of nucleotide substitutions which also feature a transition mutation bias are given in Li *et al.* (1985).

(v) The statistically significant P-Q direct repeat words are quite different in form in the human, rabbit and mouse sequences. In human the long P-Q repeats relate J_1, J_2, J_4 and J_5 (extending over the splice junction) but strangely not J_3. On the other hand, it is surprising the J_5 of human, which showed the weakest DNA similarity to the other J segments is quite similar in the P-Q alphabet. The long P-Q direct repeats of mouse are restricted to intervening J regions. These mostly present new identities not detected from straight DNA comparisons. (A 16-bp nucleotide repeat exists in mouse 47 -5′ to J_4 and 46 -5′ to J_5.)

In the S-W alphabet there is a 31-bp direct repeat in human connecting the region of the enhancer element (see Karlin and Ghandour, 1985) with a region 5′ to J_1. Both these regions have been experimentally established to be DNase hypersensitive (Chung *et al.*, 1983).

The rabbit shows no significant P-Q repeats but some S-W long repeats mainly associated with A+T-rich stretches. The P-Q direct repeats may reflect species-specific characteristics. Moreover, because of the differences among the species, it is suggested that some of these direct repeats may be a recent result of homogenization (concerted evolutionary) activity.

(vi) Dickerson (1983) found that the structure of the DNA double helix is quite irregular in the X-ray structure determination of crystalline CGCGAATTCGCG. He then suggested that the distortions were well represented by Calladine's rules which attributed the distortion to the occurrence of PQ or QP doublets which cause steric hindrance between adjacent purines in opposite strands. Nussinov *et al.* (1984) have attempted to associate rule-predicted regions of large distortion with control or action regions. From this perspective, it seems interesting to make a comparison of predicted distortions for the various significant

P-Q block identities with predicted distortions in unconserved regions and randomly generated regions. Among the P-Q block identities there often appear to be a deficit of PQ runs involving lengths of one or two bases conforming to the general Calladine predictions.

Acknowledgements

Supported in part by NIH Grants 2R01 GM10452-21, IR01-HL30856-01A1, and NSF Grant MSC82-15131. We convey our thanks to E.Blaisdell for helpful comments on the manuscript.

References

Aquadro,C.F. and Greenberg,B.O. (1983) *Genetics*, **103**, 287-312.
Bird,A.P. (1980) *Nucleic Acids Res.*, **8**, 1499-1504.
Brown,W.M. (1983) in Nei,M. and Koehn,R.K. (eds.), *Evolution of Genes and Proteins*, Sinauer Associates, Sunderland, MA, USA, pp. 62-88.
Chung,S.Y., Folson,V. and Wooley,J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2427-2431.
Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. **5**, Suppl. **3**, published by National Biomedical Research Foundation, Washington, DC.
Dickerson,R.E. (1983) *J. Mol. Biol.*, **166**, 419-441.
Fitch,W.M. (1966) *J. Mol. Biol.*, **16**, 9-16.
Gojobori,T., Li,W.-H. and Graur,D. (1982) *J. Mol. Evol.*, **18**, 379-386.
Haber,J.E. and Koshland,D.E. (1970) *J. Mol. Biol.*, **50**, 617-639.
Heiter,P.A., Maizel,J.V. and Leder,P. (1982) *J. Biol. Chem.*, **257**, 1516-1522.
Jimenez-Montano,M.A. and Zamora-Cortina,L. (1981) *Proceedings of the Seventh International Biophysics Congress*, Mexico City, August 23-28.
Karlin,S. and Ghandour,G. (1985) *Mol. Biol. Evol.*, **2**, 53-65.
Karlin,S., Ghandour,G. and Foulser,D.E. (1985) *Mol. Biol. Evol.*, **2**, 35-52.
Karlin,S., Ghandour,G., Ost,F., Tavaré,S. and Korn,L.J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 5660-5664.
Karlin,S., Ghandour,G., Foulser,D.E. and Korn,L.J. (1984) *Mol. Biol. Evol.*, **1**, 357-370.
Li,W.-H., Luo,C.-C. and Wu,C.-I. (1985) in McIntyre,R.J. (ed.), *Molecular Evolutionary Genetics*, Plenum Press, NY, (in press).
McLachlan,A.D. (1972) *J. Mol. Biol.*, **64**, 417-437.
Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) *J. Mol. Evol.*, **12**, 219-236.
Nussinov,R., Shapiro,B., Lipkin,L.E. and Maizel,J.V.,Jr. (1984) *J. Mol. Biol.*, **177**, 591-607.
Queen,C. and Baltimore,D. (1983) *Cell*, **33**, 741-748.

Received on 26 November 1984; revised on 19 February 1985