Comparative statistics for DNA and protein sequences: Multiple sequence analysis

(multiple DNA sequence block identities/dyad symmetries)

SAMUEL KARLIN AND GHASSAN GHANDOUR

Department of Mathematics, Stanford University, Stanford, CA 94305

Contributed by Samuel Karlin, March 8, 1985

ABSTRACT Concepts and methods [Karlin, S. & Ghandour, G. (1985) Proc. Natl. Acad. Sci. USA 82, 5800–5804] for the analysis of patterns and relationships are extended to multiple DNA and protein sequences. Functionalities include multiple sequence common word occurrence distributions, characterizations of high frequency shared words, and ascertainment of long block identities. Various comparisons of sequences using natural alphabets obtained from grouping nucleotides or amino acids by their chemical and functional characteristics are described. Specific applications are given to globin genes, mitochondial genomes, and a variety of mammalian viruses.

Various nucleic acid and protein sequence statistics and concepts for purposes of discerning, classifying, and assessing statistical significance of word relationships within and between sequences have been introduced (I). The previous paper concentrated on the analysis of structural repeats within a single sequence. Extensions of the concepts and methods to multiple sequences are the focus of this paper. We adhere to the terminology and notation of ref. 1.

Functionalities for Multiple Sequences

For s sequences consisting of letters from a finite alphabet, an r-of-s block identity (common word) of length k is a k-word appearing at least once in each of at least r out of the s sequences. The number of representatives of a specific k-word across the various sequences is described by the vector \( v = \{v_1, v_2, ..., v_s\} \), where \( v_i \) is the number of its appearances in the ith sequence.

(i) Multiple Sequence Count Distribution of Common Word Occurrences. We define \( f_s(v_1, v_2, ..., v_s) = f_s(v) = n_r \) as the number of k-words that are represented across the sequences according to the vector \( v \). We present the bivariate count distributions, along with permutation ranges, of common 8-words of the \( \alpha \)- and \( \beta \)-globin DNA sequences of human and mouse (Table 1). Similar results hold for the human and mouse \( \beta \)-globin and mouse \( \alpha \)- and \( \beta \)-globin comparisons (data not shown). The frequency of shared words is considerably higher than that obtained from corresponding permuted sequences. The number of words shared once between the human \( \alpha \)- and \( \beta \)-globin sequences is also significantly high but closer to the permutation range. The fact that the human and mouse globin sequences decisively share k-words for all \( k \geq 6 \) undoubtedly reflects the common function and evolution of these genes. The human \( \alpha \)- and \( \beta \)-globin sequences also show significant coincidence in k-word counts, but to a reduced degree, consistent with the precept of a more remote divergence time of the \( \alpha \)- and \( \beta \)-globin genes than the split times of the human and mouse genomes. The significant but reduced

\( \alpha \)- and \( \beta \)-globin common word counts \( (k \geq 6) \) perhaps also suggests partially differentiated structure or function between these globin genes.

(ii) Multiple Sequence High Frequency (hf) Words. A criterion based on the cumulative number of occurrences across the s sequences prescribes a high frequency word of length k as a k-word satisfying \( \Sigma_{i=1}^s v_i \geq s^2(k) \) or alternatively \( v_i \geq s^\alpha \) for \( i = 1, ..., s \).

hf shared k-words \( (k = 6-9) \) among 1g-K sequences (Table 2). All the hf shared words of \( \geq 7 \) base pairs (bp) relate to the "consensus" nonamer (GGTTTTTTGT) postulated to play a control role in variable region (V) joining region (J) rearrangement during B-cell differentiation (2–4). Among the three sequences, the mouse sequence tends to contain the most hf words of length \( \geq 8 \) bp. Note that all hf 6-words of Table 2 can be formed by a single point mutation in a string of one-letter iterations with subsequent tandem duplications (cf. ref. 1). Interestingly, the common form of the polyadenylation sequences (AATAAAA) occurs nine times in each of the human and mouse sequences and six times in the rabbit sequence. This suggests that this oligonucleotide alone cannot be a decisive signal for transcription termination.

(iii) Long Common Words. For s sequences we define \( K_{rs} \) as the length of the longest word occurring in at least r of the s sequences. With the aid of Theorem 1 in ref. 1, the statistical significance of a long r-of-s shared word can be assessed. We present in Table 3 a number of examples of the several distinct longest r-of-s common words (block identities).

(a) Virtually all the significant common oligonucleotides in the papova-, papilloma-, and the hepatitis viruses relate similar coding regions and are in alignment (i.e., spacings between successive significant block identities are about the same length for each sequence (see refs. 5 and 6 for examples of alignment maps)). In the papovaviruses, the significant long shared words emphasize DNA similarity between the two capsid proteins VP1 and VP2. In contrast, the most conserved DNA block identities for the papillomaviruses relate the 3' half of the E1 transforming gene.

(b) The most conserved DNA block identities (in terms of their number and lengths) of the four mitochondrial genomes relate the two rRNA genes, 12S and 16S. The 12S rRNA block identities are perfectly aligned in the mammalian species (human, mouse, bovine), reflecting the absence of major insertion/deletion events in the evolution of this gene among the mammalian lineages. However, in Xenopus the corresponding conserved segments are displaced \( =140 \) bp, indicating an insertion or deletion in the 5' half of the 12S rRNA gene. Accounting for this displacement, all the conserved oligonucleotides can be aligned and are otherwise evenly distributed over the gene. The 16S rRNA gene shows a substantial number of long conserved oligonucleotides in almost perfect alignment with respect to all four genomes.

Abbreviations: hf, high frequency; bp, base pair(s); V, J, and C, variable, joining, and constant regions of immunoglobulin chain.
Table 1. Bivariate count distribution of 8-word occurrences in (i) human and mouse α-globin and (ii) human α- and β-globin sequences, along with corresponding ranges over 20 random permutations

<table>
<thead>
<tr>
<th>µ/ν</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Mouse α-globin (N = 1138)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N = 1441)</td>
<td>0 —</td>
<td>924</td>
<td>20</td>
<td>0</td>
<td>7 (34)</td>
</tr>
<tr>
<td>1</td>
<td>172</td>
<td>157</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(ii) Human β-globin (N = 2165)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N = 1138)</td>
<td>0 —</td>
<td>1924</td>
<td>78</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1041</td>
<td>44</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The entry in the (µ,ν) cell is the number of 8-words that appear exactly ℜ times in the row sequence and ν times in the column sequence. The numbers in parentheses in the (µ,ν) cell correspond to the (minimum and maximum) values from 20 independent random permutations of both sequences.

suggesting no major sequence changes during mitochondrial DNA evolution within these lineages. The significant shared oligonucleotides are clustered toward the 3' half of the gene. The high conservation of the rRNAs presumably reflects the importance of their function in the translation of the mitochondrial genes. This importance is also displayed in chloroplasts, where the complete rRNA gene complex is duplicated in the opposite polarity.

Six tRNA genes contain significant block identities common to all four species. Of these, five involve the entire anticodon loop and part of the associated stem. Thus, the persistence of the anticodon loop as the most conserved region in mammals apparently extends through amphibians. The mammalian species show significant block identities for 13 tRNA, of which 12 show strong anticodon arm identity. The genes for methionine, tyrosine, and leucine show conserved segments that extend beyond the gene boundaries. Asparagine tRNA involves a highly conserved oligonucleotide that extends into the noncoding region containing the origin of light-strand replication (OL). Is this conserved segment a primary control site for replication?

With respect to all pairwise and three-species significant block identities, the most clustered tRNA genes are the most conserved. Of the 15 genes existing in clusters, 11 show significant conserved oligonucleotides including all 5 genes proximal to the OL region. Furthermore, all tRNA genes with significant oligonucleotides common to all four genomes are in tRNA gene clusters. The tRNA genes in the vicinity of the D-loop show no significant conservation.

Another highly conserved region is one extending from URF2 through the COI gene encompassing a number of tRNA genes and the region of OL.

Of the three mammals (human, bovine, mouse), mouse is most similar to Xenopus in cumulative length of significant block identities; human is the least similar. Another distinction between Xenopus and the three mammals is in the D-loop region, where only the three mammalian mitochondria display a number of aligned significant common oligonucleotides.

(c) The longest conserved block identities among the β-globin DNA sequences emphasize exon-intron splice boundaries. Actually, for the β-globin sequences, the longest block identity (17 bp) shared by the four species (human, mouse, rabbit, and chicken) overlaps the 3' splice point of the second exon. This identity extends to 35 bp between human, mouse, and rabbit. (This segment involves only two mismatches with chicken.) The second longest block identity involving all four sequences (14 bp) covers the 3' splice junction of exon 1, which extends to 27 bp when chicken is excluded.

Only the β-globin processed mRNA is available for duck and Xenopus. The concluding 8-word of exon 2, CAACTTCA, is identical over all six species and probably extends well into the intron. Similar strong identities prevail for all six species at the 5' boundaries of exons 2 and 3. A 16-bp block identity in human, mouse, and chicken starting 8 bp 5' to exon 3, and a 15-bp identity between human, mouse, and rabbit overlap the same splice junction (Table 3).

The heme regions and α- and β-globin peptide contacts tend to show strong DNA conservation, although not as strong as those in the vicinity of the splice boundaries.

The tendency for conserving DNA segments traversing splice junctions is also evident from the Ig κ gene comparisons. The longest shared oligonucleotides across the J-gene segments of human, mouse, and rabbit single out the 3' splice junctions (4). Moreover, comparison of DNA sequences for the Ig κ gene of human, mouse, rabbit, and rat highlight DNA preservation on both sides of the 5' splice junction of the constant (C) domain (data not shown).

In contrast, the longest statistically significant block identities of the α-globin sequences shared by at least four of the six species emphasize both gene termini and do not overlap any of the exon-intron splice junctions. Conservation over the 3' end of the α-globin gene is pronounced among the mammalian species compared to the avian species. Except for a single mismatch, the goat, human, and mouse sequences share a 47-bp oligonucleotide covering the end of exon 3.

There appears to be a complementarity in the patterns and distribution of the conserved block identities of the α- versus the β-globin genes. Reasons for these differences are not known. Perhaps the fact that the α-globin gene exists with two replica may impose less stringent requirements on the composition of the splice junction, for maintaining an acceptable level of transcripts.

In the α-globin comparison, the polyadenylylation signal "ATAAA" is embedded in a 13-bp oligonucleotide common to the mammalian species human, mouse, goat, and rabbit. A significant extension of this 6-word was also observed in the Ig κ gene human, mouse, and rabbit sequences (4). In the latter case, the signal was embedded in a 20-bp oligonucleotide common to the three species (one mismatch). Again, this reinforces the proposition that there is much more than the commonly cited 6-word pertinent to the control of transcription termination.

Table 2. High frequency shared words in all three [human (N = 5062), mouse (N = 5495), rabbit (N = 5235)] Ig κ gene (J-C region) sequences

<table>
<thead>
<tr>
<th>k</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTGTG (9, 11, 6); AAATAA (11, 8, 9); AAATAA (9, 9, 6); TTGTTC (8, 7, 8); TTAT (6, 6, 8)</td>
<td>GTTTGTG (4, 3, 4); TTTTGGT (5, 5, 3)</td>
<td>GTTTGGT (4, 4, 3)</td>
<td>AGTTTTGT (2, 2, 2); GGTGTTTGT (2, 3, 2)</td>
<td></td>
</tr>
</tbody>
</table>

Criteria are ≥6, ≥3, ≥2 in each sequence for 6-, 7-, 8-, and 9-words, respectively. A hf word would occur by chance with probability P = 0.01. The array (mω, mω, mω) indicates the number of occurrences (mω) in the nth sequence of the specified oligonucleotide.
Table 3. Longest r-of-s common words for selected DNA sequences

<table>
<thead>
<tr>
<th>r</th>
<th>r-of-s word</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Between SV40 and BKV; located 880 bp into the VP2 gene.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Two such oligonucleotides: (i) relates initial part of VP1 genes; (ii) relates VP1 in SV40 and BKV to a region 166 bp 5' to AGNO gene in polyoma.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>This identity relates 3' end of exon 1 extending 9 bp into intron 1 of human, mouse, and rabbit. This also involves human, mouse, and rabbit covering 3' end of exon 2 and extending 6 bp into intron 2. Starts 8 bp 5' to exon 3 in human, mouse, and chicken. Starts 4 bp 5' to exon 3 in human, mouse, and rabbit. The 16-bp and 15-bp block identities share a 12-bp identity involving all four sequences. This block identity is embedded in the 27 bp 3-sequence identity. It covers the last 12 bp in exon 2 and extends 5 bp into intron 2. Starts 36 bp upstream from 3' end of exon 2. It is in the proximity of one of the heme junctions. Between the four mammals (goat, human, mouse, and rabbit), starts 33 bp upstream from 3' end of exon 3. (Excludes mouse.) In alignment at bp 73 of exon 3. (Excludes mouse.) Starts 3 bp upstream from exon 1. Located early in exon 2.</td>
<td></td>
</tr>
</tbody>
</table>

**Human globin genes of chromosome 11** (s = 4; β (N = 2165), δ (N = 1985), γ (N = 1648), ε (N = 3919))

<table>
<thead>
<tr>
<th>r</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>All three lie in 3' half of the E1 coding region in perfect alignment.</td>
</tr>
<tr>
<td>4</td>
<td>There are 34 statistically significant oligonucleotides common to all four genomes distributed as follows: 16 S rRNA: 14 significant block identities (250 bp total) confined to two clusters in 3' half of gene. 125 S rRNA: 7 significant block identities (total 117 bp) distributed evenly over gene. Two significant block identities in each of the genes COI, URFS, and URFS2. Remaining significant block identities occur in COII gene and the following tRNA genes: Asn, Ile, Met, Trp, Ser-1, Ser-2.</td>
</tr>
</tbody>
</table>

**β-Globin genes** (s = 4; human (N = 2165), mouse (N = 1567), rabbit (N = 1827), chicken (N = 2157)) | |
| 3 | This identity relates 3' end of exon 1 extending 9 bp into intron 1 of human, mouse, and rabbit. | |
| 4 | This identity relates 3' end of exon 1 extending 9 bp into intron 1 of human, mouse, and rabbit. | |

**Mitochondrial genomes** (s = 4; human (N = 16,569), mouse (N = 16,295), bovine (N = 16,338), Xenopus (N = 17,550)) | |
| 3 | This identity relates 3' end of exon 1 extending 9 bp into intron 1 of human, mouse, and rabbit. This also involves human, mouse, and rabbit covering 3' end of exon 2 and extending 6 bp into intron 2. | |
| 4 | Starts 8 bp 5' to exon 3 in human, mouse, and chicken. | |
| 5 | Starts 4 bp 5' to exon 3 in human, mouse, and rabbit. | |

**Papovaviruses** (s = 3; SV40 (N = 5243), BKV (N = 5153), polyoma (N = 5292)) | |
| 2 | Between SV40 and BKV; located 880 bp into the VP2 gene. | |
| 3 | Relates VP1 of SV40 and BKV. | |
| 4 | Two such words between SV40 and BKV in 2nd exon of 'large T' gene. | |

**α-Globin genes** (s = 6; chicken (N = 1216), duck (N = 1145), goat (N = 1894), human (N = 2165), mouse (N = 1441), rabbit (N = 552)) | |
| 4 | Between the four mammals (goat, human, mouse, and rabbit), starts 33 bp upstream from 3' end of exon 3. | |

**Human papilloma virus**; BPV, bovine papilloma virus. †, P < 0.0001; ‡, P < 0.001; §, P < 0.01.

The longest conserved oligonucleotides common to all four human globin genes of chromosome 11 (β, δ, γ, ε) are 20 bp and 19 bp in perfect alignment, starting, respectively, 42 3' in E2 (i.e., 42 bp from the 3' end of exon 2) and 21 3' in E2. (Pairwise comparisons among the human β-globin gene family are presented in ref. 7.) When combined, these produce a 40-bp sequence with a single mismatch between the β-δ and γ-ε genes. This matching region expands to the longest block identity (58 bp) between β and δ extending 7 bp over the E2 splice junction and to a 48-bp block identity between γ and ε that overlaps the second intron by 6 bp. The other three significant block identities common to the four genes (lengths 11, 11, and 10 bp) retain perfect alignment and are located, respectively, at 97 bp 5' into E2, 19 bp 5' into E2, and 74 bp 5' into E2. The E1 and E3 exons are also related by a number of 9-bp and 8-bp common block identities.

Most three-way comparisons highlight significant block identities that are extensions of the four-sequence significant block identities. For example, the longest distinct common oligonucleotide of the β-δ-γ genes extends (in 3' direction) the 11-bp identity at 19 bp 5' into E2 to 20 bp; the longest distinct β-δ-ε block identity is 29 bp long at position 51 3' in E2. The δ-γ-ε combination features a 17-bp identity extending the 11-bp common segment of coordinates 97 bp 5' into E2. This combination, δ-γ-ε, shows a significant 12-bp common oligonucleotide starting 4 bp before E3, the longest three-sequence identity involving E3 over the splice junction.

In all cases, the significant conserved segments falling in coding regions correspond to heme-related sites and to the vicinity of contact points in the hemoglobin (α and β) tertiary structures, suggesting that high reliability of amino acid translation and concomitant preferential codon usage of these regions are possibly crucial.

The cumulative significant block identities of β with δ is 389 bp (a pairwise block identity is significant provided it exceeds 15 bp). This includes a 42-bp block identity perfectly aligned in the first intron 57 bp 3' to E1. The pronounced similarity of β and δ is attributed to gene conversion events of the region from the first through the second exon (8, 9). However, the significant block identities traverse all coding domains including all the splice boundaries. The extent of the cumulative pairwise conserved segments among the four genes is consistent with their order and the distances separating them.

When the α and β human globin family (α, ζ, ωβ; β, δ, γ, ε) DNA sequences are compared, the longest common identity block observed is 7 bp (four such oligonucleotides; these are statistically significant). Even allowing for mismatches, the degree of identity between these DNA sequences is not extensive. This supports the proposition that the α- and β-globin genes have complementary roles in hemoglobin function.

**The Use of Multiple Alphabets**

The 20 amino acids can be grouped according to structural, chemical, charge, functional, or hydrophobicity properties. Specifically, an eight-letter chemical alphabet (10) entails the following classification of amino acids: acidic (aspartic acid, glutamic acid); aliphatic (alanine, glycine, isoleucine, leucine, valine); amide (asparagine, glutamine); aromatic (phenylalanine, tryptophan, tyrosine); basic (arginine, histidine, lysine); hydroxyl (serine, threonine); imino (proline); sulfur (cysteine, methionine). In the canonical charge alpha-
bet (three letters), aspartic acid and glutamic acid are classified as acidic (−); arginine, histidine, lysine are basic (+); and the remaining amino acids are designated neutral (0). (See refs. 6 and 11–13 for discussion, examples, and further references to other alphabets.)

The use of different alphabets in sequence comparisons can contribute in several ways. It may reveal significant identities common to several amino acid classifications, suggesting that in these regions specific amino acid composition may be more crucial than general biochemical attributes. On the other hand, multiple alphabet comparisons may reveal regions of significant equivalence that would not otherwise be detected in a standard amino acid comparison. Significant identities may be demarcated that appear in some, but not in other, alphabets, suggesting similar and contrasting functional or structural properties for different regions of the sequence.

We illustrate first the use of multiple alphabets by comparing the human, mouse, and rabbit Ig κ gene DNA sequences (J–C region) in terms of the DNA R–Y alphabet distinguishing purines (R = A + G) from pyrimidines (Y = C + T). We ascertained all long common words to at least two of the three sequences exceeding the theoretical expected length by at least 2 standard deviations, where the expectations and variances are based on a corresponding independence random model (see Theorem 1 in ref. 1). The R–Y alphabet comparisons reveal a striking abundance of substantial extensions of DNA block identities and new statistically significant block identities (14). All the statistically significant R–Y block identities are in close alignment relative to boundaries of the J gene segments or the C region. Those R–Y block identities intersecting J gene segments tend to involve relatively few nucleotide mismatches, while those located in flanking or intervening regions generally contain many more transition substitutions. This is especially the case for a substantial R–Y identity block, in alignment, of 100 bp 5′ to the J4 gene segment in all three species. Also, the segment between the classical consensus nonamer (GGTTCTTG) and consensus heptamer (CAGCTGG), which does not show much nucleotide identity, reveals pronounced R–Y identity. These results lead to the hypothesis that the contrasting physical or chemical characteristics inherent to purines and pyrimidines may be more salient than the specific DNA content in mediating or avoiding certain protein binding properties, self-regulation needs, or for other functional purposes.

The human β−, δ−, γ−, ε-globin gene regions with respect to the R–Y alphabet reveal a remarkable extent of similarity. DNA (R–Y) preservation around the splice junctions plus the 5′ and 3′ ends of the genes are emphasized. The longest R–Y common block identity common to all four sequences is 61 bp starting at 52 bp upstream from the 3′ splice junction of exon 2. This block is embedded in a 64-bp identity of δ, ε, and γ, which expands to a 77-bp identity between δ and γ extending 17 bp over the splice junction. The sequences β and δ show a 76-bp R–Y block identity starting 19 bp 5′ to exon 2.

We also investigated the nature of the identities in the two-letter S–W alphabet where strong bonding bases (S = G + C) are identified versus weak bonding bases (W = A + T). Precluding long stretches of bona fide DNA identity, no significant S–W identity blocks were observed. The absence of significant S–W block identities both in coding and noncoding regions suggests that DNA bonding configuration plays a relatively minor role in control or function at any level.

The proliferation of R–Y block identities both in coding and noncoding regions suggests that DNA mutations tolerate transitions more than transversions. As synonymous codon replacements are inclined toward transitions, coding regions subject to functional constraints are expected to show a bias toward transitions. For elaborations of this analysis, see ref. 14.

Table 4 compares for three amino acid alphabets the Ig κ gene C-domain and β-globin gene between human, mouse, and rabbit. The total number of matches are recorded as well as the cumulative counts for the statistically significant block identities. For all alphabets considered, the β-globin gene shows a much greater degree of matching than do the Ig C domains. In these comparisons, the human and rabbit manifest more matched β-globin residues, while the human and mouse Ig κ C domains are more similar. Statistical tests for the homogeneity of alphabet relative frequencies for the same gene across the three species showed no significant differences for all alphabets. Comparisons of the frequencies of letters for the two different genes within each species indicated statistically significant differences for all three species for all alphabets.

The charge alphabet displays an extraordinary degree of total matching on the β-globin sequences (also for α-globin). Although the frequency distributions of the charge alphabet are the same for the β-globin and the Ig κ C domain for these species (25% charged to 75% uncharged), the charge configurations of interspecies matching for the two genes are markedly different. The Ig κ C domain involves 11 charge mismatches that are spread evenly over the domain. On the other hand, between the human and the mouse β-globin gene there were only 3 mismatches. Apparently, the specific arrangement of charges is more decisive to β-globin function and conformation than it is for the Ig κ gene C domain.

The human β- and δ-globin gene matchings exhibit 10 amino acid replacements but entail in the charge alphabet only 2 mismatches at residues 23 and 118. Comparison of the human e- and γ-globin genes in the charge alphabet reveals two highly significant block identities of lengths 71 and 55 bp covering

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Chemical</th>
<th>Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu/Ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>54 (60.4)</td>
<td>76 (71.7)</td>
</tr>
<tr>
<td>C</td>
<td>36 (34.0)</td>
<td>38 (35.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>β-globin</td>
</tr>
<tr>
<td>T</td>
<td>117 (80.1)</td>
<td>129 (88.4)</td>
</tr>
<tr>
<td>C</td>
<td>83 (56.8)</td>
<td>88 (60.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ig κ C domain</td>
</tr>
<tr>
<td>Hu/Rb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>48 (45.3)</td>
<td>66 (62.3)</td>
</tr>
<tr>
<td>C</td>
<td>16 (15.1)</td>
<td>24 (22.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>β-globin</td>
</tr>
<tr>
<td>T</td>
<td>132 (90.4)</td>
<td>137 (93.8)</td>
</tr>
<tr>
<td>C</td>
<td>119 (81.5)</td>
<td>119 (81.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ig κ C domain</td>
</tr>
<tr>
<td>Ms/Rb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>49 (46.2)</td>
<td>66 (62.3)</td>
</tr>
<tr>
<td>C</td>
<td>15 (14.2)</td>
<td>17 (16.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>β-globin</td>
</tr>
<tr>
<td>T</td>
<td>116 (79.5)</td>
<td>128 (87.7)</td>
</tr>
<tr>
<td>C</td>
<td>84 (57.5)</td>
<td>82 (56.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ig κ C domain</td>
</tr>
<tr>
<td>Hu/Ms/Rb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>40 (37.7)</td>
<td>58 (54.7)</td>
</tr>
<tr>
<td>C</td>
<td>14 (13.2)</td>
<td>17 (16.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>β-globin</td>
</tr>
<tr>
<td>T</td>
<td>113 (77.4)</td>
<td>126 (86.3)</td>
</tr>
<tr>
<td>C</td>
<td>82 (56.2)</td>
<td>81 (55.5)</td>
</tr>
</tbody>
</table>

Total (T) indicates overall matching counts. Conserved (C) is based on cumulative matches of 2-of-3 species and 3-of-3 species significant block identities. Total length for human and mouse Ig κ C domain 106, rabbit Ig κ C domain 104 residues long. All three β-globin sequences have 146 residues.
Table 5. Counts of \( g_{D,A} \) (dyad symmetry pairs of length 8) for SV40 \((N = 5243)\) and polyoma \((N = 5292)\) compared with corresponding permuted sequences

<table>
<thead>
<tr>
<th></th>
<th>SV40</th>
<th>Polyoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6</td>
<td>1 2 3</td>
</tr>
<tr>
<td>1</td>
<td>276 68 9 3 0</td>
<td>1 1 248 36 0</td>
</tr>
<tr>
<td>2</td>
<td>15 3 1 0 0</td>
<td>2 2 2 0 0</td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 0 0</td>
<td>0 0 0 0 0</td>
</tr>
<tr>
<td>Permutation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 2 3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>1</td>
<td>(195–260) (0–25) (0–3)</td>
<td>(192–229) (0–16) (0–3)</td>
</tr>
<tr>
<td>2</td>
<td>(0–4) (0–1)</td>
<td>(0–2) (0–0)</td>
</tr>
<tr>
<td>3</td>
<td>(0–0) (0–0)</td>
<td>(0–0) (0–0)</td>
</tr>
</tbody>
</table>

The entries of \((v,\mu)\) for \(v \neq \mu\) count all 8-words repeated \(v\) times that have \(\mu\) dyads and vice versa (each position is counted once). For permutation ranges, the entries are the (minimum–maximum) number observed from 20 independent random permutations.

Word Relationships and Associated Functionalities

The well-known example of dyad symmetries (inverted complements) suggests a further generalization of the repeat occurrence distribution to allow for word relationships. Let \(\Delta\) be a mapping of the alphabet \(A\) to itself. For example, DNA complementarity induces the mapping: \(\Delta\) \((A\rightarrow T, T\rightarrow A, G\rightarrow C, C\rightarrow G)\). Let \(\pi = (\pi_2)\) be a family of permutations such that \(\pi_2\) rearranges the set of letters in a k-word. We prescribe \(\pi^D\) as the inversion permutation. The composed mapping of \(\Delta\) and \(\pi^D\) is denoted by \(D\). Two words related by \(D\) are said to be in dyad symmetry relation.

Consider the word relation \(R\) composed from a one-to-one correspondence, \(\Delta\), among the alphabet letters and a family of permutations \(\pi\) applied to the positions of each k-word. Let \(w\) stand for a given k-word. Then \(w^* = R(w)\) is the k-word in \(R\)-relation to \(w\). Although the concepts and methods developed below extend to multiple sequences and multiple word relationships, we concentrate our discussion on the case of a single sequence.

(i) Count Occurrence Distribution of \(R\)-Related Pairs.

The frequency distribution of occurrences of a word relationship within a single sequence is encompassed in the two-dimensional distribution \(g_{R,A}(v, \mu)\), which enumerates for each \(k\), the numbers of \(k\)-word pairs \((w, w^*)\), in relation \(R\) where \(w\) occurs \(v\) times and its \(R\)-word, \(w^*\) occurs \(\mu\) times.

The count distributions \(g_{D,A}\) for simian virus 40 (SV40) and polyoma display the same distinction observed for repeat occurrence distributions of these two viruses: Polyoma shows a reduced number of \(D\)-related 8-words compared to SV40 (Table 5). Furthermore, comparison to the \(g_{D,A}\) permutation ranges indicates that for both viruses \(g_{D,A}(v, \mu)\) of the original sequences exceeds the range from the shuffled sets.

(ii) Multiple Sequence \(R\)-Related Occurrence Distributions.

Consider a sequence and \(r\) word relations \(R_1, R_2, ..., R_r\). We define \(f_k(v_1, v_2, ..., v_r|R_1, ..., R_r\) as the number of distinct \(k\)-words \(w\) such that there exists \(v_r\)-related words to \(w\). Table 6 presents human \(\beta\)-globin the 6-word two-dimensional count distributions of the four relations repeats (\(R\)), inserts (\(I\)), complements (\(C\)), and dyads (\(D\)). Examination of Table 6 reveals that the occurrence of unique dyad pairs (one representative each of the word and its dyad) is significantly more likely than that of unique combinations of a word and its complement or invert. This bias in favor of dyad relationships is also evident for repeated words. Moreover, the number of once-occurring words with no corresponding dyad word is significantly lower than that for invert and complement (480 compared to 595 and 602, respectively).

This work was supported in part by National Institutes of Health Grants NIH 2R01 GM10452-21 and NIH 1R01 HL30856, and National Science Foundation Grant MCS 82-15131.