Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 5660–5664, September 1983
Genetics

New approaches for computer analysis of nucleic acid sequences

(dyad symmetries/homology/self-dyad/algorithm)

SAMUEL KARLIN†, GHASSAN GHANDOUR†, FRIEDEMANN OST†‡, SIMON TAVARE§, AND LAURENCE J. KORN¶

†Department of Mathematics, Stanford University, Stanford, California 94305; ‡Department of Applied Mathematics and Statistics, Technical University of Munich, Arcisstr. 21, D-8000 Munich 2, Federal Republic of Germany; §Department of Statistics, Colorado State University, Fort Collins, Colorado 80523; and ¶Department of Genetics, Stanford University School of Medicine, Stanford University, Stanford, California 94305

Contributed by Samuel Karlin, May 5, 1983

ABSTRACT **A new high-speed computer algorithm is outlined that ascertains within and between nucleic acid and protein sequences all direct repeats, dyad symmetries, and other structural relationships. Large repeats, repeats of high frequency, dyad symmetries of specified stem length and loop distance, and their distributions are determined. Significance of homologies is assessed by a hierarchy of permutation procedures. Applications are made to papovaviruses, the human papillomavirus HPV, λ phage, the human and mouse mitochondrial genomes, and the human and mouse immunoglobulin κ-chain genes.**

Technical advances for determining the sequence of nucleic acids have been used to generate large data bases of more than 1.5×10^6 nucleotides. Analyses of this data require accurate and efficient computer programs that can rapidly search a single large sequence or many smaller sequences and identify relevant features (1–8). Currently available programs for finding all homologies execute in time essentially proportional to the square of the sequence length (n^2) and usually follow a single pair of homologues at a time. The programs of refs. 1 and 2 compare various sequences for homologies by successively sliding along each sequence, assuming all possible alignments. The matrix method (3, 5), which displays regions of homology in two dimensions, is primarily limited to pairwise sequence comparisons. In this method, large sequences or multiple sequences are usually partitioned into sections ≤ 1 kilobase (kb) in length, and the matrices are examined by eye.

Another problem underlying nucleic acid sequence analysis is the distinction between meaningful structures and chance configurations. Confidence estimates for accepting significant homology are based on criteria related to the number of bases matched, percentage of mismatches tolerated, and constraints on numbers of insertions or deletions, etc. The algorithms of refs. 9 and 10 provide ways of measuring homology between pairs of nucleic acid or protein sequences based on minimization principles involving various penalty assignments for errors. The results are limited by the specification of the penalties.

FORMAT OF SEQUENCE COMPARISONS

We have devised an algorithm of an intrinsic global character whose running time is a linear function of the input sequence size. The algorithm can find a variety of relationships between any number of nucleic acid sequences totalling up to 500 kb. The program executes in about 2 sec for sequences of length $n \approx 10{,}000$ (see Table 1). The following four categories of output data representations are used to help in the identification of pertinent patterns in nucleic acid sequences: (*i*) frequency distribution of nucleic acid sequence homologies—e.g., direct repeats (DR) and dyad symmetries (DS); (*ii*) distribution of distances among repeats of high occurrence, clusters of repeats, and dyad symmetries of stem length ≥ a specified size s, allowing a central loop distance ≤ ℓ bases; (*iii*) special attributes of the sequence—e.g., the several largest DR and DS; and (*iv*) distributions of the above for multiple sequences.

Assessment of statistical significance for these homologies is determined by various permutation and randomization procedures described in the next section. New theoretical results germane to significance evaluations are also presented. The last section contains specific results to illustrate the scope and flexibility of the program when applied to DNA sequences including λ phage (11), papovaviruses [simian virus 40 (SV40; ref. 12), polyoma (13), BK virus-Dun (14)], human papillomavirus (HPV; ref. 15), human and mouse mitochondrial genomes (Hu-Mt and Ms-Mt; refs. 16 and 17, respectively), and human and mouse κ-chain genes (Hu-Ig and Ms-Ig; ref. 18). Detailed sequence analysis of the examples will be presented elsewhere.

We elaborate now on the capabilities of the program.

Distribution of DR and DS. For a DNA or RNA sequence of n bases, the program ascertains and locates all DR and DS of any word size (a word of size k is an oligonucleotide of k consecutive bases). The program allows for a prescribed number of mismatches and/or deletions and insertions. For each given word size k, the program determines the complete frequency distribution $f_k(\cdot)$ of repeat occurrences—i.e., the number n_r of words of size k repeated r times, $r = 0,1,2,\ldots$ (see Table 7).

The frequency distribution for DS is two-dimensional $f_k(\cdot,\cdot)$, which indicates, for each pair of integers (ℓ,m), the number of DS pairs each of size k, one repeated ℓ times and its DS word repeated m times (see Table 10). The program also ascertains a hierarchy of functionals identifying all DS pairs above a prescribed word size, all DS pairs of numerous occurrences, the distribution of self-dyads (self-D; a word is a self-D if it is identical to its DS word—e.g., C-A-C-T-A-G-T-G), and all close DS of given minimum stem length and maximal loop length.

Special Attributes of the Sequence. The program records the locations of the largest distinct DR and DS and of oligonucleotides that occur with exceptionally high frequency (hf).

Comparisons Among Multiple Sequences. Multiple sequences are concatenated, and the DR patterns (sizes and locations) for the extended sequence are determined. The distribution of homologies for the extended sequence is readily converted back into comparisons between the original sequences. The algorithm takes into consideration whether the original sequences are circular or linear.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: DR, direct repeats; hf, high frequency; URF, undefined reading frame; DS, dyad symmetries; HPV, human papillomavirus; kb, kilobase(s); BKV-Dun, BK virus, Dunlop strain; bp, base pair(s); self-D, self-dyads; Hu-Mt and Ms-Mt, human and mouse mitochondrial genomes; Hu-Ig and Ms-Ig, human and mouse Ig κ-chain genes; J, joining; C, constant.

The program records for each word ≥ size s the vector number $(m_1,m_2,\ldots,m_L)$ signifying the occurrence m_α of its representatives (with their locations) in the αth sequence ($\alpha = 1,2,\ldots,L$). Similarly, in analyzing hf DR, the program highlights those words of any given size whose cumulative occurrences $\Sigma_{i=1}^{L} m_i$ exceed a specified level.

Our algorithm is not restricted to the four-letter DNA alphabet. Repeat patterns or other word relationships can be determined for amino acid sequences by the same procedures. Furthermore, categorizing amino acids by functional or charge criteria allows additional comparisons. Searches for DNA homologies can focus on two-letter alphabets, distinguishing purine versus pyrimidine or strong versus weak base pairs (bp).

OUTLINE OF ALGORITHM AND THEORETICAL RESULTS ON HOMOLOGY

In a sequence of length n, comprised of letters from an alphabet of size α ($\alpha = 4$ in the case of DNA or RNA sequences and $\alpha = 20$ for amino acid sequences; when nucleic acids or amino acids are grouped by functional characteristics, the alphabet size is appropriately reduced), if the letters are assigned the numerical representation $\{0,1,2,\ldots,\alpha-1\}$, then a word of size $k > 0$ (i.e., a succession of k letters from the sequence $\{a_1,a_2,\ldots,a_k\}$) will have the *unique* word value $\Sigma_{i=1}^{k} a_i\alpha^{k-i}$. Such word values will range between 0 and $\alpha^k - 1$. The conversion to word values is done recursively ($\leq 4n$ operations).

Utilizing the information on positions and frequencies of words of size ℓ, we move in one pass and in order down the sequence of positions of words with frequencies exceeding 1, determining for each if there exists another DR of size $\ell + D$. The value of D is chosen to be $1 \leq D \leq \ell$. This procedure is iterated to determine all DR of size $\ell_1 + D_1$, $\ell_2 + D_2$, etc. ($\ell_1 = k$).

The process of extending DR from size ℓ to size $\ell + D$ requires operations of the order $R(\ell)$, where $R(\ell)$ is the total number of DR of size ℓ. If the relative frequencies of words are not excessively skewed, $R(\ell)$ shrinks rapidly, and the total number of operations to extract all DR from a sequence of size n will tend to be linear in n.

The above procedure is extended, with modifications, to determine all strings or groups of strings that exhibit relationships such as DS or more elaborate word relationships or patterns. The actual execution time of our program (on an IBM 3081) from sequence data for identifying *all* possible DR and DS and a hierarchy of distributions is given in Table 1.

The details of the algorithm will be presented elsewhere.

Consider a randomly generated sequence of length n based on an r letter alphabet with probabilities p_i of getting the letter A_i, $i = 1,2,\ldots,r$. Let L_n be the length of the largest exact DR contained in the sequence.

THEOREM. *The length of the expected largest DR,* L_n^*, *is of precise asymptotic growth:* $L_n^* = 2\log n/\log(1/\lambda) - [1 + \log(1 - \lambda)/\log\lambda + 0.5772/\log\lambda] + \log 2/\log\lambda$ *where* $\lambda = \Sigma_{i=1}^{r} p_i^2$. *The variance* $\mathrm{Var}\, L_n \approx 1.645\,(-1/\log\lambda)^2$ *independent of* n.

It follows that the expected largest DR in a sequence of length 5,000, for a four-letter alphabet with all letters equally likely and independently generated, is of length 12.

The largest theoretically expected DR, L_n^*, and the largest observed DR [obs(L_n)] are recorded in Table 2. Note a DR of length 15 is expected by chance and found in λ phage.

EVALUATING THE SIGNIFICANCE OF SEQUENCE HOMOLOGIES

Homologies can occur by chance when a single large sequence or a number of smaller sequences are examined. To evaluate the significance of homologies within a sequence or between two or more sequences, we compare the pattern of DR and DS in the original sequence(s) with that found in 20–100 random permutations of the sequence. Significance is determined by comparing the range of outputs from the separate permutation sequences with the corresponding evaluations on the original DNA sequence. Consider a specific property of the sequence—for example, the largest DR. We order the value of this attribute for the original DNA sequence with the collection of values of the same attribute computed for the permuted sequences. Significance clearly occurs when the original value falls outside the range of the values based on the permuted sequences.

Permutations are obtained by randomly shuffling the nucleotides of the original DNA sequences. The base frequencies are unchanged for each such permutation. Other permutation constructions randomly exchange (*i*) codons in gene regions, (*ii*) bases of introns, or even (*iii*) permutations of nucleotides between analogous chromosomal segments of different species.

In addition to the permutation analysis, we generate random (independent or Markov dependent) DNA sequences of the required length. Our theoretical results for the independence or Markov-dependent cases serve as a further control.

Table 1. Execution time of the program

DNA sequence	Size, bp	Identification time, sec	
		DR	DS
SV40	5,243	1.38	1.67
Polyoma	5,293	1.02	1.29
BKV	5,153	1.34	1.52
Hu-Ig	5,019	1.00	1.10
Ms-Ig	5,473	1.01	1.12
HPV	7,811	1.12	1.50
Ms-Mt	16,295	3.09	3.81
Hu-Mt	16,569	2.49	2.76
λ phage	48,502	6.48	6.57

DR

This section is divided into three parts: (*i*) large DR and their locations are identified and distinctive attributes noted; (*ii*) hf repeats of oligonucleotides are indicated; and (*iii*) the distribution of DR in different sequences is compared.

The role of large or hf DR is unknown. Duplications and genomic rearrangements are persistent through evolutionary time. These events may modulate gene expression and amplification or may provide opportunities for partial, differentiated gene function to cope with new environmental contingencies while maintaining the original gene expression. The latter scenario conforms with the prevailing view of evolutionary development—i.e., as starting with a few primordial genes with se-

Table 2. Largest observed and expected DR lengths

DNA sequence	L_n^*	$\mathrm{Var}(L_n)^+$	obs(L_n)
SV40	11.85	0.91	72
Polyoma	11.61	0.86	16
BKV	11.88	0.91	67
Hu-Ig	11.78	0.91	18
Ms-Ig	11.94	0.91	20
HPV	12.50	0.91	16
Ms-Mt	14.27	0.99	16
Hu-Mt	14.10	0.99	15
λ phage	14.80	0.86	15

Table 3. Large DR within and between Hu-Ig and Ms-Ig

Size, bp	Oligonucleotide	Hu-Ig Position	Hu-Ig J/C location	Ms-Ig Position	Ms-Ig J/C location	D(D*)
20	T-G-G-A-A-A-T-A-A-A-A-C-G-T-A-A-G-T-A-G			1121, 1751	J_2, J_4	610
19	C-A-C-T-G-T-G-G-T-G-G-A-C-G-T-T-C-G-G	638	J_1	733	J_1	95*
18	T-G-G-A-G-A-T-C-A-A-A-C-G-T-A-A-G-T	1033, 1680	J_2, J_4			629
17	G-A-A-T-C-A-C-T-G-T-G-A-T-T-C-A-C	1301	J_3	1714	J_4	413*
16	A-G-G-T-T-T-T-T-G-T-A-A-A-G-G-G	1273	J_3	1684	J_4	411*
16	C-C-T-C-A-C-T-G-T-G-G-C-T-C-A-C	1644	J_4	2053	J_5	409*
16	A-G-A-T-T-A-C-A-G-T-T-G-A-C-C-T	3510	after J_5	3999	after J_5	489*
16	C-A-G-G-C-A-G-G-T-T-T-T-T-G-T-A			1679, 2018	J_4, J_5	323
15	A-G-G-G-A-C-C-A-A-G-G-T-G-G-A	660, 1669	J_1, J_4			994
15	G-G-G-G-A-C-C-A-A-G-C-T-G-G-A	1022	J_2	1110	J_2	88*
15	G-G-G-A-C-C-A-A-G-C-T-G-G-A-G	1023	J_2	2079	J_5	1056*
15	A-T-C-A-A-A-C-G-T-A-A-G-T-A-C	1038, 1343	J_2, J_3			290
15	G-A-C-A-G-C-A-C-C-T-A-C-A-G-C	4329	C	4782	C	453*
15	A-A-G-A-G-C-T-T-C-A-A-C-A-G-G	4440	C	4893	C	538*
14	T-G-G-A-A-A-T-C-A-A-A-C-G-T	671	J_1	766	J_1	95*
14	G-A-G-A-T-T-T-C-A-G-A-A-A-T	2885	after J_5	3279	after J_5	394*
14	A-C-C-A-C-C-A-C-C-A-C-C-A-C			215, 218	before J_1	−11
14	A-C-C-A-A-G-C-T-G-G-A-A-A-T			759, 1114	J_1, J_2	34

D, distance between repeats within a species; *D**, difference of the distances from the start position of the sequences when the homology occurs between species. Positions in the *J* or *C* regions are noted.

quential duplication and divergence.

Enhanced tandem and interspersed DR can be induced through unequal crossing-over and through intra- and interchromosomal conversions and transpositions. Transposon termini and sites of genomic insertions are associated with DR. Moreover, DNA stuttering coupled to specific DNA repair enzyme actions may bring about short frequent DR.

DR may aid in transcription. For example, several DR identified by *Alu* restriction sites are located in spacer regions of rRNA genes of *Drosophila melanogaster* coincident with a 45-bp DR overlapping the 5′ end of the 28S gene (19). It is conjectured that these *Alu* spacer DR can simulate potential binding sites for transcription, drawing RNA polymerase I to this vicinity and thereby ameliorating transcription of the actual 28S gene. Similar DR patterns (but species specific) occur in spacer regions of other *Drosophila* species and *Xenopus laevis*.

Large DR. *Hu-Ig and Ms-Ig.* Both sequences (κ-chain gene) start about 600 bp 5′ to the joining (*J*) regions and extend from the last *J* downstream (≈2.5 kb) through the constant (*C*) region. The DNA sequence is continuous for the Ms-Ig, but a segment of about 600 bp between J_5 and *C* is missing in Hu-Ig. The number of large DR is far more than expected from corresponding randomly permuted sequences (Table 3). We would expect for $n = 10{,}000$ bases, on average, one DR of size 13; a DR ≥ 15 bp would be extremely unlikely.

The distances *D* between large DR within each species are consistently about $325m \pm 25$ bp, where m is a nonnegative integer. The distances *D** from the initial position to the large DR between the human and mouse Ig regions are about $100 \pm 325m$ (that is, apart from a phase shift of 100 bases, the correspondence of multiples of 325 bp apart is persistent). The *J* regions are spaced about 300 bp apart (18), and most of the DR patterns reflect this inherent homology. There is also substantial homology between the *C* regions—e.g., the two 15-bp DR in the proximity of the *C* regions of Hu-Ig and Ms-Ig (positions 4329 and 4782 and positions 4440 and 4893).

In Ms-Ig none of the larger DR are in region J_3. The degeneration of J_3 in Ms-Ig (18) appears emphatic. Also, in Hu-Ig J_5 does not have any large oligonucleotides in common with J_1–J_4 of Ms-Ig, although two significant regions 1- and 1.5-kb 3′ to J_5 in Ms-Ig and Hu-Ig are homologous. A phase shift (insertion or deletion in the vicinity of J_5) also is suggested.

Mitochondria. The largest DR in Hu-Mt—C-A-A-A-C-T-C-A-A-A-C-T-A-C-G—is 15 bp located at positions 3,674 and 11,748. The largest DR in Ms-Mt—T-A-T-C-C-T-T-A-T-T-T-A-T-T-A-T—is 16 bp, located at positions 2,933 and 9,986. The large distances between these DR and their sizes relative to those expected by chance alone suggests no special significance. However, the occurrences of the highest DR in Ms-Mt and Hu-Mt both fall into separate undefined reading frames (URF 1 and URF 4). The lack of large DR in both mitochondrial genomes stands in sharp contrast to the papovaviruses.

Papovaviruses (SV40, Polyoma, and BKV). In SV40 there is the well-known 72-bp abutting (gap 0) enhancer segment and the 21-bp abutting "promoter" region. The next largest distinct DR are the 13-bp oligonucleotides A-A-A-A-A-C-C-A-G-A-A-A-G at positions 486 and 507 separated by a gap of 8 bp and G-A-G-G-C-A-C-A-C-G-A-G-G at positions 2,440 and 2,510 with a gap of 57 bp. SV40 contains two distinct DR of size 12—T-G-C-T-T-T-A-T-T-T-G-T at 2,600 and 2,629 (gap 17) and T-C-A-T-C-A-T-C-A-T-C-A at 2,908 and 2,911 (gap −9; a negative

Table 4. hf DR in Hu-Ig and Ms-Ig

Oligonucleotide	hf DR, no. Hu-Ig	Ms-Ig
Length 5		
A-A-A-A-A	46*	10
A-A-A-A-G	21†	9
T-T-A-A-A	21†	10
T-T-T-A-A	22†	12†
T-T-T-T-T	26†	23†
G-T-G-A-A	7	18†
Length 9		
A-A-A-A-A-A-A-A-A	7†	0
G-A-A-A-G-A-A-A-A	1	4†
A-C-C-A-C-C-A-C-C	3†	0

* Very abundant clustering involving >4 clusters.

† hf DR occurring asymmetrically in the Ms-Ig versus Hu-Ig sequences by a factor ≥ 2.5 or appearing in clusters, with each cluster defined to have at least four consecutive DR of gap length ≤ 50.

Table 5. hf DR in Hu-Mt and Ms-Mt

Word size	Repeat level	hf DR, no. Hu-Mt	Ms-Mt
6	≥20	42	63
	≥30	3	8
	≥40	0	1
8	≥ 5	65	88
	≥ 7	2	11
10	≥ 3	10	12
12	≥ 2	29	52
14	≥ 2	2	9
16	≥ 2	0	1

gap indicates overlapping oligonucleotides) so that TCA is tandemly repeated five times in the large T-antigen gene. We note that all DR of size ≥12 have distances between them of ≤60 bp. The 72-, 21- and one of the two 13-bp DR are G·C-rich. There is a significant number of moderate-to-large DR of size ≥ 11 located in the large T-antigen gene.

In polyoma virus the largest DR is of size 16, consisting of six successive GGA codons. We further note that all DR of size ≥ 11 tend to be close (four out of five with a distance of ≤0.5 kb and the fifth with a distance of <1 kb); four of the five DR are part of the large T-antigen gene.

In BKV there is a large DR of size 67 at positions 143 and 261 (gap 51) and a DR of size 51 at positions 167 and 217 (gap −1). The next largest distinct DR is of size 14 with gap 34. Again, all large DR are close, and they tend to favor G·C nucleotides.

HPV. The only significant large DR (size 16) occurs at positions 4,056 and 4,071 (gap −1) and shows a G·C bias.

λ phage. None of the large DR (involving ≥ 12 bp) have more than two representatives. The largest size, 15, is not significantly different from chance expectation.

hf DR. *Hu-Ig and Ms-Ig.* All hf DR were ascertained over the aggregate Hu-Ig and Ms-Ig sequences. For a sequence length $n = 5{,}000$, the criterion for a word of size k to qualify as a hf DR, $k = 5$ and 9, is to have at least 25 and 3 representatives, respectively. These threshold levels are determined from theoretical studies. The base composition for the Hu-Ig and Ms-Ig κ-chain gene sequences is similar.

Table 4 shows that hf DR occur more in Hu-Ig compared to Ms-Ig. Moreover, clustering of DR also favor the Hu-Ig sequence. This is in contrast to the preponderance of close DS of stem length ≥ 7 [excluding self-dyads (self-D)] in the Ms-Ig versus Hu-Ig sequences (see Table 8).

We also note the 9-bp oligonucleotide A-A-A-C-G-T-A-A-G occurring at positions 1041, 1346, 1688, 2033 (spaced 296, 333, 306 bp apart, respectively) in Hu-Ig. The same oligonucleotide is in Ms-Ig at positions 774, 1129, 1759, 2097 with gaps 346, 621, 329, respectively. Notice that the jump of 621 equals 2 (300) for the Ms-Ig, reflecting the inactivation of J_3.

Hu-Mt and Ms-Mt. The Ms-Mt has many more hf DR than does the Hu-Mt (Table 5). By contrast, Hu-Ig shows many more hf DR than does Ms-Ig.

Papovaviruses. Table 6 shows that the original sequence contains significantly more oligonucleotides for sizes 4 and 6, repeated in high numbers, compared to the corresponding cases

Table 6. hf DR in papovaviruses

Virus	Size 4 ≥30	≥40	≥60	Size 6 ≥6	≥9	Size 8 ≥3	≥4
SV40	48	25	3	42	6	11	0
P-range	(31,43)	(5,14)	(0,0)	(24,41)	(0,4)	(7,14)	(0,2)
Polyoma	46	4		13	1	5	3
P-range	(10,15)	(0,0)		(6,15)	(0,1)	(1,9)	(0,1)

Number of oligonucleotides of a given size that occur in repeats ≥ t times in the original sequence. The permutation range (P-range) is the corresponding range (minimum, maximum) determined from 20 randomly permuted sequences.

when the sequence is randomly permuted. This significance fails for word size ≥ 8. Short DR might be easier to generate (e.g., mediated by unequal crossing-over and gene conversion processes, by enzyme repair actions, and DNA stuttering mechanisms). However, a constant rate of mutation might tend to more readily disrupt the identity of longer repeats.

λ phage. The most frequent repeats of size 5 are A-A-A-A-A (147 occurrences), T-G-C-T-G (138), T-G-A-T-G (133), and T-T-T-T-T (133). The sequence T-G-A-T-G is singled out in ref. 20 in relation to the concept of translational coupling (see also ref. 11). The most frequent repeats of word size 10 are A-C-C-T-G-A-C-C-G-C, A-C-G-C-C-C-G-G-C-A, and C-T-G-A-T-G-C-A-G-G, each occurring four times. They all show a G·C bias.

It is interesting that the largest DR among the viruses tend to be in noncoding regions or in URFs and usually are G·C-rich relative to expectations.

Distribution of DR. We use examples from the papovaviruses. The sums in each row of Table 7 are the number of distinct oligonucleotides of sizes 6 and 10 that occurred in the given sequences. The total number of possible words of size 6 with a four-letter alphabet is $4^6 = 4{,}096$ and of size 10 is $4^{10} = 1{,}048{,}580$. If r_ℓ words are repeated ℓ times, then $\Sigma \ell r_\ell = n =$ sequence length. Note that hf DR words of all sizes are considerably rarer in polyoma compared to both BKV and SV40. However, the number of unique oligonucleotides in polyoma is significantly more than in SV40 and BKV, even adjusting for the large 72 and 21 DR of SV40 and the two large DR in BKV.

DS

The importance of identifying DS pairs for their potential function in secondary structure and as signals for transcriptional control is well recognized. This section presents examples, interpretations, and speculation on the bivariate distribution of DS with particular attention to self-D, close DS, and large DS.

Close DS. Table 8 indicates the numbers of close DS. The following facts are worth emphasis: (*i*) Ms-Mt compared to Hu-Mt has 50% more close DS; (*ii*) polyoma has significantly less close DS than does SV40; (*iii*) the Ms-Ig compared to Hu-Ig has ≈50% more close DS. After adjustment for genome size, λ phage exhibits significantly fewer self-D in comparison to all other examples of Table 8.

As to the significance of self-D, we discuss the output only for polyoma; the results are consistent with respect to all the virus data sets examined. The number of unique DS pairs is

Table 7. Number of DR of oligonucleotides of sizes 6 and 10

	Size 6, frequency 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Sum	Size 10, frequency 1	2	3	4	Sum
SV40	898	628	361	175	87	57	30	14	7	1	0	0	2	0	0	0	0	1	2,261	4,892	78	31	3	5,004
Polyoma	1,174	740	421	198	78	22	5	2	1	0	0	0	0	0	0	0	0	0	2,641	5,250	16	2	1	5,269
BKV	943	627	363	183	90	47	29	11	7	5	1	3	1	2	0	0	0	0	2,314	4,991	123	2	0	5,116

Table 8. Number of DS with stem length ≥ 5 and loop length ≤ 50*

Stem size	SV40	Plyma	HPV	Ms-Mt	Hu-Mt	Hu-Ig	Ms-Ig	λ
5	176	158	247	501	337	208	192	1797
6	121	84	152	327	247	111	115	832
SD	(65)	(49)	(87)	(218)	(155)	(64)	(60)	(332)
7	14	11	19	49	19	15	19	113
8	23	17	23	71	46	20	17	142
SD	(17)	(11)	(20)	(61)	(40)	(19)	(9)	(102)
9	1	0	0	4	1	1	1	11
10	5	4	7	16	13	4	1	24
SD	(5)	(4)	(6)	(13)	(11)	(4)	(1)	(20)
11	0	0	1	0	1	0	0	4
12	1	1	1	3	5	1	0	4
SD	(1)	(1)	(1)	(2)	(5)	(1)	(0)	(4)

SD, self-D; Plyma, polyoma.

*The integer without parentheses is the number of DS corresponding to the given stem length including self-D. The integer in parentheses is the number of self-D of the given stem length. For close DS of stem length exceeding 12, see text.

significantly diminished for the original genome compared to that for all the permuted sequences for words of size 6. The opposite is true for unique DS pairs of sizes 8 and 10 (Table 9). Significant orderings are not realized with respect to multiple DS pairs, where at least one of the component words occurs more than once. Self-D occurring twice may be considered a bona fide unique dyad pair because a more stable stem–loop can be formed for the different words but with neither self-D.

The percentage of self-D compared with single DS pairs was significantly less for the original genome for words of size 8. No significant difference in self-D percentage was observed for word size 6. Self-D may be selected against. Possible reasons are that self-D and associated potential secondary structures may slow transcription (pause sites) and encumber efficiency. It is reasonably documented in laboratory experiments that when short self-dyads are coerced into a stem–loop form, they generally snap back quickly to a linear configuration.

Three large self-D of size 16, 14, and 12 occur in HPV. These cannot be considered statistically significant because the same-size self-D appeared in a number of permuted sequences.

Bivariate Distribution of DS. We illustrate this concept for the cases of SV40 and polyoma (Table 10). The entries in row ℓ and column m refer to the number of occurrences of DS pairs of size 8, one word repeated ℓ times, and its dyad word repeated m times. Thus, the entry for (1,1) is the number of unique DS pairs where each word occurs once. Unique self-D are included. Observe that the original bivariate distributions exhibit more DS pairs than occur by randomly permuting the data. In particular, the number of unique DS pairs in SV40 is 276, in excess of the maximum number in the 20 permutations.

Table 9. DS and self-D in polyoma

		Original		Permutation ranges	
Size	DS type	DS, no.	Self-D, %	DS, no.	Self-D, %
6	(1,1)	170	0.09	(243,285)	(0.06,0.13)
	(2,2)	94	0.08	(109,145)	(0.09,0.15)
	(3,3)	34	0.12	(20,36)	(0.17,0.32)
8	(1,1)	248	0.06	(192,229)	(0.06,0.10)
	(2,2)	2	0.00	(0,8)	(0.00,1.0)
10	(1,1)	30	0.17	(5,26)	(0.12,0.31)

The number of DS pairs and the proportion of self-D for polyoma and 20 permuted sequences are shown with the minimum and maximum, respectively, in parentheses.

Table 10. Bivariate distributions of DS of size 8 for SV40 and Polyoma

Virus	Original			Permutation ranges		
SV40		1	2	3	1	2	3
	1	276	22	3	(195,260)	(0,14)	(0,1)
	2	46	15	1	(15,25)	(0,4)	(0,0)
	3	6	2	0	(0,3)	(0,1)	(0,0)
	4	3	1	0	(0,0)	(0,0)	(0,0)
	5	0	0	0			
	6	1	0	0			
Polyoma		1	2		1	2	3
	1	248	14		(192,229)	(0,10)	(0,1)
	2	22	2		(7,16)	(0,2)	(0,0)
					(0,3)	(0,0)	(0,0)

Large DS. *Papovaviruses.* Apart from the size 13 exact DS (gap one) at the replication point in SV40, there are two 12-nucleotide DS pairs, both involving 1-kb loop length and a multiple-size 11 DS having A-A-T-T-A-G-T-C-A-G-C at position 27 with its two dyad words at positions 112 and 146.

Polyoma had no DS pair of size ≥ 10.

λ phage large DS. The largest exact DS consists of A-G-A-A-A-G-G-A-A-A-C-G-A-C-A-G (size 16), and its dyad word occurs at positions 109 and 151 (loop 26). The large size and short loop probably portend a biological function. This potential stem loop structure is 24 nucleotides 5′ of *Nu-1* gene (11).

We thank Drs. P. Botchan, D. Brutlag, A. Buchman, D. Clayton, P. Hieter, J. Maizel, C. Queen, and Ms. I. Stratton for valuable discussions and help in providing sequences for analysis. This work was supported in part by National Institutes of Health Grant GM10452-20 and National Science Foundation Grant MCS79-24310,A2 to S.K. and by American Cancer Society Grant CD 122 to L.J.K.

1. Korn, L. J., Queen, C. L. & Wegman, M. N. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 4401–4405.
2. Brutlag, D. L., Clayton, J., Friedland, P. & Kedes, L. H. (1982) *Nucleic Acids Res.* **10**, 279–294.
3. Pustell, J. & Kafatos, F. C. (1982) *Nucleic Acids Res.* **10**, 4765–4782.
4. Dumas, J. P. & Ninio, J. (1982) *Nucleic Acids Res.* **10**, 197–206.
5. Goad, W. B. & Kanehisa, M. (1982) *Nucleic Acids Res.* **10**, 247–263.
6. Wilbur, J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
7. Gingeras, T. R. & Roberts, R. J. (1980) *Science* **209**, 1322–1328.
8. Staden, R. (1980) *Nucleic Acids Res.* **8**, 817–825.
9. Sellers, P. H. (1974) *SIAM J. Appl. Math.* **26**, 787–793.
10. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
11. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. (1982) *J. Mol. Biol.* **162**, 729–773.
12. van Heuverswyn, H. & Fiers, W. (1979) *Eur. J. Biochem.* **100**, 51–60.
13. Soeda, E., Arrand, J. R., Smolar, N., Walsh, J. E. & Griffin, B. E. (1980) *Nature (London)* **283**, 445–453.
14. Seif, I., Khoury, G. & Dhar, R. (1979) *Cell* **18**, 963–977.
15. Danos, O., Katinka, M. & Yaniv, M. (1982) *EMBO J.* **1**, 231–236.
16. Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. & Young, I. G. (1981) *Nature (London)* **290**, 457–465.
17. Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. & Clayton, D. A. (1981) *Cell* **26**, 167–180.
18. Hieter, P. A., Maizel, J. V., Jr., & Leder, P. (1982) *J. Biol. Chem.* **257**, 1516–1522.
19. Dover, G. (1982) *Nature (London)* **299**, 111–117.
20. Oppenheim, D. S. & Yanofsky, C. (1980) *Genetics* **95**, 785–795.