J Mol Evol (1985) 21:278–288

Journal of
Molecular Evolution

Markov Chain Analysis Finds a Significant Influence of Neighboring Bases on the Occurrence of a Base in Eucaryotic Nuclear DNA Sequences Both Protein-Coding and Noncoding

B. Edwin Blaisdell

Linus Pauling Institute of Science and Medicine, PaloAlto, California 94305, USA

Summary. Sixty-four eucaryotic nuclear DNA sequences, half of them coding and half noncoding, have been examined as expressions of first-, second-, or third-order Markov chains. Standard statistical tests found that most of the sequences required at least second-order Markov chains for their representation, and some required chains of third order. For all 64 sequences the observed one-step second-order transition count matrices were effective in predicting the two-step transition count matrices, and 56 of 64 were effective in predicting the three-step transition count matrices. The departure from random expectation of the observed first- and second-order transition count matrices meant that a considerable sample of eucaryotic nuclear DNA sequences, both protein coding and noncoding, have significant local structure over subsequences of three to five contiguous bases, and that this structure occurs throughout the total length of the sequence. These results suggested that present DNA sequences may have arisen from the duplication, concatenation, and gradual modification of very early short sequences.

Key words: Higher-order Markov chains — Prediction of following DNA bases

Introduction

Recent observation (Blaisdell 1983a) has found in a collection of 30 eucaryotic nuclear DNA sequences that the coding and noncoding subsets are both nonrandom, but in different ways. Coding sequences exhibit an excess of runs of length 1 and 2 and a deficit of long runs of lengths 5 to 10 of both of two base classes: weak hydrogen bonding, W (A or T), and strong hydrogen bonding, S (C or G). Noncoding sequences exhibit a deficit of runs of length 1 and 2 and an excess of long runs of lengths 5 to 10 of both of two base classes of purine, R (A or G), and pyrimidine, Y (C or T). These respective kinds of nonrandomness were found in DNA sequences whose coding part coded for proteins of widely different function, for the same protein in widely different eucaryotic species, and, in the same species, for related sequences that diverged a long time ago and that now show large differences in base and (if coding) amino acid sequence. Further examination (Blaisdell 1983b) has found that the nonrandomness in coding sequences has been produced in substantial part by the occupants of codon sites three being unlike, in the W–S sense, their neighbors on both sides for up to two sites away. This paper examines more generally, in both coding and noncoding sequences, the influence of neighboring bases on the occurrence of a base.

In 1961 Kornberg and colleagues (Josse et al. 1961; Swartz et al. 1962) developed a method for analyzing for the relative numbers of the 16 possible pairs of adjacent bases (doublets) in the total genomic double-stranded DNA, and found that these numbers were significantly different from those calculated assuming random occurrence of the observed fractions of the individual four bases. The pattern of differences was about the same for 12 animal and plant DNAs and about the same for six bacterial DNAs but the patterns of the two sets were different from one another. The observed frequency of CG doublets was much less than expected in the total genome of animals and plants, but somewhat more than expected in bacteria. On the other hand,

GC doublets were much fewer than expected in bacteria and about as frequent as expected in animals and plants. TA doublets were much fewer than expected in both sets of organisms. Over the subsequent 15 years such analyses were extended to many more organisms by several authors and the stability of the nonrandomness patterns was confirmed in several classes of organisms. Elton (1975) summarized results on several short single-stranded RNAs (ribosomal, transfer, and phage) and introduced the use of first-order Markov chain analysis for studying the first base immediately following each of the four bases. He found that the doublet nonrandomness patterns in these short single-stranded sequences were different from those in the total genomic double-stranded sequences.

Salser (1977) suggested that the deficiency in CG doublets in animal genomes is due to the extensive methylation of C in this doublet, which could make such Cs hotspots for mutation. In 1978 Gilbert and coworkers (Coulondre et al. 1978) found that methylated C residues were in fact hotspots for mutation to T in the lacI gene of *Escherichia coli*. Jukes (1978) found that four mammalian messenger RNAs were like total genomic double-stranded DNA in having far fewer than the expected number of CG doublets. Bird (1980) noticed that the extent of CG-doublet deficiency was strongly correlated with the extent of C methylation in CG doublets in the genomes of several species and that in fact the deficiency in CG was about equal to the sum of the surpluses in TG and CA doublets, as would be expected if 5-methyl cytosine deaminates to thymidine. Nussinov (1980, 1981) gave the ratios of observed to expected (based on random base distribution) frequencies of the 16 base doublets in a larger collection of 44 sequences containing a total of 28,000 eucaryotic bases and 43,000 procaryotic bases. She pointed out noteworthy patterns but without using Markov chain analysis; in coding-sequence patterns, she did not distinguish among sites within the codon.

Erickson and Altman (1979), using information-theory analyses developed by Kullback et al. (1962), found that although the total sequence of the virus MS2 showed only slight first-order Markov chain dependence, it showed very significant second-order dependence ($P = 0.002$). This second-order dependence was great in the A gene and the noncoding regions, but small in the coat and replicase regions. They also found that the identity of the occupant of codon site 3 was dependent on the identities of the occupants of site 2 preceding and of site 1 succeeding site 3, especially in the coat and replicase genes. This finding has been confirmed in 30 other gene sequences by Blaisdell (1983b), who attributes the dependency to the conservation of excess short W and S runs. There is evidence of some errors in their

Table 1. Identification of genes

Code	Organism	Gene	Source
1	human	alpha 2 globin	Proudfoot and Maniatis (1980)
2	human	beta globin	Lawn et al. (1980)
3	human	A gamma globin	Slightom et al. (1980)
4	human	delta globin	Spritz et al. (1980)
5	human	epsilon globin	Baralle et al. (1980a,b)
6	mouse	alpha globin	Nishioka and Leder (1979)
7	mouse	beta globin major	Konkel et al. (1979)
8	mouse	beta globin minor	Konkel et al. (1979)
9	rabbit	beta globin	van Ooyen et al. (1979)
10	chick	beta globin	Richards et al. (1979)
11	human	immunoglobulin kappa constant	Hieter et al. (1980)
12	mouse	immunoglobulin kappa constant	Altenburger et al. (1981)
13	mouse	immunoglobulin gamma 1 constant	Takahashi et al. (1980)
14	mouse	immunoglobin kappa variable	Nishioka and Leder (1980)
15	mouse	immunoglobin gamma 2 variable	Sakano et al. (1980)
16	human	peproinsulin	Ullrich et al. (1980), Bell et al. (1980a,b)
17	rat	preproinsulin	Lomedico et al. (1979)
18	chick	preproinsulin	Perler et al. (1980)
19	human	cortico-lipotropin	Chang et al. (1980)
20	rat	prolactin	Gubbins et al. (1980)
21	human	alpha 2 interferon	Goeddel et al. (1980)
22	human	beta interferon	Lawn et al. (1981)
23	yeast	glyceraldehyde-3-phosphate dehydrogenase	Holland and Holland (1979)
24	yeast	N-(5'-phosphoribosyl)-anthranilate isomerase	Tschumper and Carbon (1980)
25	mouse	alpha amylase	Young et al. (1981)
26	yeast	actin	Ng and Abelson (1980)
27	sea urchin	histone H2B	Sures et al. (1978)
28	sea urchin	histone H3	Sures et al. (1978)
29	chick	ovalbumin (fragment)	Robertson et al. (1979)
30	French bean	phaseolin	Sun et al. (1981)
31	human	Interspersed repetitive Alu at 6000 bases 3' to insulin gene	Bell et al. (1980b)
32	human	Interspersed repetitive Alu at 2000 bases 5' to G gamma globin gene	Pan et al. (1981)
33	silkworm	fibroin	Tsujimoto and Suzuki (1981)
34	human	1920 bases 5' to epsilon globin	Baralle et al. (1980b)

Table 2. Sample first-order transition matrices for rabbit beta globin sequences

	T12				T13				T14				TSQ1				TCB1				TX1			
	T	C	A	G	T	C	A	G	T	C	A	G	T	C	A	G	T	C	A	G	T	C	A	G
Coding																								
T	17	27	7	58	24	30	27	28	31	24	20	34	30	26	22	31	27	26	22	34	27	25	23	34
C	42	25	31	4	27	20	21	34	26	24	22	30	23	24	21	35	26	24	21	31	25	24	21	31
A	16	18	28	30	21	19	16	36	26	22	20	24	22	21	21	28	23	22	19	28	23	21	19	28
G	34	33	25	43	36	34	28	37	25	34	30	46	34	32	28	41	34	32	28	42	34	31	28	42
Noncoding																								
T	161	76	75	92	161	77	90	76	160	82	86	76	140	79	101	84	139	80	103	82	139	80	103	82
C	91	49	84	8	82	51	42	57	78	54	54	46	82	46	60	44	80	46	59	47	80	46	59	47
A	89	62	82	68	95	52	101	53	94	58	81	68	102	60	78	61	103	59	77	61	103	59	77	61
G	63	45	60	71	66	52	67	54	72	39	79	49	79	47	62	51	82	47	61	49	82	47	61	49

For explanation of table, see text. T12, observed one-step order 1; T13, observed two-step order 1; T14, observed three-step order 1; TSQ1, predicted two-step order 1; TCB1, predicted three-step order 1; TX1, predicted limit order 1. Row labels represent 5′ elements of base pairs; column labels, 3′ elements

reported values; for example, their statistic T1 should be less than or equal to S1, but this is not true of some of their reported values.

Lipman and Wilbur (1983) studied the occurrence of doublets in the three possible codon phases in mitochondria, procaryotes, and nuclear DNA using the methods of information-theory analysis given by Gatlin (1972). Gatlin's method wastes information, since permutation of a row of the transition matrix does not change the D2 statistic used by Lipman and Wilbur. The method of Kullback et al. does not have this fault. In mitochondria the observations were in satisfactory agreement with the predictions of a model in which bases in site 3 in synonymous codons were assigned at random. In procaryotes this model failed, but the observations were in satisfactory agreement with the predictions of a model in which the distribution of bases in site 3 of synonymous-codon sets is not random but is the same in all synonymous sets. In eucaryotes both of these models failed, and the distribution of bases in codon site 3 was found to depend on the bases in both preceding site 2 of the same codon and succeeding site 1 of the following codon, in agreement with the finding of Blaisdell (1983b) by another method.

Almagor (1983) ran computer simulations of zeroth- and first-order Markov chains for two long sequences and determined the distributions of the identities of bases in position 3 relative to the identities of the bases in positions 1 and 2.

This paper describes the results of a general Markov chain study of the influence of neighboring bases on the occurrence of a base. The study was carried out separately for coding and noncoding base sequences in eucaryotic nuclear DNA. First-, second-, and third-order Markov chains were studied. The first-order one-step transition count matrices are very different from those calculated assuming random occurrence of the four bases (mutual independence of neighbors), as had been found earlier for the total double-stranded genome (Josse et al. 1961). The second-order one-step transition count matrices (which show the dependence of the base occupant of a site on the doublet occupying the immediately preceding two sites) are also very different from those calculated on the assumption of random occurrence of the four bases. However, the observed second-order one-step transition count matrix gives good predictions for the two- and three-step transitions. The departures from random expectation of the observed first- and second-order transition count matrices indicate that a significant local influence of neighboring bases on the identity of a base occurs throughout the length of a DNA sequence, whether coding or noncoding.

Results and Discussion

The sequences studied (Table 1) are the same set of about 30 eucaryotic nuclear protein-coding DNA sequences used previously (Blaisdell 1983a, 1983b). For this study results are given separately for a partition into four subsets. A first partition was made into coding and noncoding sequences because of the finding of different nonrandomness patterns for these two classes (Blaisdell 1983a). To avoid the possibility of overall results being dominated by the large number of globins in the sample, each of these subsets was further partitioned into two subsets, for a total of four subsets: (1) a coding subset consisting of 1 of 10 globins (rabbit beta), 1 of 5 immunoglobulins (mouse kappa constant), one of three insulins (human), one of two interferons (human beta), one of two histones (sea urchin H3), and all of the other coding sequences (this subset is hereafter called C13 because it consists of 13 coding sequences); (2)

Table 3. Sample second-order transition matrices for rabbit beta globin sequences

	T123				T124				T125				TSQ2				TCB2				TX2			
	T	C	A	G	T	C	A	G	T	C	A	G	T	C	A	G	T	C	A	G	T	C	A	G
Coding																								
TT	4	5	2	6	6	1	3	7	4	4	1	8	4	5	4	4	5	4	4	4	4	4	4	5
TC	7	7	12	1	10	5	4	8	6	8	6	7	8	6	6	8	6	6	5	10	7	6	6	8
TA	3	2	2	0	1	3	0	3	0	2	3	2	2	2	1	2	2	2	1	2	2	2	1	2
TG	10	16	11	21	14	15	13	16	14	14	9	21	16	15	13	14	14	13	12	19	14	14	12	18
CT	5	8	4	25	10	14	10	8	9	11	7	15	9	12	10	11	11	11	9	11	10	10	9	13
CC	15	5	4	0	7	4	5	9	9	1	5	9	6	5	4	10	5	6	5	8	6	6	5	7
CA	5	7	11	8	8	6	7	10	8	7	4	12	7	7	5	12	8	7	6	10	8	7	6	10
CG	2	0	1	1	1	0	0	3	1	2	0	1	1	1	1	2	1	1	1	1	1	1	1	1
AT	3	5	1	7	3	6	3	4	7	3	3	3	4	4	4	4	4	4	4	4	4	4	3	5
AC	5	7	4	2	7	3	5	3	3	7	3	5	7	3	4	4	4	4	3	7	4	4	4	6
AA	6	5	4	13	7	4	5	12	11	8	6	3	6	5	5	11	7	7	6	8	7	7	6	9
AG	7	2	7	14	9	9	7	5	8	3	5	14	7	8	6	9	7	7	6	10	7	7	6	9
GT	5	9	0	20	5	9	11	9	11	6	9	8	7	9	8	9	10	8	8	8	8	8	7	11
GC	15	6	11	1	3	9	7	14	8	8	8	9	8	7	7	12	7	8	7	11	8	8	7	10
GA	1	4	11	9	5	6	4	10	7	5	7	6	6	4	5	10	6	6	5	8	6	6	5	8
GG	15	15	6	7	12	10	8	13	2	15	16	10	12	10	9	13	10	10	9	14	11	10	9	13
Noncoding																								
TT	73	32	24	32	72	32	32	25	61	33	39	28	65	31	35	30	58	32	38	33	55	32	41	33
TC	34	11	27	4	29	21	5	21	27	20	14	15	27	16	13	20	27	15	20	15	26	15	19	16
TA	24	15	21	15	25	8	32	10	24	14	28	19	24	13	25	13	25	15	19	16	26	15	19	15
TG	30	19	18	25	34	21	17	20	31	11	33	17	26	20	26	20	31	18	24	29	32	18	23	19
CT	34	17	9	31	34	23	16	18	39	26	13	13	36	17	20	18	32	18	22	19	31	18	23	19
CC	19	16	13	1	21	11	9	9	14	9	11	15	18	11	9	11	18	10	12	10	17	10	12	10
CA	27	16	19	22	22	19	27	16	25	17	19	23	26	14	28	15	28	17	21	18	29	17	21	17
CG	2	2	1	3	2	1	2	3	2	1	3	2	2	2	2	2	2	2	2	2	3	2	2	2
AT	31	15	24	18	28	13	28	20	39	11	19	19	34	17	20	17	31	17	22	18	30	17	23	18
AC	23	10	27	2	21	15	11	15	23	17	13	9	22	13	11	16	22	12	16	12	21	12	16	13
AA	26	17	26	13	29	17	19	17	26	14	29	13	26	15	27	14	28	16	21	17	28	16	21	17
AG	15	10	24	19	16	13	23	16	22	17	18	11	18	15	19	16	22	13	19	14	23	13	17	14
GT	23	12	17	11	27	10	14	12	21	11	15	16	25	12	15	12	22	12	16	13	22	12	16	13
GC	15	12	17	1	12	4	17	12	14	8	16	7	16	10	9	10	16	9	11	9	16	9	11	9
GA	12	14	16	18	19	8	23	10	19	13	15	13	18	10	21	11	19	12	16	13	21	12	15	12
GG	16	14	17	24	14	17	25	15	17	10	25	19	19	16	20	16	23	14	19	15	24	14	18	14

For explanation of table, see text. T123, observed one-step order 2; T124, observed two-step order 2; T125, observed three-step order 2; TSQ2, predicted two-step order 2; TCB2, predicted three-step order 2; TX2, limit order 2. Row labels represent 5′ elements of base triplets; column labels, 3′ elements

the complementary coding subset consisting of the nine other globins, four other immunoglobins, two other insulins, one other interferon, and one other histone, called C17; (3) N15, consisting of the noncoding counterparts of the members of C13 together with one of two middle repetitive Alu regions and a silkworm fibroin noncoding sequence; and (4) N19, consisting of the complementary noncoding subset consisting of the noncoding counterparts of C17 together with the other middle repetitive Alu region and a 1920-base noncoding sequence 5′ to the human epsilon globin gene.

Tables 2 and 3 present samples, using rabbit beta globin, of the transition count matrices the statistical properties of which for all sequences will be considered. In Table 2, for first-order Markov chains, the entries in the 4 × 4 matrices are the counts of all base pairs in the sequence; the 5′ element of each pair is given by the row label (the first column of the table) and the 3′ element of each pair is given by the column heading of the respective matrices. This means that a row in the matrix gives the distribution of the counts of the four bases in positions 3′ to the base specified by the row label and that a column in the matrix gives the distribution of the counts of the four bases in positions 5′ to the base specified by the column label. In matrix T12, the 3′ element of the pair is in position 2 relative to the 5′ element in position 1; in matrix T13, the 3′ element of the pair is in position 3 relative to the 5′ element in position 1; and similarly for matrix T14.

The matrix TSQ1 gives the counts predicted for T13 if the sequence is assumed to be the expression of a Markov chain of order 1 with transition count matrix equal to the observed T12. Let t12 be the transition matrix corresponding to T12 and ob-

tained from it by dividing each element by the sum of the elements in its row. Then the two-step matrix TSQ1 is given by

$$\mathrm{TSQ1}(i, j) = \left[\sum_k \mathrm{t12}(i, k)\mathrm{t12}(k, j)\right] \sum_k \mathrm{T12}(i, k), \qquad i, j, k = 1, 2, 3, 4$$

A first-order Markov process is defined as one in which the occupant of site k depends only on the occupant of site k − 1 and not on the occupants of sites 0, 1, 2, . . . , k − 2.

The matrix TCB1 gives the counts predicted for T14 if the sequence is assumed to be the expression of a Markov chain of order 1 with transition count matrix equal to the observed T12. Then the three-step matrix TCB1 is given by

$$\mathrm{TCB1}(i, j) = \left[\sum_{k,l} \mathrm{t12}(i, k)\mathrm{t12}(k, l)\mathrm{t12}(1, j)\right] \sum_k \mathrm{T12}(i, k), \quad i, j, k, l = 1, 2, 3, 4$$

The matrix TX1 gives the counts expected if the occupant of position k is independent of the occupants of all other positions. This means, for example, that the doublets TT, TC, TA, and TG of row T are distributed proportionally to the total numbers of T, C, A, and G, respectively, in the sequence, and that the sum of the numbers of TT, TC, TA, and TG equals the total number of T residues in the sequence. That is,

$$\mathrm{TX1}(i, j) = \text{sum of row } i \times \text{sum of column } j / \text{sum of all elements.}$$

Therefore the transition count matrix is symmetric.

The counts in matrices TSQ1, TCB1, and TX1 are rounded from the seven-digit values obtained from single-precision computer calculation.

Table 3 presents a sample, using rabbit beta globin, of the second-order transition count matrices the statistical properties of which for all sequences will be considered. The entries in the 16 × 4 matrices are the counts of all triples of bases in the sequence; the 5′ doublet of each triple is given by the row label (the first column) and the 3′ singlet is given by the column heading. In matrix T123, the 3′ singlet of the triple is in position 3 relative to the 5′ doublet in positions 1 and 2 of the triple, and similarly for T124 and T125.

Matrix TSQ2 gives the counts predicted for T124 by the two-step application of the observed transition count matrix T123; that is, it is the matrix that would be expected if the sequence were the expression of a second-order Markov chain. TSQ2 is given by

$$\mathrm{TSQ2}(i, j) = \left[\sum_k \mathrm{t123}(i, k)\mathrm{t123}(k + 4m, j)\right] \sum_k \mathrm{T123}(i, k),$$

$$i = 1, 2, 3, \ldots, 16, \qquad j, k = 1, 2, 3, 4, \qquad m = (i - 1) \bmod 4.$$

A second-order Markov chain is defined as one in which the occupant of site k depends only on the occupants of sites k − 1 and k − 2 and not on the occupants of sites 0, 1, 2, . . . , k − 3.

Similarly, matrix TCB2 gives the counts predicted for T125 by the three-step application of T123:

$$\mathrm{TCB2}(i, j) = \left[\sum_{k,l} \mathrm{t123}(i, k)\mathrm{t123}(k + 4m, 1)\,\mathrm{t123}(1 + n(k), j)\right] \sum_k \mathrm{T123}(i, k),$$

$$i = 1, 2, 3, \ldots, 16, \qquad j, k, l = 1, 2, 3, 4, \qquad m = (i - 1) \bmod 4, \qquad n = (0, 4, 8, 12).$$

Matrix TX2 gives the counts expected if the occupants of singlet position 3 are assigned at random as described above for the first-order analysis.

Casual inspection of Table 2 finds that observed T12 is quite different from the predicted (independent, zeroth order) matrix TX1 based on a random base distribution, but that T13 and T14 and TSQ1, TCB1, and TX1 are much alike. Similarly, in Table 3 T123 is quite different from TX2, but T124 bears some resemblance to TSQ2. (For a clear similarity, look at row GG of the transition matrices.

The significance of the apparent difference between T12 and TX1 may be determined from the statistic

$$\mathrm{S1} = \sum_{i,j} [\mathrm{T12}(i, j) - \mathrm{TX1}(i, j)]^2 / \mathrm{TX1}(i, j)$$

which is asymptotically distributed as chi squared with nine degrees of freedom (Anderson and Goodman 1957). Its significance may also be determined from the statistic

$$\mathrm{L1} = \sum_{i,j} 2\mathrm{T12}(i, j) \ln[\mathrm{T12}(i, j)/\mathrm{TX1}(i, j)]$$

which is likewise asymptotically distributed as chi squared with nine degrees of freedom (Kullback et al. 1962). For these finite sequences I have found S1 to be approximately equal to L1. From here on I shall use the first alternative, S1, which can be regarded as the squared Euclidian length of the vector of weighted elements of the differences of the

Table 4. Significance levels of Markov chain order determinations: $-10 \log P$

Sequence code	Order 1	Order 2	Order 3
		Subset C13	
9	126	16	
12	53	3	
16	97	9	
19	23	35	
20	94	14	
22	102	14	
23	74	235	
24	20	14	
25	30	0	
26	36	133	
28	34	11	
29	48	4	
30	58	16	
		Subset C17	
1	32	59	
2	125	28	
3	119	6	
4	122	19	
5	98	19	
6	83	12	
7	115	6	
8	123	5	
10	89	8	
11	45	3	
13	182	1	
14	64	9	
15	43	11	
17	47	22	
18	46	28	
21	151	0	
27	29	14	
		Subset N15	
9	119	10	
12	257	9	14
16	302	34	32
19	45	7	
20	139	22	8
22	23	9	
23	4	2	
24	5	5	
25	38	9	
26	25	23	
28	18	4	
29	172	66	7
30	2	4	
31	160	2	4
33	28	16	17
		Subset N19	
1	45	5	
2	148	9	4
3	132	22	
4	138	40	14
5	249	10	0
6	222	26	17
7	195	21	
8	166	36	
10	132	16	
11	127	7	
13	133	8	
14	136	72	
15	31	16	
17	175	31	
18	300	7	
21	133	10	
27	70	31	
32	34	5	
34	155	31	

two matrices regarded as a point in 16-dimensional space. These statistics test the null hypothesis that the sequence is independent (zeroth order) within the hypothesis that the sequence is a first-order Markov chain. Significance levels as $-10 \log P$ are given in the columns headed "order 1" in Table 4. Note that $-10 \log P = 20$ corresponds to $P = 0.01$, and $-10 \log P = 70$ corresponds to $P = 0.0000001$. The null hypothesis is rejected at the level $P = 0.01$ for all but three (codes 23, 24, and 30 of subset N15) of the 64 sequences, and in 37 of 64 cases at a much more significant level, $P = 0.0000001$. This means that in nearly all sequences, both coding and noncoding, the identity of the occupant of site k is strongly influenced by that of the occupant of site $k - 1$.

Similarly, the null hypothesis that the sequence is a first-order Markov chain within the hypothesis that the sequence is a second-order Markov chain may be tested with the statistic

$$L2 = \sum_{ijk} 2T(ijk)\ln\{T(ijk) \times T(\cdot j\cdot)/ [T(ij\cdot) \times T(\cdot jk)]\}$$

(Kullback et al. 1962) and its χ^2-like analog

$$S2 = \sum_{ijk} T(\cdot j\cdot) \times T(ij\cdot) \times [T(ijk)/T(ij\cdot) - t(\cdot jk)/T(\cdot j\cdot)]^2/T(\cdot jk)$$

where commas have been omitted from the index strings for simplicity, dot means summation over the index replaced by the dot, and $T = T123$. The statistic L2 is asymptotically distributed as chi squared with 36 degrees of freedom and has been found to be approximately equal to S2. In Table 3 index i is to be associated with the first character of the row label, j with the second character of the row label, and k with the column-heading character. Significance levels of this statistic are given in the columns headed "Order 2" in Table 4. In general, these significances are less than those for the rejection of

independence (zeroth order). However, in 43 of 64 cases, the hypothesis that the sequence is adequately represented by a first-order Markov chain is rejected at the significance level P = 0.1. This means that in these cases, for both coding and noncoding sequences, the identity of the occupant of site k is substantially influenced by those of the pair of occupants in sites k − 2 and k − 1.

The null hypothesis that the sequence is a second-order Markov chain within the hypothesis that the sequence is a third-order Markov chain may be tested with the statistic

$$L3 = \sum_{ijkl} 2T(ijkl)\ln\{T(ijkl) \times T(\cdot jk\cdot)/[T(ijk\cdot) \times T(\cdot jkl)]\}$$

(Kullback et al. 1962) and its χ^2-like analog

$$S3 = \sum_{ijkl} T(\cdot jk\cdot) \times T(ijkl) \times [T(ijkl)/T(ijk\cdot) - T(\cdot jkl)/T(\cdot jk\cdot)]^2/T(\cdot jkl)$$

L3 is asymptotically distributed as chi squared with 144 degrees of freedom and has been found to be approximately equal to S3. Since calculation of this statistic requires the counting of 256 4-tuples, it has been done only for sequences longer than 1400 bases, all noncoding, to insure an average cell count greater than 5. These values are given in the columns headed "Order 3" in Table 4. In 5 of 10 cases the null hypothesis that the sequence is adequately represented by a second-order Markov chain is rejected at the significance level P = 0.1. This means that in these five cases, the identity of the occupant of site k is substantially influenced by the identities of the occupants of sites k − 3, k − 2, and k − 1.

Tests S1, S2, and S3 (or L1, L2, and L3) permit a parallel determination of the minimum order, 1, 2, or 3, of Markov chain necessary to represent the parallel global population of doublets, triplets, or quadruplets in a sequence. Further, more local validation of the effectiveness of a Markov chain of the chosen order as a representation of the sequence can be found in additional properties of the Markov chain. Since in the observed sequences each base is eventually followed by every base (i.e., the chain is regular), Markov chain theory proves that transition count matrices for 1, 2, . . . , k steps approach, as k increases without limit, the random transition count matrix TX (in which, in all rows, columns T, C, A, and G are proportional to the total counts of T, C, A, and G residues in the sequence). In fact, for these observed sequences TX is obtained exactly to the nearest unit after only a few steps; that is, the distribution of the bases only a few positions beyond a given singlet, doublet, or triplet is random. The average squared weighted distance between any two transition count matrices A and B is calculated as

$$S = (2/n)\sum_{ij} [A(ij) - B(ij)]^2/[A(ij) + B(ij)]$$

where n, the number of elements in a matrix, is 16, 64, and 256 for first-, second-, and third-order matrices, respectively. Although S has the form of a chi square statistic, significance-level probabilities cannot be determined from the usual tables because the observed sequences have been found not to be independent according to the statistic S1 (S. Karlin, personal communication). Values of S for matrix pairs (A,B) of interest are given in Tables 5, 6, and 7 for first-, second-, and third-order matrices, respectively.

Consider transition count matrices for first-order Markov chains, such as those shown in Table 2. According to the theorem stated, T13 is expected to be closer to TX1 than is T12, and in turn T14 is expected to be closer than is T13. The numbers of sequences for which these statements are true are given in Table 8, lines 1 and 2. It is also expected that T13 will be closer to TSQ1, the matrix predicted for two steps from T12, than to TX1, the matrix predicted for a very large number of steps, and that T14 will be closer to TCB1, the matrix predicted for three steps, than to TX1. The numbers of sequences for which these statements are true are given in Table 8, lines 3 and 4. Corresponding counts for second- and third-order Markov chains are also given in Table 8.

The S1 test found that 61 of the 64 observed sequences required at least first-order Markov chains for their representation (Table 4, column "Order 1"), and for these same 61 of 64 T13 is closer to the limiting matrix TX1 than is T12 (Table 8, line 1). In turn, for 47 of 64 sequences T14 is closer to TX1 than is T13 (Table 8, line 2). The lower number in this case is probably due to the facts that T13 is already very close and that inevitable statistical fluctuation (Almagor 1983) obscures the somewhat greater closeness of T14. Of greater significance are the values in lines 3 and 4 of Table 8, which show that the observed two-step transition count matrix T13 is closer to TSQ1, the two-step transition count matrix predicted by the two-step application of the observed one-step transition count matrix T12, than to the limiting matrix TX1 (line 3), and similarly for T14 and TCB1 (line 4). The expected relation holds true for 47 of the 64 studied sequences for the first of these comparisons, and for 40 of 64 for the second. The failure to achieve complete agreement with first-order expectations is probably due to the fact that most of the sequences require at least second-order Markov chains for their representation and, to a lesser extent, to statistical fluctuation.

Table 5. Average weighted squared distances (× 10) between pairs of first-order transition matrices

Coding						Noncoding					
	Matrix pair						Matrix pair				
Sequence code	T12 TX1	T13 TX1	T14 TX1	T13 TSQ1	T14 TCB1	Sequence code	T12 TX1	T13 TX1	T14 TX1	T13 TSQ1	T14 TCB1
	Subset C13						Subset N15				
9	49	5	3	5	4	9	47	18	11	17	11
12	26	6	14	8	18	12	90	8	4	6	4
16	40	15	17	20	20	16	104	14	6	13	6
19	14	6	13	6	13	19	23	12	10	11	11
20	39	8	5	8	5	20	53	21	9	20	9
22	42	12	9	9	9	22	15	9	6	8	7
23	33	56	25	71	28	23	6	2	6	3	7
24	14	9	4	9	4	24	7	8	3	9	3
25	18	6	4	6	4	25	21	8	8	6	8
26	37	22	9	26	9	26	16	25	21	26	25
28	19	6	5	6	6	28	13	12	9	15	9
29	24	4	3	4	3	29	64	35	6	31	6
30	28	8	5	8	5	30	5	11	5	11	6
						31	60	16	24	13	23
						33	17	6	17	6	18
	Subset C17						Subset N19				
1	18	11	11	13	11	1	23	4	5	3	5
2	49	4	3	6	4	2	56	13	4	15	4
3	48	7	8	6	8	3	51	13	4	11	4
4	48	4	4	6	4	4	53	27	4	27	4
5	41	6	9	4	10	5	88	15	8	10	9
6	36	8	9	8	10	6	79	23	19	19	19
7	46	6	5	4	5	7	71	18	8	14	10
8	48	5	3	5	3	8	62	24	15	20	16
10	38	7	8	6	8	10	51	14	9	12	9
11	23	7	8	4	8	11	50	13	11	8	9
13	67	8	6	6	7	13	52	12	2	13	3
14	29	6	8	4	8	14	53	26	18	21	17
15	22	8	6	9	6	15	18	16	9	16	10
17	24	10	11	11	11	17	64	10	8	6	8
18	23	13	11	23	13	18	103	13	8	13	8
21	58	10	9	6	10	21	52	6	4	6	4
27	17	6	10	6	11	27	36	12	3	9	3
						32	19	7	12	8	12
						34	59	14	22	13	25

For abbreviations, see Table 2

Test S2 found that 43 of the 64 observed sequences required at least second-order Markov chains for their representation at the P = 0.1 level of significance. Nevertheless, for all 64 of sequences T124, the observed second-order transition count matrix, is closer to TSQ2, the two-step second-order transition count matrix predicted by the two-step application of T123, the observed one-step second-order transition matrix, than to the limiting matrix TX2 (Table 8, line 7). A similar agreement with second-order Markov chain predictions for the three-step transition count matrices T125 and TCB is observed for 56 of 64 sequences (Table 8, line 8). Lines 7 and 8 of Table 8 show that the ability of the observed second-order one-step transition count matrix T123 to predict the observed two- and three-step transition count matrices T124 and T125 is a better determinant of second-order character than is the standard statistic S2. It is better because it is more sensitive; for example, in Table 4, subset C13, codes 12, 25, and 29, the S2 significance levels are 0.5, 0.9, and 0.4, respectively, but all three show power to predict T124 (Table 6), and power to predict is more satisfying intuitively. The observed ordering of the distances of the observed second-order one-, two-, and three-step transition count matrices from the limiting k-step transition count matrix is that expected for most of the 64 sequences (Table 8, lines 5 and 6).

Test S3 found that 5 of the 10 noncoding sequences selected for their lengths (greater than 1400 bases) required at least third-order Markov chains

Table 6. Average weighted squared distances (× 100) between pairs of second-order transition matrices

Coding						Noncoding					
	Matrix pair						Matrix pair				
Sequence code	T123 TX2	T124 TX2	T125 TX2	T124 TSQ2	T125 TCB2	Sequence code	T123 TX2	T124 TX2	T125 TX2	T124 TSQ2	T125 TCB2
	Subset C13						Subset N15				
9	213	63	88	45	91	9	189	131	91	84	84
12	117	83	97	55	88	12	294	89	59	67	61
16	181	123	116	70	91	16	366	117	95	72	88
19	145	116	100	92	94	19	119	109	94	86	102
20	180	84	88	61	83	20	236	111	109	64	109
22	183	73	81	42	77	22	105	75	69	53	66
23	408	250	122	105	102	23	69	63	86	53	86
24	113	70	56	38	53	24	80	61	61	41	59
25	83	67	72	55	69	25	119	75	78	67	77
26	313	163	92	97	89	26	131	186	128	128	100
28	117	67	63	52	58	28	91	116	102	84	98
29	113	66	69	50	70	29	316	147	73	58	70
30	150	80	48	72	52	30	67	94	78	67	75
						31	244	133	153	81	122
						33	92	80	83	67	81
	Subset C17						Subset N19				
1	198	94	111	48	92	1	105	67	66	53	67
2	219	61	80	52	81	2	220	84	91	50	88
3	183	78	116	56	114	3	222	102	98	59	89
4	214	61	78	53	75	4	247	122	136	45	119
5	188	70	123	39	113	5	278	94	55	52	50
6	169	75	97	53	97	6	311	141	142	86	123
7	180	80	95	64	95	7	267	119	83	69	81
8	183	63	81	47	80	8	272	131	114	69	92
10	164	81	116	53	116	10	213	92	103	50	94
11	116	78	72	58	64	11	189	89	86	58	73
13	220	84	92	59	88	13	200	75	67	52	66
14	144	86	77	67	73	14	297	180	155	69	105
15	138	86	64	58	58	15	130	97	78	63	67
17	155	103	95	58	73	17	270	108	111	69	109
18	167	109	98	48	78	18	322	109	97	70	91
21	180	94	92	64	88	21	198	63	77	44	77
27	114	73	77	52	75	27	200	80	61	47	55
						32	108	100	77	83	69
						34	264	138	147	77	113

For abbreviations, see Table 3

for their representation at the P = 0.1 level of significance. Nevertheless, for 10 of 10 of these sequences T1235, the observed two-step third-order transition count matrix, is closer to TSQ3, the two-step third-order transition matrix predicted by the two-step application of T1234, the observed one-step third-order transition count matrix, than to TX3, the limiting third-order transition matrix (Table 8, line 10). The observed ordering of the distances of T1234 and T1235 from TX3 is that expected for all 10 of these 10 sequences (Table 8, line 9).

In conclusion, Table 4 shows by means of standard tests of statistical significance that in most of this varied collection of 64 eucaryotic nuclear DNA sequences, both coding and noncoding, first-order Markov chains cannot explain the observed distribution of base triplets, and therefore Markov chains of at least second order are necessary. For a subset of noncoding DNA sequences longer than 1400 bases, test S3 shows at the P = 0.1 significance level that in 5 of the 10 sequences second-order Markov chains cannot explain the observed distribution of base quadruplets, and therefore Markov chains of at least third order are necessary. Table 8 shows the even more satisfying result that for all 64 sequences the observed second-order transition count matrix T123 is effective in predicting the bases occupying sites k + 3 and k + 4 with respect to any triplet occupying sites k, k + 1, and k + 2 for all k. These results mean that the members of this sample of eucaryotic nuclear DNA sequences both coding and

Table 7. Average weighted squared distances between pairs of transition matrices for noncoding sequences of length >1400, 10 × order 1, 100 × order 2, 100 × order 3

Sequence code	Matrix pair								
	T12 TX1	T13 TX1	T13 TSQ1	T123 TX2	T124 TX2	T124 TSQ2	T1234 TX3	T1235 TX3	T1235 TSQ3
					Subset N15				
12	90	8	6	294	89	67	140	78	50
16	104	14	13	366	117	72	208	56	53
20	53	21	20	236	111	64	123	95	63
29	64	35	31	316	147	58	139	99	53
31	60	16	13	244	133	81	73	73	52
33	17	6	6	92	80	67	119	98	60
					Subset N19				
2	56	13	4	220	84	50	106	92	73
4	53	27	27	247	122	45	139	105	65
5	88	15	10	278	94	52	125	66	41
34	59	14	13	264	138	77	143	107	80

T1234, observed one-step order 3; T1235, observed two-step order 3; TSQ3, predicted two-step order 3; TX3, limit order 3. Other abbreviations as in Tables 2 and 3

Table 8. Counts of sequences for which intermatrix-distance comparisons shown hold true

Statement	Sequence subset				
	S13	S17	N15	N19	All
First order					
1. T12–TX1 ≥ T13–TX1	13	17	12	19	61
2. T13–TX1 ≥ T14–TX1	10	9	12	16	47
3. T13–TX1 ≥ T13–TSQ1	9	11	11	16	47
4. T14–TX1 ≥ T14–TCB1	8	10	9	13	40
Second order					
5. T123–TX2 ≥ T124–TX2	13	17	12	19	61
6. T124–TX2 ≥ T125–TX2	7	6	12	12	37
7. T124–TX2 ≥ T124–TSQ2	13	17	15	19	64
8. T125–TX2 ≥ T125–TCB2	10	16	12	18	56
Third order					
9. T1234–TX3 ≥ T1235–TX3					10[a]
10. T1235–TX3 ≥ T1235–TSQ3					10[a]

Abbreviations as in Tables 2, 3, and 7

[a] Out of 10 sequences of length >1400 bases (see Table 7)

noncoding show significant local structure over subsequences of three to five contiguous bases throughout their total lengths. This suggests that present-day DNA sequences may have arisen from the duplication, concatenation, and gradual modification of very early short sequences, as has been proposed by Zuckerkandl (1975) and Ohno and Epplen (1983).

Acknowledgments. I am grateful to S. Karlin for very helpful comments on an early version of the manuscript. I thank E. Zuckerkandl for asking me to examine the statistical properties of DNA sequences, and C. Gorham for preparation of the manuscript.

References

Almagor H (1983) A Markov chain analysis of DNA sequences. J Theor Biol 104:633–645

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

Anderson TW, Goodman LA (1957) Statistical inference about Markov chains. Ann Math Stat 28:89–109

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

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

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

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

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Blaisdell BE (1983a) A prevalent persistent nonrandomness that distinguishes coding and noncoding eucaryotic nuclear DNA sequences. J Mol Evol 19:122–133

Blaisdell BE (1983b) Choice of base at silent codon site 3 is not selectively neutral in eucaryotic structural genes: It maintains excess short runs of weak and strong hydrogen bonding bases. J Mol Evol 19:226–236

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

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775–780

Elton RA (1975) Doublet frequencies in sequenced nucleic acids. J Mol Evol 4:323–346

Erickson JW, Altman G (1979) A search for patterns in the nucleotide sequence of the MS2 genome. J Math Biol 7:219–230

Gatlin L (1972) Information theory and the living system. Columbia University Press, New York

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

Gubbins EJ, Maurer RA, Lagrimini M, Erwin CR, Donelson JE (1980) Structure of the rat prolactin gene. J Biol Chem 255: 8655–8662

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

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

Josse J, Kaiser AD, Kornberg A (1961) Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J Biol Chem 236: 864–875

Jukes TH (1978) Codons and nearest neighbor nucleotide pairs in mammalian messenger RNA. J Mol Evol 11:121–127

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

Kullback S, Kupperman M, Ku HH (1962) Tests for contingency tables and Markov chains. Technometrics 4:573–608

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

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

Lipman DJ, Wilbur WJ (1983) Contextual constraints on synonymous codon choice. J Mol Biol 163:363–376

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

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

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

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

Nussinov R (1980) Some rules in the ordering of nucleotides in the DNA. Nucleic Acids Res 8:4545–4562

Nussinov R (1981) The universal dinucleotide asymmetry rules in DNA and amino acid codon choice. J Mol Evol 17:237–244

Ohno S, Epplen JT (1983) The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc Natl Acad Sci USA 80:3391–3395

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

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

Richards RJ, Shine J, Ullrich A, Wells JRE, Goodman HM (1979) Molecular cloning and sequence analysis of adult chicken beta globin cDNA. Nucleic Acids Res 7:1137–1146.

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

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

Salser W (1977) Globin messenger—RNA sequences—analysis of base-pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 42:985–1103

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

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

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

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

Swartz MN, Trautner TA, Kornberg A (1962) Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J Biol Chem 237:1961–1967

Takahashi N, Kataoka T, Honjo T (1980) Nucleotide sequences of class-switch recombination region of the mouse immunoglobulin gamma 2b-chain gene. Gene 11:117–127

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

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

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

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

Zuckerkandl E (1975) The appearance of new structures and functions in proteins during evolution. J Mol Evol 7:1–57

Received March 2, 1984/Revised December 4, 1984