Volume 15 Number 6 1987 Nucleic Acids Research

Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis

Gregory J.Phillips, Jonathan Arnold and Robert Ivarie

Department of Genetics, University of Georgia, Athens, GA 30602, USA

Received September 9, 1986; Revised and Accepted February 18, 1987

ABSTRACT

Several statistical methods were tested for accuracy in predicting observed frequencies of di- through hexanucleotides in 74,444 bp of E. coli DNA. A Markov chain was most accurate overall, whereas other methods, including a random model based on mononucleotide frequencies, were very inaccurate. When ranked highest to lowest abundance, the observed frequencies of oligonucleotides up to six bases in length in E. coli DNA were highly asymmetric. All ordered abundance plots had a wide linear range containing the majority of the oligomers which deviated sharply at the high and low ends of the curves. In general, values predicted by a Markov chain closely followed the overall shape of the ordered abundance curves. A simple equation was derived by which the frequency of any nucleotide longer than four bases in the E. coli genome (or any genome) can be relatively accurately estimated from the nested set of component tri- and tetranucleotides by serial application of a 3rd order Markov chain. The equation yielded a mean ratio of 1.03±0.94 for the observed-to-expected frequencies of the 4,096 hexanucleotides. Hence, the method is a relatively accurate but not perfect predictor of the length in nucleotides between hexanucleotide sites. Higher accuracy can be achieved using a 4th order Markov chain and larger data sets. The high asymmetry in oligonucleotide abundance means that in the E. coli genome of 4.2 X 10^6 bp many relatively short sequences of 7-9 bp are very rare or absent.

INTRODUCTION

Analysis of prokaryotic and eukaryotic DNA sequences has uncovered distinct patterns in the frequencies of their component oligonucleotide sequences. For example, the dinucleotide TA is infrequent in prokaryotic DNA but not in DNA from higher eukaryotes while CG is rare in higher eukaryotic but not prokaryotic DNA (1-4). In spite of the nonrandomness of oligonucleotide frequencies in genomes, most estimates of oligomer frequencies use a random distribution of mononucleotides, especially in computer programs (5-7) to provide a baseline for comparison. As an alternative, we have evaluated several statistical methods in their ability to predict observed frequencies of oligomers up to 6 bases in length and report the

results here. A Markov chain model was clearly the best predictor of all methods and its accuracy was measured in several ways.

The random model was entirely inadequate in predicting frequencies of virtually all sequence lengths. Hence, computer programs should incorporate a Markov chain model to predict frequencies of oligonucleotides using, at a minimum, the observed frequencies of tri- and tetranucleotides as a starting point. Blaisdell (8) has recently reported the use of a Markov chain to identify similarities among eukaryotic DNA sequences which were undetectable by algorithms requiring sequence alignment. Markov chain analysis reported here should also be useful in designing highly specific probes for screening gene libraries and in restriction mapping of small genomes and chromosomes.

MATERIALS AND METHODS

Sequence Analysis

All E. coli DNA sequence data were retrieved from Genbank (Fall 1984 update). Genes and their flanking regions used are: L11 and β operons (12,337 bp), rrnB operon (7,508 bp), trp operon (7,335 bp), ATP synthase operon (7,152 bp), lac operon (5,808 bp), frd operon (5,482 bp), malK-lamB operon (3,799 bp), tar, tap (3,465 bp), ilvG, ilvE (2,488 bp), thrA (2,463 bp), ecoRI (2,334 bp), motA, motB (2,212 bp), tufB (1,937 bp), fnr (1,651 bp), deoC (1,538 bp), recA (1,390 bp), araC (1,335 bp), fol (1,200 bp), trpR (1,043 bp), str (1,016 bp), lexA (951 bp). The total number of nucleotides was 74,444 which included both coding and noncoding regions of DNA. Given the size of the data set, variations greater than 5% were statistically significant at a confidence level of 0.001 or less.

Sequence editing and computing were done on a Digital PDP-11/34A in the RSX-11M operation system (V4.0) (Digital Equipment Corporation), using a software package DNASEQ developed for DNA sequence analysis on the PDP-11/34A (9). Data sets were created by compiling DNA sequences from coding strands head-to-tail with spaces between genes to exclude nucleotide combinations not found in the E. coli genome. In all cases, sequences were compiled by moving a single base per step in the 5' to 3' direction on the transcribed strand.

Likelihood Ratio Test and Residual Values

Values of a nucleotide sequence predicted by a Markov chain were compared to observed values by a likelihood ratio (LR) test (10). The test statistic, G, is the sum of the squared residuals (e^2) calculated as: e^2 =

2 [(ob) ln(ob/ex)-(ob-ex)], when ob > 0, where ob is the observed frequency and ex, the expected frequency; e^2 = 2 [ex], when ob = 0. The residual is given the sign of (ob-ex), and is a useful indicator of fit because it indicates whether a sequence is present at frequency higher (positive value) or lower (negative value) than that predicted by a particular model. The magnitude of e indicates the extent of deviation from the expected frequencies and has been rounded to the nearest integer in the accompanying table.

Ordered Abundance Curves And Complement Ratio Plots.

Two graphical methods have been used to depict the statistical results. In the first, observed frequencies of oligomers of a given length were plotted from highest to lowest abundance along with Markov predicted values for comparison. Sequences denoted numerically on the X-axis are available on 5¼ inch diskettes (if supplied) in CP/M. In the second method, influence of codon usage on observed frequencies has been evaluated in part by plotting the ratio of the frequency of an X-mer to its complement in coding DNA. Since only coding DNA was analyzed, a ratio of one indicates that the X-mer and its complement occurred with equal frequency in either coding or noncoding DNA. Ratios greater than or less than one indicate that a X-mer occurred at a frequency higher or lower, respectively, than its complement in coding DNA.

Restriction Enzyme Analysis.

E. coli K12 DNA (strain AB1157) was digested with restriction enzymes HaeIII, Sau3A1, EcoRI, and MaeI under optimal reaction conditions, electrophoresed on 0.8% agarose gels, and visualized by ethidium bromide staining.

RESULTS AND DISCUSSION

The Markov Chain Rule Most Accurately Predicts Oligonucleotide Frequencies

To find a statistical method that most accurately described the levels of short sequences in E. coli DNA, frequencies of subsequences were used in various ways to estimate the frequency of larger sequences. For example, expected frequency of the tetranucleotide CTAG, p(CTAG), can be estimated several ways depending on the amount of data available. First, from base composition (Table 2A), the frequency is the product of the frequency of the four constituent mononucleotides, p(N) where N = T, C, A, G:

$$p(CTAG|C,T,A,G) = p(C)\cdot p(T)\cdot p(A)\cdot p(G). \quad [1]$$

Although the most widely used, this is the least accurate method (Table 1 and Fig. 2C). Second, if the frequency of dinucleotides is known, the

Table 1. Comparison of various random sequence models to the Markov chain rule in predicting tetranucleotide levels in E. coli genes.[a]

gene	bp	Equation 1		Equation 2		Equation 3		Equation 5		Equation 6	
		G^{189}	(α)	G^{177}	(α)	G^{129}	(α)	G^{177}	(α)	G^{144}	(α)
recA	1390	423.21	(0.00)	341.89	(0.00)	284.19	(0.00)	382.88	(0.00)	205.054	(0.000)
thyA	1160	302.61	(0.00)	257.16	(0.00)	284.25	(0.00)	361.43	(0.00)	143.88	(0.506)
dam	1134	348.42	(0.00)	301.90	(0.00)	244.70	(0.00)	325.42	(0.00)	164.22	(0.119)
fol	1200	400.61	(0.00)	318.32	(0.00)	256.73	(0.00)	358.94	(0.00)	155.45	(0.243)
lexA	952	361.62	(0.00)	302.98	(0.00)	233.63	(0.00)	333.08	(0.00)	134.05	(0.713)
atp5	817	403.48	(0.00)	319.41	(0.00)	257.49	(0.00)	279.71	(0.00)	152.970	(0.289)
ssb	764	515.77	(0.00)	412.41	(0.00)	331.30	(0.00)	365.83	(0.00)	197.21	(0.000)
rnh	757	470.84	(0.00)	366.90	(0.00)	389.26	(0.00)	383.21	(0.00)	159.64	(0.173)
rplK	428	412.13	(0.00)	374.92	(0.00)	284.87	(0.00)	393.83	(0.00)	175.93	(0.036)

[a] Equations for estimating tetranucleotides are given in the Methods section. G is the likelihood ratio statistic (see Methods) and the subscript denotes degrees of freedom; G measures concordance between observed and expected frequencies. The significance of departure from expected values is denoted by α; $\alpha > 0.05$ indicates a good fit between observed and expected values.

expected frequency of CTAG can be calculated as the product of two dinucleotides:

$$p(CTAG|CT,AG) = p(CT)\cdot p(AG). \quad [2]$$

Third, if trinucleotide frequencies are known, frequency can be calculated as the product of a mononucleotide and a trinucleotide:

$$p(CTAG|CTA,G) = p(CTA)\cdot p(G) \text{ or} \quad [3]$$

$$p(CTAG|C,TAG) = p(C)\cdot p(TAG). \quad [4]$$

Fourth, both the mononucleotide and dinucleotide frequencies can be used, for example:

$$p(CTAG|C,TA,G) = p(C)\cdot p(TA)\cdot p(G) \quad [5]$$

Fifth, a Markov chain rule (11) can be applied whereby the frequency of

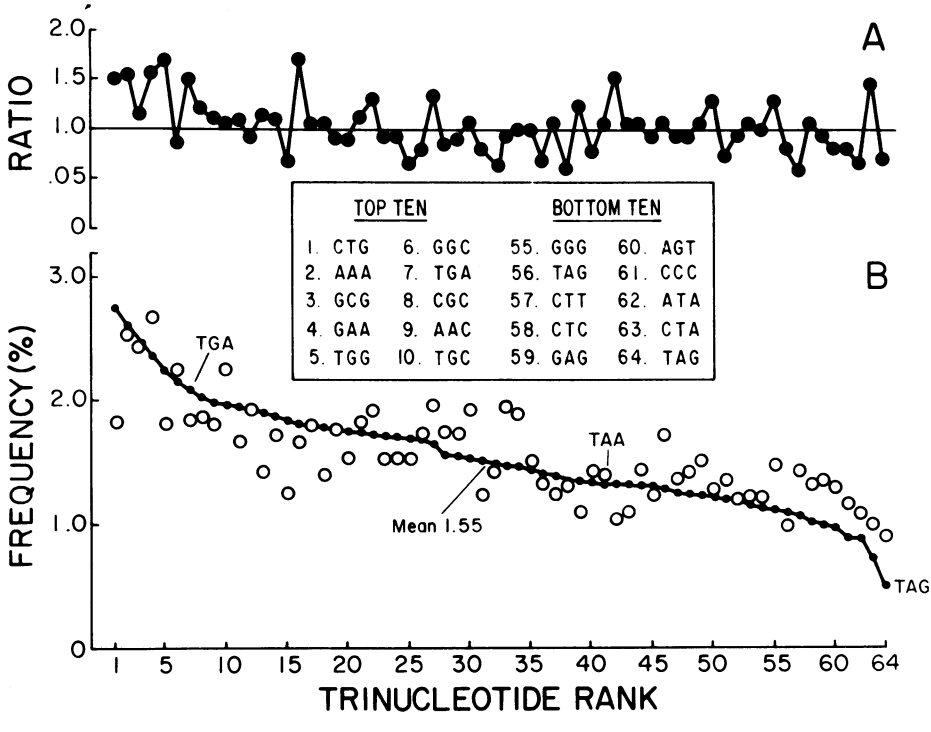

Figure 1. Ordered abundance (B) and complement ratio curves (A) for trinucleotides.

Data were plotted from highest to lowest abundance along with values determined by a 1st order Markov chain (open circles). Ratios of each frequency to its complement's frequency were also plotted immediately above the ordered abundance curve. In this and subsequent figures, the 10 most and least abundant oligomers have been listed in the insert. The rank of the 3 stop codons are also noted in B as is the mean value of abundance for trinucleotides.

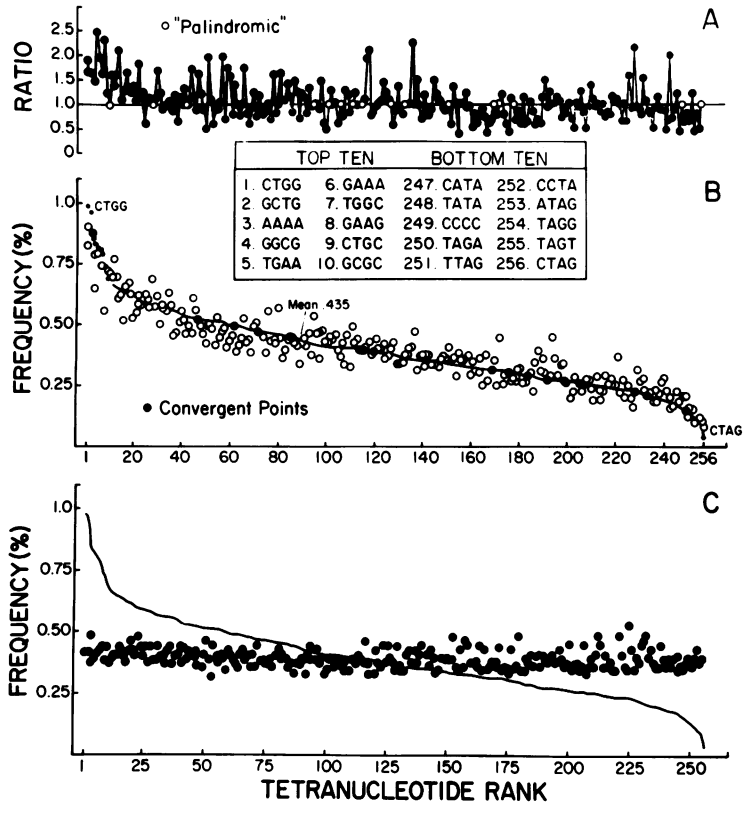

Figure 2. Ordered abundance (B) and complement ratio (A) curves for tetranucleotides.

Data were plotted as described in the legend to Figure 1 with values determined by a 2nd order Markov chain plotted in B as open circles. In C, the ordered abundance curve (solid line) is replotted against the values calculated solely on the basis of mononucleotide frequencies (closed circles). The upper end of the ordered abundance curve in B is represented by small filled circles instead of the smooth line to give a more accurate representation of the values for those abundant tetramers. Also note that "convergent points" denote where the observed and Markov predicted frequencies were the same.

CTAG is estimated from the transition probability of CTA adjacent to TAG at TA dinucleotides:

$$p(CTAG|CTA,TAG) = \frac{p(CTA)\cdot p(TAG)}{p(TA)}. \qquad [6]$$

This is a 2nd order Markov chain and uses transition probabilities between trinucleotides. A 3rd order Markov chain uses transitions between tetranu-

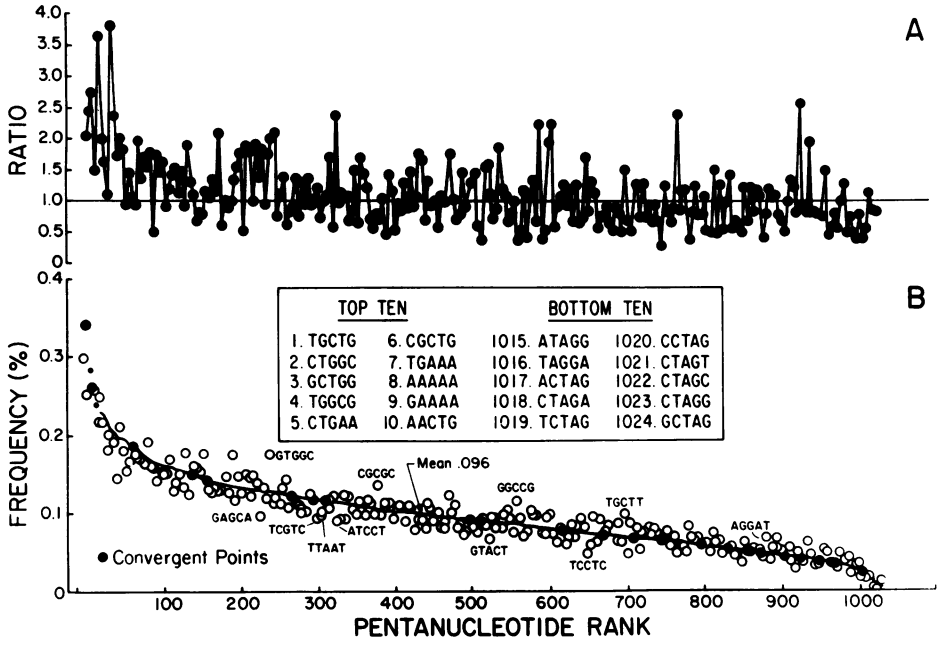

Figure 3. Ordered abundance (B) and complement ratio (A) curves for pentanucleotides.

Frequencies for each 4th pentanucleotide have been plotted as described in the legend to Figure 1. Some of the pentamers not well predicted by a 3rd order Markov chain have been noted on the abundance curve. As in Figure 3, points at the upper end of the curve have not been connected to the line.

cleotides to predict pentanucleotides, providing the frequency of tetranucleotides is known:

$$p(CTAGC|CTAG,TAGC) = \frac{p(CTAG) \cdot p(TAGC)}{p(TAG)}. \quad [7]$$

To evaluate these various methods before analyzing large data sets, several genes of E. coli were analyzed up to tetranucleotides and the results are summarized in Table 1. A 2nd order Markov chain best predicted the occurrence of tetranucleotides on a gene-by-gene basis from di- and trinucleotides. A 1st order Markov rule was generally inadequate in summarizing the composition of trinucleotides in the sequences used, but was better overall than the other methods (data not shown). Although Almagor (12) and Blaisdell (13) have used a Markov chain to predict some oligonucleotide frequencies in eukaryotic DNA and have come to a similar conclusion, their studies were not as extensive as that reported here.

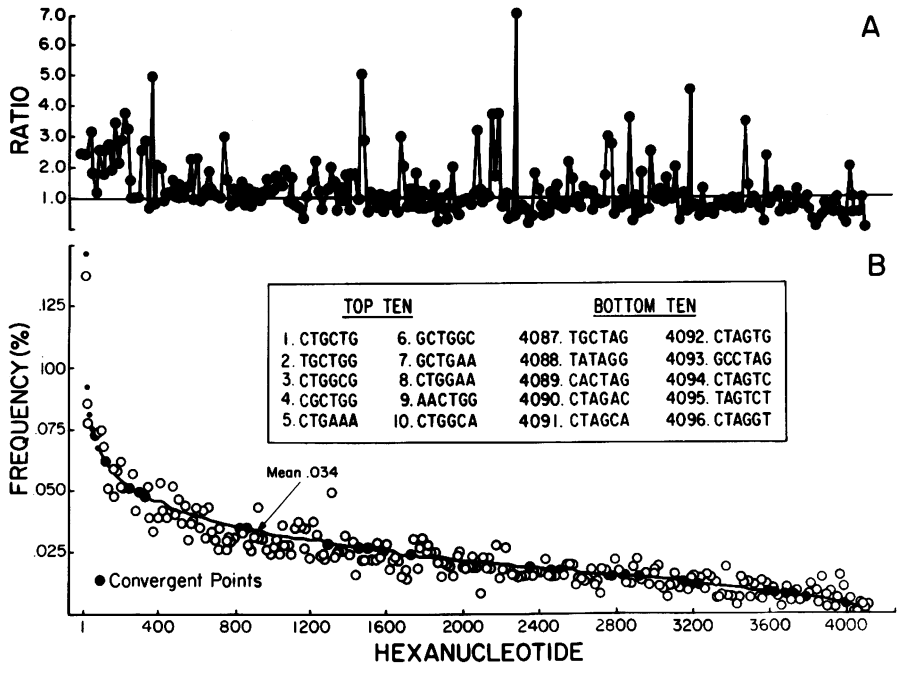

Figure 4. Ordered abundance (B) and complement ratio (A) curves for hexanucleotides.

Frequencies of each 16th hexanucleotide have been plotted as described in the legend to Figure 1. Points at the extreme upper end of the abundance curve have not been connected to the line.

Ordered Abundance Curves and Markov Predicted Values

A more direct measure of the capacity of the Markov chain to predict frequencies can be seen on the ordered abundance plots for tri- through hexanucleotides (Figures 1-4) in which values predicted by the Markov chain rule have also been plotted for comparison. Frequencies calculated on the basis of mononucleotide frequencies, e.g., a purely random model, have also been plotted for tetranucleotides in Figure 2C. All ordered abundance curves have the same overall "S"-shape in that the majority of the sequences fell on a broad linear slope such that immediate neighbors could be predicted with high accuracy. At the high and low end of the curves, a relatively sharp break in slope occurred with a wider range for the highly abundant sequences. This general shape also applied to the di- (not shown) and trinucleotides (Fig. 1). Frequencies calculated on the basis of mononucleotides did not remotely fit the ordered abundance curve for

Table 2: Mononucleotide and dinucleotide frequencies (%) in 74,444 bp of E. coli DNA.

A. Mononucleotide frequencies (%)

T	C	A	G
23.5	24.5	24.6	26.6

B. Dinucleotide frequencies (%)

	T			C			A			G		
1st	ob[a]	ex[b]	(res)[c]	ob	ex	(res)	ob	ex	(res)	ob	ex	(res)
T	6.00	5.52	(+6)	5.55	5.73	(-2)	4.29	5.77	(-18)	7.63	6.26	(+14)
C	5.56	5.73	(-2)	5.38	5.95	(-7)	5.94	5.99	(-1)	7.56	6.51	(+11)
A	5.90	5.77	(+1)	5.61	5.99	(-5)	7.89	6.03	(+20)	5.15	6.55	(-16)
G	6.01	6.26	(-3)	7.89	6.51	(+14)	6.41	6.54	(-1)	6.31	7.14	(-8)

[a] observed frequencies (ob) from E. coli data set of 74,444 bases.
[b] expected frequencies (ex) as predicted from mononucleotide frequencies.
[c] residual value (res) from a likelihood ratio test comparing observed and expected values (degrees of freedom = 9 see text).

tetranucleotides. By contrast, Markov chain values closely followed the overall shape of the curve, even though the fit was imperfect. The fit was much better at the penta- and hexanucleotide levels (Fig. 3 and 4).

The ability of the Markov chain to estimate oligonucleotide frequencies also improved as the sequence length increased as can be seen by analysis of residual values. For example, as noted by others for E. coli coding sequences (1-4) (Table 2B), virtually every dinucleotide occurred at a frequency different from that predicted by mononucleotides or a zero order Markov chain. No dinucleotide had a residual of zero and only five had residuals in the range ±2. Similarly, frequencies of nearly all trimers (Table 3) deviated from those expected by a 1st order Markov chain. Only 8 had residuals of zero; 11 had values of ±1, 27 had values of ±2-6, and the remaining 18 in excess of ±6. However, by the tetramer level, 2nd order Markov chain predicted 177 of 256 tetranucleotides with residuals of ±1 and 211 tetranucleotides with residuals of ±2. A 3rd and 4th order Markov chain was quite accurate in predicting penta- and hexanucleotides frequencies, respectively. For 1024 pentanucleotides, only 8 had residual values larger than ±2 and 65 had residuals of ±2 leaving 941 with residuals of 0 to ±1. Likewise, only 89 of 4096 hexanucleotides had residuals of ±2; 4 had residuals of ±3; and 2 had residuals of -5. Hence, 4,001 hexanucleotides were well predicted having residuals ranging from 0 to ±1.

We conclude, therefore, that the Markov chain is an adequate but not perfect predictor of higher order nucleotide sequences in the E. coli genome. Furthermore, the accuracy of specific frequencies is improved as

Table 3. Observed frequencies (%)[a] of trinucleotides and those expected by a 1st order Markov chain.

3rd:	T			C			A			G		
	ob	ex	(res)	ob	ex	(res)	ob	ex	(res)	ob	ex	(res)
1st-2nd												
TT	1.70	1.53	(+3)	1.49	1.42	(+1)	1.34	1.10	(+6)	1.48	1.95	(9)
TC	1.39	1.26	(+2)	1.20	1.21	(0)	1.40	1.35	(+1)	1.56	1.72	(-3)
TA	1.34	1.03	(+8)	1.10	1.00	(+3)	1.36	1.39	(0)	0.51	0.90	(-12)
TG	1.27	1.42	(-9)	1.98	2.25	(-5)	2.09	1.83	(+5)	2.25	1.80	(+9)
CT	1.07	1.42	(-8)	1.03	1.32	(-7)	0.72	1.01	(-8)	2.75	1.80	(+17)
CC	1.15	1.22	(-1)	0.88	1.16	(-7)	1.35	1.31	(+1)	1.95	1.66	(+6)
CA	1.34	1.43	(-1)	1.20	1.36	(-3)	1.57	1.92	(-7)	1.83	1.25	(+13)
CG	1.88	1.71	(+3)	2.14	2.24	(-1)	1.75	1.82	(-1)	1.78	1.78	(0)
AT	1.47	1.50	(0)	1.80	1.39	(+9)	0.89	1.08	(-5)	1.73	1.91	(-3)
AC	1.23	1.28	(-1)	1.56	1.23	(+7)	1.26	1.37	(-2)	1.57	1.74	(-3)
AA	1.47	1.89	(-8)	1.98	1.81	(+3)	2.60	2.45	(+1)	1.82	1.66	(+3)
AG	0.97	1.30	(-5)	1.72	1.53	(+4)	1.32	1.24	(+1)	1.14	1.24	(-1)
GT	1.77	1.54	(+4)	1.26	1.42	(-4)	1.35	1.10	(+6)	1.65	1.96	(-6)
GC	1.80	1.80	(0)	1.69	1.73	(0)	1.92	1.92	(0)	2.47	2.43	(0)
GA	1.73	1.54	(+4)	1.32	1.46	(-3)	2.37	2.67	(+5)	0.99	1.34	(-8)
GG	1.90	1.42	(+10)	2.04	1.86	(+3)	1.24	1.51	(-6)	1.09	1.48	(-8)

[a] Sequences in 74,444 bases of E. coli; DNA observed frequency (ob) and that expected (ex) by a 1st order Markov chain; residuals (res) by likelihood ratio test (degrees of freedom = 36).

the sequence length increases. Those oligonucleotides that were poorly predicted by Markov chain analysis are discussed in a separate communication (14).

MaeI Digestion of E. coli DNA

It was important to establish whether the oligonucleotide frequencies observed in the 74,444 bp sample of E. coli DNA reflected the frequencies in the 4.2 x 10^6 bp genome. CTAG is the recognition site for MaeI and the rarest tetranucleotide in the analyzed sequences; it occurred 27 times in 74,444 bp whereas the most abundant tetranucleotide, CTGG, occurred 726 times (Fig. 2). Accordingly, E. coli DNA was digested with MaeI, Sau3A, (recognition site: GATC), HaeIII (GGCC) and EcoRI (GAATTC). The frequency of CTAG (0.00036) was approximately 10-fold less than that of GATC (.0041) and GGCC (.0031); this is about the difference in frequency with which DNA is cut by restriction enzymes recognizing 6-base and 4-base sites. Figure 5 shows that E. coli DNA was infrequently cut by MaeI, but extensively cut

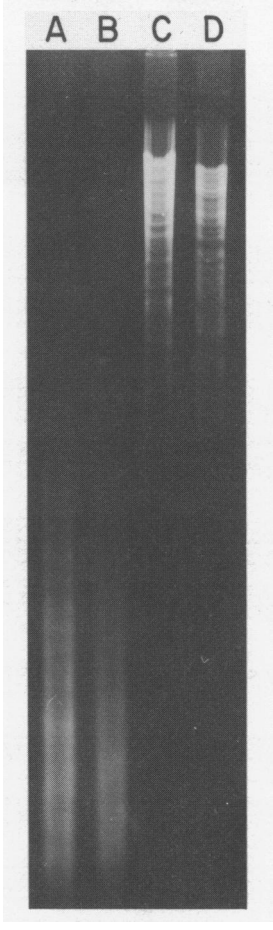

Figure 5. Digestion of E. coli DNA with various restriction enzymes that cut at high and low frequency recognition sites.
Genomic DNA was digested with HaeIII (lane A), Sau3A (lane B), MaeI (lane C), and EcoRI (lane D); 1 μg of each sample was electrophoresed on 0.8% agarose gels and bands detected by EtBr staining.

by Sau3A and HaeIII. In fact, MaeI cut DNA less frequently than did the 6-base recognition enzyme, EcoRI. Thus, frequencies of sequences found in the Genbank sample of 74,444 bp are likely to represent the overall genomic frequencies of those oligonucleotides relatively accurately.

Estimating the Frequency of Longer Oligonucleotides

One immediate application of the results of this analyis involves estimating the frequency of relatively long oligonucleotide sequences. A 3rd order Markov chain relatively accurately predicted most pentanucleotide

Table 4. Observed frequencies (%) and rank[a] of tetranucleotides in 74,444 bases of E. coli DNA.

	T	C	A	G		T	C	A	G
TTT	.508 (53)	.407 (100)	.383 (120)	.399 (107)	ATT	.421 (92)	.322 (159)	.343 (147)	.377 (125)
C	.384 (119)	.387 (110)	.329 (155)	.375 (128)	C	.397 (109)	.380 (123)	.511 (49)	.508 (52)
A	.438 (87)	.359 (132)	.405 (101)	.136 (251)	A	.296 (178)	.183 (241)	.296 (179)	.104 (253)
G	.308 (171)	.462 (75)	.438 (88)	.269 (194)	G	.320 (160)	.413 (96)	.494 (58)	.514 (47)
TCT	.254 (208)	.305 (173)	.169 (246)	.659 (11)	ACT	.230 (225)	.256 (206)	.179 (243)	.565 (30)
C	.286 (182)	.233 (224)	.278 (189)	.405 (102)	C	.325 (157)	.264 (200)	.404 (105)	.559 (32)
A	.293 (181)	.302 (175)	.347 (143)	.460 (77)	A	.329 (154)	.258 (204)	.397 (108)	.278 (188)
G	.388 (116)	.463 (74)	.399 (106)	.312 (166)	G	.389 (115)	.488 (61)	.347 (142)	.348 (140)
TAT	.343 (146)	.494 (59)	.153 (248)	.353 (136)	AAT	.359 (133)	.404 (104)	.285 (186)	.421 (91)
C	.235 (219)	.310 (168)	.237 (217)	.318 (162)	C	.459 (78)	.462 (76)	.504 (56)	.558 (33)
A	.313 (165)	.359 (131)	.443 (85)	.225 (229)	A	.512 (48)	.619 (18)	.844 (3)	.638 (14)
G	.090 (255)	.178 (244)	.138 (250)	.098 (254)	G	.345 (145)	.595 (22)	.504 (57)	.379 (124)
TGT	.475 (69)	.270 (192)	.268 (196)	.273 (190)	AGT	.286 (184)	.181 (242)	.201 (234)	.297 (177)
C	.550 (37)	.480 (67)	.396 (111)	.554 (35)	C	.375 (127)	.357 (134)	.455 (79)	.535 (40)
A	.577 (27)	.446 (84)	.812 (5)	.268 (197)	A	.335 (150)	.234 (223)	.474 (70)	.262 (201)
G	.650 (13)	.775 (7)	.452 (81)	.388 (117)	G	.325 (156)	.377 (126)	.246 (214)	.206 (233)
CTT	.301 (176)	.312 (167)	.234 (220)	.215 (230)	GTT	.466 (71)	.443 (86)	.380 (122)	.485 (65)
C	.256 (205)	.242 (215)	.252 (209)	.269 (193)	C	.349 (138)	.183 (240)	.310 (169)	.411 (97)
A	.225 (228)	.249 (211)	.210 (232)	.036 (256)	A	.385 (118)	.310 (170)	.430 (90)	.226 (227)
G	.409 (98)	.723 (9)	.649 (12)	.971 (1)	G	.249 (212)	.381 (121)	.523 (44)	.508 (51)
CCT	.261 (202)	.190 (238)	.120 (252)	.570 (29)	GCT	.320 (161)	.266 (198)	.252 (210)	.959 (2)
C	.191 (237)	.146 (249)	.199 (236)	.347 (141)	C	.337 (149)	.241 (216)	.475 (68)	.636 (15)
A	.316 (163)	.260 (203)	.269 (195)	.510 (50)	A	.405 (103)	.391 (114)	.542 (39)	.538 (26)
G	.487 (62)	.516 (46)	.419 (93)	.526 (42)	G	.618 (19)	.683 (10)	.587 (24)	.589 (23)
CAT	.371 (129)	.486 (64)	.155 (247)	.332 (151)	GAT	.392 (113)	.413 (95)	.286 (183)	.633 (16)
C	.265 (199)	.432 (89)	.235 (218)	.278 (187)	C	.270 (191)	.348 (139)	.285 (185)	.417 (94)
A	.305 (174)	.486 (63)	.530 (41)	.234 (221)	A	.339 (148)	.519 (45)	.795 (6)	.724 (8)
G	.330 (152)	.586 (25)	.408 (99)	.508 (54)	G	.201 (235)	.364 (130)	.254 (207)	.170 (245)
CGT	.557 (34)	.448 (83)	.395 (112)	.483 (66)	GGT	.455 (80)	.353 (135)	.488 (60)	.609 (20)
C	.524 (43)	.547 (38)	.507 (55)	.573 (28)	C	.349 (137)	.306 (172)	.563 (31)	.815 (4)
A	.466 (73)	.450 (82)	.624 (17)	.211 (231)	A	.347 (144)	.189 (239)	.332 (72)	.248 (213)
G	.609 (21)	.553 (36)	.314 (164)	.294 (180)	G	.324 (158)	.330 (153)	.234 (222)	.229 (226)

[a] rank of each tetranucleotide in the ordered abundance curve shown in Figure 2 is given in parentheses.

Table 5. Observed frequencies (%)[a] of the 64 palindromic hexanucleotides and those predicted by a 3rd order Markov chain.

Sequence (Enzyme)		Frequency ob	Frequency ex	Ratio ob/ex	Reciprocal of Frequency ob	Reciprocal of Frequency ex	Sequence (Enzyme)		Frequency ob	Frequency ex	Ratio ob/ex	Reciprocal of Frequency ob	Reciprocal of Frequency ex
1. TTTAAA	DraI	.0349	.0304	1.15	2,863	3,294	33. ATTAAT		.0269	.0222	1.21	3,722	4,495
2. TTCGAA	AsuII	.0363	.0403	0.900	2,757	2,481	34. ATCGAT	ClaI	.0336	.0337	0.995	2,978	2,964
3. TTATAA		.0161	.0110	1.46	6,204	9,079	35. ATATAT		.0121	.0113	1.07	8,272	8,220
4. TTGCAA		.0161	.0340	0.473	6,204	2,937	36. ATGCAT	NsiI	.00537	.0193	.278	18,611	5,170
5. TCTAGA	XbaI	.00403	.00085	4.76	24,815	118,165	37. ACTAGT	SpeI	.00269	.00066	4.08	37,222	151,927
6. TCCGGA	BspMII	.0148	.0287	0.515	6,768	3,485	38. ACCGGT	Cfr101	.0510	.0594	.859	1,959	1,683
7. TCATGA		.0228	.0256	0.892	4,379	3,908	39. ACATGT	AflIII	.0161	.0207	0.779	2,604	4,834
8. TCGCGA	NruI	.0242	.0465	0.520	4,136	2,151	40. ACGCGT	MluI	.0336	.0513	0.654	2,978	1,949
9. TATATA		.00806	.00584	1.38	12,407	17,114	41. AATATT	SspI	.0309	.0225	1.37	3,237	4,439
10. TACGTA	SnaBI	.0242	.0284	0.852	4,136	3,523	42. AACGTT		.0416	.0389	1.07	2,401	2,572
11. TAATTA		.0134	.0196	0.685	7,444	5,102	43. AAATTT		.0296	.0202	1.46	3,384	4,940
12. TAGCTA		.0121	.0194	0.622	8,272	5,148	44. AAGCTT	HindIII	.0188	.0228	0.825	5,317	4,389
13. TGTACA		.00403	.00465	0.867	24,815	21,516	45. AGTACT	ScaI	.00672	.0113	0.595	14,889	8,862
14. TGCGCA	MstI	.0363	.0393	0.922	2,757	2,543	46. AGCGCT	HaeII	.0175	.0451	0.387	5,726	2,216
15. TGATCA	BclI	.0336	.0337	0.996	2,978	2,967	47. AGATCT	BglII	.0202	.0245	0.822	4,963	4,081
16. TGGCCA	BalI	.0202	.0245	0.822	4,963	4,081	48. AGGCCT	StuI	.0121	.0167	0.726	8,272	6,004
17. CTTAAG	AflII	.0121	.0150	0.805	8,272	6,659	49. GTTAAC	HpaI	.0376	.0234	1.61	2,659	4,281
18. CTCGAG	XhoI	.00269	.0143	0.188	37,222	6,990	50. GTCGAC	SalI	.0255	.0378	0.676	3,918	2,649
19. CTATAG		.00269	.00372	0.722	37,222	26,875	51. GTATAC	AccI	.00940	.00596	1.58	10,635	16,767
20. CTGCAG	PstI	.0296	.0308	0.960	3,384	3,249	52. GTGCAC	ApaLI	.0161	.0180	.895	6,204	5,551
21. CCTAGG	AvrII	.00269	.00052	5.19	37,222	190,882	53. GCTAGC	NheI	.00269	.00125	2.15	37,222	80,047
22. CCCGGG	SmaI	.0148	.0313	0.472	6,768	3,196	54. GCCGGC	NaeI	.00672	.0564	0.119	14,889	1,773
23. CCATGG	NcoI	.0202	.0297	0.678	4,963	3,365	55. GCATGC	SphI	.0108	.0216	0.497	9,306	4,621
24. CCGCGG	SacII	.0202	.0417	0.483	4,963	2,399	56. GCGCGC	BssHII	.0497	.0504	0.986	2,012	1,984
25. CATATG	NdeI	.0121	.0138	0.877	8,272	7,256	57. GATATC	EcoRV	.0349	.0277	1.26	2,863	3,612
26. CACGTG		.00537	.0276	0.195	18,611	3,623	58. GACGTC	AatII	.0202	.0352	0.572	4,963	2,838
27. CAATTG		.0175	.0182	0.762	5,726	5,510	59. GAATTC	EcoRI	.0269	.0113	2.39	3,722	8,884
28. CAGCTG	PvuII	.0564	.0669	0.843	1,772	1,494	60. GAGCTC	SacI	.00537	.0214	0.251	18,611	4,676
29. CGTACG		.0215	.0154	1.40	4,653	6,496	61. GGTACC	KpnI	.0108	.0183	0.586	9,306	5,454
30. CGCGCG		.0416	.0422	0.986	2,401	2,369	62. GGCGCC	NarI	.00806	.0607	0.133	12,407	1,648
31. CGATCG	PvuI	.0309	.0325	.950	3,237	3,074	63. GGATCC	BamHI	.00806	.0254	0.317	12,407	3,939
32. CGGCCG	XmaIII	.00940	.0300	.314	10,635	3,335	64. GGGCCC	ApaI	.00537	.0108	0.499	18,611	9,282

[a] Observed (ob) and expected (ex) frequencies in 74,444 bases.

levels from component tri- and tetranucleotides. By using observed levels of the 64 tri- and 256 tetranucleotides (Tables 3 and 4), the frequency of a much longer sequence can be estimated by a simple calculation. Although greater accuracy can be achieved by using tetra- and pentamers or penta- and hexamers, the current data set was statistically unreliable above tetramers because CTAG-containing pentamers and hexamers were too infrequent. A larger data base will overcome this limitation.

Assume that the frequency of the hexanucleotide CCTAGG were sought. By applying a 4th order Markov chain:

$$p(CCTAGG) = \frac{p(CCTAG)\cdot p(CTAGG)}{p(CTAG)}.$$

Then by applying a 3rd order Markov chain to the two pentanucleotides:

$$p(CCTAGG) = \frac{\left[\frac{p(CCTA)\cdot p(CTAG)}{p(CTA)}\right]\cdot\left[\frac{p(CTAG)\cdot p(TAGG)}{p(TAG)}\right]}{p(CTAG)} = \frac{p(CCTA)\cdot p(CTAG)\cdot p(TAGG)}{p(CTA)\cdot p(TAG)}.$$

Thus, the frequency of the hexanucleotide is obtained multiplying the frequencies of overlapping set of tetranucleotides and dividing by frequencies of two overlapping transition trinucleotides.

Likewise, the frequencies of the decamers, AACCTAGGGT and AACCTGGGGT, each containing the lowest and highest abundant tetranucleotide in E. coli DNA, are:

$$p(AACCTAGGGT) = \frac{p(AACC)\cdot p(ACCT)\cdot p(CCTA)\cdot p(CTAG)\cdot p(TAGG)\cdot p(AGGG)\cdot p(GGGT)}{p(ACC)\cdot p(CCT)\cdot p(CTA)\cdot p(TAG)\cdot p(AGG)\cdot p(GGG)}$$

$$= 5.2 \times 10^{-8}$$

$$p(AACCTGGGGT) = \frac{p(AACC)\cdot p(ACCT)\cdot p(CCTG)\cdot p(CTGG)\cdot p(TGGG)\cdot p(GGGG)\cdot p(GGGT)}{p(ACC)\cdot p(CCT)\cdot p(CTG)\cdot p(TGG)\cdot p(GGG)\cdot p(GGG)}$$

$$= 1.8 \times 10^{-6}$$

Note that these two 10-base sequences differ by only one base yet differ by 353-fold in frequency because of the abundance asymmetries. Thus, it is unlikely that this or any other CTAG-containing decamer would occur in the E. coli genome, whereas 7-8 CTGG-containing decamers would be expected in the coding strand. To estimate a decamer in the noncoding strand as well, the frequency of the decamer's complement must also be calculated to obtain a value for the whole genome. These calculations point out how limiting

the genome size of E. coli is with respect to diversity of sequences much longer than 10 bases, given the observed abundance asymmetries.

Accuracy of a Third Order Markov Chain in Predicting Hexanucleotide Frequencies

To estimate how accurately the foregoing method predicted oligonucleotide sequences, the observed frequencies of the 4,096 hexanucleotides were compared to those expected by application of a 3rd order Markov chain. A sample of the data are given in Table 5 for the 64 possible palindromes, most of which are restriction enzyme sites. Also, the reciprocal of observed frequencies are listed which represents the average spacing in nucleotides between sites. In the total data set, the mean ratio of observed to expected frequencies was 1.03 ± 0.94. Increasing the data set to 213,557 bases did not change the mean nor significantly alter the standard deviation (data not shown). A mean of 1.02 ± 0.93 was also found for the palindromic hexanucleotides.

One of the most practical applications of the method will be to estimate sizes of restriction enzyme fragments during cloning and library construction. In this regard, the relative accuracy of the method is shown by the fact that 46 of the 64 hexanucleotide palindromes (or 72%) gave observed-to-expected ratios within the range 0.5-2.0. In practical terms, this means that the method can predict restriction fragment lengths to within half to twice the expected length 72% of the time. Of 4,096 hexanucleotides, 3,748 (84.9%) had ratios within 0.5-2.0.

We conclude, therefore, that overall the 3rd order Markov chain is a relatively accurate predictor of observed frequencies, at least up to the hexanucleotide level starting from a relatively small data set. Furthermore, increasing the data set does not necessarily improve accuracy, and 50-100 kb of known DNA sequence is probably close to a typical data base for many organisms at present. Greater accuracy can be achieved by using a higher order Markov chain should the data base be large enough to allow it.

Codon Usage and Ordered Abundance Curves

The E. coli genome is largely coding DNA. Hence, codon usage has had a significant influence over the observed di- through hexanucleotide frequencies. However, no one has quantitatively measured the effect. One estimate is shown here by the complement ratio plots. That is, the most abundant oligonucleotides had ratios greater than one while the least abundant oligomers had ratios less than one. Hence, the coding strand is enriched for oligomers in high abundance and the noncoding strand is

enriched for low abundance sequences. In a separate communication, the rank of oligonucleotides in ordered abundance curves and linear regression analysis have been used to measure directly the effect of codon usage on observed frequencies. A Markov chain has also been useful in identifying over- and underabundant oligonucleotides in the E. coli genome whose frequencies cannot be accounted for by codon usage.

ACKNOWLEDGMENTS

The authors thank Barny Whitman for a discussion that got us started, Fred Blattner for his constructive comments on the work, and Suzette Lay for her expert typing of the manuscript, especially the tables. This work was supported largely by a U.S. Army Research Office Training grant (D-AAG29-83-G-0111) to J. A. and partially by an NIH grant (CA-34066) to R.I.

REFERENCES

1. Nussinov, R. (1980) Nucleic Acids Res. 8, 4545-4562.
2. Nussinov, R. (1981) J. Biol. Chem. 256, 8458-8462.
3. Nussinov, R. (1981) J. Mol. Evol. 17, 237-244.
4. Nussinov, R. (1984) J. Mol. Evol. 20, 111-119.
5. Smith, T.F., Waterman, M.S., and Burks, C. (1985) Nucleic Acids Res. 13, 645-656.
6. Elleman, T.C. (1978) J. Mol. Evol. 11, 143-161.
7. Sankoff, D. (1973) J. Mol. Biol. 77, 159-164.
8. Blaisdell, B.E. (1986) Proc. Natl. Acad. Sci. USA. 83, 5155-5159.
9. Arnold, J., Eckenrode, V.K., Lemke, K., Phillips, G.J., and Schaeffer, S.W. (1986) Nucleic Acids Res. 14, 239-254.
10. Anscombe, F.J. (1982) in Computing in Statistical Science Through APL. Ch. 12. Contingency Tables and Pearson-Plackett Distributions. Springer-Verlag, New York.
11. Basawa, I.V., and B.L.S. Prakash-Rao (1980) in Statistical Inference for Stochastic Processes. Ch. 4. Discrete Markov Chains. Academic Press, New York.
12. Almagor, H. (1983) J. Theor. Biol. 104, 633-645.
13. Blaisdell, B.E. (1985) J. Mol. Evol. 21, 278-288.
14. Phillips, G.J., Arnold, J., and Ivarie, R. (1987) Nucleic Acids Res., submitted.