Volume 17 Number 14 1989 Nucleic Acids Research

The sequence asymmetry of the *Escherichia coli* chromosome appears to be independent of strand or function and may be evolutionarily conserved

Allen C.Rogerson

Biology Department, St Lawrence University, Canton, NY 13617, USA

Received March 24, 1989; Revised and Accepted June 6, 1989

ABSTRACT

I have examined potential determinants of the asymmetric distribution of nucleotide sequences in the genome of *Escherichia coli* as cataloged in GenBank release 44. I have used the frequency of occurrence of all possible tetranucleotides in a given sequence catalog or derivative as a comparative measure of asymmetry. The GenBank-cataloged strand and its complement show statistically similar (not complementary) distributions. The distribution is statistically similar in comparisons between the protein coding subset and the total genome, the coding subset and selected non-coding genes, the coding subset and the remainder of the DNA, and the coding subset and stable RNA sequences. I have compared the distribution in the genome of *E. coli* with the distributions found in the cataloged genomes of *Salmonella typhimurium*, *Bacillus subtilis*, and of coliphages lambda and T7. The distribution summed in both strands of the cataloged DNA differs statistically only in comparisons with lytic bacteriophage T7 because only the two strands of T7 show statistically dissimilar distributions. Despite similarities in tetranucleotide distribution, the pattern of codon complementarity in *B. subtilis* is different than that documented for *E. coli*. Thus, sequence asymmetry does not seem related to specific DNA function or to documented similarities or differences in codon bias. The sequence asymmetry of the *E. coli* genome may thus reflect a hitherto unsuspected pattern impressed on both strands of DNA which is or can be packaged into bacterial genomes.

INTRODUCTION

The statistical distribution of bases in the genome of *Escherichia coli* is not random or uniform, but seems to reflect underlying influences. The overall asymmetry has been ascribed to the summation of many possible causes, but the localized genetic information contained in a given region of DNA is typically assumed to be a dominant influence on the distribution of bases within that region (1−4).

For example, within protein coding regions the pattern generated by the unequal utilization of codons (codon bias) has been implicated as a major influence on the overall nucleotide asymmetry (5). There is no *a priori* reason to assume an influence of codon bias on non-protein coding regions. Genes for stable RNAs (predominantly ribosomal and transfer RNAs) contain regions encoding extensive base-paired secondary structure (6,7). The distribution of bases in these regions should be largely determined by the unique structural properties of these regions, and should thus differ from the distribution in protein-coding regions.

Because of base complementarity, there seemed little reason to suspect that the asymmetry in complementary strands should be similar.

There is a testable consequence which follows if the nucleotide sequence asymmetry of the *E. coli* genome is primarily determined by local function. The asymmetry in different functional parts of the genome should be different.

Further, if codon bias is a major determining factor in determining the overall nucleotide asymmetry of the *E. coli* genome (5), there should be an association between codon bias and nucleotide asymmetry. Organisms with similar coding biases should have similar overall nucleotide asymmetries, and those with different codon biases should have different nucleotide asymmetries.

Markov chain analysis has shown that the use of mono-, di- and tri-nucleotide frequencies does not accurately predict the higher-order asymmetries of the genome (2). At a minimum, the use of a third-order Markov chain utilizing a nested set of tri-and tetra-nucleotide frequencies has been shown to be required for accurate reconstruction of higher order asymmetries (2). Thus, as the distribution of tetranucleotides embeds the distribution of mono-, di- and tri-nucleotides, and together with the tri-nucleotide frequencies is a relatively accurate predictor of higher-order asymmetries, I have chosen to concentrate on the distribution of tetranucleotides as an indicator of the asymmetric distribution of nucleotides in general.

I here report my findings from a systematic examination of the tetranucleotide asymmetry of the *E. coli* genome and various definable subsets with reference to aspects of the asymmetries from the genomes of other bacteria (*Salmonella typhimurium*, *Bacillus subtilis*) and two coliphages (lambda and T7) for purposes of comparison.

MATERIALS AND METHODS

1. Comparison of nucleotide asymmetries. To compare the nucleotide sequence asymmetry of two catalogs or sets of nucleotide sequence data I first determined the frequencies of each of the 256 different possible sequentially overlapping tetranucleotides in each catalog (see below) and then plotted the frequencies against each other. If the two catalogs have the same nucleotide sequence asymmetry (exact similarity is unlikely due to statistical variations), the frequencies of each pair of tetranucleotides will be the same. Thus, frequencies of identically asymmetric sequences will plot on a line equidistant between the two axes with a slope of one. If the frequencies in the two catalogs differ, the points will deviate from the line. In similar catalogs, the points will tend to be close to the line; in catalogs with widely differing frequencies the points will deviate widely.

The Pearson product-moment correlation coefficient, 'r,' is a proper measure of the similarity or dissimilarity of two such sets of data. I have therefore used this statistic to measure the deviation of tetranucleotide frequencies from a perfect correlation, and, consequently, as a measure of the difference in nucleotide asymmetry between two sequence catalogs. Because the product-moment correlation coefficient represents a summary of deviations I have presented, in most cases, the graphical data as well.

2. Data and Program Sources. Data from GenBank was either decoded from encoded data supplied on GenBank floppy disk release 44.0 (BBN Laboratories, Inc., Cambridge, MA) or obtained directly from GenBank via BioNet (T7, Lambda). Data was decoded and converted to ASCII format utilizing programs written in Turbo Pascal (Borland, Inc., Scotts Valley, CA). Intermediate processing utilized an ASCII text processor (XYWrite III, XY Quest, Bedford, MA). ASCII genome files were processed, the complementary sequence generated as indicated, and the resultant converted to frequency tables using programs written in Turbo Pascal.

3. Strand referents. The GenBank sequences are written to correspond primarily to the sense strand, i.e., the strand similar to m-RNA or to the encoded stable RNA. However, they do contain antisense or complementary sequences. I have, therefore, when referring

Table 1. Characteristics of the Sequences Analyzed.

Sequence	Number of Nucleotides[1]	Correlation Between Strands[2]	Composition, %[3] A	C	G	T
E. coli, Total Genome:	456207	0.68	25.1	24.5	26.3	24.1
E. coli, Subsets of Genome:						
a: Open Reading Frames	343267	0.61	24.7	24.4	26.8	23.9
b: Non-Coding genes	17903	0.34	26.3	23.9	26.9	22.9
c: Residual (Total -a -b)	95037	0.80	25.7	24.5	24.5	25.2
E. coli Cataloged Stable RNA	7924	0.34	23.4	24.9	31.8	19.9
S. typhimurium	27937	0.78	24.7	25.5	26.6	23.2
B. subtilis	53232	0.69	29.2	20.4	24.3	26.2
Coliphage *lambda*	48494	0.77	25.4	23.4	26.4	24.7
Coliphage *T7*	39928	0.13[4]	27.1	22.6	25.8	24.5

[1]The number of nucleotides in the cataloged strand as analyzed; computer rounding errors may cause slight deviations from actual numbers cataloged.

[2]The correlation coefficient, r, between the tetranucleotide distribution in the cataloged strand and its complement. All are significant, $P < 0.001$, save that for T7.

[3]Composition of the cataloged strand calculated from the data.

[4]Not significant, $P < 0.001$.

to the strands of DNA as cataloged referred to the 'cataloged' strand to mean the primarily sense strand, the strand cataloged in GenBank, and the 'non cataloged' strand to refer to the complement, usually considered the antisense strand. Within defined open reading frames I refer to 'sense' and 'antisense' strands.

4. Tetranucleotide cataloging. To catalog overlapping tetranucleotides of a given length the Pascal program utilized a 'sliding window' four bases long. The 'window' was moved down the sequence one base at a time and the tetranucleotide 'covered' by the 'window' was added to the tally of that particular tetranucleotide, and, when needed, that of the complementary tetranucleotide. Each GenBank sequence was analyzed separately, without concatenation. This process loses some information from the beginning and end of sequences, but does not introduce artificial sequences at the point of concatenation. The classification table started with AAAA. The last position was incremented alphabetically, then the next to last position followed by the last, etc. generating AAAA, AAAC, AAAG, AAAT, AACA, AACC .... TTTG, TTTT. This resulted in the generation of a table listing the 256 tetranucleotides in one column, and their frequency in the adjacent column(s).

Table 2. The 20 Most Frequent and 20 Least Frequent Tetranucleotides from E. coli (cataloged strand only) in this study. [1]

RANK	WHOLE GENBANK CATALOG SEQ.	FREQ. [3]	DEFINED OPEN READING FRAMES SEQ.	FREQ [3]	REMAINDER OF SEQUENCES[2] SEQ.	FREQ.[3]
1	CTGG	91.4	CTGG	111.2	TTTT	101.6
2	AAAA	90.7	GCTG	108.5	AAAA	98.1
3	GCTG	90.0	GGCG	92.6	AAAT	68.8
4	GGCG	80.3	TGAA	88.0	GAAA	68.2
5	TGAA	79.0	AAAA	86.3	TGAA	63.7
6	GAAA	78.5	GAAA	84.5	ATTT	63.3
7	TGGC	72.1	TGGC	82.5	TAAA	61.9
8	GAAG	70.8	GAAG	82.3	TTAA	61.7
9	GCGC	69.9	CTGC	76.1	TTTA	61.2
10	AAAG	66.3	GCGC	75.3	TTAT	60.8
11	CTGA	66.0	TGGT	72.5	GCGC	60.7
12	CTGC	65.7	CTGA	71.5	GGCG	59.6
13	GATG	63.0	AAAG	70.7	GCTG	58.8
14	GCCG	62.0	TCTG	69.8	AAAG	58.8
15	TCTG	61.9	GCCG	68.9	ATAA	58.4
16	AAAC	61.7	GATG	67.7	CGCC	58.4
17	GCAG	61.5	GCGT	67.5	AAAC	58.3
18	TGGT	60.8	GCAG	66.7	CTGG	57.9
19	CAGC	60.6	GGTG	66.6	AATT	57.9
20	GCGT	60.5	CGGT	66.5	TTTC	57.6
237	GTAG	20.0	CTTG	17.3	AGTC	23.5
238	CATA	18.9	CCCA	17.1	GACT	23.4
239	CTAA	18.9	GTAG	17.1	CTCC	23.1
240	TATA	18.7	GGAG	16.8	TTAG	23.0
241	GGAC	18.7	GTCC	16.5	TACT	22.8
242	CCTC	18.6	CTAA	15.7	GTAC	22.5
243	CCCT	18.5	CCCT	14.7	ACTC	22.3
244	ACTA	18.2	TAAG	14.5	ATAG	21.5
245	TAGC	18.0	TAGC	13.8	CTCG	21.5
246	GTCC	17.5	CATA	12.9	CCTC	20.7
247	TCTA	17.3	TATA	12.8	CTAC	19.8
248	GAGG	15.3	TTAG	10.5	GTCC	19.1
249	TTAG	15.1	GAGG	10.1	ACTA	18.2
250	CCCC	14.2	TAGA	8.6	TAGA	17.6
251	TAGA	12.0	CCTA	8.5	GGAC	17.2
252	ATAG	11.9	CCCC	8.1	TAGT	16.9
253	TAGT	10.3	TAGT	6.4	TCTA	16.4
254	CCTA	10.3	TAGG	6.2	TAGG	14.6
255	TAGG	9.3	ATAG	6.2	CCTA	13.2
256	CTAG	2.6	CTAG	1.4	CTAG	4.7

[1] Frequencies are for example only; compare data and analysis in ref. 2
[2] Total minus Open Reading Frames; sum of non-coding genes and remainder in Table 1
[3] Frequency expressed as: (occurrence of nucleotide/total nucleotides)*10,000

5. Analysis of Tetranucleotide Frequencies. Final analysis of the sorted sequences was carried out using Lotus 123 (Lotus Development Corporation, Cambridge, MA). Frequencies are reported as the number of occurrences of a given sequence divided by

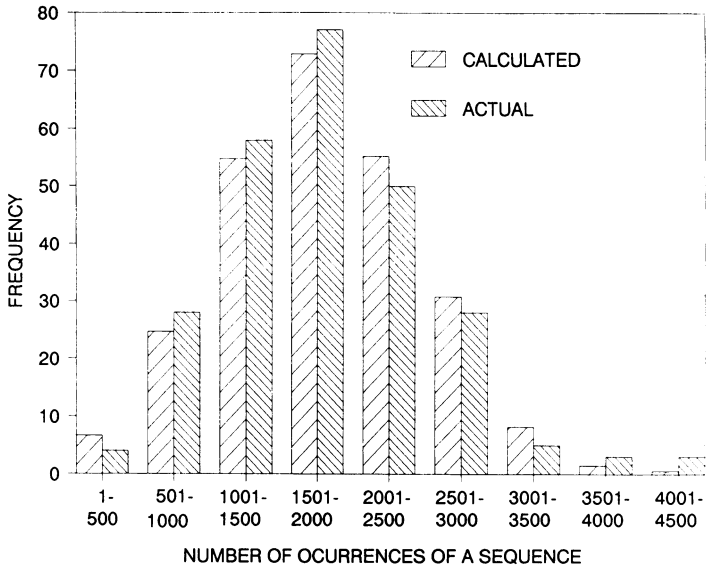

Figure 1. A histogram plotting frequency of occurrence of tetranucleotides (abscissa) against the number of occurrences (ACTUAL) and comparing that distribution to the expectation of normally distributed data (CALCULATED).

the total number of sequences multiplied by 10,000, an arbitrary factor yielding decimal results in the range 1 – 1000. Statistical routines were based on those presented by Sokal and Rolf (8). Table 1 shows the nucleotide composition of the sequences analyzed, the correlation coefficients (r) between the tetranucleotide frequencies in the two strands (see below), and the nucleotide composition as calculated from the GenBank data. Table 2 exemplifies the most frequent and least frequent tetranucleotides from portions of some of the *E. coli* data sets examined in this study. Similar data is analyzed in more detail in reference 2.

6. Random nucleotide sequences. To examine randomly selected, non-overlapping tetramers, the 'window' was moved randomly between 5 and 10 nucleotides down the sequence being analyzed. The tetranucleotides thus selected were alternately cataloged into two data sets, so that there were two data sets for comparison.

7. Generation of a Table with Randomized Tetranucleotide Frequencies. To generate a tetranucleotide frequency table similar to that of *E. coli* but as if from a randomized sequence, I assigned each tetranucleotide frequency derived from the cataloged strand of *E. coli* a random number, and then sorted the frequency column, but not the tetranucleotide identification column, according to the order of the random numbers. The tetranucleotide identification column was then assigned to the now randomized frequencies, and a table of complements was used to generate the frequencies of the opposite strand. This procedure generates a table with the same frequency distribution as that from *E. coli*, but with random frequencies assigned to each tetranucleotide.

8. Analysis within defined open reading frames. To determine sequences within open reading frames, the GenBank Annotation File was decoded for the occurrence of open reading frames in the cataloged strand only. The DNA was examined for the occurrence of an

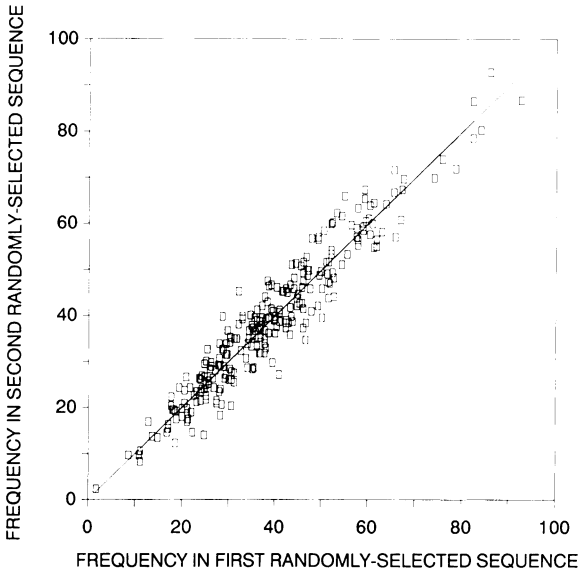

Figure 2. Comparison of the frequencies of two randomly-selected tetranucleotide sets. There are 256 datapoints, each represents one possible tetranucleotide. The frequency (as number in class divided by the total number of nucleotides multiplied by 10,000) from the first randomly selected set is plotted against the frequency from the second randomly selected set. The line is a referent and represents the expectation for a correlation of +1.

ATG, GTG or TTG to be sure that the reading frame began in the correct frame (some continue out-of-phase at the nucleotide sequence start). Sequences were then cataloged from the confirmed start to the GenBank-cataloged terminator. Sequences with more than one possible initiator and/or terminator were read once between the most distal signals.

RESULTS

1. The Distribution of Tetranucleotides is Normal. Proper interpretation of the significance of product-moment correlation coefficients demands that the variables have a normal distribution (8). Figure 1 is a histogram plotting tetranucleotide frequency classes from the *E. coli* genome cataloged in GenBank against number of occurrences. Superimposed on this plot is the intrinsic expectation of a normal distribution of this data. By the Kolmogorov–Smirnov test for goodness of fit to an intrinsic hypothesis (8), the data fits a normal distribution, $P < 0.01$. Thus, parametric statistical methods can be applied to this data set. Other measures and analysis of the statistical distribution of very similar data have previously been presented (2,5).

2. Selecting overlapping short sequences does not seem to generate artifacts due to overlap. Selection of sequentially overlapping tetranucleotide sequences might bias the resulting frequencies because of sequence intersection. For example, the frequency of the tetranucleotide TTTT might be influenced by the frequency of the pentanucleotide TTTTT. To confirm that the selection of overlapping sequences did not introduce an artifact I selected non-overlapping sequences in two data sets (see *Materials and Methods*). Figure 2 shows that these two data sets were similar. The correlation coefficient between the sets is 0.96 indicating similarity, $P < 0.001$. They were also similar to the overlapping data set minus

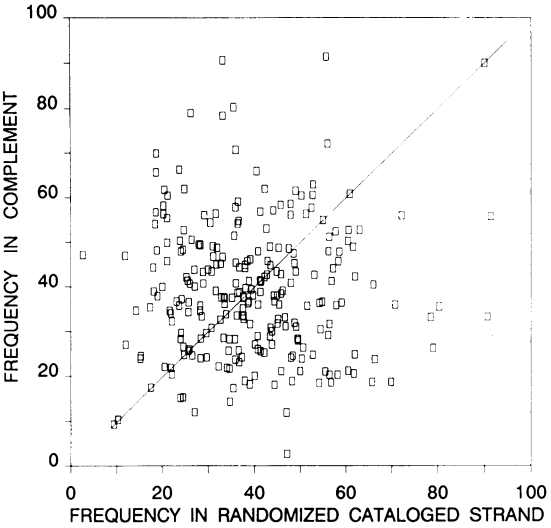

Figure 3. The expectation of the distribution of complementary strands of DNA. The figure was generated by randomizing the actual distribution of the tetranucleotide frequencies from *E. coli*, and generating the frequencies in the opposite strand from the randomized first strand (see text for more details). The frequency of sequences from the randomized, cataloged strand is on the abscissa, from its complement, on the ordinate. There are 256 data points, each representing the two frequencies of one tetranucleotide. See legend to Figure 2 for other details.

the randomly selected frequencies, $P < 0.001$, with correlation coefficients of 0.97 and 0.98. Thus, the strategy of selecting all possible overlapping tetranucleotides yields essentially the same distribution as a non-overlapping, randomized sampling, and the use of overlapping sequences does not in itself generate a systematic asymmetry. Further, examination of Figure 2 gives some idea of the differences expected when comparing two data sets with nearly identical distributions.

3. Comparison of the Two Strands of E. coli DNA. My first comparison was between the asymmetry in the GenBank-cataloged *E. coli* sequences and in the strand complementary to those sequences. There was little reason to believe that the two complementary strands would have similar nucleotide sequence asymmetry except for an indication of similarity in the distribution of codons and their complements (13). It seemed necessary to generate some sort of expectation of an artificial sequence as a standard by which to judge the actual data. A sequence generated from the probability of occurrence of the individual nucleotides has been previously shown not to model the actual asymmetric distribution of nucleotides in the *E. coli* genome (2). Therefore, I utilized the table of the distribution of bases from *E. coli* as a starting point to generate a random assortment of tetranucleotides (see Materials and Methods).

Figure 3 shows the comparison of such a randomized sequence and its complement. The correlation coefficient is 0.05, not significant, $P < 0.001$. Other passes at randomization produce similar results, with correlation coefficients of the same magnitude. Thus, the expectation was that the two strands of *E. coli* DNA should have different asymmetries.

That this is not the case in the actual, unrandomized genome of *E. coli* is shown graphically in Figure 4 where it can be seen that there is a positive correlation between the tetranucleotide sequences of the two strands. The correlation coefficient is +0.68,

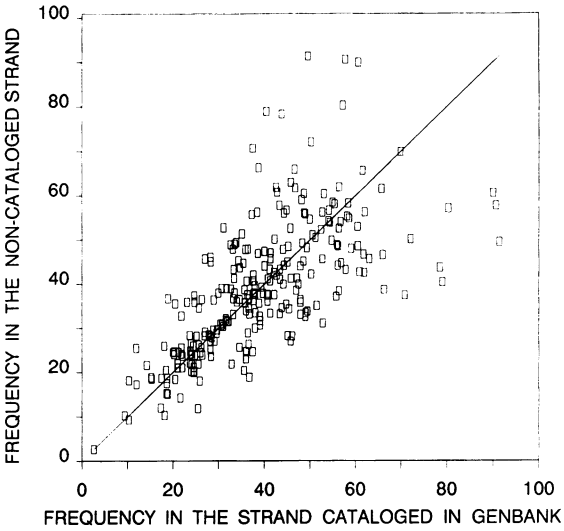

Figure 4. The distribution of tetranucleotide frequencies from the cataloged genome of *E. coli*. The frequency of sequences from the cataloged strand is on the abscissa, from its complement, on the ordinate. There are 256 data points, each representing the two frequencies of one tetranucleotide. The central line is a referent showing the expectation if the two strands were perfectly identical. See legend to Figure 2 for other details.

demonstrating similarity between the tetranucleotide composition of each strand, $P < 0.001$.

Examination of the points of Figure 4 revealed a symmetry around the line representing a correlation of +1 (the central line of the figure, the 'line of perfect correlation'). This symmetry is a result of the complementarity of the two compared strands. Any sequence in one strand must be represented by a complementary sequence in the opposite strand. A consequence is that any palindromic sequences fall on the 'line of perfect correlation,' and may therefore tend to inflate the correlation coefficient. However, removing the palindromes reduces the correlation coefficient by only a small amount, from 0.68 to 0.66 (238 degrees of freedom), and thus the contribution of palindromes is small.

4. Protein-coding Sequences from E. coli. The analyses thus far presented are of the entire *E. coli* genomic data set as represented in GenBank release 44. Most of this data is for coding DNA, but because of the heterogeneity of DNA functions in the data set it is impossible to assign a precise relationship between nucleotide asymmetry and specific DNA function. I therefore examined a subset of the data set whose function was solely to code for proteins. I utilized the GenBank annotation table to define sections of the data precisely within coding regions, starting with the first defined initiation codon and proceeding, in reading frame, to the most distal defined terminator. The reading frame of coding fragments (sequences without an initiator and/or terminator) could not be confirmed by these criteria, and they were therefore not included. Because there are only a small number of genes with coding oriented in the non-cataloged strand I selected only genes whose coding strand corresponded to the cataloged DNA strand.

I report here and below two comparisons between sequence sets. I report the comparison between the tetranucleotide distribution in the cataloged strands, and the comparison between the aggregate tetranucleotide distributions of both strands. I note that a comparison of the

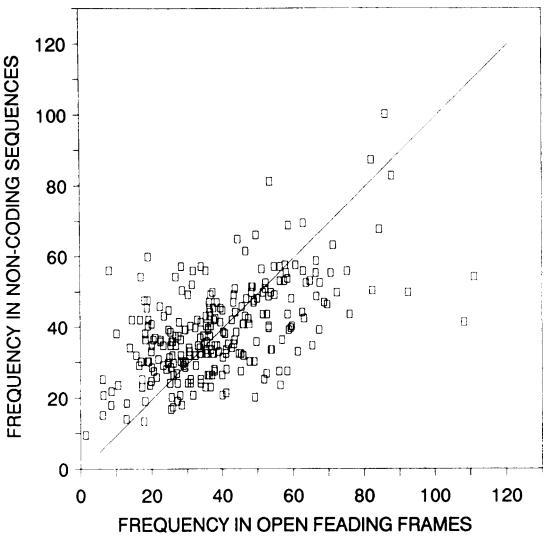

Figure 5. The frequency of tetranucleotides from the defined open reading frames from the cataloged strand of the *E. coli* genome (abscissa) versus the frequency of tetranucleotides from selected non-coding regions from the cataloged strand. See text, Table 4 and legend to Figure 2 for other details.

distributions in the non-cataloged strands is a mirror image of comparisons of the cataloged strands, and thus produces exactly similar coefficients of correlation.

The coding subset showed a significant correlation ($r = 0.97$, $P < 0.001$) with the parental (total) set in both comparisons. It also had a high correlation between its two strands ($r = 0.61$). Graphically, the comparisons are very similar to those shown in Figures 2 and 4. These observations were hardly unexpected given that the defined coding subset was 75% of the cataloged sequences, and thus dominates the genomic data set. I have used the defined coding subset as a benchmark for other comparisons with sequences derived from *E. coli*.

5. Sequences Without Coding Function from E. coli. It was not possible to automatically select DNA sequences without any coding function as such sequences are not systematically annotated. The remainder of the DNA after removing the coding subset represented an assortment of partial coding frames, structural RNA sequences, and other functions, to say nothing of areas of undefined or no function. I therefore manually selected a subset of genes without protein coding function. The selected sequences are shown in table 5. This set showed nucleotide asymmetry correlating significantly with the benchmark coding set (Tables 1 and 3.) Figure 5 shows that the correlation is still obvious, although skewed. Both strands of the set correlated well with each other ($r = 0.34$, significant, $P < 0.001$).

6. Remaining Sequences from E. coli. The tetranucleotide asymmetry of the undefined remainder of the data set, the total set minus the coding set minus the manually-selected non-coding set also correlated well ($P < 0.001$) with the coding set ($r = 0.81$, comparison of cataloged strands, and $r = 0.87$,comparison of both strands, Table 3), and showed a very high correlation ($P < 0.001$) between the two strands ($r = 0.80$, Table 1). Thus, the tetranucleotide asymmetry of the *E. coli* genome seems to be relatively invariant throughout definable subsets of the GenBank catalog.

Table 3. Correlation Coefficients between Tetranucleotide Frequencies from E. coli Open Reading Frames and other E. coli Tetranucleotide Data Sets:

Open Reading Frames With:	Correlation (r) Between Cataloged Strands	Correlation (r) Between Both Strands
Entire E. coli Genome	0.97	0.97
Selected E. coli Non-coding Sequences:	0.63	0.67
Residual (Total *minus* Open Reading Frames *minus* Non-coding)	0.81	0.87
Structural RNA Data Set	0.24	0.22

Table 4. Sequences selected from the Genomic Catalog of E. coli as having no open reading frames.

Sequence(s)	Length, bases
part of rRNA operon rrnA	463
part of rRNA operon rrnB	7508
parts of rRNA operon rrnD	1425
part of rrnF or G for asp-tRNA-1, 23S and 5S rRNA	600
parts of rrnX coding for ile-tRNA-1, ala-tRNA-1B, 5S RNA,asp-tRNA, trp-tRN, 23S - 5S spacer, 545bp more	1023
asp-tRNA and trp-tRNA genes	200
pheU gene	312
leuV gene	699
supB-E gene cluster	1100
ser-tRNA-2am gene	150
ser-tRNA-2 sup60D gene	287
ser-tRNA-2 supH gene	150
tRNA operon for arg, his, leu, pro	646
trp-tRNA Su+7 gene	149
tyrG promoter region	317
tyrT locus	1949

Table 5. Sequences in the E. coli structural RNA Catalog

Sequence	Length,bases
RNA component of RNAase P	375
5S RNA sequences (multiple)	440
16S RNA sequences (multiple)	1874
23S rRNA sequences (multiple)	2290
ala-tRNA-1a,1b	152
cys-tRNA	74
asp-tRNA-1	77
glu-tRNA-1,2	152
phe-tRNA	76
gly-tRNA-1,2,3 and mutations	225
his-tRNA-1	77
ile-tRNA-1,2	153
lys-tRNA	76
leu-tRNA-1,2,5	261
initiator met-tRNA-f	77
met-tRNA-m	77
asn-tRNA	76
gln-tRNA-1,2	150
arg-tRNA-1,2	153
ser-tRNA-1,2,3,I,V	447
amber supressor ser-tRNA-am	90
thr-tRNA-ggt	76
val-tRNA-1,2a,2b	230
trp-tRNA also ts and uga supressor	76
tyr-tRNA-1,2	170

7. Structural RNA Sequences from E. coli. If the asymmetry is truly uniform throughout the genome of *E. coli*, then it should also be represented in independently cataloged segments of the genome. A stringent test thus becomes examination of the asymmetry of the *E. coli* structural RNA cataloged in the structural RNA catalog, a section of the GenBank sequences distinct from the genomic DNA sequences. These sequences are identified in Table 5. These gene fragments representing structural RNAs also show significant correlations with the tetranucleotide asymmetry of the coding DNA ($r = 0.24$, cataloged strand, and $r = 0.22$, both strands, significant, $P < 0.001$, Table 3). Figure 6 shows that his data set is also skewed from the reference, coding set. The tetranucleotide distribution in the non-cataloged strand also correlates significantly with that in the cataloged strand ($r = 0.34$, Table 1). This confirms that the tetranucleotide asymmetry seems to

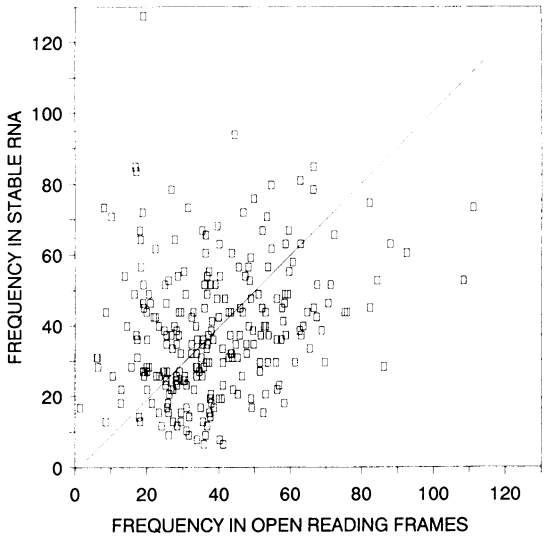

Figure 6. Abscissa, see Figure 5. Ordinate, the frequency of tetranucleotides from DNA corresponding to stable RNA. See text, Table 5 and legend to Figure 2 for other details.

be derived from the same distribution pattern in various subsets of the genome, that is, the asymmetry does not seem to vary significantly with DNA function.

8. Comparison of E. coli sequences with Coliphage sequences. To further explore the distribution of the asymmetry I compared the tetranucleotide asymmetry of the entire *E. coli* genomic data set with that of other organisms. Table 6 summarizes the comparisons with other organisms. I first compared the asymmetry of *E. coli* with that of two coliphages, the temperate phage lambda and the lytic coliphage T7. The codon biases of both phages seem to be similar to the codon bias of *E. coli* (9,10). The tetranucleotide asymmetry of the temperate phage is similar in both strands ($r = 0.77$, significant, $P < 0.001$), and is otherwise very similar to that of *E. coli* (Table 6 and Figure 7). However, while the

Table 6. Correlation Coefficients between Tetranucleotide Frequencies from E. coli and Those From Other Organisms:

E. coli with:	Correlation (R) Between Cataloged Strands	Correlation (R) Between Both Strands
S. typhimurium	0.88	0.89
B. subtilis	0.53	0.52
Coliphage *lambda*	0.85	0.87
Coliphage *T7*	0.22	0.05 [1]

[1] Not significant, $P < 0.001$

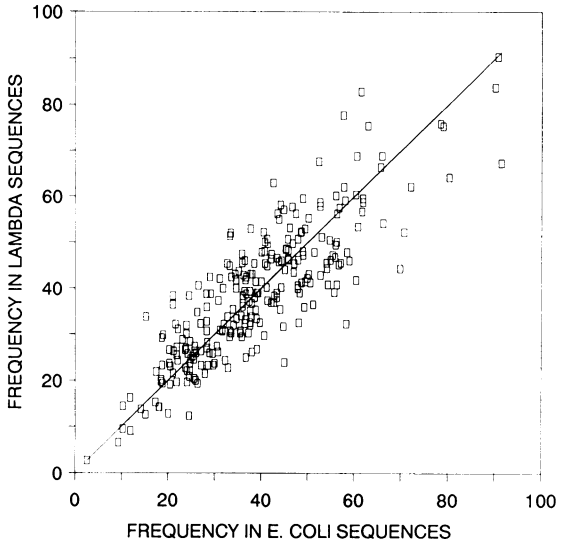

Figure 7. The frequency of the tetranucleotides from the cataloged strand of *E. coli* (abscissa) versus those from the cataloged strand of coliphage lambda. See text and legend to Figure 2 for other details.

tetranucleotide distribution of the cataloged strand of the lytic phage correlates significantly ($r = 0.22$, $P < 0.001$, Figure 8) with that of the cataloged strand of *E. coli* it is the lowest correlation between cataloged strands reported here (compare Figures 3 and 8), and may be due solely to the similarity in codon bias. Otherwise, the distributions from

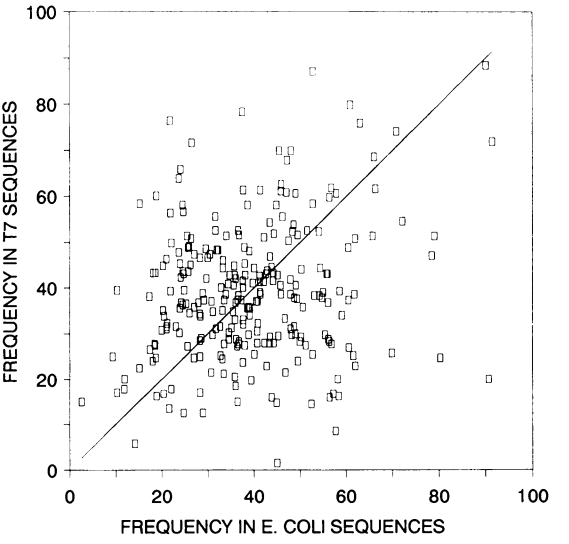

Figure 8. The frequency of the tetranucleotides from the cataloged strand of *E. coli* (abscissa) versus those from the cataloged strand of coliphage T7. See text and legend to Figure 2 for other details.

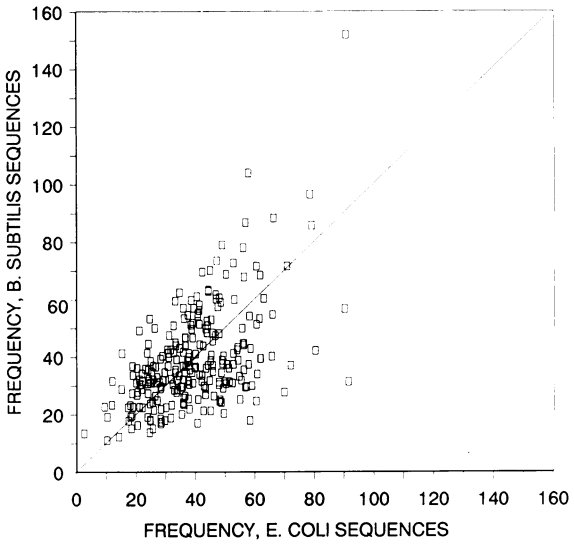

Figure 9. The frequency of the tetranucleotides from the cataloged strand of *E. coli* (abscissa) versus those from the cataloged strand of the gram positive bacterium *B. subtilis*. See text and legend to Figure 2 for other details.

T7 are not similar to those of *E. coli*. The distribution in the two strands of T7 are not alike ($r = 0.13$, not significant, Table 1), and the correlation between the distribution in both strands of T7 and that of both strands of *E. coli* is not significant ($r = 0.05$), unlike all other comparisons reported here.

9. Comparison of E. coli with S. typhimurium. *S. typhimurium* is a gram negative enteric bacterium closely related to *E. coli* by rRNA comparison (11). The tetranucleotide distribution pattern of the cataloged sequences of the two organisms is also similar ($r = 0.88$, cataloged strand, and 0.89, both strands, Table 6).

10. Comparison of E. coli with Bacillus subtilis. *B. subtilis* is a gram-positive bacterium not closely related to *E. coli* by rRNA homology (11), and with a codon usage pattern much more uniform than that of *E. coli* (12). Comparison of the tetranucleotide asymmetry of *B. subtilis* with that of *E. coli* also reveals significant similarities ($r = 0.53$ and $r = 0.52$, Table 6, cataloged strands compared in figure 9) and a similarity between the two DNA strands of the cataloged *B. subtilis* genome ($r = 0.69$).

11. Codon and Codon Complement Distribution in B. subtilis. The distribution of codons in *E. coli* and the complements to the codons shows a significant similarity (13). To comparatively investigate this similarity I examined the distribution of codons and their complements in *B. subtilis*. I generated the codon bias of *B. subtilis* coding regions from the GenBank data (essentially identical to the codon bias reported in 12), and compared their distribution to that of their complements. The distribution of the complements to the codons in open reading frames of the cataloged *B. subtilis* genome do not correlate significantly ($r = 0.24$, not significant assuming the codon frequencies are parametrically distributed, $P < 0.01$, see Figure 10) with the distribution of codons. Thus, this particular feature of the *E. coli* coding bias is difficult to explain as either the obligatory cause or consequence of the similarity of the tetranucleotide asymmetries of the two strands. Further,

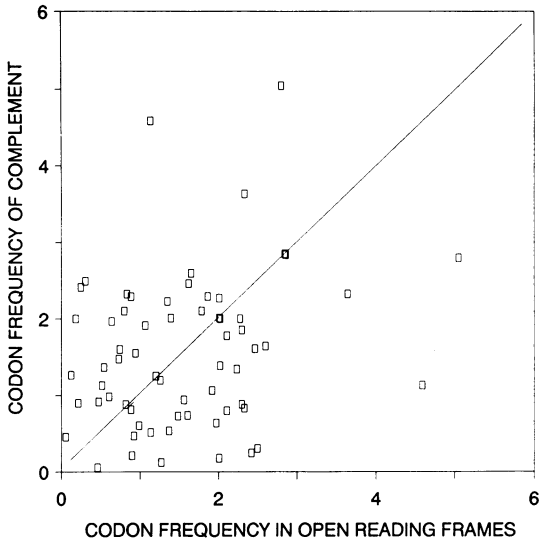

Figure 10. The frequency of the codons in *B. subtilis* versus the frequency of their complements. Frequency is in simple percentage. See text and reference 13.

reportedly different codon biases reside within statistically similar tetranucleotide asymmetries, again demonstrating a lack of correlation between codon bias and tetranucleotide asymmetry.

DISCUSSION

The analyses presented above reveal an unanticipated sameness in the tetranucleotide asymmetry of functionally different subsets of the GenBank 44.0-cataloged genome of *E. coli*, a similarity which encompasses both strands of the DNA and extends to the asymmetries of the cataloged genomes of *S. typhimurium*, *B. subtilis* and temperate coliphage lambda, but not to lytic coliphage T7.

It is important to reiterate that the asymmetric distribution of tetranucleotides is a good measure of other, higher-order asymmetries as revealed by Markov Chain analysis (2) as well as by selected comparisons of the sequences examined here at the penta- and hexa-nucleotide levels (data not reported here). Thus, these findings are probably good indicators of nucleotide sequence asymmetry patterns in general.

Previous work had suggested a correlation between the coding bias of *E. coli* and the higher-order asymmetries (5). This work also shows that the asymmetry exists in coding regions, and that therefore a correlation can be drawn between codon bias and nucleotide asymmetry. However, this contribution shows that it cannot be concluded that the correlation implies that codon bias directly causes nucleotide asymmetry, as the general pattern of asymmetry is not restricted to coding regions.

Moreover, there seems to be little general association between the codon bias and the tetranucleotide asymmetry of a particular genome. The tetranucleotide asymmetry pattern of temperate coliphage lambda is similar to that of *E. coli*; the pattern of lytic coliphage T7 is not. Both phages have codon utilization patterns like that of *E. coli* (the lambda

comparison is not published but can be reproduced by comparing data in refs. 9 and 10).

In contrast, the tetranucleotide asymmetry of *B. subtilis* is similar to that of *E. coli* whereas the codon utilization pattern is quite different. I show here, moreover, that *B. subtilis* does not share an unusual property of the *E. coli* codon set, which is that the frequency distribution of the complements of the codons correlate well with the frequency distribution of the codons. Here, then, similarity in tetranucleotide distribution does not associate with similarities in codon bias. Thus, there does not seem to be a simple causal relationship between nucleotide asymmetry and codon bias.

Similarity does not presage identity, and it should be obvious that there are differences in distribution of nucleotides within different functional parts of the genome, although these differences are within the background of a high overall correlation. For example, examination of the comparison of the two strands of *E. coli* (Figure 4) reveals divergence of the frequencies of some of the more highly represented frequencies, a divergence further detailed by examining frequencies of exemplary tetranucleotides and their complementary sequences (or lack of a complement of similar frequency) in Table 2 or in tables in reference 2. Non-coding (Figure 5) and structural sequences (Figure 6) also deviate in their frequency distributions from those of the coding sequences in ways demanding further analysis. The shift in frequency between *E. coli* and *B. subtilis* (Figure 9) also reveals intriguing differences.

It should be noted that generating an artificial DNA sequence from a given codon bias will reproduce the tetranucleotide asymmetry of the genome from which the codon bias was originally derived (not shown). However, the codon bias information is not a parametric data set, but is rather a sampling of non-overlapping short sequences with their frequencies of occurrence. Reconstruction of a sequence in this manner is analogous to the construction of a Markov chain, but utilizing actual short sequences and their frequencies instead of overlapping, statistical frequencies.

This analysis shows that the tetranucleotide asymmetry of the *E. coli* genome seems to be a consistent feature of both strands of the DNA and of all functional subsets of the genome, and a related asymmetry is found in the genomes of *S. typhimurium* and *B. subtilis*. A related asymmetry also seems to be a feature of the genome of the temperate coliphage lambda, but not of the lytic coliphage T7. The unifying factor from the factors examined here thus seems to be whether or not the DNA can be found as an integral part of a bacterial genome. The asymmetry seems to be conserved enough to exist in a statistically related form in both gram negative and gram positive bacteria, and hence to appear conserved through bacterial evolution. Its biological basis may thus be the result of a hitherto undescribed interaction between bacterial DNA and the bacterial cytoplasm. The implications of such a background motif in chromosomal DNA for theories of evolution and for the analysis of patterns could be profound. We are working to broaden the comparative analysis.

ACKNOWLEDGEMENTS

I thank K. McKnight, L. Lindahl, J. Zengel, N. Smith, C. Wu, R. Selander, D. Daniels and P. Hope for suggestions, help and advice. GenBank was purchased with a grant from Dean A. Rembert. This work was facilitated by a leave in the laboratory of L. Lindahl at the University of Rochester.

REFERENCES

1. Bernardi,G. and Bernardi,G. (1986) J. Mol. Evol. **24**, 1–11
2. Philips,G.J., Arnold,J. and Ivarie,R. (1987) Nucleic Acids Res. **15**, 2611–2626
3. Yarus,M. and Folley,L.S. (1985) J Mol. Biol. **182**, 529–540
4. McClelland,M., Jones,R., Patel,Y. and Nelson,M. (1987) Nucleic Acids Res. **15**, 5985–6005
5. Philips,G.J., Arnold,J. and Ivarie,R. (1987) Nucleic Acids Res. **15**, 2627–2638
6. Woese,C.R., Gutell,R., Gupta,R. and Noller,H.F. (1983) Microbiological Revs. **47**, 621–669
7. Staden,R.(1980) Nucleic Acids Res. **8**, 817–825
8. Sokal,R.R. and Rohlf,F.J. (1981) **Biometry**, 2ed. W.H. Freeman, NY, NY.
9. Sharp,P.M.,Rogers,M.S., and McConnell,D.J. (1985) J. Mol. Evol. **21**, 150–160
10. Daniels,D.L., Sanger,F. and Coulson,A.R. (1983) Cold Spring Harbor Symp. Quant. Biol. **XLVII**, 1009–1024
11. Woese,C.R. (1987) Microbiological Revs. **51**, 221–271
12. Ogasawara,N. (1985) Gene **40**, 145–150
13. Alff-Steinberger,C. (1984) Nucleic Acids Res. **12**, 2235–2241

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.