Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 3699–3703, May 1989
Genetics

Nonrandom utilization of codon pairs in *Escherichia coli*

GEORGE A. GUTMAN AND G. WESLEY HATFIELD

Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, CA 92717

Communicated by Masayasu Nomura, January 9, 1989 (received for review October 3, 1988)

ABSTRACT **We have analyzed protein-coding sequences of *Escherichia coli* and find that codon-pair utilization is highly biased, reflecting overrepresentation or underrepresentation of many pairs compared with their random expectations. This effect is over and above that contributed by nonrandomness in the use of amino acid pairs, which itself is highly evident; it is much weaker when nonadjacent codon pairs are examined and virtually disappears when pairs separated by two or three intervening codons are evaluated. There appears to be a high degree of directionality in this bias: any codon that participates in many nonrandom pairs tends to make both over- and underrepresented pairs, but preferentially as a left- or right-hand member. We show a relationship between codon-pair utilization patterns and levels of gene expression: genes encoding proteins expressed at high levels tend to contain more abundant, but more highly underrepresented, codon pairs, relative to genes expressed at low levels. The nonrandom utilization of codon pairs may be a consequence of their effects on translational efficiency, which in turn may be related to the compatibility of adjacent aminoacyl-tRNA isoacceptors at the A and P sites of a translating ribosome.**

The protein-coding regions of genes in all organisms are subject to a wide variety of functional constraints, some of which depend on the requirement for encoding a properly functioning protein, as well as appropriate translational start and stop signals. However, several features of protein-coding regions have been discerned that are not readily understood in terms of these constraints; two important classes of such features are those involving codon usage and codon context.

It has been known for a considerable time that codon utilization is highly biased and varies considerably among various organisms (1–3). Codon-usage patterns have been shown to be related to the relative abundance of tRNA isoacceptors (4–6) and genes encoding proteins of high versus low expression show differences in their codon preferences (6, 7). The possibility that biases in codon usage can alter peptide elongation rates has been widely discussed, although direct effects of codon choice on translation have been difficult to demonstrate (8, 9). Other constraints on codon-usage patterns have been proposed, including optimization of the fidelity (10, 11) and kinetic efficiency (12) of translation.

Apart from the nonrandom use of codons, considerable evidence has accumulated that codon/anticodon recognition is influenced by sequences outside the codon itself, a phenomenon termed codon context. There exists a strong influence of nearby nucleotides on the efficiency of suppression of nonsense codons (see refs. 13–15) as well as missense codons (16). Clearly, the abundance of suppressor activity in natural bacterial populations (17), as well as the use of a termination codon to encode selenocysteine (18), require that termination be context-dependent. Similar context effects have been shown to influence the fidelity of translation (19, 20) as well as the efficiency of translation initiation (21–23).

Statistical analyses of protein-coding regions of *Escherichia coli* have demonstrated yet another manifestation of codon context. The presence of a particular codon at one position has been shown to correlate with the frequency of occurrence of certain nucleotides in neighboring codons, and these context constraints differ markedly for genes expressed at high versus low levels (24–26). In addition, Nussinov (27) has demonstrated biases in dinucleotide frequencies within coding regions that imply the existence of constraints between codons.

It has been suggested that "suppression context" effects may be mediated through interactions between adjacent aminoacyl-tRNA molecules on the surface of ribosomes during the process of peptide elongation (13–15). Such interactions could also represent an important factor limiting or regulating gene expression by varying translational rates. If interactions between tRNA molecules binding to adjacent codons are generally important, then several consequences ought to follow. First, there should be a substantially nonrandom pattern of utilization of the 3721 (61^2) possible pairs of codons in protein-coding DNA sequences. Second, this pattern should be related (although possibly in complex ways) to the structure and abundance of tRNA isoacceptors and to the level of expression of different genes. Furthermore, these patterns might be expected to differ between organisms and, therefore, to have important consequences for the expression of foreign genes in genetically engineered expression systems. In this report we present an analysis of the pattern of codon-pair utilization in protein-coding genes of *E. coli*.

METHODS

Data base. The source of DNA sequences was the GenBank data base (Release 40.0, February 1986). Two hundred thirty-seven chromosomal *E. coli* protein-coding regions greater than 100 base pairs long were used for analysis (212 contained a termination codon) encompassing a total of 235,920 nucleotides and 78,403 codon pairs.

Computer Analysis of Codon Pairs. The set of programs we refer to as CODPAIR was written in TURBO PASCAL (Borland International, Scotts Valley, CA) running in an MS-DOS environment. For *each sequence* in the data base, CODPAIR enumerates the total number of codons, the frequency of each codon, and the number of occurrences of each codon pair. The expected frequency of each codon pair is then calculated as the product of the frequencies of its two component codons, assuming they are used randomly; the expected number of occurrences of each codon pair is the product of its expected frequency and the total number of codon pairs in the sequence. CODPAIR adds the values for observed and expected occurrences of codon pairs to two global tables and then goes on to the next sequence. (Although our initial approach was to calculate the expected values by using global codon usage frequencies, the use of locally evaluated frequencies, as described above, is more conservative. If, for example, a sequence has an unusually

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

high proportion of rare codons, it will also tend to have a high proportion of rare codon pairs made up of these codons, simply by virtue of their local abundance.)

The result of these calculations is a list of 3721 codon pairs, each with an expected and observed number of occurrences, together with a value for χ^2 (χ_1^2):

$$\chi_1^2 = (\text{observed} - \text{expected})^2/\text{expected}.$$

The sum of these χ^2 values is that for a distribution with 3720 degrees of freedom. The expected value of χ^2 is equal to the number of degrees of freedom, N, with a variance of $2N$ (40).

To remove the contribution by nonrandomness of amino acid pairs, a new value (χ_2^2) was calculated in the following manner. For each group of codon pairs encoding the same amino acid pair (i.e., 400 groups), the sums of the expected and observed values were tallied; if amino acid pairs were utilized in a random fashion, these two values would be equal. Therefore, each of the expected values within the group was multiplied by the factor (sum observed/sum expected), so that the sums of the expected and observed values within the group were now equal. The new χ^2, χ_2^2, was evaluated using these expected values. This manipulation reduces the number of degrees of freedom by 400, resulting in a distribution with 3320 degrees of freedom. We omitted from these calculations all codon pairs for which the expected value was less than 3, to avoid high χ^2 values generated simply by very low expected values. This correction results in a loss of 507 records, leaving 2813 degrees of freedom; the sum of the χ^2 values for this distribution (χ_2^2, calculated as described above) has an expected mean of 2813 with a standard deviation (SD) of 75.

The distribution of amino acid pair occurrences was also evaluated; the sums of the expected and observed values for codon pairs corresponding to each of the 400 amino acid pairs were used to calculate a χ^2 value for this distribution with 399 degrees of freedom.

RESULTS

Utilization of Codon Pairs in *E. coli*. An example of the output of CODPAIR is shown in Table 1 (see also Table 2). It is evident that although many codon pairs occur at levels close to those expected, some are highly overrepresented (e.g., CTG-GCA or CTG-GCG) or underrepresented (CTG-GAC or CTG-GCC). Examples can also be seen of the effect of correcting χ_1^2 for amino acid pair nonrandomness to yield χ_2^2. For CTG-GCG, χ_2^2 is very much smaller than χ_1^2 (although still extremely high at 54); for AAC-CCG, the very large value of 41 for χ_1^2 is reduced to only 2.9 for χ_2^2. In the ensuing discussion we shall deal only with this latter value, χ_2^2. Although this is a conservative approach, it is important to note that the difference between χ_1^2 and χ_2^2 does not necessarily imply that part of the codon pair nonrandomness is the *consequence* of amino acid pair nonrandomness; the direction of causation is indeterminate, and it is equally possible that the amino acid pair nonrandomness is driven by selection on codon pairs.

Table 1. Sample output of codon pair analysis

aa1	aa2	Cod1	Cod2	Exp.	Obs.	χ_1^2	χ_2^2
Asn	Gln	AAC	CAA	24.4	14	4.4	2.5
Asn	Gln	AAC	CAG	67.1	50	4.3	1.3
Asn	His	AAC	CAC	23.9	20	0.6	0.1
Asn	His	AAC	CAT	17.8	19	0.1	0.6
Asn	Pro	AAC	CCA	14.8	15	0.0	3.1
Asn	Pro	AAC	CCG	51.8	98	41.3	2.9
Asn	Pro	AAC	CCC	5.7	5	0.1	1.8
Asn	Pro	AAC	CCT	10.3	13	0.7	0.7
Leu	Glu	CTG	GAA	213.0	271	15.8	17.8
Leu	Glu	CTG	GAG	86.0	78	0.7	0.5
Leu	Asp	CTG	GAC	115.6	67	20.4	23.2
Leu	Asp	CTG	GAT	131.8	147	1.8	0.8
Leu	Ala	CTG	GCA	93.5	159	45.8	27.3
Leu	Ala	CTG	GCG	155.1	272	88.2	54.2
Leu	Ala	CTG	GCC	98.9	48	26.2	36.1
Leu	Ala	CTG	GCT	92.7	115	5.4	1.1

aa1 and aa2, left and right amino acids, respectively; Cod1 and Cod2, left and right codons of a pair, respectively; Exp. and Obs., expected and observed number of occurrences of each codon pair, respectively; χ_1^2, χ^2 based on the indicated expected and observed values; χ_2^2, χ^2 calculated so as to remove any contribution by nonrandom association of amino acid pairs. These data represent two groups of eight consecutive codon pairs selected from a complete listing.

As illustrated in Fig. 1, the sum of the χ^2 values for the *E. coli* data base (ECO) is 12,105, which is 124 SD higher than its expected mean. Although overrepresented pairs represent a minority of all pairs (45%), they account for almost 60% of the total χ^2 value.

Two kinds of controls were evaluated. First, a "jumbled" data base was generated by randomizing the order of the codons (excluding the initiating AUG and termination codon) in each sequence. Analysis of this randomized data base (RAND, Fig. 1) yields a χ^2 value of 2701, which is within 2 SD of its expected mean. A second control involved evaluating codon pairs separated by one, two, or three intervening codons. As seen in Fig. 1, separation by a single codon (+1) reduces the χ^2 value derived from ECO to only 4031; this is considerably lower than the original 12,105, but still 16 SD away from its expected mean. Separation by two or three intervening codons (+2, +3) reduces the χ^2 value to 3062 and 3047, respectively, only 3 SD away from the expected mean. Thus, the influence driving nonrandomness in codon pair utilization acts only over a very short distance.

A strong correlation is evident between the sums of χ^2 values for the overrepresented and underrepresented pairs made by any given codon; this holds for codons as left-hand members (correlation coefficient, $r = 0.86$) as well as right-

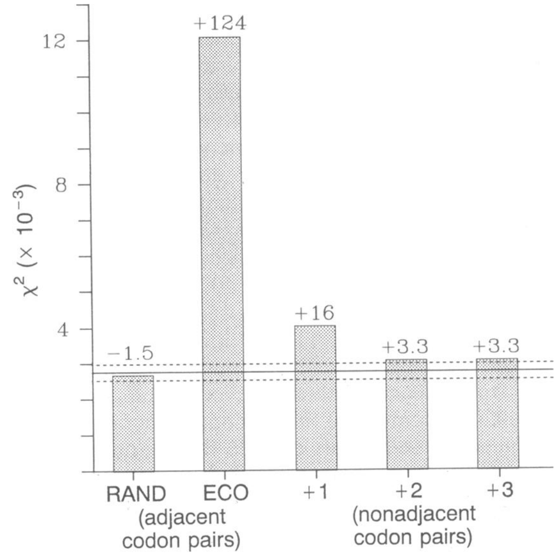

FIG. 1. Sums of χ^2 values generated by CODPAIR. The numbers above each bar represent the number of standard deviations by which that value differs from its expected mean. The total χ^2 value has an expected mean of 2813 with a standard deviation of 75; the horizontal lines represent this mean (solid line) ± 3 SD (broken lines). ECO, *E. coli* data base; RAND, randomized *E. coli* data base; +1, +2, or +3, results of evaluating nonadjacent codon pairs in the *E. coli* data base, separated by 1, 2, or 3 intervening codons, respectively.

hand members ($r = 0.84$). However, the correlation is much less striking when comparing codons as left- versus right-hand members of a pair ($r = 0.32 - 0.51$), and no evident relationship exists between the values of χ^2 for a codon pair and its reverse counterpart ($r = 0.10$). There is little relationship between the abundance of a codon pair (number of occurrences) and its degree of nonrandom representation (χ_2^2) for either overrepresented ($r = 0.26$) or underrepresented ($r = 0.15$) pairs. In general, therefore, any codon that participates in many nonrandom pairs tends to make both over- and underrepresented pairs, but does so preferentially as a left- or a right-hand member. The high degree of directionality in this bias may be important in understanding its biochemical basis (see *Discussion*).

Codon Pair Utilization and Codon Context. Shpaer (26) has reported that the codon AAA (lysine) is preferentially followed by GXX, and AAG (lysine) is followed by CXX, where X is an unknown base. These results are amply confirmed by our own findings (χ^2 = 36 and 88, respectively); in addition, we find a highly significant deficit in pairs of the form AAG-GXX (χ^2 = 80), but a random representation of AAA-CXX (χ^2 = 1.5). Although our data are, therefore, quite consistent with those of Shpaer (26), our results more closely define the nature of the nonrandom patterns. For example, it is evident (as seen in Table 2) that the high degree of overrepresentation of the 16 pairs of the form AAA-GXX is due mainly to only 4–6 pairs; 1 pair is markedly underrepresented (AAA-GCT; χ^2 = 7.2), and others are close to random. Similar results hold for pairs of the form AAG-CXX (data not shown). The same situation exists for at least some of the results of Yarus and Folley (25). For example, their finding of an excess of pairs of the form GCC-XGX is confirmed by our results (χ^2 = 64; data not shown); however, while 8 of these 16 codon pairs are substantially overrepresented (χ^2 ranges from 9 to 88), several occur at levels close to their random expectations, and one (GCC-GGC) is grossly underrepresented (χ^2 = 39).

Thus, although analyses of codon context may correctly describe the "average" behavior of sequences adjacent to particular codons, representation of individual codon pairs may differ markedly from this average.

Nonrandom Utilization of Amino Acid Pairs. Table 3 lists the six most highly overrepresented and underrepresented amino acid pairs in our data base, together with their expected, observed, and χ^2 values. The sum of the χ^2 values derived from the amino acid pair distribution is 1219, more than 29 SD away from its expected mean, and contributed roughly equally by over- and underrepresented pairs. Thus, there exists a highly significant degree of nonrandomness in the occurrence of amino acid pairs in *E. coli* proteins.

Table 2. Codon AAA (lysine) preceding GXX

Cod2	Exp.	Obs.	χ_2^2
GAA (Glu)	154.8	137*	1.7
GAC (Asp)	81.1	105	4.0
GAG (Glu)	58.8	125	77.2
GAT (Asp)	91.0	113	2.7
GCA (Ala)	71.3	71	0.1
GCC (Ala)	64.1	78	2.2
GCG (Ala)	99.7	153	24.8
GCT (Ala)	72.1	51*	7.2
GGA (Gly)	15.8	11*	1.3
GGC (Gly)	94.1	113	4.8
GGG (Gly)	22.1	40	15.8
GGT (Gly)	104.6	94*	0.6
GTA (Val)	44.9	39*	1.6
GTC (Val)	37.6	43	0.3
GTG (Val)	66.8	104	15.3
GTT (Val)	81.1	86	0.0
Total	1159.9	1363	

Expected, observed, and χ_2^2 values for all codon pairs consisting of AAA (lysine) followed by a codon beginning with G (see Table 1). Pairs of this form are found, overall, more frequently than expected (χ^2 = 35.7), although several pairs (marked by an asterisk) are underrepresented (AAA-GCT significantly), and several more are observed close to their random expectation. Abbreviations are as in Table 1.

Table 3. Most highly nonrandom amino acid pairs

aa1	aa2	Exp.	Obs.	χ^2
	Most highly overrepresented			
Asn	Pro	126.5	202	45.1
Trp	Gln	38.8	68	22.0
Ser	Gly	332.7	415	20.4
Ile	Asn	185.1	246	20.0
Glu	Gln	230.6	296	18.5
Gln	Gln	171.7	226	17.2
	Most highly underrepresented			
Gly	Pro	236.1	152	30.0
Ile	Gln	196.0	123	27.2
Phe	Met	74.6	38	18.0
Glu	Asp	292.8	226	15.2
Leu	Gln	336.4	266	14.7
Ile	Ile	291.3	227	14.2

Six most highly overrepresented and underrepresented amino acid pairs in the *E. coli* data base are presented. aa1 and aa2, left and right amino acids of a pair, respectively; Exp. and Obs., expected and observed number of occurrences of each pair, respectively; χ^2, χ^2 value based on the indicated expected and observed values. The sum of the χ^2 values for the entire distribution (399 degrees of freedom) is 1219, or 29 SD higher than its expected mean (see *Results*).

Termination Codons. As has been noted (28), the overall use of the three termination codons by *E. coli* is highly biased; among 212 sequences that include stop codons, we find that 152 (72%) use TAA, 49 (23%) use TGA, and only 11 (5%) use TAG. Some unusual features of the nearest neighbors of stop codons are shown in Table 4. Several codons are overrepresented to a significant degree, most notably GGG (glycine) and the two lysine codons (AAA and AAG) adjacent to TAA, and the GCC (alanine) and TCC (serine) codons next to TGA.

Several amino acids are also represented in a significantly nonrandom fashion adjacent to particular stop codons. Lysine is highly overrepresented next to TAA, while isoleucine is underrepresented, and proline and threonine are *never* seen adjacent to a TAA stop codon. Serine, on the other hand, is

Table 4. Codons and amino acids preceding termination codons

Codon	Amino acid	Terminator	Exp.	Obs.	*P*
GCT (Ala)		TAA	3.1	8	0.014
GGG (Gly)		TAA	1.2	7	0.00025
AAA (Lys)		TAA	6.0	15	0.0014
AAG (Lys)		TAA	1.9	9	0.00016
TAC (Tyr)		TAA	2.2	6	0.025
GCC (Ala)		TGA	1.1	5	0.0054
TCC (Ser)		TGA	2.8	9	0.0024
	Ile	TAA	8.9	3	0.023
	Lys	TAA	7.8	24	0.0000025
	Arg	TAA	8.6	13	0.097*
	Pro	TAA	6.1	0	0.0022†
	Thr	TAA	7.9	0	0.00037†
	Ser	TGA	2.8	9	0.0024

Codons and amino acids that are substantially over- or underrepresented immediately adjacent to a termination codon. *P*, Poisson probability. Other abbreviations are as in Table 1.

*The overrepresentation of arginine—TAA is not statistically significant and is presented only for comparison with lysine above.

†Although proline occurs three times preceding TGA, threonine is never found before any stop codon.

seen more frequently than expected next to TGA. Therefore, in spite of the limited amount of data available, there is an evident lack of homogeneity both in the use of termination codons and in the codons and amino acids immediately preceding them.

Codon-Pair Utilization in Proteins Expressed at High Versus Low Levels. Coding regions for *E. coli* proteins expressed at high versus low levels are known to differ in both codon usage and codon context (7, 25). Therefore, we have examined the utilization of codon pairs in representatives of these two classes of proteins. We have arbitrarily defined abundant pairs as those occurring 90 times or more in the data base, and nonabundant pairs 15 times or less (each category representing about 18% of all pairs). For underrepresented pairs (observed < expected), we have chosen a lower limit of 3 for χ^2_2, and for overrepresented pairs (observed > expected), we have chosen a lower limit of 12 (each comprising about 11% of all pairs).

Fig. 2 shows the proportion of abundant versus nonabundant (rare) pairs, as well as overrepresented versus underrepresented pairs, in the two classes of proteins. Codon pairs that are abundant in the data base as a whole are used almost three times more frequently than rare ones in highly expressed proteins but are used less frequently than rare ones in genes of low expression; this may be, at least in part, a reflection of the known avoidance of rare codons in genes for proteins expressed at high levels. Highly underrepresented pairs are used almost twice as frequently as overrepresented ones in highly expressed genes, whereas in poorly expressed genes overrepresented pairs are used more frequently. Thus, proteins expressed at high levels tend to favor more abundant, but more highly *under*represented, codon pairs; conversely, proteins expressed at low levels favor less abundant, but more highly *over*represented, pairs.

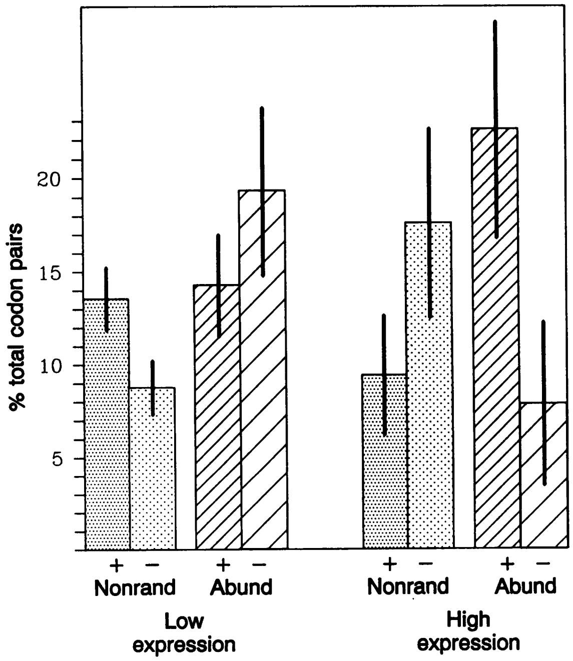

FIG. 2. Evaluation of frequencies of various categories of codon pairs (expressed as percent of total) within genes encoding proteins expressed at high (four bars to the right) or low (four bars to the left) levels [from Yarus and Folley (25)]. Nonrand+ and Nonrand−, codon pairs that are highly overrepresented (observed > expected, $\chi^2_2 \geq 12$) or underrepresented (observed < expected, $\chi^2_2 \geq 3$) in the *E. coli* data base, respectively; Abund+ and Abund−, codon pairs that are used abundantly (≥90 times) or rarely (≤15 times) in the *E. coli* data base, respectively. Vertical bars represent ± 1 SD.

DISCUSSION

Codon Pairs. Our analyses show that many codon pairs are highly over- or underutilized in *E. coli* compared with random expectations and that this effect is in addition to any contribution by nonrandomness in codon use or in amino acid pair utilization. We have also shown that the pattern of codon-pair utilization differs between genes whose protein products are expressed at high versus low levels. Powerful constraints must, therefore, exist on the utilization of codon pairs, and these constraints may be related to translational efficiency.

Several groups have lent support to the idea that codon usage patterns may relate to differences in peptide elongation rates (9, 29) or to the efficiency of translation initiation (30). Others have argued for the central importance of constraints based on the fidelity (10, 11) or kinetic efficiency (12) of translation. The constraints on codon pairs we describe may be important in clarifying such relationships; elongation rates (or other features of the translation process) could be varied independently of the known relationship between codon use and tRNA abundance, at the same time placing minimal constraints on the structure of the encoded protein.

It is clear that the relationship among codon choice, tRNA abundance, and the translation process cannot at this time yield to any simple interpretation. For example, the work of Holmes *et al.* (31) showed in *E. coli* that although the the most abundant leucine tRNA isoacceptor (anticodon CAG) does, in fact, correspond to the most abundant leucine codon (CUG), the tRNA that is actually utilized at these positions *in vivo* under normal growth conditions is a minor isoacceptor that recognizes CUG by nonconventional base pairing (anticodon GAG); only at rapid growth rates, when the minor tRNA is fully utilized, does the major species participate significantly. Certainly, much still needs to be known to explain these relationships.

Amino Acid Pairs and Termination Codons. We have also examined the pattern of utilization of amino acid pairs and of termination codons in the *E. coli* data base. Amino acid pairs are utilized in a highly nonrandom fashion, which may be a consequence of constraints on codon pairs or may be due to global constraints on protein structure within *E. coli*; these possibilities are certainly not mutually exclusive.

Doolittle (32) has evaluated dipeptide frequencies in a protein sequence data base that includes prokaryotes, eukaryotes, and viruses, and he has identified a number of pairs that are significantly overrepresented (notably most homodimers) or underrepresented. However, the most highly nonrandom pairs in our *E. coli* data base and in Doolittle's global data base (22) produce nonoverlapping lists, and we find few of the 20 homodimers significantly overrepresented in *E. coli*. These differences appear to reflect a high degree of species specificity in the constraints on amino acid pair utilization.

Although our evaluation of termination codons is limited by their relatively small numbers, we have found several examples of nonrandom use of codons and of amino acids immediately preceding termination codons. The nonrandom occurrence of codons preceding terminators may relate to the complex phenomenon of termination efficiency. On the other hand, the bias in amino acid occurrence at the C-terminal position may simply be a consequence of the codon-pair bias, since there is no reason to believe this position is of particular significance in protein structure.

Significance of Codon Pair Nonrandomness. The codon pair nonrandomness we have described acts only over a short distance and is highly directional. Both of these features are consistent with the view that the effect is mediated through interactions between aminoacyl-tRNAs bound to adjacent codons on the translation complex, a concept that has been invoked to explain the context effect on termination suppression (13–15). In fact, variation in chemical modification of tRNAs can markedly affect the efficiency of termination suppression (13, 33–35) and peptide elongation rates (36). The relationship between specific codon pairs and rates of translation initiation and elongation, as well as the influence of tRNA modification, ought to be fruitful areas for investigation.

One area in which our ability to control translation efficiency is of special importance is the expression of heterologous genes in *E. coli* or other organisms. If patterns of codon-pair utilization turn out to differ substantially between organisms, our approach might provide rules for modifying coding sequences for the purpose of altering translation efficiency or regulating peptide folding in genetically engineered expression systems. Another increasingly important problem is our ability to identify protein-coding regions in open reading frames of DNA sequences, particularly given the current exponential increase in published nucleotide sequences and the planned project for sequencing the entire human genome. Various approaches have been described that take advantage of nonrandom features of codon utilization (37, 38) as well as other statistical features of protein-coding regions (39), but none has proven completely satisfactory. Application of the nonrandom features we have described, for codon pairs and amino acid pairs, could contribute significantly to such analyses.

The coevolution of protein-coding regions with the protein synthetic machinery has led to complex species-specific relationships between the structure and abundance of tRNA isoacceptors and the pattern of utilization of codons and codon pairs. Elucidation of these relationships will clearly have many important practical and theoretical consequences.

We thank Shriram Nene (Washington University, Saint Louis) for excellent programming, Dr. Howard Tucker (University of California, Irvine) for assistance in statistical analysis, Dr. Walter Fitch (University of Southern California, Los Angeles) for helpful discussions and advice during the development of this work, and Dr. Russell Doolittle (University of California, San Diego) for sharing the results of his unpublished analysis of dipeptide usage. This work was supported in part by grants from the U.S. Public Health Service [AI-21366 (to G.A.G.) and GM-24330 (to G.W.H.)].

1. Wilson, J. T., Wilson, L. B., Reddy, V. B., Cavallesco, C., Gosh, P. K., de Riel, J. K., Forget, B. G. & Weissman, S. M. (1980) *J. Biol. Chem.* **255**, 2807–2815.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9**, r43–r74.
3. Tate, V., Finer, M., Boedtker, H. & Doty, P. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1039–1049.
4. Post, L. E. & Nomura, M. (1980) *J. Biol. Chem.* **255**, 4660–4666.
5. Ikemura, T. (1981) *J. Mol. Biol.* **146**, 1–21.
6. Grosjean, H. & Fiers, W. (1982) *Gene* **18**, 199–209.
7. Gouy, M. & Gautier, C. (1982) *Nucleic Acids Res.* **10**, 7055–7074.
8. Pedersen, S. (1984) *EMBO J.* **3**, 2895–2898.
9. Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. & Humphrey, G. (1984) *Nucleic Acids Res.* **17**, 6663–6671.
10. Holm, L. (1986) *Nucleic Acids Res.* **14**, 3075–3087.
11. McPherson, D. T. (1988) *Nucleic Acids Res.* **16**, 4111–4120.
12. Kurland, C.G. (1987) *Trends Biochem. Sci.* **12**, 126–128.
13. Bossi, L. & Roth, J. R. (1980) *Nature (London)* **286**, 123–127.
14. Bossi, L. (1983) *J. Mol. Biol.* **164**, 73–87.
15. Miller, J. H. & Albertini, A. M. (1983) *J. Mol. Biol.* **164**, 59–71.
16. Murgola, E. J., Pagel, F. T. & Hijazi, K. A. (1984) *J. Mol. Biol.* **175**, 19–27.
17. Marshall, B. & Levy, S. B. (1980) *Nature (London)* **286**, 524–525.
18. Zinoni, F., Birkmann, A., Leinfelder, W. & Bock, A. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3156–3160.
19. Johnston, T. C., Borgia, P. T. & Parker, J. (1984) *Mol. Gen. Genet.* **195**, 459–465.
20. Carrier, M. J. & Buckingham, R. H. (1984) *J. Mol. Biol.* **175**, 29–38.
21. Ganoza, M. C., Fraser, A. R. & Neilson, T. (1978) *Biochemistry* **17**, 2769–2775.
22. Manderschied, U., Bertram, S. & Gassen, H. G. (1978) *FEBS Lett.* **90**, 162–166.
23. Taniguchi, T. & Weissman, C. (1978) *Nature (London)* **275**, 770–772.
24. Lipman, D. J. & Wilbur, W. J. (1983) *J. Mol. Biol.* **163**, 363–376.
25. Yarus, M. & Folley, L. S. (1985) *J. Mol. Biol.* **182**, 529–540.
26. Shpaer, E. G. (1985) *J. Mol. Biol.* **188**, 555–564.
27. Nussinov, R. (1981) *J. Mol. Biol.* **149**, 125–131.
28. Kohli, J. & Grosjean, H. (1981) *Mol. Gen. Genet.* **182**, 430–439.
29. Varenne, S. & Lazdunski, C. (1986) *J. Theor. Biol.* **120**, 99–110.
30. Liljenstrom, H. & von Heijne, G. (1987) *J. Theor. Biol.* **124**, 43–55.
31. Holmes, W. M., Goldman, E., Miner, T. A. & Hatfield, G. W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 1393–1397.
32. Doolittle, R. F. (1988) in *Prediction of Protein Structure and the Principles of Protein Conformation*, ed., Fasman, G. (Plenum, New York), in press.
33. Colby, D. S., Schedl, P. & Guthrie, C. (1976) *Cell* **9**, 449–463.
34. Laten, H., Gorman, J. & Bock, R. M. (1978) *Nucleic Acids Res.* **5**, 4329–4342.
35. Bradley, D., Park, J. V. & Soll, L. (1981) *J. Bacteriol.* **145**, 704–712.
36. Diaz, I., Pedersen, S. & Kurland, C. G. (1987) *Mol. Gen. Genet.* **208**, 373–376.
37. Staden, R. (1984) *Nucleic Acids Res.* **12**, 551–567.
38. Gribskov, M., Devereux, J. & Burgess, R. R. (1984) *Nucleic Acids Res.* **12**, 539–549.
39. Fickett, J. W. (1982) *Nucleic Acids. Res.* **10**, 5303–5318.
40. Cramer, H. (1946) *Mathematical Methods of Statistics* (Princeton Univ. Press, Princeton, NJ).