Chapter 3

# CORRELATION BETWEEN CODON USAGE AND tRNA CONTENT IN MICROORGANISMS

**Toshimichi Ikemura**

## TABLE OF CONTENTS

# I. INTRODUCTION

The following few sentences are from the opening speech of Prof. O. Maaløe[1] at the Alfred Benzon Symposium IX, entitled *Control of Ribosome Synthesis:* "As you know, the ribosomes account for nearly 1/3 of the mass of rapidly growing *E. coli* cells, and at least 3/4 of the flow of energy and matter made available by a host of different enzymes pass over the ribosomes on its way into stable macromolecules. This alone shows the general importance of our subject." Although the exact meaning of "the flow of energy and matter ... pass over the ribosomes" is not clear, it is apparent that great amounts of energy and mass are required for translation. This is closely related to the subject matter of this chapter. The choice among synonymous codons in both prokaryotic and eukaryotic genes is clearly nonrandom, although it does not affect the nature of proteins synthesized. On the basis of codon usages in a total of 161 genes compiled, Grantham and his colleagues[2,3] proposed that each type of genome has a particular coding strategy ("genome hypothesis"), and synonymous codons are used differently by different kinds of organisms, i.e., an organism-specific codon-choice pattern. The accumulation of DNA sequence data on diverse organisms has made it clear that codon-choice patterns of genes in a single unicellular organism are actually similar to each other regardless of gene function, even though the patterns are somewhat related to amounts of protein produced from the genes. This is consistent with the genome hypothesis. However, in higher eukaryotes, as in higher vertebrates, codon-choice patterns of different genes in a single organism often differ significantly.[4] Some genes are extremely GC-rich at the third codon position, e.g., 95% G+C, while others are rather AT-rich (e.g., 30% G+C). Thus, it is not generally possible to identify organism-specific codon-choice patterns in higher vertebrates such as mammals. The obvious diversity in the third codon G+C% was found related to long-range G+C% mosaic structures of the genome that are designated "isochores" which are possibly related to chromosomal banding patterns.[4-10]

Codon usage of a wide range of organisms has been extensively studied from various respects and has been reviewed elsewhere.[11-15] Thus, in this chapter, we restrict the subject matter to the correlation of codon usage of three microorganisms to tRNA content and related topics. The cellular tRNA content of *E. coli, Salmonella typhimurium,* and *Saccharomyces cerevisiae* was measured by Hatfield et al.,[16-18] and the organism-specific codon-choice patterns of these microorganisms were attributed to the availability of tRNA isoacceptors within a cell.[16-21] In these studies, the following four deductions listed were made, which, based on codon usage data[22,23] accumulated after the publications, are reexamined in this chapter.

1. There is a close correlation between tRNA content and codon usage in most protein genes sequenced for these organisms. The exceptional cases are largely confined to genes with low expression.

2. There are clear similarities in codon choice among different genes of one microorganism, regardless of gene function. We called this organism-specific codon choice the "codon dialect" of the organism. The codon dialect is related to the specific tRNA isoacceptor population of an organism.
3. The extent of bias in codon choice is related to the protein production level of each gene. Codon usage in genes for abundant protein molecules is always more dependent on tRNA content than that in moderately or poorly expressed genes.
4. Foreign-type genes such as those of transposons, plasmids, and viruses often have quite different codon patterns than those of host organisms, and thus, the above deductions are not necessarily applicable to them.

## II. CODON USAGE IN GENES OF *E. COLI*, *S. TYPHIMURIUM*, AND *S. CEREVISIAE*

Codon usage in several highly expressed genes of two Enterobacteriaceae, *E. coli* and *S. typhimurium*, and of the yeast, *S. cerevisiae*, is listed in Table 1. Choices among synonymous codons are evidently biased, and there is a clear similarity of codon choice patterns among the genes of each organism, regardless of gene function, i.e., the codon dialect for unicellular organisms.[4] The characteristics of *E. coli* codon choice (described as the *E. coli* dialect) differ considerably from those of the *S. cerevisiae* dialect, but are similar to those of the *Salmonella* dialect. This is true for a major portion of genes within these organisms, because the codon usage pattern summed for all available genes of each organism has these characteristics (Table 1). The extent of codon bias (which is described as the accent of codon dialect) of the summed pattern is more moderate than that for highly expressed genes. Highly expressed genes almost always have a strong accent, but those with moderate or low expression have only a moderate accent.[16-21] Dialectal codon choice patterns for these three microorganisms were attributed mainly to the availability of tRNA isoacceptors.

## III. CELLULAR tRNA CONTENTS OF *E. COLI*, *S. TYPHIMURIUM*, AND *S. CEREVISIAE*

Using two-dimensional polyacrylamide gel electrophoresis, tRNAs can be separated with a high degree of resolution.[24,25] Most tRNA isoacceptor molecules of an organism can be separated by improved methods.[26] The separation pattern for *S. typhimurium*[18] is shown in Figure 1. Since cells were labeled by $^{32}$P-orthophosphate for several generations, the radioactivity of each gel spot indicates the cellular content of tRNA molecules, after correction for RNA chain length. Separated tRNA molecules were assigned to known tRNAs by RNA fingerprinting, and relative amounts were measured based on the number

## TABLE 1
## Codon Usage in *E. coli*, *S. typhimurium*, and *S. cerevisiae* Genes

| | | | ECO | | | | | | STY | | | | | YSC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | groEL | gap | ompC | tufA | fba | Sum. | ompA | rpoB | prsA | Sum. | | G3PD | pyk1 | ENO | EF1A | adh1 | Sum. |
| ARG | CGA | | 0 | 0 | 0 | 0 | 0 | 2.7 | 0 | 1 | 0 | 3.3 | | 0 | 0 | 0 | 0 | 0 | 2.2 |
| | CGC | * | 5 | 4 | 1 | 2 | 2 | 23.2 | 2 | 31 | 10 | 24.3 | | 0 | 0 | 0 | 0 | 0 | 2.0 |
| | CGG | | 0 | 0 | 0 | 0 | 0 | 4.3 | 0 | 0 | 0 | 5.3 | | 0 | 0 | 0 | 0 | 0 | 1.1 |
| | CGU | * | 17 | 8 | 12 | 21 | 4 | 25.7 | 10 | 57 | 15 | 21.9 | | 0 | 2 | 1 | 0 | 0 | 7.4 |
| | AGA | | 0 | 0 | 0 | 0 | 0 | 1.2 | 0 | 0 | 0 | 2.3 | * | 11 | 22 | 13 | 18 | 8 | 23.7 |
| | AGG | | 0 | 0 | 0 | 0 | 0 | 0.8 | 0 | 0 | 0 | 1.6 | | 0 | 0 | 0 | 0 | 0 | 7.5 |
| LEU | CUA | | 0 | 0 | 0 | 0 | 0 | 2.6 | 0 | 1 | 1 | 4.0 | | 0 | 0 | 0 | 0 | 3 | 11.9 |
| | CUC | | 1 | 0 | 1 | 0 | 0 | 9.9 | 0 | 14 | 3 | 10.0 | | 0 | 0 | 0 | 0 | 0 | 4.2 |
| | CUG | * | 38 | 19 | 24 | 27 | 26 | 57.7 | 21 | 102 | 18 | 54.6 | | 0 | 0 | 0 | 0 | 0 | 8.6 |
| | CUU | | 0 | 0 | 1 | 1 | 0 | 9.1 | 0 | 6 | 1 | 9.8 | | 0 | 0 | 0 | 0 | 0 | 9.7 |
| | UUA | | 1 | 1 | 1 | 0 | 0 | 9.3 | 1 | 0 | 0 | 12.1 | | 0 | 3 | 2 | 3 | 2 | 24.4 |
| | UUG | | 1 | 0 | 0 | 0 | 1 | 11.2 | 1 | 4 | 3 | 12.0 | * | 21 | 32 | 36 | 21 | 19 | 32.0 |
| SER | UCA | | 0 | 0 | 0 | 0 | 0 | 5.2 | 0 | 0 | 2 | 6.2 | | 0 | 1 | 0 | 0 | 0 | 15.6 |
| | UCC | | 5 | 9 | 8 | 3 | 8 | 9.7 | 6 | 32 | 7 | 11.4 | * | 12 | 13 | 17 | 6 | 7 | 14.8 |
| | UCG | | 0 | 0 | 0 | 0 | 1 | 8.0 | 0 | 4 | 1 | 8.9 | | 0 | 0 | 0 | 0 | 0 | 6.7 |
| | UCU | | 11 | 6 | 6 | 7 | 11 | 9.7 | 9 | 23 | 2 | 8.4 | * | 13 | 13 | 13 | 14 | 14 | 24.7 |
| | AGC | | 1 | 0 | 2 | 0 | 5 | 15.3 | 2 | 18 | 5 | 16.7 | | 0 | 0 | 0 | 0 | 0 | 7.3 |
| | AGU | | 0 | 0 | 1 | 0 | 0 | 6.4 | 0 | 0 | 0 | 6.6 | | 0 | 0 | 0 | 1 | 0 | 11.5 |
| THR | ACA | | 0 | 0 | 0 | 1 | 0 | 5.0 | 1 | 3 | 0 | 5.0 | | 0 | 0 | 0 | 0 | 0 | 15.5 |
| | ACC | * | 25 | 15 | 12 | 16 | 10 | 25.6 | 15 | 36 | 11 | 25.0 | * | 12 | 28 | 12 | 14 | 9 | 14.3 |
| | ACG | | 0 | 0 | 0 | 0 | 0 | 12.7 | 0 | 11 | 1 | 16.6 | | 0 | 0 | 0 | 0 | 0 | 6.7 |
| | ACU | * | 8 | 12 | 12 | 13 | 7 | 9.6 | 7 | 11 | 3 | 7.4 | * | 12 | 10 | 8 | 14 | 5 | 22.2 |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRO | CCA | | 3 | 1 | 3 | 1 | 1 | 7.9 | 2 | 6 | 2 | 5.7 | | 12 | 24 | 13 | 23 | 10 | 21.1 |
| | CCC | | 0 | 0 | 0 | 0 | 0 | 4.0 | 0 | 0 | 1 | 5.4 | | 0 | 0 | 0 | 0 | 1 | 5.8 |
| | CCG | * | 10 | 8 | 1 | 19 | 13 | 25.9 | 17 | 40 | 9 | 25.4 | | 0 | 0 | 0 | 0 | 0 | 4.2 |
| | CCU | | 1 | 0 | 0 | 0 | 1 | 6.0 | 3 | 10 | 1 | 7.6 | | 0 | 1 | 0 | 0 | 2 | 12.8 |
| ALA | GCA | * | 25 | 4 | 8 | 5 | 9 | 20.0 | 7 | 12 | 8 | 11.9 | | 0 | 0 | 0 | 0 | 0 | 15.0 |
| | GCC | | 3 | 1 | 0 | 1 | 2 | 24.3 | 3 | 12 | 11 | 27.4 | * | 7 | 12 | 10 | 7 | 16 | 15.6 |
| | GCG | * | 13 | 2 | 3 | 8 | 8 | 35.7 | 3 | 38 | 5 | 40.5 | | 0 | 0 | 0 | 0 | 0 | 5.3 |
| | GCU | * | 34 | 25 | 18 | 13 | 12 | 17.3 | 22 | 15 | 10 | 13.6 | * | 25 | 31 | 48 | 30 | 19 | 28.0 |
| GLY | GGA | | 0 | 0 | 0 | 0 | 0 | 5.7 | 0 | 0 | 0 | 6.3 | | 1 | 0 | 0 | 0 | 0 | 8.9 |
| | GGC | * | 31 | 14 | 19 | 21 | 8 | 32.7 | 22 | 47 | 12 | 36.2 | * | 0 | 0 | 0 | 0 | 3 | 8.9 |
| | GGG | | 0 | 0 | 0 | 1 | 0 | 9.3 | 0 | 3 | 0 | 10.9 | | 0 | 0 | 0 | 0 | 0 | 5.1 |
| | GGU | * | 28 | 19 | 29 | 19 | 23 | 28.3 | 16 | 55 | 8 | 19.7 | * | 25 | 34 | 35 | 42 | 41 | 34.1 |
| VAL | GUA | * | 14 | 8 | 9 | 10 | 12 | 11.6 | 10 | 17 | 5 | 11.8 | | 0 | 0 | 0 | 0 | 0 | 9.9 |
| | GUC | | 2 | 1 | 2 | 0 | 1 | 14.7 | 2 | 22 | 8 | 18.4 | * | 15 | 24 | 20 | 19 | 17 | 14.7 |
| | GUG | * | 12 | 4 | 2 | 4 | 3 | 26.6 | 3 | 33 | 9 | 25.9 | | 0 | 0 | 0 | 1 | 0 | 9.5 |
| | GUU | * | 30 | 21 | 12 | 24 | 14 | 20.1 | 12 | 38 | 13 | 14.8 | * | 22 | 24 | 15 | 26 | 19 | 26.4 |
| LYS | AAA | * | 37 | 26 | 17 | 18 | 19 | 36.2 | 15 | 52 | 6 | 35.3 | | 1 | 1 | 3 | 3 | 4 | 38.2 |
| | AAG | | 3 | 1 | 0 | 5 | 6 | 11.1 | 5 | 27 | 2 | 11.8 | * | 25 | 36 | 32 | 46 | 20 | 35.2 |
| ASN | AAC | * | 18 | 17 | 32 | 7 | 18 | 24.7 | 16 | 48 | 14 | 23.2 | * | 12 | 26 | 21 | 16 | 11 | 25.8 |
| | AAU | | 1 | 1 | 0 | 0 | 0 | 13.7 | 2 | 3 | 3 | 17.9 | | 0 | 0 | 0 | 0 | 0 | 31.6 |
| GLN | CAA | | 0 | 0 | 1 | 0 | 2 | 12.9 | 2 | 6 | 3 | 12.2 | | 5 | 10 | 9 | 12 | 9 | 29.3 |
| | CAG | * | 16 | 5 | 20 | 8 | 13 | 30.8 | 16 | 52 | 5 | 32.7 | | 0 | 0 | 0 | 0 | 0 | 10.6 |
| HIS | CAC | | 1 | 5 | 1 | 10 | 11 | 11.3 | 3 | 16 | 2 | 9.2 | | 8 | 7 | 10 | 6 | 10 | 8.3 |
| | CAU | | 0 | 1 | 0 | 1 | 2 | 11.4 | 2 | 4 | 2 | 11.2 | | 0 | 0 | 0 | 5 | 1 | 12.4 |
| GLU | GAA | * | 43 | 13 | 11 | 30 | 19 | 44.8 | 9 | 85 | 13 | 40.7 | * | 12 | 28 | 28 | 30 | 20 | 49.0 |
| | GAG | | 3 | 2 | 0 | 7 | 3 | 19.2 | 3 | 34 | 1 | 21.5 | | 2 | 0 | 0 | 1 | 0 | 17.5 |
| ASP | GAC | | 25 | 18 | 23 | 20 | 13 | 23.0 | 16 | 62 | 11 | 22.9 | | 16 | 19 | 23 | 16 | 14 | 22.5 |
| | GAU | | 10 | 7 | 9 | 4 | 8 | 31.9 | 9 | 33 | 15 | 33.8 | | 9 | 13 | 7 | 8 | 2 | 37.1 |

## TABLE 1 (continued)
## Codon Usage in *E. coli, S. typhimurium*, and *S. cerevisiae* Genes

| | | | ECO | | | | | | STY | | | | | YSC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | groEL | gap | ompC | tufA | fba | Sum. | ompA | rpoB | prsA | Sum. | | G3PD | pyk1 | ENO | EF1A | adh1 | Sum. |
| TYR | UAC | * | 7 | 6 | 24 | 8 | 10 | 13.7 | 10 | 32 | 3 | 12.5 | * | 11 | 15 | 9 | 8 | 13 | 16.4 |
| | UAU | | 0 | 2 | 5 | 2 | 3 | 14.0 | 5 | 10 | 1 | 15.6 | | 0 | 0 | 1 | 0 | 0 | 16.5 |
| CYS | UGC | | 3 | 3 | 0 | 2 | 4 | 6.4 | 1 | 3 | 3 | 5.7 | | 0 | 1 | 0 | 1 | 0 | 3.7 |
| | UGU | | 0 | 0 | 0 | 1 | 0 | 4.5 | 1 | 4 | 1 | 4.1 | | 2 | 6 | 1 | 6 | 8 | 7.6 |
| PHE | UUC | * | 7 | 10 | 17 | 13 | 11 | 18.6 | 8 | 30 | 6 | 15.8 | * | 10 | 15 | 14 | 16 | 8 | 20.0 |
| | UUU | | 0 | 1 | 2 | 1 | 2 | 17.5 | 3 | 15 | 5 | 20.5 | | 0 | 0 | 1 | 1 | 0 | 23.2 |
| ILE | AUA | | 0 | 0 | 0 | 0 | 0 | 2.4 | 0 | 0 | 0 | 4.6 | | 0 | 0 | 0 | 0 | 0 | 13.0 |
| | AUC | * | 30 | 18 | 10 | 26 | 20 | 28.3 | 14 | 64 | 16 | 27.0 | * | 11 | 26 | 9 | 14 | 12 | 18.7 |
| | AUU | | 3 | 1 | 0 | 3 | 1 | 26.6 | 3 | 20 | 9 | 27.5 | * | 9 | 11 | 11 | 16 | 9 | 30.9 |
| MET | AUG | | 22 | 9 | 4 | 10 | 8 | 26.9 | 5 | 36 | 9 | 25.8 | | 7 | 11 | 10 | 8 | 7 | 21.4 |
| TRP | UGG | | 0 | 3 | 4 | 1 | 4 | 12.7 | 5 | 4 | 0 | 11.0 | | 3 | 1 | 5 | 6 | 5 | 10.2 |
| TER | UAA | | 1 | 1 | 1 | 1 | 1 | 1.8 | 1 | 1 | 0 | 1.9 | | 1 | 1 | 1 | 1 | 1 | 1.0 |
| | UAG | | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0.10 | | 0 | 0 | 0 | 0 | 0 | 0.4 |
| | UGA | | 0 | 0 | 0 | 0 | 0 | 0.7 | 0 | 0 | 1 | 0.7 | | 0 | 0 | 0 | 0 | 0 | 0.6 |

*Note:* The name of each species is listed at the top of the column in an abbreviated form: ECO (*E. coli*), STY (*S. typhimurium*), and YSC (*S. cerevisiae*). Usages for individual genes are listed by the actual number of codons. Codon usage summed for each organism is listed as frequency/1000 (Sum.): in *E. coli*, 937 intrinsic genes were summed; in *S. typhimurium*, 130 genes; and in *S. cerevisiae*, 581 genes. Optimal codons of each organism were deduced by combining the predictions from Rules 1–4. An example of this deduction for *E. coli* arginine is as follows. The most abundant *E. coli* arginine isoacceptor responds to CGU, CGC, and CGA (Table 2). Therefore, Rule 1 predicts the preference "CGU,CGC,CGA>CGG,AGA,AGG". Because this tRNA has inosine in the anticodon, Rule 3 predicts "CGU,CGC>CGA". Then the combination of these preferences is "CGU,CGC>CGA,CGG,AGA,AGG," and thus, both CGU and CGC are the optimal codons in *E. coli* for arginine. The optimal codons of these organisms deduced previously[17,18,20,21] are specified by asterisks. No asterisks were used for amino acids whose isoacceptors have not yet been quantified.

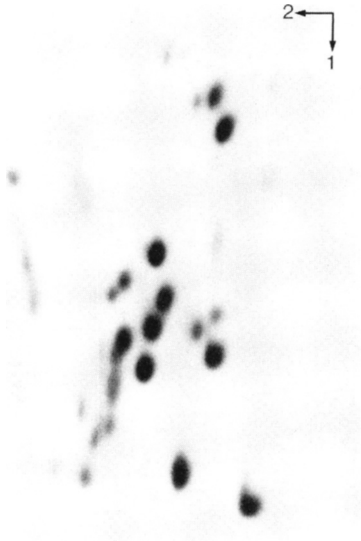

**FIGURE 1.** Two-dimensional gel electrophoresis of $^{32}$P-labeled *S. typhimurium* tRNAs.[18] Electrophoresis in the first dimension was from top to bottom on a 10% (w/v) polyacrylamide gel, and that in the second dimension was from right to left on a 22% (w/v) polyacrylamide gel containing 7 *M* urea. The conditions for electrophoresis were presented previously.[26]

of molecules in cells. The amount of each tRNA measured previously by these procedures from *E. coli*[16,21] and *S. typhimurium*,[18] as well as the codons recognized by each tRNA, is listed in Table 2. Eight tRNA species were newly quantified (see the legend in Table 2). The amount of tRNA$^{Leu}_{1}$ was normalized to 1.0. Clearly, the tRNA contents of the two Enterobacteriaceae are quite similar, indicating that the populations of tRNA molecules have been conserved during evolution. It should be pointed out that these tRNA populations differ greatly from that of *S. cerevisiae*.[17,21]

Recently, Komine et al.[27] mapped all the tRNA genes in the *E. coli* genome, and determined the exact gene copy numbers (Table 2). Figure 2 shows that the contents of individual *E. coli* tRNAs are closely related to the gene copy number (correlation coefficient r = 0.73). This observation confirms our previous proposal that the gene number is a major factor in determining *E. coli* tRNA content.[16] The amounts of tRNAs encoded within ribosomal RNA operons designated as spacer or distal tRNAs (e.g., Ala1B, Trp, Thr1+3) appear somewhat higher than that expected from the copy number. This may be related to the strong expression of rRNA operons. Dispensable tRNAs appear to be present in lower amounts than what would be expected from the corresponding gene copy number, e.g., RNA$^{Ser}_{2}$, tRNA$^{Gly}_{1}$, tRNA$^{Thr}_{2}$.

## TABLE 2
## Content and Codon Recognition of tRNAs in *E. coli* and *S. typhimurium* tRNAs

| | | | No. of tRNA genes | tRNA content | |
|---|---|---|---|---|---|
| **aa** | **tRNA** | **Codon** | ***E. coli*** | ***E. coli*** | ***S. typhimurium*** |
| Leu | 1 | CUG | 4 | 1.00 | 1.0 |
| | 2 | CUU,CUC | 1 | 0.30 | 0.2 |
| | 4 | UUA,UUG | 1 | 0.25 | 0.2 |
| | 5(6) | UUG | 1 | 0.20 | 0.2 |
| | 3 | CUA,CUG | 1 | Minor | Minor |
| Val | 1 | GUA,GUG,GUU | 4 | 1.05 | 0.9 |
| | 2A+2B | GUC,GUU | 2 | 0.40 | 0.2 |
| Gly | 3 | GGU,GGC | 4 | 1.10 | 0.9 |
| | 2 | GGA,GGG | 1 | 0.15 | 0.2 |
| | 1 | GGG | 1 | 0.10 | (0.1) |
| Ala | 1 | GCU,GCA,GCG | 3 | 1.00 | 1.0 |
| | 2 | GCC | 2 | (0.3) | 0.3 |
| Arg | 2(1) | CGU,CGC,CGA | 4 | 0.90 | 0.7 |
| | CGG | CGG | 1 | Minor | (Minor) |
| | AGR | AGA,AGG | 1 | Minor | (Minor) |
| | AGG | AGG | 1 | Minor | (Minor) |
| Ile | 1 | AUU,AUC | 3 | 1.00 | 1.0 |
| | 2 | AUA | 1 | 0.05 | (0.05) |
| Lys | | AAA,AAG | 2 | 1.00 | 0.9 |
| Glu | 2 | GAA,GAG | 4 | 0.90 | 0.9 |
| Asp | 1 | GAU,GAC | 3 | 0.80 | 1.0 |
| Thr | 1+3 | ACU,ACC | 2 | 0.80 | 0.6 |
| | 2 | ACG | 1 | (0.1) | 0.1 |
| | 4 | ACA,ACG | 1 | (0.1) | 0.1 |
| Asn | | AAU,AAC | 3 | 0.60 | 0.5 |
| Gln | 2 | CAG | 2 | 0.40 | 0.4 |
| | 1 | CAA | 2 | 0.30 | 0.3 |
| Tyr | 1+2 | UAU,UAC | 3 | 0.50 | 0.3 |
| Ser | 3 | AGU,AGC | 1 | 0.25 | 0.2 |
| | 1 | UCU,UCA,UCG | 1 | 0.25 | 0.3 |
| | 5 | UCU,UCC | 2 | 0.25 | 0.2 |
| | 2 | UCG | 1 | (0.05) | 0.05 |
| His | | CAU,CAC | 1 | 0.40 | 0.3 |
| Trp | | UGG | 1 | 0.30 | 0.2 |
| Pro | 3 | CCG,CCA,CCU | 1 | (0.3) | 0.3 |
| | 1 | CCG | 1 | (0.3) | 0.3 |
| | 2 | CCC,CCU | 1 | Minor | Minor |
| Phe | | UUU,UUC | 2 | 0.35 | 0.2 |
| Cys | | UGU,UGC | 1 | Minor | (Minor) |
| Met | m | AUG | 2 | 0.30 | 0.4 |
| | f | AUG,GUG,UUG | 3 | 0.50 | 0.3 |

**TABLE 2 (continued)**
**Content and Codon Recognition of tRNAs in *E. coli* and *S. typhimurium* tRNAs**

*Note:* The relative abundances of individual tRNAs were measured on the basis of molecular numbers in cells using two-dimensional polyacrylamide gel electrophoresis. tRNA in column two denotes the specific tRNA designation for the isoacceptor(s). The amount of $\text{tRNA}_1^{\text{Leu}}$ is normalized to 1.0. The tRNA contents of the two organisms were measured by different gel systems.[16,18] Because of the general similarity of the two tRNA populations, contents of several tRNAs not quantified by the respective separation system are tentatively estimated from data of the other organism and listed in brackets. Gene copy numbers of *E. coli* tRNAs are included according to the data of Komine et al.[27] Based on their RNA fingerprints, eight gel spots in our previous work[16,18] can be newly assigned to known tRNAs.

# IV. CORRELATION BETWEEN CODON USAGE AND tRNA CONTENT

## A. FREQUENCY OF tRNA USAGE

To examine the correlation between frequency of codon usage and tRNA content, the frequency of tRNA usage, i.e., anticodon usage, was determined as follows. The usage frequency of a tRNA responding to a single codon was defined as the occurrence of the codon itself, and that of a tRNA responding to multiple codons as the total occurrence of the corresponding codons.[16] The content of tRNA and frequency of its usage calculated for the summed codon usage are roughly proportional (see Figure 3a), indicating a close correlation between the two variables.[10,16,20,28,29] Biological significance of this correlation is discussed later in this chapter. The degree of the correlation was previously found to depend on the expressivity of genes [a regression line ($y = a + bx$) was drawn in the graph in Figure 3]. High slope values as well as negative y intercepts of the regression line were found for genes that encode abundant protein species. This is clear from the example of *groEL* in Figure 3b. High slope values indicate strong dependence of tRNA use on its content, and negative y intercept has been attributed to "upward concavity" of functions deduced from curvilinear regression.[16,17,20,21] To confirm the "upward concavity", the data points in Figure 3b were analyzed by quadratic curve fitting as shown by the dotted line.

## B. FREQUENCY OF ISOACCEPTING tRNA USAGE

By comparing synonymous codon choices with tRNA isoacceptor content, the correlation between codon usage and tRNA content can be examined without the influence of amino acid composition of individual protein species. In Figure 4, both content and usage frequency of the most abundant isoacceptor of each amino acid are normalized to 1.0. Clearly, the most abundant

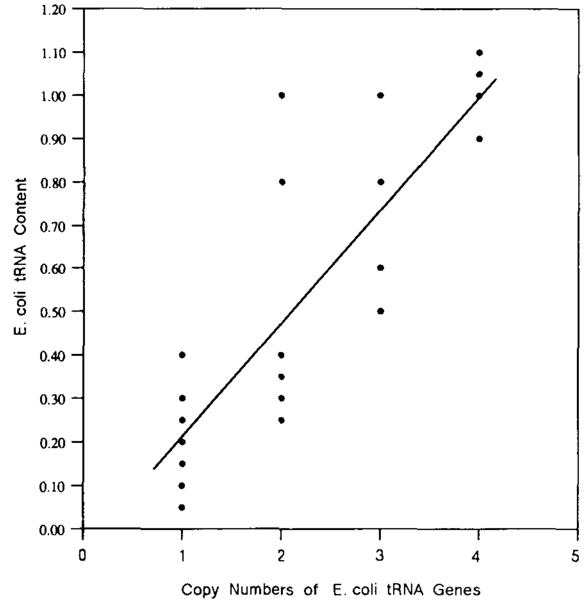

**FIGURE 2.** Correlation between gene copy number and tRNA content of *E. coli*. The line corresponds to the regression line (correlation coefficient r = 0.73). The data from Table 2 are shown graphically. Redundant tRNAs (Val 2A+2B, Metf 1+2, or Tyr 1+2) were treated as a single tRNA species.

isoacceptor is always used at the highest frequency. The line in each graph indicates the function when the frequency of isoacceptor use is directly proportional to content. In modestly expressed genes, the data points were found to be represented roughly by this linear function.[16-21] This notion was confirmed here by analysis of the summed codon usage pattern (Figure 4a). Data points of three highly expressed genes are situated far away from the line, and are drawn as probable concave curves (dashed lines; Figure 4 b–d). The dependence of isoacceptor usage on its content is thus much greater than that expected from the proportionality of the two values. That is, the highly expressed genes selectively use codons recognized by the most abundant isoacceptor but almost completely avoid codons for other isoacceptors.[11] Essentially the same results were obtained for most of highly expressed *E. coli*, *S. typhimurium*, and *S. cerevisiae* genes.[16-21] Based on these findings, we proposed that codon choice in these organisms is constrained by tRNA availability, and this is particularly evident for highly expressed genes. It should be noted again that the population of *S. cerevisiae* tRNA isoacceptors is very different from that of *E. coli*.[4,21] The differential isoacceptor population is a cause of the organism-specific codon-choice pattern.

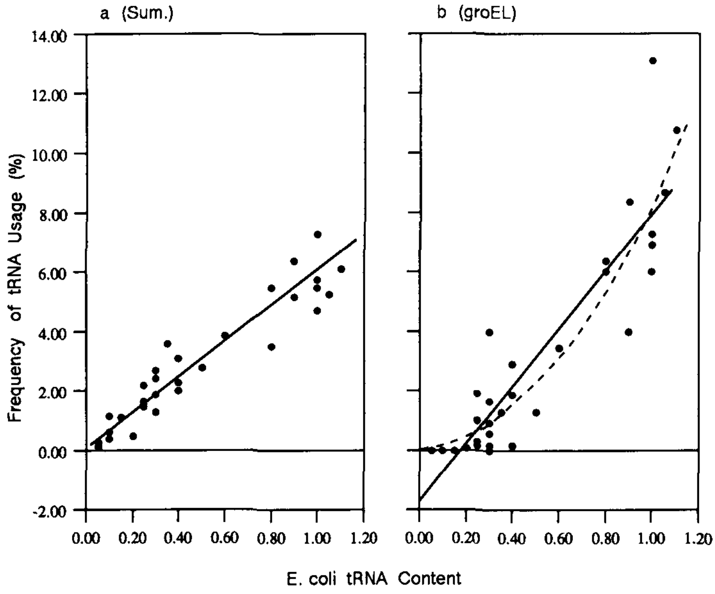

**FIGURE 3.** Correlation between tRNA content and frequency of codon usage summed for 937 *E. coli* intrinsic genes (a) and that of *groEL* (b). The data from Table 1 are shown graphically.

## C. PREFERENCE AMONG MULTIPLE CODONS RECOGNIZED BY A SINGLE tRNA

Figure 4 shows that the codon choice in *E. coli* genes is strongly constrained by tRNA availability. This observation provided the basis for **Rule 1** in a previous report.[21] The following definite constraints on the choices among codons recognized by a single tRNA (Rules 2–4) were defined:[20,21]

- **Rule 2** — Introduction of a thiolated uridine or of 5-carboxymethyl uridine at the anticodon wobble position leads to a preference for an A-terminated codon over a G-terminated codon.[30] This rule is mainly applicable for highly expressed genes, and is affected by the "context effect",[31] most likely due to the modified nucleotide. For example, AAA is preferred if G is 3′ to the Lys codon, i.e., AAA-G, while AAG is used quite often if C is 3′ to the Lys codon, i.e., AAG-C. General features of the context effect have also been reported.[32]
- **Rule 3** — Introduction of inosine at the anticodon wobble position produces preference for U- and C-terminated codons over an A-terminated codon, which leads to purine-purine wobble pairing. The tRNA isoacceptor populations as well as the modification patterns in the anticodon and anticodon loop often differ for taxonomically distant organ-

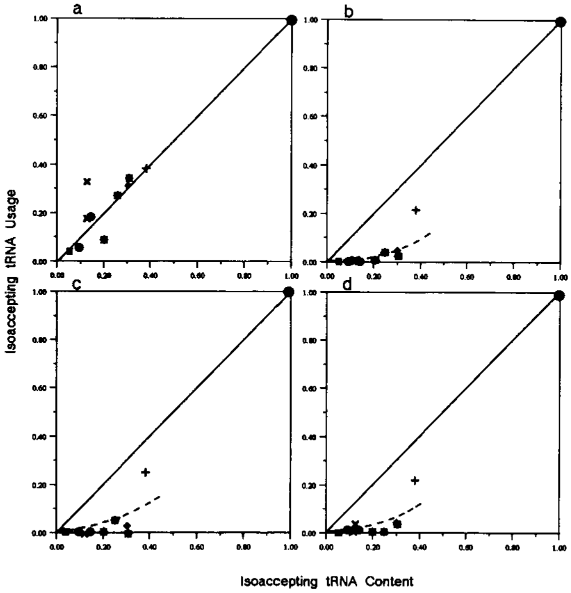

FIGURE 4. Correlation between tRNA isoacceptor content and frequency of tRNA isoacceptor usage in decoding *E. coli* mRNAs. Both the amount and the frequency of use of the most abundant tRNA for each amino acid are normalized to 1.0. The straight line in each graph is that predicted if the use of isoacceptors is proportional to its content. The most abundant isoacceptor of each amino acid is indicated by (•). Other isoacceptors are Leu (*), Gly (·), Thr (X), Ala (♦), Val (+) and Ile (■). (a) Sum. in Table 1; (b) *groEL;* (c) *gap;* and (d) *tufA*. Details as well as results for other *E. coli, S. typhimurium,* and *S. cerevisiae* genes have been presented previously.[16-21]

isms. This is also a factor responsible for organism-specific codon-choice patterns.[20,21]

- **Rule 4** — Codons of the (A/U)-(A/U)-pyrimidine type, i.e., AAX, AUX, UAX, and UUX where X is a pyrimidine, would support an optimal interaction strength between a codon and an anticodon when the third letter of the codon is C.[33,34] This rule is mainly applicable for highly expressed genes, but not for genes with moderate or low expression, showing its weaker constraint. For example, compare the summed codon patterns of Phe, Tyr, and Asn with those of highly expressed genes (Table 1). This rule may possibly be related to fidelity in translation.[35]

## V. CODONS OPTIMAL FOR AN ORGANISM'S TRANSLATION SYSTEM

### A. OPTIMAL CODONS

Codon choices in *E. coli, S. typhimurium,* and *S. cerevisiae* genes were found to conform well to expectations based on Rules 1–4, and exceptional cases were mostly confined to genes of low expression.[4,21] However, it is not an easy task to understand this conclusion directly from the data in Table 1 or those in the Codon Compilation Database.[22,23] The codon that is translated by the most abundant isoacceptor of each amino acid (Rule 1) and also conforms to Rules 2–4 was previously designated as the "optimal codon", i.e., the codon optimal for translational efficiency of the respective organism as discussed in detail in our previous papers.[20,21] The optimal codons of the Enterobacteriaceae and *S. cerevisiae* are indicated by an asterisk in the first column below the respective organism in Table 1. Each optimal codon corresponds strikingly well to the preferred codon of each organism. We proposed the four rules and the optimal codons derived from them when only small numbers of *E. coli* and *S. cerevisiae* gene sequences had been determined.[16,17,19,20] It has become clear that these are applicable for a major portion of *E. coli, S. typhimurium,* and *S. cerevisiae* genes (refer to the summed codon patterns; Table 1). This again supports the idea that the synonymous codon choices in genes of these organisms are under constraints related to translation efficiency. The difference in the spectrum of the optimal codons used by *E. coli* and *S. cerevisiae* is due to that in the actual population of isoacceptors and in the modified nucleotides at the anticodon wobble position.[4,21]

## B. FREQUENCY OF OPTIMAL CODON USAGE

Figure 5 shows the distribution of optimal (o) and nonoptimal (X) codons in *E. coli* and *S. typhimurium* genes.[18] For Met and Trp, a single codon corresponds to the amino acid and is indicated by the symbol "–". Highly expressed genes, i.e., the *E. coli* ribosomal protein gene, *rplA*, and the *S. typhimurium* membrane protein gene, *ompA*, use mostly optimal codons, showing their codon choices to be highly constrained by tRNA availability. In moderately expressed genes, such as amino acid synthesis genes, e.g., *E. coli thrA* and *S. typhimurium trpA*, optimal codon usage clearly decreases. To examine this feature quantitatively, the frequency of use of optimal codons was defined as

$$F_{op} = \frac{\text{the numbers of o}}{\text{sum of the numbers of o and X}}$$

We previously listed the $F_{op}$ calculated for all available *E. coli* and *S. cerevisiae* genes, and found that this parameter is high for highly expressed genes and low for weakly expressed genes.[4,18,21] The cellular content of a wide variety of *E. coli* protein molecules has been measured by Neidhardt and his colleagues[36] by two-dimensional gel electrophoresis. Their cellular content is thought to correspond to the degree of gene expression, although the stability of protein molecules has to be taken into consideration. It is thus possible to examine the correlation between the frequency of optimal codon use ($F_{op}$) and gene expressivity. For all *E. coli* genes whose nucleotide sequences and relative protein contents are known (Table 3), the $F_{op}$ and protein content

```
a)  E. coli

rplA  (ribosome protein L1) ; Fop=0.93
-ooooXo-ooooXoo ooooo oooooooooXoooooooo o ooooXoo ooo  oooooooo
oo oooo oooXoXoXoooooooooooooooo-o oo ooXooo-oX ooXo o o-ooooooo
oooooooo-XoooooooXooooooooooooooooooo ooooo oooooo X o ooooooooo
oooooooooooooooooXoo o oo-oooooo oooo o oo

thrA ; Fop=0.61
-XoXXoooX XoXoooXoooX Xoo XXXoXooXoXX XXXoooo ooo-XoooX oo oXXXo
  XooXXXoXXXXXXXXooXooXoXXoXoX ooXXXXo Xo oX XXXo o  oooooX ooXX
- oXX-XooXoooo ooooo oXooooooX oXo oX XXX oooXoo oXoo  -oo-ooooX
oXXooooooXXooo  o oooooX XoX   XX-X ooXXXo  ooooX oXXXX - oooo-X
X ooooooX XooXoXoXooooX oXoXoXXXoXoXXXoo o o oXoXXoX XoXo-o-o o
 ooX-oX-Xo-oooXXoo- oXoX oooXXX   oo o o oXX   oXooXo-ooXooXoooo
oXoXooooooXoXXo ooo oXooXoXo oooXXooXoXXooXXXooX  oo o XooXo  o
ooooooo o-ooXo oooooXoXoXooooooooXXoXooX -oXXo o XoX ooXo XooXoX
o oXXooo-oXoooXXoXoXXXXoXXoXoooX ooooXXoo o  oooo XXX oooooo oXX
oooXXoo  - oo oXoXoooo XoooXX oooXoXXooXXooXXoXXoo oX-oo oXX o X
 XoooXX oo- o XooXooXo-oXooo oX  X o- ooooXXXoooXXoooXoo XoXoXo
oXoXXoXXo oXoX-oXo XX  XXXooooXXo oXooXoXooXX o oX ooXXXoo oX oo
ooooXoooXoXoX  XXooooXoooXXoooX oXoXoXXo oXooX -XXXX

b)  S. typhimurium

ompA ; Fop=0.88
-oXXoooXooooooooooooXooo oo-ooooooo- oo  oooX X ooo ooXoooooXooo
ooooXooXo-oo -Xoo-oooo ooXooXooooooXoooooXXoo  o XXooooo-o-oo oX
 oXXooo oo   ooo oooooooXXoooXooooooooo-ooooo Xooooooo oooo ooo
oooooXoooooooooXooooooooX oooX  ooooooo ooXooooooo oooooo oo oo
 o oXooooo ooo  oooooo Xoooo oo ooo ooXo  oo ooo-oo ooooooo  ooo
XooXoo  ooo oooXooooooo ooooooo

trpA ; Fop=0.64
-ooooXXXXXXo oXoooXXXooooo XoXoo ooXX XoX XoX ooooXoXo  ooX oXooo
XooXoXoXooXXooo Xo-oooXooo ooXoXooX-ooXoooXooX ooXXo ooooo  ooXo
Xoooo XXoooooXX Xoooooo oXXo   XoooXo ooooooooX o oXooooooooXoX
 XXXXXoXo XoXoXooooo  oooo oXooXXooooo o XXXXXoXooXo Xoo-XoXXX XX
 X-oXX oo
```

**FIGURE 5.** Occurrence of optimal (o) and nonoptimal (X) codons in *E. coli* and *S. typhimurium* genes.[18] Codons of Met and Trp are indicated by a dash (–). Blank spaces are used for several amino acids either because the contents of their tRNA isoacceptors have not been clarified or because no criteria were deduced. (a) *E. coli* genes; and (b) *S. typhimurium* genes. Results for other *E. coli* genes have been presented previously.[20,21]

are plotted in Figure 6. The definite correlation between the two values shows optimal codon frequency in each gene to be actually related to its production level. A similar type of correlation between codon bias and protein production has also been noted by other groups.[37-39] Experimental evidence which shows that codon choice affects translational efficiency has been reported by many groups.[40-45]

Figure 7a and 7b show histograms of $F_{op}$ for *E. coli* intrinsic genes and foreign-type genes (genes for plasmids, transposons, enterotoxins, and restriction-modification systems), respectively. $F_{op}$ of foreign-type genes is significantly lower than those of intrinsic genes, showing that the former genes do not necessarily use the *E. coli* dialect. *E. coli* genes with low $F_{op}$ were found to be mostly confined to regulatory genes with low expression.[4,21] *S. cerevisiae* genes with high $F_{op}$ correspond to genes with high expression,[17,21] and those with very low $F_{op}$ values often correspond to nuclear genes encoding mitochondrial proteins, e.g., PET122, MRS3, PET54, or mating-related factors, e.g., STE4, HML.

## VI. OTHER CONSTRAINTS ON CODON CHOICES

To explain the nonrandom patterns of synonymous codon choices in the three microorganisms, four constraints related to translation efficiency (**Rules**

## TABLE 3
## Frequency of Optimal Codon Usage ($F_{op}$) and Numbers of Protein Molecules per *E. coli* Genome

| | | | No. of molecules | |
|---|---|---|---|---|
| Gene | Locus | $F_{op}$ | Average | Rich |
| tufA | ECOSTR3 | 0.93 | 60567 | 86606 |
| rpsA | ECORPSA | 0.92 | 11504 | 19003 |
| dnaK | ECODNAK | 0.89 | 6078 | 8968 |
| atpA | ECOUNC#6 | 0.87 | 6143 | 4786 |
| aceE | ECOACE#2 | 0.86 | 5403 | 10844 |
| lpd | ECOACE#4 | 0.85 | 6152 | 5645 |
| fabB | ECOFABB | 0.85 | 21043 | 28285 |
| atpD | ECOUNC#8 | 0.85 | 5408 | 4650 |
| ssb | ECOSSB | 0.84 | 1452 | 1725 |
| rpoB | ECORPLRPO#5 | 0.83 | 1551 | 2173 |
| carB | ECOCARAB#2 | 0.82 | 1009 | 321 |
| sucB | ECOGLTA#7 | 0.82 | 3747 | 243 |
| valS | ECOVALS | 0.82 | 550 | 859 |
| glyS | ECOGLYS#2 | 0.80 | 1130 | 1576 |
| trxA | ECOTRXA | 0.80 | 2269 | 2254 |
| metK | ECOMETK | 0.79 | 3578 | 4211 |
| sdhA | ECOGLTA#4 | 0.79 | 2013 | 158 |
| ilvE | ECOILVE#2 | 0.78 | 4227 | 2666 |
| glnS | ECOGLNS | 0.78 | 794 | 1041 |
| pheT | ECOTHRINF#6 | 0.77 | 890 | 1138 |
| pheS | ECOTHRINF#5 | 0.77 | 1050 | 1350 |
| nusB | ECONUSB | 0.77 | 684 | 839 |
| purC | ECOPURCA#2 | 0.77 | 6167 | 2755 |
| sucA | ECOGLTA#6 | 0.76 | 1616 | 250 |
| tyrS | ECOTYRS | 0.76 | 1563 | 1649 |
| livJ | ECOLIVHMGF#1 | 0.75 | 14662 | 1254 |
| grpE | ECOGRPE | 0.74 | 2236 | 2905 |
| glpK | ECOGLYK | 0.73 | 14652 | 2045 |
| hisS | ECOHISS | 0.73 | 1027 | 1414 |
| ppc | ECOPPCG | 0.72 | 903 | 351 |
| trxB | ECOTRXB | 0.72 | 1496 | 1726 |
| rpoA | ECORPOA#1 | 0.71 | 4664 | 6377 |
| gor | ECOGOR | 0.71 | 536 | 517 |
| thrS | ECOTHRINF#1 | 0.71 | 553 | 855 |
| lon | ECOLON | 0.71 | 1319 | 816 |
| folA | ECOFOLA | 0.68 | 7459 | 9388 |
| polA | ECOPOLA#1 | 0.67 | 283 | 432 |
| hisP | ECOHISMP | 0.64 | 282 | 211 |
| tyrA | ECOPHEAB#4 | 0.61 | 222 | 120 |

*Note:* The relative numbers of *E. coli* protein molecules analyzed in three different growth conditions (acetate-, glycerol-, and rich-medium) were obtained from VanBogelen et al.,[36] and the number of protein molecules per genome was calculated with their equation. Average (column 4) corresponds to the average values observed with the three growth conditions, and Rich (column 5) corresponds to those observed with the rich-medium. Proteins that gave multiple spots in their gel system were omitted from the analysis. Locus (column 2) corresponds to the GenBank Locus name.

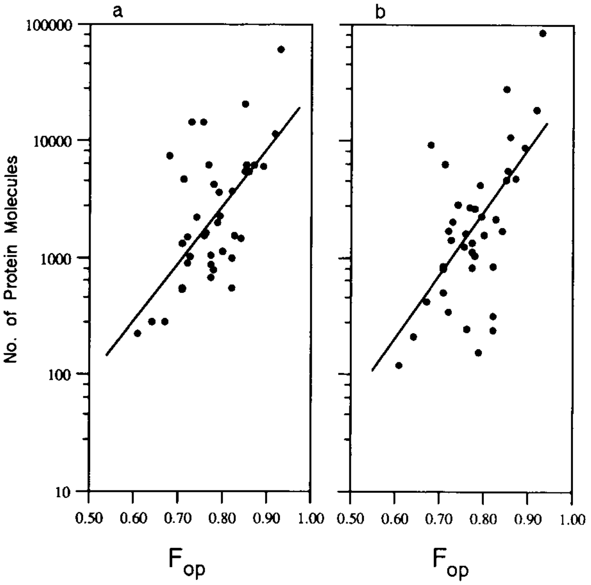

FIGURE 6. Correlation between frequency of optimal codon use ($F_{op}$) and number of *E. coli* protein molecules per genome. The data of Table 3 are shown graphically. (a) Average and (b) Rich.

1–4) are proposed. When a wide range of microorganisms is concerned, other factors must be considered and these have been summarized previously.[21] For organisms with a base-biased genome, the bias of the genome G+C% is an important factor; the G+C% becomes **Rule** 5.[46-54] Figure 8 shows histograms of G+C% at the third codon position of genes of a GC-rich microorganism, *Streptomyces* (70–75% G+C; see graph a), and of an AT-rich microorganism, *Clostridium* (30–35% G+C; see graph b), as well as of *S. cerevisiae* (ca. 40% G+C; see graph c) and *E. coli* (ca. 50% G+C; see graph d). The effect from the genome G+C% is clearly less evident for *S. cerevisiae* and *E. coli* than for the G+C%-biased organisms, *Streptomyces* and *Clostridium* (see graphs a and b). In the foreign-type *E. coli* genes (graph e), there is a significant number of AT-rich ones, and the distribution is much wider than that of the intrinsic genes (graph d). This may be related to their evolutionary origin. The third codon G+C% of higher vertebrates is known to be distributed in a wide range; a histogram of 1882 human genes is listed (see graph f). This wide distribution was attributed to the long range G+C% mosaic structures of the genome (as discussed below).

Codon choices of individual microorganisms are determined by combined effects of translational efficiency (Rules 1–4) and genome G+C% (Rule 5). It should be noted that T- and C-terminated codons are often recognized by a G-starting anticodon, and A- and G-terminated codons by a U-starting anticodon.

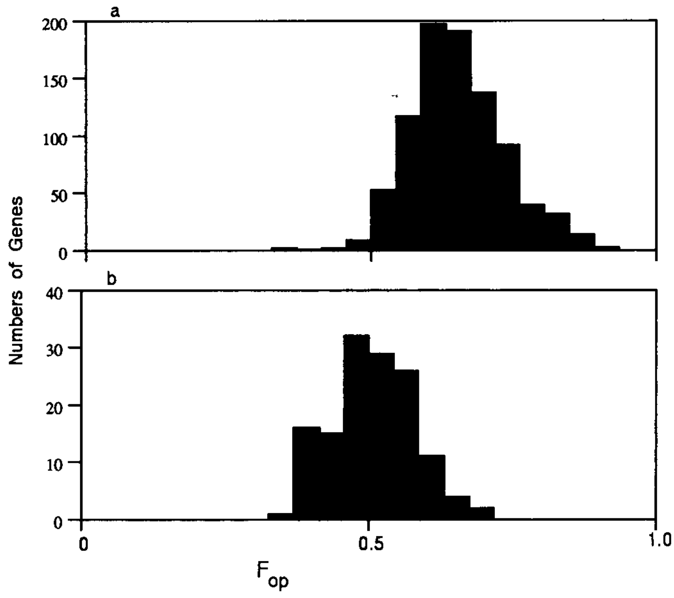

**FIGURE 7.** Histograms for $F_{op}$ of *E. coli* genes: (a) 937 intrinsic genes are represented and (b) 133 foreign-type genes.

Because of this wobble pairing, constraints due to tRNA content and genome G+C% are fairly compatible with each other. In organisms with an evidently biased genome, e.g., ≥70% or ≤30% G+C, these two factors appear usually harmonious, since tRNA content has been well adapted to the spectrum of biased codons.[15,53] For organisms with moderate genome G+C%, e.g., 40–60% G+C, however, there are cases where the two factors direct the opposite preference. For example, AAG is the optimal Lys codon of *S. cerevisiae* (the CUU anticodon tRNA is the most abundant $\text{tRNA}^{\text{Lys}}$ isoacceptor),[17] while the genome G+C% is about 40%. By examining such cases for enterobacterial genes, it was found that the bias of codon choices due to the genome G+C% is evident for low or modestly expressed genes, but not for highly expressed genes.[46,47,55] That is, the contribution of base composition increases as gene expressivity decreases. This can also be observed by analyzing codon choices of *S. cerevisiae* genes for such amino acids encoded by two codons as Lys, Phe, Asn, and Tyr, in which constraints due to translational efficiency cause a tendency toward using G- or C-terminated codons, i.e., the tendency of using such codons is opposite to the genome G+C%. In Figure 9, the two values, G+C% at the third codon position of the four amino acids and $F_{op}$ (used for indices of gene expressivity), are plotted for individual *S. cerevisiae* genes. A clear correlation (r = 0.81) can be seen: genes with low $F_{op}$ (thus with low expression) use AT-rich codons conforming to its genome G+C%, while those with high $F_{op}$ use GC-rich codons conforming to translation efficiency.

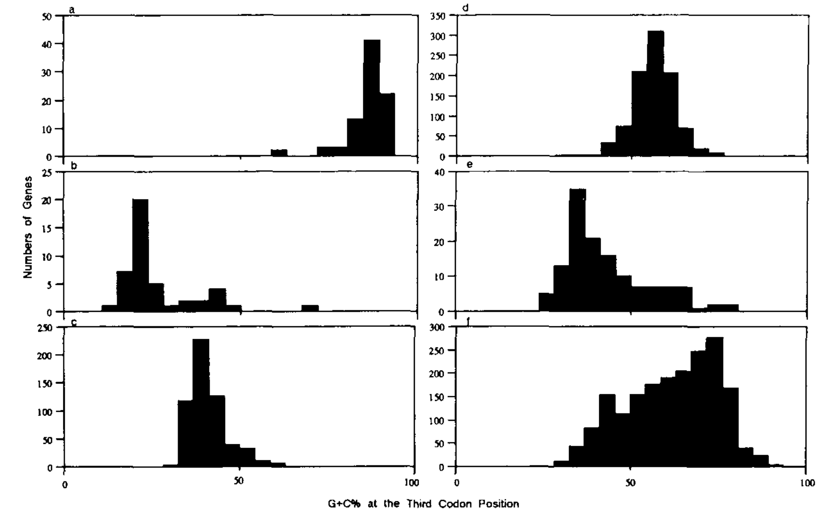

FIGURE 8. Histograms for G+C% at the third codon position in genes of different organisms. The G+C% at the third position was calculated using codon usage data compiled by our group.[23] (a) *Streptomyces*, 84 genes; (b) *Clostridium*, 44 genes; (c) *S. cerevisiae*, 560 genes; (d) and (e) *E. coli* intrinsic (937 genes) and foreign-type (133 genes), respectively; and (f) Human, 1882 genes. Since the number of genes differed significantly among organisms, the scale in the graph was changed accordingly.

# VII. MECHANISMS FOR THE CORRELATION BETWEEN CODON USAGE AND tRNA CONTENT

## A. MOLECULAR MECHANISMS

A close correlation between codon usage and tRNA population should have resulted from the accumulation of a great number of mutations and successive base substitutions in both protein and tRNA genes during evolution. Figure 3a shows that the *E. coli* tRNA population and its summed codon spectrum are closely correlated. One clear factor that may be deduced from this correlation is the constraints placed on the tRNA content by the amino acid composition of cellular proteins.[16,20] However, the correlation between synonymous codon usage and tRNA isoacceptor content (Figure 4) should not be due to the amino acid composition. For organisms with strong G+C%-biased genomes, the adaptation of the isoacceptor content to the biased codon spectrum should be a major mechanism used in producing the correlation.[15,53] In the case of organisms with moderate G+C%, e.g., *E. coli*, 50% G+C, we proposed that the codon usage was adapted to the tRNA isoacceptor content.[16-21] The similarity in codon-choice patterns among different genes of one organism with a moderate genome G+C% is considered as evidence for the constraint imposed by the tRNA content on codon choice, although the adjustment of the tRNA content to the codon spectrum may also be an important factor. If only the adjustment of isoacceptor content to the codon spectrum is considered, one

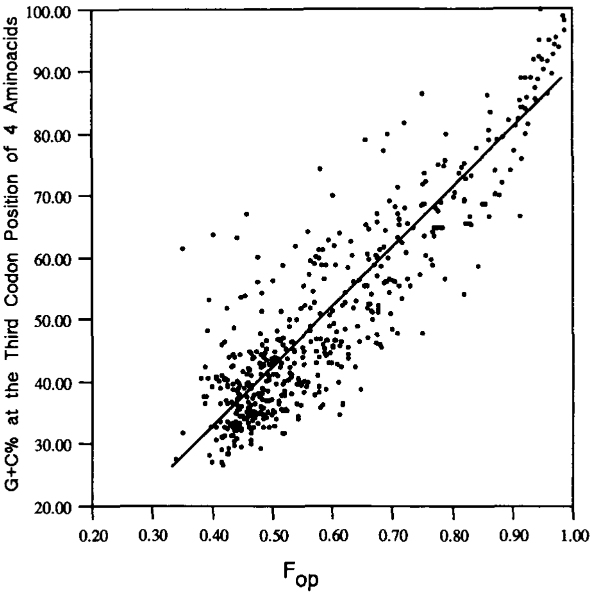

**FIGURE 9.** Correlation between *S. cerevisiae* $F_{op}$ and G+C% at the third codon position for four amino acids (Lys, Phe, Asn, and Tyr).

must postulate an unknown mechanism by which codon-choice patterns are roughly equalized among different genes (codon dialect) of the organism with the moderate G+C% genome.

On the basis of the fact that cellular processes of protein synthesis require large amounts of energy and mass, a molecular mechanism was proposed by which the codon choice has been constrained during evolution by the tRNA content. This can be further explained regarding the correlation between the extent of codon bias (accent of dialect) and protein production level. Highly expressed genes use mostly the codons optimal for translation efficiency. As protein production decreases, so does optimization (Table 3, Figure 6). This correlation can be explained on the basis of the following two viewpoints.

First, as pointed out at the beginning of this chapter, cellular processes of protein synthesis require large amounts of energy and mass. Based on analysis of codon usages of *E. coli* and *S. cerevisiae,* we proposed that a codon dialect should reflect an organism's strategy for producing large amounts of proteins with a minimal load.[4,21] The amount of protein production is known to be closely related with the amount of mRNA production. If codons translated by minor isoacceptors are frequently used in a highly expressed gene, ribosomes perform the uneconomical task of finding the proper tRNA present in small quantities for many mRNA molecules of this gene. Energy (GTP) seems to be consumed during this waiting period for "proofreading": a certain level of GTP hydrolysis is involved in rejecting tRNAs incorrectly bound to ribosomes, especially near-cognate tRNAs whose anticodon sequences are similar to and confused with that of the proper tRNA.[56-58] If a synonymous mutation from an optimal codon (recognized by a major tRNA) to a nonoptimal codon (recognized by a minor tRNA) occurs, the collision frequency of ribosomes increases with incorrect, but near-cognate tRNAs, e.g., major isoacceptors, at the respective codon; and therefore, the level of GTP hydrolysis would also increase. The resultant loss of GTP energy and of productive ribosome working time would result in phenotypic effects such as decrease in growth rate and/or viability. This should become important as the protein production levels (and mRNA levels) increase, which would result in the obvious codon bias.[4,21]

Second, a codon choice pattern may be involved in determining the amount of protein produced from a gene. In addition to the regulatory portions of a gene such as the promoter or operator, the coding region may also be involved in determining gene expressivity. For example, there are significant numbers of *E. coli* genes in which nonoptimal codons characteristically cluster at the beginning portion of the gene.[20,59-61] This may be related to a regulatory mechanism of gene expression in which the translation efficiency in the beginning portion of a gene is involved, e.g., autogeneous repression. Ribosomal frameshifting[62] may also be related to the clustered occurrence of nonoptimal codons.

Both of these views are related in part to each other. However, we think that the first possibility is more important in establishing the general correlation between codon bias and gene expressivity, since gene expressivity can be regulated more efficiently at the transcriptional level, except in cases of translation-transcription coupled regulation. In the following section, the first possibility will be discussed from an evolutionary viewpoint. The correlation between codon bias and gene expressivity was discussed from various view points by many groups.[37-39,55,63] Dynamic aspects of protein synthesis, e.g., tRNA cycling, has also been discussed on a quantitative basis.[64-66]

## B. EVOLUTIONARY VIEWPOINTS

In the previous section, a synonymous change from the optimal to nonoptimal codon is postulated to be slightly deleterious and that in the opposite direction to be slightly advantageous. The level of the fitness change should decrease as the protein production decreases (because of the decrease in amounts of mRNA). If the absolute value of fitness change is below a certain level, the mutation can be regarded as neutral,[67] even when it causes change in translation efficiency. The proportion of mutations that can be regarded as neutral or nearly neutral should be larger for genes with low gene expressivity than for genes with high expressivity. The present DNA sequence may represent an equilibrium or near equilibrium between the selective force and random drift of a neutral mutation, and $F_{op}$ (extent of codon bias) may be an index of this balancing point.[20,21] The mathematical evaluation of codon usage and tRNA content from an evolutionary view has been published by other groups.[68,69]

We previously proposed that the evolutionary view can be examined by analyzing the rate of synonymous substitution (nucleotide substitution that does not cause amino acid replacement).[4,18] Table 2 shows that tRNA populations of *E. coli* and *S. typhimurium* are essentially the same. We can, therefore, examine the effect of this common constraint imposed by the tRNA population on the rate of synonymous substitution between these organisms. That is, we can examine whether this constraint accelerates or decelerates the substitution rate. By analyzing all pairs of protein genes thus far available for both organisms, this constraint was found to decelerate, rather than accelerate, the synonymous substitution rate; i.e., synonymous substitution in moderately or poorly expressed genes is clearly higher (e.g., by severalfold) than that of highly expressed genes.[4,18] This is consistent with the prediction of the view based on the neutral theory of evolution.[67] A similar conclusion has been reported by analyzing more than 20 pairs of enterobacterial genes.[70] In the case of synonymous substitution between taxonomically distant organisms, changes in the tRNA population during evolution should be taken into account. The effects of such change on codon usage have been discussed previously.[16,20,69]

## VIII. CODON CHOICES IN GENES OF OTHER UNICELLULAR ORGANISMS AND DISTINCTION FROM HIGHER EUKARYOTES

### A. UNICELLULAR ORGANISMS

Since the same conclusions concerning codon choice were drawn for the prokaryotes, *E. coli* and *S. typhimurium,* and the eukaryote, *S. cerevisiae,* these observations should be true for a wide range of organisms.[21] When codon usage in various prokaryotes was examined, the synonymous codon-choice patterns of *E. coli* were found to be fairly similar to those of other Enterobacteriaceae, but significantly different from those of *Anabaena, Bacillus, Clostridium, Halobacterium,* and *Streptococcus.*[22,23] The latter organisms are all taxonomically distant from Enterobacteriaceae. These variations should be mainly due to differences in genome G+C% and tRNA isoacceptor populations.

### B. DISTINCTION FROM HIGHER EUKARYOTES

The general characteristics of codon usage of unicellular organisms are significantly different from those of higher eukaryotes, especially those of higher vertebrates.[4] When codon choice patterns of higher vertebrate genes (even those of one organism) are compared, they are often shown to be quite different, and thus, the organism-specific codon-choice pattern (codon dialect) is usually difficult to detect. For example, by extensively searching the GenBank database, we calculated codon usage in about 1900 human sequences (see Figure 8f). The highest G+C% in the third codon position was 97%, and that of about 250 sequences was 80% or more. The lowest G+C% was 27%, and that in about 150 sequences was 40% or less.[10] This evident diversity in the third codon G+C% was attributed to long range G+C% mosaic structures of the genome that are possibly related to chromosomal banding patterns.[5-10] Regarding codon choice patterns, the distinction between unicellular microorganisms and multicellular higher eukaryotes, rather than the distinction between prokaryotes and eukaryotes, is emphasized.[4] The reason for this is as follows. In unicellular organisms, most, if not all, genes are expressed in individual cells. To maintain the efficiency of translation processes, e.g., to save GTP energy used in the proofreading process, these genes have an analogous codon-choice pattern (codon dialect) that fits the tRNA population. In the case of higher eukaryotes, each organism is composed of an enormous number of cells. Many of the cells are highly differentiated, and a restricted spectrum of genes is expressed in individual cells. The target of Darwinian selection is the organism instead of individual cells. Thus, a fitness change caused by a synonymous change in an ordinary gene should be extremely smaller than that for unicellular organisms. A synonymous change to a codon with higher tRNA availability may save a certain amount of GTP energy and effective working time of ribosomes *in a restricted portion of cells*. However, the contribution of this change *in overall fitness of an organism* should be extremely small, so it would presumably be counted as a neutral mutation. This contrasts with the case of unicellular organisms where the cell itself is the direct target of selection. Codon usage in ordinary genes of multicellular organisms is therefore thought to be less stringently constrained by tRNA availability than that in genes of unicellular organisms. When some constraints from factors other than translation efficiency exist, e.g., long range G+C% mosaic structures of the genome, their codon choices presumably follow the constraints. The evident diversity of the third codon G+C% and the correlation of this diversity with the G+C% mosaic structures may reflect the weaker constraint imposed by tRNA content.[4]

## REFERENCES

1. **Maaløe, O.,** Past, present and future trends, in *Control of Ribosome Synthesis (Alfred Benzon Symposium IX)*, Kjeldgaard, N. C. and Maaloe, O., Eds., Munksgaard, Copenhagen, 1976, 15.
2. **Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A.,** Codon catalog usage and the genome hypothesis, *Nucl. Acids Res.*, 8, r49, 1980.
3. **Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R.,** Codon catalog usage is a genome strategy modulated for gene expressivity, *Nucleic Acids Res.*, 9, r43, 1981.
4. **Ikemura, T.,** Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, 2, 13, 1985.
5. **Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F.,** The mosaic genome of warm-blooded vertebrates, *Science*, 228, 953, 1985.

6. **Bernardi, G.,** The isochore organization of the human genome, *Annu. Rev. Genet.,* 23, 637, 1989.
7. **Aota, S. and Ikemura, T.,** Diversity in G+C content at the third position of codons in vertebrate genes and its cause, *Nucleic Acids Res.,* 14, 6345, 8702 (Erratum), 1986.
8. **Ikemura, T. and Aota, S.,** Global variation in G+C content along vertebrate genome DNA: possible correlation with chromosome band structures, *J. Mol. Biol.,* 203, 1, 1988.
9. **Ikemura, T., Wada, K., and Aota, S.,** Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions, *Genomics,* 8, 207, 1990.
10. **Ikemura, T. and Wada, K.,** Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers: relation between nucleotide sequence data and cytogenetic data, *Nucleic Acids Res.,* 19, 4333, 1991.
11. **Post, L. E. and Nomura, M.,** DNA sequences from the *str* operon of *Escherichia coli, J. Biol. Chem.,* 255, 4660, 1980.
12. **Li, W.-H., Luo, C.-C., and Wu, C.-I.,** Evolution of DNA sequences, in *Molecular Evolutionary Genetics,* MacIntyre, R. J., Ed., Plenum Press, New York, 1, 1985.
13. **Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., and Wright, F.,** Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens;* A Review of the Considerable Within-Species Diversity, *Nucleic Acid Res.,* 16, 8207, 1988.
14. **Andersson, S. G. E. and Kurland, C. G.,** Codon preferences in free-living microorganisms, *Microbiol. Rev.,* 54, 198, 1990.
15. **Osawa, S. and Jukes, T. H.,** Evolution of the genetic code as affected by anticodon content, *Trends Genet.,* 4, 191, 1988.
16. **Ikemura, T.,** Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.,* 146, 1, 1981.
17. **Ikemura, T.,** Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes, *J. Mol. Biol.,* 158, 573, 1982.
18. **Ikemura, T.,** Codon usage, tRNA content, and rate of synonymous substitution, in *Population Genetics and Molecular Evolution,* Ohta, T. and Aoki, K., Eds., Japan Sci. Soc. Press, Tokyo, 1985, 385.
19. **Ikemura, T.,** The frequency of codon usage in *E. coli* genes: correlation with abundance of cognate tRNA, in *Genetics and Evolution of RNA polymerase, tRNA and Ribosomes,* Osawa, S., Ozeki, H., Uchida, H., and Yura,T., Eds., University of Tokyo Press, Tokyo, Elsevier/North Holland, Amsterdam, 1980, 519.
20. **Ikemura, T.,** Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *J. Mol. Biol.,* 151, 389, 1981.
21. **Ikemura, T. and Ozeki, H.,** Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents, *Cold Spring Harbor Symp. Quant. Biol.,* 47, 1087, 1983.
22. **Wada, K., Aota, S., Tsuchiya, R., Ishibashi, F., Gojobori, T., and Ikemura, T.,** Codon usage tabulated from the GenBank genetic sequence data, *Nucleic Acids Res.,* 18, 2367, 1990.
23. **Wada, K., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T., and Ikemura, T.,** Codon usage tabulated from the GenBank genetic sequence data, *Nucleic Acids Res.,* 19, 1981, 1991.
24. **Ikemura, T. and Dahlberg, J. E.,** Small ribonucleic acids of *Escherichia coli.* I. Characterization by polyacrylamide gel electrophoresis and fingerprint analysis, *J. Biol. Chem.,* 248, 5024, 1973.
25. **Ikemura, T. and Nomura, M.,** Expression of spacer tRNA genes in ribosomal RNA transcription units carried by hybrid ColE1 plasmid in *E. coli, Cell,* 11, 779, 1977.

26. **Ikemura, T.**, Purification of RNA molecules by gel techniques, *Methods Enzymol.*, 180, 14, 1989.
27. **Komine, Y., Adachi, T., Inokuchi, H., and Ozeki, H.**, Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12, *J. Mol. Biol.*, 212, 579, 1990.
28. **Varenne, S., Buc, J., Lloubes, R., and Lazdunski, C.**, Translation is a non-uniform process; effect of tRNA availability on the rate of elongation of nascent polypeptide chains., *J. Mol. Biol.*, 180, 549, 1984.
29. **Holm, L.**, Codon usage and gene expression, *Nucleic Acids Res.*, 14, 3075, 1986.
30. **Nishimura, S.**, Modified nucleosides and isoaccepting tRNA, in *Transfer RNA*, Altman, S., Ed., MIT Press, Cambridge, MA, 1978, 168.
31. **Shpaer, E. G.**, Constraints on codon context in *Escherichia coli* genes: their possible role in modulating the efficiency of translation, *J. Mol. Biol.*, 188, 555, 1986.
32. **Yarus, M. and Folley, L. S.**, Sense codons are found in specific contexts, *J. Mol. Biol.*, 182, 529, 1985.
33. **Grosjean, H., Sankoff, D., Min Jou, W., and Cedergren, R. J.**, Bacteriophage MS2 RNA: a correlation between stability of the codon: anticodon interaction and the choice of code words, *J. Mol. Evol.*, 12, 113, 1978.
34. **Grosjean, H. and Fiers, W.**, Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes, *Gene*, 18, 199, 1982.
35. **Parker, J.**, Errors and alternatives in reading the universal genetic code, *Microbiol. Rev.*, 53, 273, 1989.
36. **VanBogelen, R. A., Hutton, M. E., and Neidhardt, F. C.**, Gene-protein database of *Escherichia coli* K-12: edition 3, *Electrophoresis*, 11, 1131, 1990.
37. **Bennetzen, J. L. and Hall, B. D.**, Codon selection in yeast, *J. Biol. Chem.*, 257, 3026, 1982.
38. **Gouy, M. and Gautier, C.**, Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.*, 10, 7055, 1982.
39. **Sharp, P. M. and Li, W.-H.**, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res.*, 15, 1281, 1987.
40. **Robinson, M., Lilley, R., Little, S. Emtage, J. S., Yarranton, G., Stephens, P., Millican, A., Eaton, M., and Humphreys, G.**, Codon usage can affect efficiency of translation of genes in *Escherichia coli*, *Nucleic Acids Res.*, 12, 6663, 1984.
41. **Pedersen, S.**, *Escherichia coli* ribosomes translate in vivo with variable rate, *EMBO J.*, 3, 2895, 1984.
42. **Bonekamp, F., Andersen, H. D., Christensen, T., and Jensen, K. F.**, Codon-defined ribosomal pausing in *Escherichia coli* detected by using the *pyrE* attenuator to probe the coupling between transcription and translation, *Nucleic Acids Res.*, 13, 4113, 1985.
43. **Bonekamp, F. and Jensen, K. F.**, The AGG codon is translated slowly in *E. coli* even at very low expression levels, *Nucleic Acids Res.*, 16, 3013, 1988.
44. **Williams, D. P., Regier, D., Akiyoshi, D., Genbauffe, F., and Murphy, J. R.**, Design, synthesis and expression of a human interleukin-2 gene incorporating the codon usage bias found in highly expressed *Escherichia coli* genes, *Nucleic Acids Res.*, 16, 10453, 1988.
45. **Sørensen, M. A., Kurland, C. G., and Pedersen, S.**, Codon usage determines translation rate in *Escherichia coli*, *J. Mol. Biol.*, 207, 365, 1989.
46. **Nichols, B. P., Blumenberg, M., and Yanofsky, C.**, Comparison of the nucleotide sequence of *trpA* and sequences immediately beyond the Trp operon of *Klebsiella aerogenes, Salmonella typhimurium* and *E. coli*, *Nucleic Acids Res.*, 9, 1743, 1981.
47. **Nichols, B. P., Miozzari, G. F., vanCleemput, M., Bennet, G. N., and Yanofsky, C.**, Nucleotide sequences of the *trpG* regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens*, *J. Mol. Biol.*, 142, 503, 1980.

48. **Yanofsky, C. and vanCleemput, M.,** Nucleotide sequence of *trpE* of *Salmonella typhimurium* and its homology with the corresponding sequence of *Escherichia coli, J. Mol. Biol.,* 155, 235, 1982.
49. **Wada, A. and Suyama, A.,** Third letters in codons counterbalanced the (G+C)–content of their first and second letters, *FEBS Lett.,* 188, 291, 1985.
50. **Bernardi, G. and Bernardi, G.,** Compositional constraints and genome evolution, *J. Mol. Evol.,* 24, 1, 1986.
51. **Muto, A. and Osawa, S.,** The guanine and cytosine content of genomic DNA and bacterial evolution, *Proc. Natl. Acad. Sci. U.S.A.,* 84, 166, 1987.
52. **Sueoka, N.,** Directional mutation pressure and neutral molecular evolution, *Proc. Natl. Acad. Sci. U.S.A.,* 85, 2653, 1988.
53. **Osawa, S., Ohama, T., Yamao, F., Muto, A., Jukes, T. H., Ozeki, H., and Umesono, K.,** Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codons sets, *Proc. Natl. Acad. Sci. U.S.A.,* 85, 1124, 1988.
54. **Filipski, J.,** Evolution of DNA sequence contribution of mutational bias and selection to the origin of chromosome compartments, in *Advances in Mutagenesis Research,* Vol. 2, Obe, G., Ed., Springer-Verlag, New York, 1990, chap. 1.
55. **Shields, D. C. and Sharp, P. M.,** Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases, *Nucleic Acids Res.,* 15, 8023, 1987.
56. **Hopfield, J. J.,** Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity, *Proc. Natl. Acad. Sci. U.S.A.,* 71, 4135, 1974.
57. **Thompson, R. C. and Stone, P. J.,** Proofreading of the codon-anticodon interaction on ribosomes, *Proc. Natl. Acad. Sci. U.S.A.,* 74, 198, 1977.
58. **Thompson, R. C., Dix, D. B., Gerson, R. B., and Karim, A. M.,** A GTPase reaction accompanying the rejection of Leu-tRNA$_2$ by UUU-programmed ribosomes, *J. Biol. Chem.,* 256, 81, 1981.
59. **Nomura, M., Yates, J. L., Dean, D., and Post, L. E.,** Feedback regulation of ribosomal protein gene expression in *Escherichia coli:* structural homology of ribosomal RNA and ribosomal protein mRNA, *Proc. Natl. Acad. Sci. U.S.A.,* 77, 7084, 1980.
60. **Burns, D. M. and Beacham, I. R.,** Rare codons in *E. coli* and *S. typhimurium* signal sequences, *FEBS Lett.,* 189, 318, 1985.
61. **Chen, G. T. and Inouye, M.,** Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes, *Nucleic Acids Res.,* 18, 1465, 1990.
62. **Weiss, R. B., Dunn, D. M., Atkins, J. F., and Gesteland, R. F.,** Ribosomal frameshifting from –2 to +50 nucleotides, *Prog. Nucl. Acid Res. Mol. Biol.,* 39, 159, 1990.
63. **Dix, D. B. and Thompson, R. C.,** Codon choice and gene expression: synonymous codons differ in translational accuracy, *Proc. Natl. Acad. Sci. U.S.A.,* 86, 6888, 1989.
64. **von Heijne, G. and Blomberg, C.,** The concentration dependence of the error frequencies and some related quantities in protein synthesis, *J. Theor. Biol.,* 78, 113, 1979.
65. **Gouy, M. and Grantham, R.,** Polypeptide elongation and tRNA cycling in *Escherichia coli:* a dynamic approach, *FEBS Lett.,* 115, 151, 1980.
66. **Varenne, S. and Lazdunski, C.,** Effect of distribution of unfavorable codons on the maximum rate of gene expression by an heterologous organism, *J. Theor. Biol.,* 120, 99, 1986.
67. **Kimura, M.,** *The Neutral Theory of Molecular Evolution,* Cambridge University Press, Cambridge, 1983.
68. **Bulmer, M.,** Coevolution of codon usage and transfer RNA abundance, *Nature (London),* 325, 728, 1987.
69. **Shields, D. C.,** Switches in species-specific codon preferences: the influence of mutation biases, *J. Mol. Evol.,* 31, 71, 1990.
70. **Sharp, P. M. and Li, W.-H.,** The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias, *Mol. Biol. Evol.,* 4, 222, 1987.