Analysis of Codon Usage Patterns of Bacterial Genomes Using the Self-Organizing Map

Huai-Chun Wang,† Jonathan Badger,* Paul Kearney,* and Ming Li**

*Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada; and †Institute of Basic Medical Sciences, Beijing, China

Codon usage varies both between organisms and between different genes in the same organism. This observation has been used as a basis for earlier work in identifying highly expressed and horizontally transferred genes in *Escherichia coli.* In this work, we applied Kohonen's self-organizing map to analysis of the codon usage pattern of the *Escherichia coli, Aquifex aeolicus, Archaeoglobus fulgidus, Haemophilus influenzae* Rd., *Methanococcus jannaschii, Methanobacterium thermoautotrophicum,* and *Pyrococcus horikoshii* genomes for evidence of highly expressed genes and horizontally transferred genes. All of the analyzed genomes had a clear category of horizontally transferred genes, and their apparent percentages ranged from 7.7% to 21.4%. The apparent percentage of highly expressed genes ranges from 0% to 11.8%. A clustering of average codon usage of main gene categories of the seven genomes showed an interesting mixing of gene classes in four thermophilic/hyperthermophilic organisms, *A. aeolicus, A. fulgidus, M. thermoautotrophicum,* and *P. horikoshii,* which suggests possible origins of their horizontally transferred genes as well as the need for adaptation to a specific environment. Further classification of the three gene categories in *E. coli* and *H. influenzae* according to gene function revealed that genes involved in communication (such as regulation and cell process) and structure (cell structure and structural proteins) are more likely to be horizontally transferred than are genes involved in information (transcription, translation, and related processes) and in some groups of energy (such as energy metabolism and carbon compound catabolism).

Introduction

The choice of synonymous codons within a genome is not random. In bacteria, the codon usage of genes is correlated with their expression level (Ikemura 1981, 1985). Another source of variation in codon usage is horizontal gene transfer, as the transferred (or ''alien'') genes tend to have a codon usage different from that of the host organism (Médigue et al. 1991; Lawrence and Ochman 1997; Karlin, Mrazek, and Campbell 1998; Badger 1999; Mathe et al. 1999). Therefore, comparative analysis of intraspecies codon usage may provide insights into genomic evolution (Nakamura, Gojobori, and Ikemura 1999). Following the pioneering work of Grantham et al. (1980), many multivariate statistical methods have been applied to the analysis of codon usage of various organisms. In essence, these procedures consider gene sequences as points in multidimensional space, with each dimension representing the frequency of a given codon. In this work, we apply an unsupervised neural network algorithm, the self-organizing map (SOM) (Kohonen 1982) to the analysis of codon usage patterns. The SOM is an effective software tool for the visualization of high-dimensional data. It converts complex nonlinear statistical relationships between high-dimensional data items into simple geometric relationships that can be viewed in two dimensions (Kohonen 1997). We show that the SOM, which efficiently clusters and projects data, is an excellent tool for analyzing codon usage.

Key words: codon usage, self-organizing map, genome, gene function, horizontal gene transfer.

Address for correspondence and reprints: Huai-chun Wang, Institute of Basic Medical Sciences, 27 Taiping Road, Beijing 100850, China. E-mail: wanghc@nic.bmi.ac.cn.

Mol. Biol. Evol. 18(5):792–800. 2001

Materials and Methods

Sequence and Codon Usage Data

The protein-coding sequences of seven complete genomes (those of *Aquifex aeolicus* [Deckert et al. 1998], *Archaeoglobus fulgidus* [Klenk et al. 1997], *Escherichia coli* [Blattner et al. 1997], *Haemophilus influenzae* [Fleischmann et al. 1995], *Methanococcus jannaschii* [Bult et al. 1996], *Methanobacterium thermoautotrophicum* [Smith et al. 1997], and *Pyrococcus horikoshii* [Kawarabayasi et al. 1998]) were retrieved from the NCBI FTP server. As short genes may have a limited codon sampling which could possibly distort our results, genes shorter than 75 codons (approximately 4% of the data) were excluded from the analysis. During the calculation of codon usage, the three stop codons and the two codons encoding Met and Trp were excluded. The codon usage (normalized for each amino acid) was then tabulated for the remaining 59 codons. The G+C content of each sequence was also tabulated. The tabulation program is available from the authors on request.

Functional Groups of Bacterial Genes

The functional classification of the *E. coli* protein-coding genes (Riley 1993) was obtained from http://www.genetics.wisc.edu/html/orftables/index.html. Some groups of similar functions were merged for our analysis, leaving the following 20 functional groups: amino acid biosynthesis and metabolism (A); biosynthesis of cofactors, prosthetic groups and carriers (B); cell structure (C); structural proteins (C1); energy metabolism (E); carbon compound catabolism (E1); phage, transposon, or plasmid (F); central intermediary metabolism (I); fatty acid and phospholipid metabolism (L); membrane proteins (M); nucleotide biosynthesis and metabolism (N); other known genes (O); cell processes (including adaptation, protection, and putative chaperones)

(P); DNA replication (R); transcription, RNA processing, and degradation (S); translation, posttranslational modification (T); hypothetical, unclassified, unknown (U); transport and binding proteins and putative transport proteins (X); putative enzymes (Y); regulatory function and putative regulatory proteins (Z). The original group R in Riley's (1993) scheme contains not only DNA replication genes, but also DNA recombination, restriction-modification, and repair genes as well. Because previous work has suggested that such genes may be of alien origin (Lawrence and Ochman 1997; Mrazek and Karlin 1999), they were removed from this group and added to the O group (other known genes). Sixteen of the 20 functional groups can be further grouped as four superclasses (Tamames et al. 1996): Communication (containing P and Z), Energy (containing A, B, E, E1, I, L, N, and X), Information (containing R, S, and T), and Structure (containing C, C1, and M).

The functional classification of *H. influenzae* genes was obtained from http://www.tigr.org/tdb/CMR/ghi/htmls/SplashPage.html. It has a scheme similar to that of *E. coli,* except that it lacks the four categories of carbon compound catabolism (E1), structural proteins (C1), membrane proteins (M), and putative enzymes (Y), and there is an additional group (denoted group P1) which includes functions of protein secretion and trafficking, protein folding, stabilization, modification, repair, and degradation. Thus, for *H. influenzae,* the functional superclasses are as follows: Communication (containing P, P1, and Z), Energy (containing A, B, E, I, L, N, and X), Information (containing R, S, and T), and Structure (containing only C).

Self-Organizing Map

The SOM is an unsupervised neural network method that is particularly useful for data visualization (Kohonen 1997). The SOM algorithm simultaneously finds a representative set of reference vectors of the training data and positions them on a regular two-dimensional grid of neurons. It can be thought of as a flexible net that is spread into the data ''cloud.'' Because the net is two-dimensional, it can easily be visualized. The mapping from the input space onto the grid of neurons is learned from the training data samples by a simple stochastic learning process whereby the SOM neurons (the reference vectors) are adjusted by small steps with respect to the input vectors. A thorough description of the algorithm can be found elsewhere (Kohonen 1997; Marabini and Carazo 1994). The following is a summary of the method:

0. Convert the gene sequences into normalized vectors of codon usage $\mathbf{x}(t)$. Each component of input vectors is scaled with the following formula so that its mean becomes 0 and its variance becomes one.

$$x_i^{\mathrm{new}} = (x_i^{\mathrm{old}} - o_i)/s_i,$$

where x_i^{old} is the original value of component i of the data vector $\mathbf{x}$, o_i is the mean of values of x_i, and s_i is their standard deviation. The scaling is used to ensure that no component has excessive influence on the learning results just because of its greater variance or larger absolute value (Vesanto 1999).

1. Set up the map size and initialize the reference vectors $\mathbf{m}_j$ associated with each output node to random values.
2. Present one input vector $\mathbf{x}(t)$.
3. Compute the distance d_j between the input vector $\mathbf{x}(t)$ and each output node j.
4. Select the winner output node j that is associated with the reference vector $\mathbf{m}_j(t)$ that minimizes d_j.
5. Update reference vectors for node j and its neighborhood by the function

$$\mathbf{m}_j(t + 1) = \mathbf{m}_j(t) + \alpha(t)h_{cj}(t)\ (\mathbf{x}(t) - \mathbf{m}_j(t)),$$

where $\alpha(t)$ is the learning rate and h_{cj} is a neighborhood kernel. $\alpha(t)$ is a decreasing function that controls the magnitude of the changes with the time $(0 < \alpha(t) < 1)$. The neighborhood kernel is a function that determines the area of effect the input vector has on the map. It is greatest for the winner neuron j, the reference vector of which is closest to the input vector, and monotonically decreases the farther away neuron i is from neuron j on the map grid.

6. Go to step 2 and repeat the cycle until convergence.

For our analysis, we used the SOM Toolbox (http://www.cis.hut.fi/projects/somtoolbox), a newly developed MATLAB-based (The MathWorks, Inc.) SOM.

Visualization of the Map

The SOM Toolbox provides several ways to display a map. The two methods used in this paper are the U-matrix (unified distance matrix) (Ultsch and Siemon 1990) and component planes. The U-matrix uses color to show the distances between neighboring map units: longer distances are represented by shades of yellow and red, while shorter distances are represented by shades of blue. Therefore, clusters of genes with similar codon usages appear as blue areas, while areas in which the codon usage is changing rapidly between adjacent units appear as yellow and red areas. Because horizontally transferred genes and highly expressed genes have different-from-normal codon usage, they are expected to be separated on the U-matrix.

Clusters of genes can be made easier to recognize by labeling the map with known gene names or functional classes. As genes encoding ribosomal proteins are known to be highly expressed, these genes were marked on the map to serve as a landmark for clusters of highly expressed genes in the projection. Similarly, as genes such as insertion sequences, transposases, restriction-modification endonucleases, and flagella-related genes are often thought to be horizontally transferred, they can serve as landmarks for presumed alien genes. For the *E. coli* genome, we also labeled those genes that were previously classified according to different expression levels (Sharp and Li 1986): very highly expressed genes

(45 ribosomal protein genes and 10 other genes from Sharp and Li's data); highly expressed genes (15 genes in Sharp and Li's data, of which 4 were not found in the current genome sequence); moderate codon bias (57 genes, of which 12 were not found), and low codon bias (58 genes, of which 6 were not found). The genes which were not found probably had their locus names changed over the years.

A component plane consists of the values of one vector component (representing one codon) in all map units and provides an idea of the spread of values of that component. Thus, by showing component planes of all synonymous codons for one amino acid, it is easy to see which codons are more frequently used than the others. An inspection of all component planes indicates which components are correlated. Correlations between codons are revealed as similar patterns in identical positions of the component planes. By comparing the U-matrix with component planes, it is possible to identify which components contribute strongly to a cluster observed in the U-matrix.

Shannon Uncertainty

Another way to measure codon usage is through the information theoretical notion of Shannon uncertainty, or entropy. Shannon uncertainty can be thought of as a measure of randomness; a fair die has a higher Shannon uncertainty than a loaded die. In terms of codon usage, unbiased codon usage has a higher Shannon uncertainty than biased usage. The advantage of using Shannon uncertainty is that it allows a complex source of bias to be represented by a single statistic. The Shannon uncertainty H for M possible outcomes is given by the following formula (Shannon 1948):

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i,$$

where P_i is the probability of the i^{th} outcome. To apply Shannon uncertainty to codon usage, we calculated the average index of the 20 amino acids as follows (Badger 1999):

$$H' = -\sum_{a=1}^{20} f_a \sum_{ca=1}^{na} f_{ca} \log_2 f_{ca},$$

where f_a is the frequency of codons encoding amino acid a, n_a is the number of synonymous codons for amino acid a, and f_{ca} is the frequency of codon c among synonymous codons of amino acid a. The term $-\Sigma\, f_{ca}\log_2 f_{ca}$ measures the codon bias for codons encoding amino acid a. H' is the average codon bias among all 20 amino acids. When the frequency of every codon is 1/64, $H' = 1.76$ bits.

Results

Gene Categories in *E. coli*

A self-organizing map of the codon usage of 4,135 *E. coli* genes is shown in figure 1. Two areas (the top quarter and the lower right corner) have clearly different hues from the rest of the map. By drawing borderlines separating the yellow/red areas and surrounding blue areas, we obtained three major categories of genes. The lower right corner (marked "V" for very highly expressed genes and "H" for highly expressed genes) was identified as a class of genes with highly biased codon usage that presumably represents highly expressed genes. It contains 357 genes, including all 45 ribosomal genes longer than 75 codons, 18 amino acyl tRNA synthetases, 5 chaperonins, all very highly expressed genes, and most highly expressed genes that were previously identified (Sharp and Li 1986). It also contains 23 genes that were previously identified as having moderate codon bias (Sharp and Li 1986). None of genes of low codon bias in Sharp and Li's data were found in this cluster. The Shannon uncertainty for this class of genes is 1.25 bits, 0.26 bits down from the average of 1.51 bits.

The top cluster contains 779 genes, of which 461 are of unidentified function, 24 are transposases (labeled

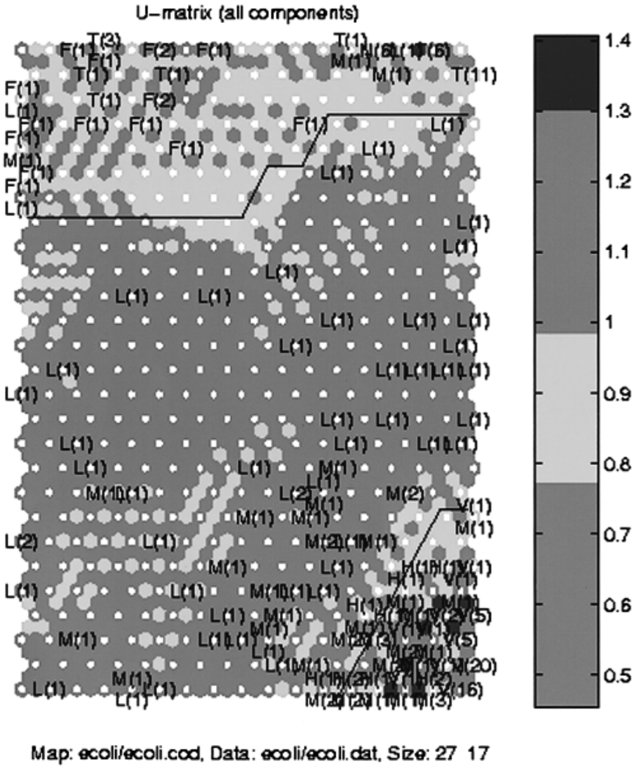

FIG. 1.—U-matrix visualization of a self-organizing map of *Escherichia coli* codon usage (see data supplement at http://www.molbiolevol.org for color figures). Hexagons having white dots indicate locations of map units and are colored according to the medians of the surrounding hexagon. Hexagons without dots show the distances between two neighboring map units. The color scale for distance is shown as a color bar on the right. Borderlines between the classes are highlighted. The lower right corner is presumably a class of very highly (V) and highly (H) expressed genes. It also contains some genes previously identified (Sharp and Li 1986) as being of moderate codon bias (M). The bright area of upper seven rows was identified as a class of presumed alien genes by checking sequence annotations. Some known alien genes are labeled on this area as F (genes of bacteriophage origin), N (insertion sequences), and T (transposases). The large blue area represents a class of normally expressed genes and genes previously identified as being of moderate (M) codon bias and low (L) codon bias (Sharp and Li 1986). A number in brackets is the number of labeled genes in a given unit.

''T'' on the map), 6 are insertion sequences (labeled ''N'' on the map), and 16 are of apparent bacteriophage origin. Four hundred sixty-three of the 779 genes have previously been identified as alien genes (Lawrence and Ochman 1998), which suggests that this area is a cluster of alien genes. The Shannon uncertainty for this class of genes is 1.60 bits, 0.09 bits up from the average. An inspection of the locations of the 799 putative alien genes on the *E. coli* chromosome revealed that 485 were in 150 sets of gene clusters of at least two contiguous genes each. The two largest adjacent clusters (*E. coli* genes b1137–b1148, b1154, b1155, and b1157–b1169) contain 27 almost-contiguous genes (with only 6 genes missed). This adjacency suggests that these are alien genes transferred as genomic fragments. This phenomenon has been observed in *H. influenzae* and Synechocystis (Mrazek and Karlin 1999). The remainder of the genes form a third class, genes with normal codon usage, containing 2,999 genes. Within this class are the majority of the genes previously classified as being of moderate or low bias by Sharp and Li (1986) (labeled ''M'' and ''L,'' respectively). Their codon usage is similar to the average codon usage of the genome, and the Shannon uncertainty of this class (1.49 bits) also is very close to the average of 1.51 bits.

Although the axes of a U-matrix have no intrinsic biological meaning, the vertical axis was found to be correlated with G+C content and average gene length, with genes of lower G+C content and shorter length found near the top of the map, and increasing G+C content and length toward the lower part (figure not shown). This corresponds with the fact that the presumed alien genes (on the upper side of the map) have a lower average G+C content (46.0%) and a much shorter average length (246 amino acids) than the genes with normal codon usage (52.9% G+C and 349 amino acids) and highly expressed genes (52.3% G+C and 338 amino acids). Note, however, that not all alien genes have reduced G+C content compared with nonalien genes. Some alien genes have very high G+C contents (Lawrence and Ochman 1997; Karlin, Mrazek, and Campbell 1998). The horizontal axis was found to have an increasing average G+C content on the left, maintaining a high G+C content in the middle and decreasing in G+C content on the right. There is no simple trend in average gene length along this axis.

From component planes of synonymous codons, one can identify which codons are more commonly used to encode an amino acid. Figure 2 shows the component planes of the six synonymous codons of arginine. It can be observed that CGC and CGT are the two most commonly used codons for coding arginine, while AGA and AGG are the least common. A comparison of images of component planes and the U-matrix shows that CGT (arginine), TCT (serine), GGT (glycine), TTC (phenylalanine), AAC (asparagine), TAC (tyrosine), and ATC (isoleucine) mainly occur in the class of presumed highly expressed genes. AGA (arginine), AGG (arginine), CTA (leucine), ACA (threonine), and ATA (isoleucine) occur mainly in the class of presumed alien genes. These codons contribute strongly to the formation of their respective categories.

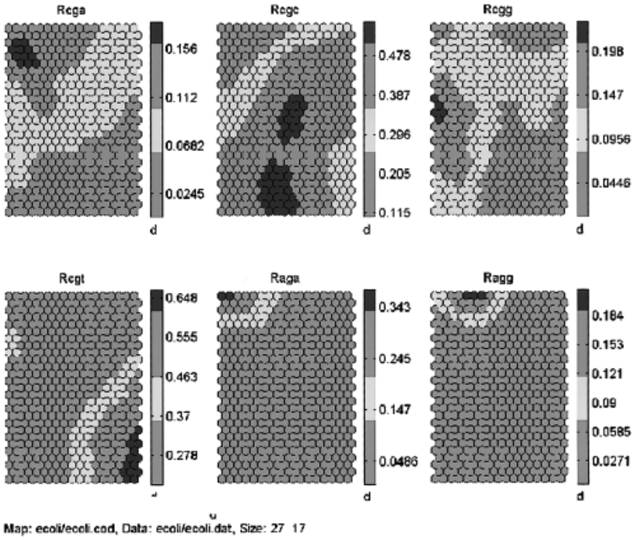

FIG. 2.—Component planes of the six synonymous codons of arginine (R) from an *Escherichia coli* codon usage map. Each hexagon represents one map unit, and its color (scaled on the color bar) indicates the value of the component in that unit (the frequency of the corresponding synonymous codon) (see data supplement at http://www.molbiolevol.org for color figures). Analogous hexagons on different images correspond to the same map unit. The planes indicate that CGC and CGT are the two most common codons for arginine, and AGA and AGG are the least common. As the lower right corner represents the cluster of genes with high codon bias (see fig. 1), and CGT has highest value exclusively in this corner, CGT contributes strongly to the formation of this class. Similarly, AGA and AGG contribute strongly to the formation of the class of genes with unusual codon usage.

Gene Categories in *A. aeolicus*, *A. fulgidus*, *H. influenzae*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii*

The SOM was used in a similar fashion to investigate the codon usage pattern of six other genomes: two eubacteria (*A. aeolicus* and *H. influenzae*) and four archaea (*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii*). Areas of presumed highly expressed genes were identified by observing where the ribosomal proteins clustered, and areas of presumed alien genes were identified by observing where genes believed to be of alien origin (e.g., insertion sequences, transposes, restriction-modification enzymes, and flagellar proteins) clustered. Usually, the bulk of the genes in this latter class are of unknown function. The Shannon uncertainty of the genes in a cluster also provides evidence for its identification, as highly expressed genes are expected to have below-average Shannon uncertainty, while presumed alien genes are expected to have above-average Shannon uncertainty. Three genomes (*H. influenzae*, *M. jannaschii*, and *M. thermoautotrophicum*) were found to contain a class of presumed highly expressed genes, and all of the genomes were found to contain a class of presumed alien genes. The identified gene categories of the six genomes, as well as that of *E. coli*, are summarized in table 1.

Table 1
Gene Categories Identified with the Self-Organizing Map

	All Genes				Presumed Class of High Expression				
Organism	No.[a]	Length[b]	GC%[c]	Shannon Uncertainty	No.[a]	%[d]	Length[b]	GC%[c]	Shannon Uncertainty
AA.	1,509 (1,522)	320	43.7	1.46	—	0.0	—	—	—
AF	2,249 (2,407)	295	49.4	1.50	—	0.0	—	—	—
EC	4,135 (4,289)	329	51.9	1.51	357	8.6	338	52.3	1.25
HI	1,629 (1,709)	319	38.8	1.36	193	11.8	294	40.1	1.21
MJ	1,654 (1,715)	292	31.9	1.22	165	10.0	268	36.1	1.15
MT	1,761 (1,868)	298	50.6	1.48	182	10.3	282	50.1	1.37
PH	2,032 (2,064)	280	42.3	1.53	—	0.0	—	—	—

	Presumed Class of Alien Genes					Class of Normal Expression				
Organism	No.	%	Length	GC%	Shannon Uncertainty	No.	%	Length	GC%	Shannon Uncertainty
AA.	239	15.8	293	38.7	1.50	1,270	84.2	325	44.5	1.44
AF	408	18.1	231	43.8	1.57	1,841	81.9	309	50.3	1.48
EC	779	18.8	246	46.0	1.60	2,999	72.5	349	52.9	1.49
HI	286	17.6	266	38.0	1.44	1,150	70.6	337	38.7	1.34
MJ	127	7.7	223	29.4	1.26	1,362	82.3	302	31.7	1.21
MT	377	21.4	243	45.4	1.55	1,202	68.3	319	52.0	1.46
PH	225	11.1	142	44.4	1.70	1,807	89.0	297	42.2	1.52

Note.—Organism abbreviations: AA, *Aquifex aeolicus*; AF, *Archaeoglobus fulgidus*; EC, *Escherichia coli*; HI, *Haemophilus influenzae* Rd.; MJ, *Methanococcus jannaschii*; MT, *Methanococcus thermoautotrophicum*; PH, *Pyrococcus horikoshii.*

[a] The first number is the number of all genes greater than 75 codons in the genome; the number in parentheses is the number of all protein-coding genes in the genome (including those of less than 75 codons).

[b] Average sequence length.

[c] Average percentage of G+C content.

[d] Percentage of genes in the class.

Clustering Genomes

An SOM of the average codon usage for the major gene categories (table 1) of the seven organisms is shown in figure 3. It can be seen that the two proteobacteria, *E. coli* and *H. influenzae*, form a codon usage cluster, as do the thermophilic organisms (with the exception of *M. jannaschii*). The gene categories from *A. aeolicus*, *A. fulgidus*, *M. thermoautotrophicum*, and *P. horikoshii* are intermingled on the map, suggesting that horizontal transfer events may have occurred between these organisms.

Interpretation of SOM Classes: Functional Classification of the *E. coli* and *H. influenzae* Gene Categories

For *E. coli*, each of the 4,135 genes in the three categories (highly expressed, normally expressed, and alien genes) was assigned to one of the 20 functional groups (A, B, C, C1, E, E1, F, I, L, M, N, O, P, R, S, T, U, X, Y, and Z; see *Materials and Methods*). The largest group was U (genes of unknown function), as it comprised 59.2% of the presumed alien cluster, 33.9% of the normal usage cluster, and 13.2% of the presumed highly expressed cluster. This cluster was excluded from further analysis, leaving 1,981 genes of normal codon usage, 310 presumed highly expressed genes, and 318 presumed alien genes. Figure 4 shows the distribution of the 19 functional groups across the three gene categories. Genes with normal codon usage, making up the largest of the three categories, are the majority in all functional groups except groups F (phage, transposon, or plasmid) and T (translation and posttranslational modification), in which presumed alien genes and presumed highly expressed genes, respectively, are in the majority. Groups C (cell structure), C1 (structural proteins), P (cell processes), Y (putative enzymes), and Z (regulatory proteins) also include larger-than-expected numbers of the presumed alien genes. Groups T, N (nucleotide biosynthesis and metabolism), and E (energy metabolism) are the most avoided by presumed alien genes.

The above classification was further generalized by considering four superclasses of gene functions: Communication, Energy, Information, and Structure (Tamames et al. 1996). Figure 5 shows the amount of each superclass in the three gene categories. The genes with normal usage, making up the majority of the genome, were the largest contributor to each superclass. The presumed highly expressed genes were found in higher-than-expected amounts in the Energy and Information superclasses and were rarely found in the Communication and Structure superclasses. Likewise, the presumed alien genes were found in excess in Communication, Energy, and Structure and rarely in Information.

In *H. influenzae*, 31% of the genes were of unidentified function. The remaining 1,123 functionally assigned genes (822 genes with normal codon usage, 163

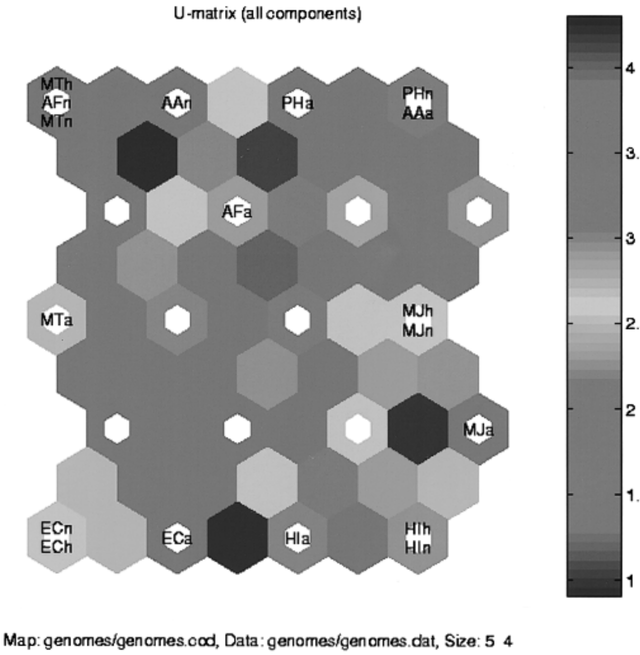

FIG. 3.—A self-organizing map of the average codon usage of the main gene categories of the seven organisms visualized with a U-matrix. Organisms are abbreviated by two capital letters (see table 1 for organism abbreviations). The small letters behind the organism abbreviations represent gene categories: ''a'' for the class of genes with unusual codon usage (presumably alien genes), ''h'' for the class of genes with highly biased codon usage, and ''n'' for the class of genes with normal codon usage (see data supplement at http://www.molbiolevol.org for color figures).

presumed highly expressed genes, and 138 presumed alien genes) had a functional distribution similar to that of *E. coli* (figure not shown).

Discussion

Comparison with Other Methods

Using factorial correspondence analysis (FCA), Médigue et al. (1991) found that the two-dimensional projection of the codon usage data of the 782 available *E. coli* coding sequences resembled a ''rabbit's head,'' with two ''ears'' on each side of a large ''head.'' The head was identified as a cluster of genes with normal codon usage, with the two ears being a cluster of genes with highly biased codon usage (presumably highly expressed genes) and a cluster of genes with unusual codon bias (presumably alien genes). This image was not changed when the principal component analysis (PCA) was used to project the codon usage of the complete *E. coli* genome (figure not shown). However, as boundaries of clusters in PCA projections are vague, one has to determine them manually (and somewhat arbitrarily). Although cluster boundaries in FCA and PCA can be inferred with clustering algorithms, such as the *k*-means method (Médigue et al. 1991; Mathe et al. 1999), the result of the classification depends on *k*, the number of clusters there are in the data. When this number is unknown for a data set, one has to determine it empirically. The clustering algorithm in SOM and the *k*-means method are closely related, but in the SOM the number of reference vectors can be chosen to be very large, irrespective of the number of clusters. The clusters found by SOM can be observed using the U-matrix visualization described in *Materials and Methods.* Table 2

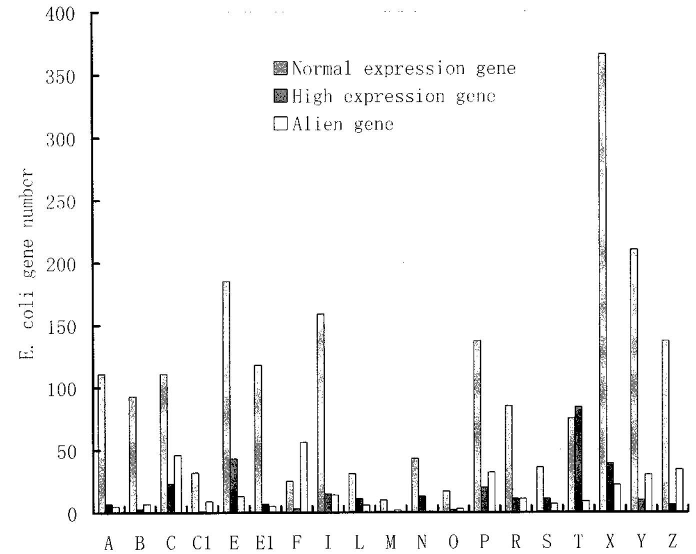

FIG. 4.—*Escherichia coli* gene categories grouped according to function (genes of unidentified function were excluded). Each block of the three bars represents, from left to right, genes with normal codon usage, genes with highly biased codon usage, and genes with unusual codon usage, respectively. The bar height represents the number of genes. The codes of the functional groups are as follows: A, amino acid biosynthesis and metabolism; B, biosynthesis of cofactors, prosthetic groups and carriers; C, cell structure; C1, structural proteins; E, energy metabolism; E1, carbon compound catabolism; F, phage, transposon, or plasmid; I, central intermediary metabolism; L, fatty acid and phospholipid metabolism; M, membrane proteins; N, nucleotide biosynthesis and metabolism; O, other known genes; P, cell processes; R, DNA replication; S, transcription, RNA processing, and degradation; T, translation and posttranslational modification; X, transport and binding proteins; Y, putative enzymes; Z, regulatory functions.

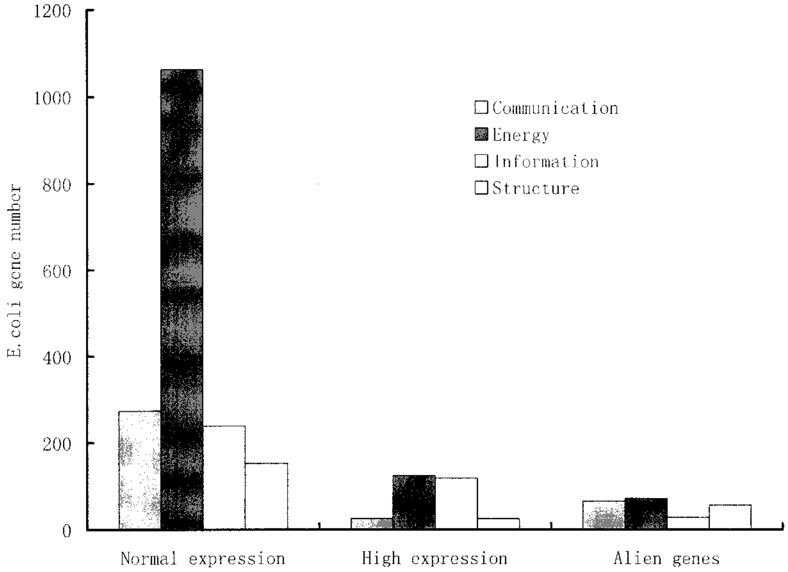

FIG. 5.—*Escherichia coli* genes grouped according to the four superclasses of functions (Communication, Energy, Information, and Structure). Within each of the three categories, the four bars represent the four superclasses, respectively. The bar height represents the number of genes.

shows the numbers of putative alien genes detected by SOM, *k*-means, and the method of Lawrence and Ochman (1998) and the number of shared genes between the three methods.

The three sets of results have 413 genes in common. As expected, the results of SOM and the *k*-means method are more similar to each other than either are to the results obtained with the method of Lawrence and Ochman. The difference between SOM/*k*-means and the method of Lawrence and Ochman probably arises from several reasons. First, as Lawrence and Ochman (1998) pointed out, their list of 755 alien genes represents the minimal amount of DNA acquired through horizontal transfer; therefore, some alien genes may have missed by their method. Second, the analyzed data are different. Here, we considered only codon usage, while Lawrence and Ochman (1998) combined G+C content, codon bias, and biological information. Third, for the SOM and *k*-means methods here genes shorter than 75 codons were excluded from the analysis, resulting in 155 fewer genes examined. Finally, the SOM method is more sensitive to different codon usage patterns of a genome than the PCA method, as demonstrated by the fact that although PCA did not find a cluster of genes with unusual codon usage in either the *M. jannaschii* or the *P. horikoshii* genome (Badger 1999), SOM did find such clusters, suggesting that horizontal transfer does occur in these organisms.

Karlin, Campbell, and Mrazek (1998) introduced a new method to measure the codon bias of one group of genes relative to that of another group of genes based on a modification of the existing Codon Adaptation Index (CAI; Sharp and Li 1987). They used the method to detect alien genes and highly expressed genes in several bacterial genomes, including those of *H. influenzae* and *M. jannaschii* (Mrazek and Karlin 1999). In the *H. influenzae* genome, Mrazek and Karlin (1999) identified 48 alien genes. Since there is some discrepancy between their gene-naming system and the one we used, we found only 38 of these genes. Nevertheless, a comparison of these 38 presumed alien genes with the presumed alien genes identified with SOM reveals 30 genes in common. Therefore, most (79%) of the presumed alien genes detected by Mrazek and Karlin are within the alien genes cluster identified in this study. In the *M. jannaschii* genome, Mrazek and Karlin identified 23 alien genes, of which only 7 (30%) were identified by SOM. However, of the 22 highly expressed genes identified by Mrazek and Karlin, 19 were found by SOM (86%). It should, however, be noted that Karlin et al.'s method considers only genes of >200 codons, which excludes many shorter alien genes. For example, in a 1.43-Mb *E. coli* contig, they identified only 16 alien genes, while Lawrence and Ochman (1997) identified 228 alien genes (Karlin, Mrazek, and Campbell 1998). However, the above comparisons demonstrate that both the alien genes and the highly expressed genes detected by Mrazek and Karlin (1999) are generally subsets of the respective gene categories identified by SOM.

In addition to the main methodological difference between this work and other works concerning detection of alien genes in bacterial genomes, we also applied an information measurement of codon usage based on

Table 2
The Numbers of Putative Alien Genes Detected by the Self-Organizing Map (SOM), the *k*-Means Method, and the Method of Lawrence and Ochman (1998) (L&O) and the Number of Shared Genes Between the Three Methods

	SOM	*k*-Means	L&O
SOM	779	597	463
k-means.		852	497
L&O			755

Shannon uncertainty. Although the ‘‘effective number of codons’’ of a gene (Wright 1990) also quantifies how far the codon usage of a gene departs from equal usage of synonymous codons, it takes no account of amino acid composition. The Shannon uncertainty of codon usage calculates combined entropy of 61 synonymous codons and averages over the 20 amino acids. It appears this index has effectively distinguished the major gene categories within each of the seven genomes (table 1). In all cases, the highly expressed gene class shows a decrease in Shannon uncertainty compared with the average codon usage of the genome, which indicates that this class is more biased in codon usage than the genome average. The index for the alien gene class is increased above the genome average, indicating that codon bias is decreased in this class. This is expected for a group of genes that have been horizontally transferred from many different lineages. The normally expressed gene class has a Shannon uncertainty close to the genome average, since the bulk of the genes of a genome are in the normal class. However, it seems meaningless to compare Shannon uncertainty across genomes, since there is no connection between this index for the genomes and their phylogenetic relationship (see, e.g., in table 1; the Shannon uncertainty of the *E. coli* genome is more different from that of *H. influenzae* than from that of *A. aeolicus*).

Class Structures of Codon Usage Map

In addition to the major class structures (genes with normal codon usage, presumed highly expressed genes, and presumed alien genes) found from the U-matrix, there are several subclusters within each of the classes, suggesting that all categories can be further divided. Such subclasses in the presumed alien gene category indicate multiple origins of horizontally transferred genes and their different ameliorating stages to the host genome (Lawrence and Ochman 1997, 1998). As the codon usage of horizontally transferred genes approaches the host usage over time, it is expected that earlier horizontally transferred genes should appear near the border between the clusters of alien genes and genes with normal codon usage.

Similarly, there exist subclasses in the clusters of genes with highly biased codon usage, as very highly expressed genes (such as the ribosomal proteins) tend to have more biased usage than do moderately highly expressed genes (such as tRNA synthetase and RNA polymerase). On the *E. coli* codon usage map (fig. 1), some genes that were previously identified as having moderate codon bias (Sharp and Li 1986) are found to locate in the area of highly expressed genes. This discrepancy is caused by the criteria that were used to define different levels of codon bias. As Sharp and Li (1986) noted, insufficient data were available then to classify these genes accurately with regard to level of gene expression. The smaller variation within the category of normally expressed genes may be derived from a combination of many factors, such as different expression levels, sequences of different functions, and different G+C contents.

Functional Classification of Gene Categories

The functional classification of gene categories identified with SOM gives further insight into codon usage patterns of the *E. coli* and *H. influenzae* genomes. The distributions of different functions among highly expressed genes, normally expressed genes, and putative alien genes are very different from each other. Of particular interest, with a different approach, Jain, Rivera, and Lake (1999) recently found that operational genes (those involved in housekeeping) are more likely to be horizontally transferred than informational genes (those involved in transcription, translation, and related processes). Our work supports this proposal. However, in Jain, Rivera, and Lake’s (1999) work, several functional groups (energy metabolism, transport proteins, and replication) were not included in their list of operational or informational genes. Our result supports the hypothesis that energy metabolism, replication, and transport protein genes are relatively less often horizontally transferred (fig. 4). However, some replication-related genes, such as those for DNA recombination, restriction, and modification, tend to have unusual codon usage and thus may also be horizontally transferred. It will be interesting to find out whether functional distribution of alien genes shows the same trend in other bacterial genomes, especially the archaebacteria.

Horizontal Gene Transfer in the Seven Genomes

The fact that a cluster of genes with unusual codon usage was found by SOM in all seven bacterial genomes suggests that horizontal gene transfer in prokaryotes is a widespread phenomenon (Doolittle 1999; Jain, Rivera, and Lake 1999). The mixing of gene classes of the four hyperthermic organisms (*A. aeolicus, A. fulgidus, M. thermoautotrophicum,* and *P. horikoshii*) is surprising but understandable, as both convergent evolution and horizontal transfer between these organisms is expected because they live in similar environments. In cases of horizontal transfer, perhaps a cluster analysis of the codon usage information in the CUTG database (Nakamura, Gojobori, and Ikemura 1999) could be used to identify the possible source organisms of the transferred genes.

Acknowledgments

We thank J. Vesanto for help in using the MATLAB-based SOM programs, and G. von Heijne, D. Hickey, H. Chipman, and the two anonymous reviewers for critical comments. This work was supported by NSERC Research Grant OGP0046506, CITO and the Steacie supplement grant (to M.L. and P.K.), and Beijing 248 Great Creative Project 9550213300 (to H.-C.W.).

LITERATURE CITED

BADGER, J. 1999. Exploration of microbial genomic sequences via comparative analysis. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Blattner, F. R., G. Plunkett III, C. A. Bloch et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1474.

Bult, C. J., O. White, G. J. Olsen et al. (40 co-authors). 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273**:1058–1073.

Deckert, G., P. V. Warren, T. Gaasterland et al. (15 co-authors). 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature **392**:353–358.

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. Science **284**:2124–2129.

Fleischmann, R. D., M. D. Adams, O. White et al. (40 co-authors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**:496–512.

Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Pave. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–r62.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **146**:1–21.

———. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA **96**:3801–3806.

Karlin, S., A. M. Campbell, and J. Mrazek. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**:185–225.

Karlin, S., J. Mrazek, and A. M. Campbell. 1998. Codon usages in different gene classes of the *Escherichia coli* genome. Mol. Microbiol. **29**:1341–1355.

Kawarabayasi, Y., M. Sawada, H. Horikawa et al. (30 co-authors). 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. **5**:55–76.

Klenk, H. P., R. A. Clayton, J. F. Tomb et al. (51 co-authors). 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature **390**:364–370.

Kohonen, T. 1982. Self-organized formation of topologically correct feature map. Biol. Cybern. **43**:59–69.

———. 1997. Self-organizing maps. 2nd extended edition. Springer, Berlin.

Lawrence, J. G., and H. Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44**:383–397.

———. 1998. Molecular archaeology of *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95**:9413–9417.

Marabini, R., and J. M. Carazo. 1994. Pattern recognition and classification of images of biological macromolecules using artificial neural networks. Biophys. J. **66**:1804–1814.

Mathe, C., A. Peresetsky, P. Dehais, M. van Montagu, and P. Rouze. 1999. Classification of *Arabidopsis thaliana* gene sequences: clustering of coding sequences into two groups according to codon usage improves gene prediction. J. Mol. Biol. **285**:1977–1991.

Médigue, C., T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. 1991. Evidence of horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.

Mrazek, J., and S. Karlin. 1999. Detecting alien genes in bacterial genomes. Ann. N.Y. Acad. Sci. **870**:314–329.

Nakamura, Y., T. Gojobori, and T. Ikemura. 1999. Codon usage tabulated from the international DNA sequence databases; its status 1999. Nucleic Acids Res. **27**:292.

Riley, M. 1993. Functions of the gene products of *Escherichia coli.* Microbiol. Rev. **57**:862–952.

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Tech. J. **27**:379–423, 623–656.

Sharp, P. M., and W.-H. Li. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for ‘rare’ codons. Nucleic Acids Res. **14**:7737–7749.

———. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

Smith, D. R., L. A. Doucette-Stamm, C. Deloughery et al. (37 co-authors). 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J. Bacteriol. **179**:7135–7155.

Tamames, J., G. Casari, C. Ouzounis, and A. Valencia. 1996. Genomes with distinct function composition. FEBS Lett. **389**:96–101.

Ultsch, A., and H. P. Siemon. 1990. Kohonen’s self-organizing feature maps for exploratory data analysis. Pp. 305–308 *in* Proceedings of the International Neural Network Conference 1990. Kluwer, Dordrecht, The Netherlands.

Vesanto, J. 1999. SOM-based data visualization methods. Intelligent Data Anal. **3**:111–126.

Wright, F. 1990. The ‘effective number of codons’ used in a gene. Gene **87**:23–29.

B. Franz Lang, reviewing editor

Accepted November 20, 2000