J. Mol. Biol. (1991) 222, 851–856

# Evidence for Horizontal Gene Transfer in *Escherichia coli* Speciation

**C. Médigue**

*Atelier de BioInformatique, Section Physique-Chimie*
*Institut Curie, 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05, France*

**T. Rouxel, P. Vigier, A. Hénaut**

*Centre de Génétique Moléculaire du CNRS*
*91198 Gif sur Yvette Cedex, France*

**and A. Danchin**

*Régulation de l'Expression Génétique, Institut Pasteur*
*28 rue du Docteur Roux, 75724 Paris Cedex 15, France*

*(Received 16 May 1991; accepted 23 September 1991)*

After extracting more than 780 identified *Escherichia coli* genes from available data libraries, we investigated the codon usage of the corresponding coding sequences and extended the study of gene classes, thus obtained, to the nature and intensity of short nucleotide sequence selection, related to constraints operating at the nucleotide level. Using Factorial Correspondence Analysis we found that three classes ought to be included in order to match all data now available. The first two classes, as known, encompass genes expressed either continuously at a high level, or at a low level and/or rarely; the third class consists of genes corresponding to surface elements of the cell, genes coming from mobile elements as well as genes resulting in a high fidelity of DNA replication. This suggests that bacterial strains cultivated in the laboratory have been fixed by specific use of antimutator genes that are horizontally exchanged.

*Keywords:* Factorial Correspondence Analysis; codon usage; nucleic acids folding; evolution; mutator genes

Rapid DNA sequencing techniques permit significant statistical analysis of the gene content of whole organisms such as bacteria. For instance, more than 1500 kb† of non-duplicated sequences of the *Escherichia coli* genome, i.e. more than 30% of the total, are currently available in data libraries (Médigue *et al.*, 1991). It has long been known that *E. coli* coding sequences could be split into two well-defined classes according to their codon usage. As this was obtained from a limited set of genes (Gouy & Gautier, 1982; Blake & Hinds, 1984), the question arises: would two classes still account for the larger set of genes known at present?

† Abbreviations used: kb, $10^3$ bases or base-pairs; CDS, coding sequence.

### (a) *Clustering of* E. coli *CDSs in three classes*

Coding sequences can best be described by the usage of codons specifying each amino acid residue of the polypeptide they encode. Accordingly, each CDS is represented as a point in a 61-dimensional space, each dimension corresponding to the relative frequency of each of the 61 codons. The set of CDSs is displayed as a cloud of points in the space of codon frequencies. Using the $\chi^2$ distance between each CDS, Factorial Correspondence Analysis allows calculation of the two-dimensional projection of the cloud of points yielding maximum scattering (Hill, 1974; Lebart *et al.*, 1984). As a consequence, genes that have a similar codon usage will appear as a neighbour (but the converse is not necessarily true). In order to make this graphical representation



0022–2836/91/240851–06 $03.00/0

**Table 1**
*Codon usage in the three classes of* E. coli *genes*

| | T-- | I | II | III | | C-- | I | II | III |
|---|---|---|---|---|---|---|---|---|---|
| Phe | TTT | 55·09 | 29·08 | 67·14 | Leu | CTT | 9·70 | 5·56 | 19·00 |
| | TTC | 44·91 | 70·92 | 32·86 | | CTC | 10·40 | 8·03 | 9·04 |
| Leu | TTA | 10·99 | 3·44 | 20·09 | | CTA | 3·09 | 0·83 | 6·81 |
| | TTG | 13·02 | 5·47 | 15·05 | | CTG | 52·79 | 76·67 | 29·99 |
| Ser | TCT | 13·26 | 32·41 | 19·63 | Pro | CCT | 13·71 | 11·23 | 28·30 |
| | TCC | 15·02 | 26·56 | 11·34 | | CCC | 11·19 | 1·63 | 16·26 |
| | TCA | 10·83 | 4·79 | 22·09 | | CCA | 18·63 | 15·25 | 31·50 |
| | TCG | 16·88 | 7·39 | 10·60 | | CCG | 56·47 | 71·89 | 23·94 |
| Tyr | TAT | 54·42 | 35·23 | 69·60 | His | CAT | 56·80 | 29·77 | 61·69 |
| | TAC | 45·58 | 64·77 | 30·40 | | CAC | 43·20 | 70·23 | 38·31 |
| TER | TAA | — | — | — | Gln | CAA | 33·40 | 18·65 | 37·06 |
| | TAG | — | — | — | | CAG | 66·60 | 81·35 | 62·94 |
| Cys | TGT | 40·90 | 38·85 | 55·71 | Arg | CGT | 38·99 | 64·25 | 26·05 |
| | TGC | 59·10 | 61·15 | 44·29 | | CGC | 43·23 | 32·97 | 21·94 |
| TER | TGA | — | — | — | | CGA | 5·52 | 1·07 | 12·80 |
| Trp | TGG | 100·00 | 100·00 | 100·00 | | CGG | 8·97 | 0·80 | 13·62 |
| | A-- | I | II | III | | G-- | I | II | III |
| Ile | ATT | 51·20 | 33·49 | 47·57 | Val | GTT | 23·74 | 39·77 | 34·33 |
| | ATC | 44·37 | 65·94 | 26·65 | | GTC | 22·48 | 13·45 | 18·95 |
| | ATA | 4·43 | 0·57 | 25·78 | | GTA | 14·86 | 19·97 | 21·78 |
| Met | ATG | 100·00 | 100·00 | 100·00 | | GTG | 38·92 | 26·81 | 24·94 |
| Thr | ACT | 14·85 | 29·08 | 26·83 | Ala | GCT | 14·52 | 27·54 | 22·86 |
| | ACC | 46·83 | 53·60 | 24·45 | | GCC | 27·62 | 16·14 | 23·67 |
| | ACA | 10·52 | 4·67 | 27·93 | | GCA | 19·63 | 24·01 | 31·27 |
| | ACG | 27·81 | 12·65 | 20·80 | | GCG | 38·23 | 32·30 | 22·19 |
| Asn | AAT | 40·87 | 17·25 | 64·06 | Asp | GAT | 62·83 | 46·05 | 70·47 |
| | AAC | 59·13 | 82·75 | 35·94 | | GAC | 37·17 | 53·95 | 29·53 |
| Lys | AAA | 75·44 | 78·55 | 72·21 | Glu | GAA | 68·33 | 75·35 | 66·25 |
| | AAG | 24·56 | 21·45 | 27·79 | | GAG | 31·67 | 24·65 | 33·75 |
| Ser | AGT | 13·96 | 4·52 | 18·73 | Gly | GGT | 32·91 | 50·84 | 31·79 |
| | AGC | 30·04 | 24·33 | 17·61 | | GGC | 43·17 | 42·83 | 24·51 |
| Arg | AGA | 1·75 | 0·62 | 15·63 | | GGA | 9·19 | 1·97 | 24·75 |
| | AGG | 1·54 | 0·29 | 9·96 | | GGG | 14·74 | 4·36 | 18·95 |

Groups of genes are established from Factorial Correspondence Analysis (Hill, 1974; Lebart *et al.*, 1984) and strong cluster determination (Diday, 1971; Delorme & Hénaut, 1988). The classes are comprised of 502, 191 and 89 CDSs, respectively. Relative synonymous codon usage values are presented for each class.

easier to analyse, it is necessary to use a second method that automatically clusters the objects (here, the CDSs) that are close to one another: in a first step, one parts the objects into $k$ groups by a dynamic clustering method; then, in a second step, objects that are always clustered together in the course of the different partition processes are selected (Diday, 1971; Delorme & Hénaut, 1988). A total of 782 *E. coli* CDSs, each corresponding to a unique sequence, have been included in the study. Analysis reveals that the best hypothesis is a clustering of CDSs into three well-separated classes, as shown on the two-dimensional projection of the cloud of points (Fig. 1) by its "rabbit head" shape. The three classes thus obtained comprise a different number of genes: 502 CDSs (class I); 191 CDSs (class II); 89 CDSs (class III).

As seen in the codon usage Table for each class (Table 1), there exists a gradient in codon bias between the classes. The relevant nature of the clustering into three classes is, however, prominent, due to the fact that the gradient always goes in the same direction independent of the codon considered. The bias is very strong in class II, intermediate in class I and weak in class III. For example, in class II, CTA is used in less than 1% of all leucine codons, which CTG is used in 76% of all cases (this corresponds to a major leucine tRNA, $tRNA_1^{Leu}$). A significant bias also exists in class I against codon CTA in favour of codon CTG, but the relative frequency of the bias is somewhat smaller. Finally, in class III codon CTA is rarely used (but as much as codon CTC) and the "frequent" codon CTG is used in only 30% of all leucine codons. In the same way, the second and first classes show significant bias against a few codons (mainly ATA, AGA and AGG) that are not discriminated against in class III; the bias is also strong for several codons only in class II: TTA, TCA, CCC, ACA, CGA, CGG, GGA and GGG have a relative frequency smaller than 5% (Table 1). In general the codon usage characterizing the third class is different from that of classes I and II for the distribution of codons is quite even. For instance the "rare" codon ATA for isoleucine is

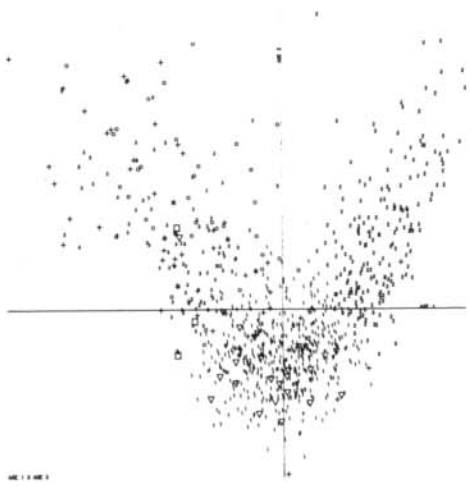

**Figure 1.** Factorial Correspondence Analysis of *E. coli* and related genes for a clustering into 3 classes. Each CDS is represented as a point in a 61-dimensional space, each dimension corresponding to the relative frequency of 1 of the 61 codons. The set of CDSs appears as a cloud of points in the space of codon frequencies. Genes that have a similar codon usage will therefore appear as neighbours. Factorial Correspondence Analysis allows us to calculate projections of the cloud of points on planes corresponding to maximum scattering (Hill, 1974; Lebart *et al.*, 1984). Clustering is then performed by a 2nd method that automatically clusters the objects that are close to one another. Strong cluster determination proceeds as 2 successive steps (Delorme & Hénaut, 1988): (1) one builds up different ways to separate the objects into $k$ groups by a dynamic clustering method; (2) the objects that are always clustered together in the different partitions are selected. Here, genes are labelled 1, 2 or 3, according to the class to which they belong. Genes having a special meaning have been labelled using different symbols: *mut* genes, large open circles; IS genes, stars; plasmid genes, plus; lambda, triangles for head and tail genes, or small circles for lysis/lysogeny genes.

used in 26% of cases. Moreover, no codon has a relative frequency smaller than 7%.

In addition to their partition according to codon usage, these three classes of *E. coli* genes can also be clearly distinguished by their biological properties. As shown in Table 2, class I contains almost all genes of intermediary metabolism, such as those permitting synthesis of coenzymes and amino acids, or catabolism of carbon sources other than glucose, with the noticeable exception of genes involved in the core of carbon assimilation (glycolysis, TCA cycle and fatty acids sythesis, etc.). It also contains genes specifying gene regulation (activators and repressors). Finally it contains genes responsible for DNA metabolism. Thus, class I comprises those genes that maintain a low or intermediary level of expression, but can potentially be expressed at a very high level (e.g. the lactose operon). In contrast class II contains genes that are constitutively expressed at a high level. As Table 2 shows (and as expected), most of these genes are involved in the translation machinery (ribosomal proteins, except protein S13, and most tRNA synthetases), in the folding of proteins and in the transcription apparatus (RNA polymerase, except, perhaps unexpectedly, the α-subunit), as well as genes coding for DNA binding proteins present at high level in the cytoplasm (regulatory proteins such as H1 or CAP, but also the only 2 class II proteins involved in recombination: Ssb and RecA). In addition, the

**Table 2**
*Gene repartition among the three classes*

| | Class | | |
|---|---|---|---|
| | I | II | III |
| Central metabolic routes: catabolism of glucose | | | |
| Phosphotransferase system for glucose | | 4 | |
| and other sugars | 6 | 4 | |
| Intermediary carbohydrate metabolism: | | | |
| glycolysis, pentose phosphate pathway | 2 | 9 | |
| TCA cycle | 7 | 10 | |
| Energy production | | | |
| Pathways of electrons to oxygen | | 2 | |
| Nitrate reductase | 8 | | |
| ATP synthase | 3 | 6 | 1 |
| Catabolism of carbon from sources other than glucose | | | |
| Sugars, polyols and carboxylates | 30 | 6 | 2 |
| Acetate and fatty acids | 3 | | |
| Biosynthesis of amino acids | 59 | 5 | 3 |
| Biosynthesis and conversions of nucleotides | 15 | 10 | |
| Purines | 6 | 8 | |
| Pyrimidines | 9 | 2 | |
| Biosynthesis of lipids | | | |
| Biosynthesis of fatty acids | | 3 | |
| Biosynthesis of phospholipids | 7 | | 1 |
| Biosynthesis of coenzymes and prosthetic groups | 29 | | 2 |
| Biosynthesis of proteins | | | |
| Ribosomal proteins | 1 | 45 | |
| RNA polymerase and transcription factors | 2 | 7 | |
| tRNA synthetases and translation factors | 6 | 19 | |
| Chaperonins | | 5 | |
| DNA binding regulatory genes | 46 | 6 | 2 |
| Cell envelope and septation | 27 | 5 | |
| Biosynthesis of DNA | | | |
| DNA polymerization | 10 | | |
| Topoisomerases, helicases, ligase | 3 | | |
| Modified DNA editing | 14 | | |
| Antimutator genes | | | 3 |
| Recombination and repair | 15 | 2 | |
| Motility and chemotaxis | 11 | | |
| Transport and permeation | 42 | 15 | 5 |
| Sexuality and gene transfer | | | |
| Fimbriae, major pilus | 3 | 1 | 18 |
| Insertion sequences | | | 8 |
| Restriction endonucleases | 2 | | 7 |
| Lambdoïd lysogeny (including *dic* genes) | | | 25 |
| Lambda head and tail | 24 | | 3 |
| Plasmids (*mot*, *tra* etc.) | 13 | 1 | 31 |
| Total | 397 | 175 | 111 |

For each column the total number of genes takes into account only those *E. coli* genes that have an identified function, as well as genes from lambda and plasmids, when corresponding to identified functions.

genes of the core intermediary metabolism are also members of the class (genes coding for the glucose, mannose or *N*-acetyl-glucosamine phosphoenol-pyruvate dependent phosphotransferase system, phosphoglucose isomerase, phosphofructokinase, triosephosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase (GAPDH): all genes currently known to be involved in this part of intermediary metabolism, except minor species such as the second phosphofructokinase, or the second GAPDH, which are members of class I). Finally purine biosynthesis stems mainly from class II genes, whereas the pyrimidine counterpart is found in class I, suggesting a significant difference in the phylogenic pathways that led to the two nucleotide families. Up to this point the clustering of *E. coli* genes accords with the codon usage clusters described by Gouy & Gautier (1982). However, as seen in Figure 1 it has been necessary to introduce a third class in order to describe properly the codon usage of all *E. coli* genes. Genes in this latter class code for fimbriae, flagellae and pili, integration host factors (*hip* and *himA*), genes controlling cell division (*dicABC*, structurally related to temperate phages (Bejar *et al.*, 1988)), several outer membrane or periplasmic proteins genes and several catabolic operons (threonine degradation, *β*-glucoside degradation, fucose degradation). In addition, it comprises genes coded by insertion sequences and genes that behave as mutators when inactivated (*mutH*, *mutT* and *mutD*).

This prompted us to repeat the analysis, adding genes of temperate bacteriophage lambda, plasmids and transposons to the initial set. Many of the new genes belong to class III. The case of bacteriophage lambda is particularly interesting in this respect: genes involved in the lysis/lysogeny balance (i.e. *c*I, *c*II, *c*III, *cro, xis, int*...) are members of class III, whereas genes involved in the construction of the head and tail of the phage are members of class I. From the physiological point of view it should be noted that many genes found in class III can be expressed at a fairly high level, and sometimes continuously (Table 2), but that their codon usage does not reflect the average distribution of specific tRNAs availability.

### (b) *Analysis of the oligonucleotide composition of the three classes*

It has been accepted for some time (Ikemura, 1981; Lipman & Wilbur, 1983) that codon choice allows adaptation of gene sequences to specific features of the translation apparatus. In addition it is also reasonable to infer that further constraints operate directly at the nucleotide level, for example associated to the existence of DNA–protein interactions or to the formation of secondary structures in transcripts. In this context, code degeneracy can be visualized as a means to satisfy overlapping constraints. We have therefore extended the study of gene classes to the nature and intensity of short nucleotide sequence selection.

For all CDSs of the three classes, the observed frequency of an oligonucleotide has been measured for di-, tri-, tetra- and pentanucleotides. In order to study constraints that are not related to the reading frame, only cases corresponding to oligonucleotides overlapping two contiguous codons have been taken into account. The comparison, with a control constituted by sequences of the same codon composition but in which the codon succession is randomized, enabled us to identify oligonucleotides for which an important deviation between observed and calculated frequency could be observed. The corresponding constraints, which can neither be related to the succession of amino acid residues nor to the use of the code adopted for each class of genes, tend to avoid or privilege specific oligonucleotides in a coding sequence (Hénaut *et al.*, 1985). They are presumably related to the complexity of the cognate RNA secondary structure but also to the intensity of the selective pressure acting on nucleotide sequences such as in DNA or in RNA.

The average amino acid composition of proteins coded by the genes is the same in the three classes but a small bias is observable in their G+C content (53% in classes I and II, and 47% in class III). Since transcription operates on a single strand, frequencies of two oligonucleotides found in a given CDS should not be equal when they are complementary, except if a significant part of DNA or RNA is involved in stem and loop structures. Therefore, in order to divide oligonucleotides into three classes according to the ability of their corresponding nucleic acid to form secondary structure, the correlation between the observed frequency of an oligonucleotide, overlapping two contiguous codons, and the observed frequency of its complement in the same strand, has been measured for di-, tri-, tetra- and pentanucleotides. Obviously, secondary structure formation is only pertinent for the longest oligonucleotides (tetra- and pentanucleotides), where a strong bias can be seen (Table 3A). In this analysis, genes from class II are immediately sorted out from those of the other classes, because the observed correlation appears to be significantly lower. In contrast, genes of classes I and III exhibit frequencies corresponding to long oligonucleotides that are strongly correlated to that of the complementary counterparts. This indicates that nucleic acids of classes I and III can be folded easily into a secondary structure.

For a given oligonucleotide, the difference in frequency between a real sequence and its theoretical counterpart reflects the intensity of the selective pressure exerted on the oligonucleotide, independent of protein or translational constraints (Hénaut *et al.*, 1985). Classes I and II contain a significant number of oligonucleotides submitted to such selective pressure in comparison to class III (Table 3B). Though the frequency of some oligonucleotides is biased in all three classes simultaneously (CTGG is overrepresented, whereas CTAG is underrepresented), when an oligonucleotide is biased in two classes, it is always biased in class I

**Table 3**

*Statistical analysis of the selective pressure acting on oligonucleotides present in the coding strand of all genes from classes I (502), II (191) and (III) (89)*

| | I | II | III | |
|---|---|---|---|---|
| A. *Spearman's rank correlation* | | | | |
| Di- | 0·657 | 0·829 | −0·371 | † |
| Tri- | 0·762 | 0·453 | 0·522 | ns |
| Tetra- | 0·661 | 0·418 | 0·724 | ‡ |
| Penta- | 0·689 | 0·457 | 0·700 | ‡ |
| B. *Comparison to theoretical background* | | | | |
| 10% Di- | 2/16 | 2/16 | 1/16 | ns |
| 10% Tri- | 10/64 | 7/64 | 2/64 | § |
| 10% Tetra- | 36/256 | 29/256 | 12/256 | † |
| 10% Penta- | 118/1024 | 111/1024 | 78/1024 | † |
| 5% Penta- | 66/1024 | 58/1024 | 30/1024 | ‡ |

In A, the Spearman's rank correlation between the observed frequency of a given oligonucleotide and its complement for dinucleotides (6 pairs), trinucleotides (32 pairs), tetranucleotides (120 pairs) and pentanucleotides (512 pairs) is given. A strong correlation for the longest oligonucleotides indicates that a significant part of a nucleic acid can be part of stem and loop structures. In B, for each oligonucleotide (16 di-, 64 tri-, 256 tetra- and 1024 pentanucleotides) the observed number is compared to the theoretical background, made of sequences generated from polypeptides having sequences identical with that of the corresponding protein and in which the overall codon composition is identical with the original composition, but where the succession of codons is random (Hénaut *et al.*, 1985). Only cases corresponding to oligonucleotides overlapping 2 contiguous codons are taken into account. A Pearson's distance $(1/N[(\text{obs.}-\text{cal.})^2/\text{cal.}])$ allows evaluation of the constraints operating independently of the coding frame, and which tend to act in favour or against specific oligonucleotides. The significance of Pearson's distance is analysed by retaining, in each class, only those oligonucleotides that correspond to values higher than a given threshold (10%, but also 5% in the case of pentanucleotides). As can be seen classes I and II are submitted to a selection pressure significantly stronger than class III.

ns, not significant.

† Significant at the 1% level.

‡ Significant at the 0·1% level

§ Significant at the 5% level.

(for instance TTCC is overrepresented in classes I and II, while TTGG is underrepresented in classes I and III).

This observation is further substantiated by the result of analysing the position of the most frequently used oligonucleotides along selected DNA sequences for which all genes belong to a given class (class I, or class II, or class III). The procedure for identifying frequent oligonucleotides was performed independently of the frequency of the complement in the same strand, in order to take into account the selection pressure operating on individual sequences. It concerned seven tetranucleotides in each class (as compared to a total of 256 tetranucleotides): CTGG, GGCG, GAAG, TTCC, CCAG, CAGG and TATC in class I; AGAG, CGGT, TTCC, GGTA, GGCG, CTGG and CAGG, in class II; and CTGG, TATT, CAGT, CCGG, ACCA, AATA and CGTC in class III. It appears that for class I (and class II) the distribution is quite uneven

**Table 4**

*Frequency of the overrepresented set of seven tetranucleotides from each class of genes (SS1, SS2, SS3, for classes I, II and III, respectively) in coding and non-coding regions extracted from sequences comprised only of genes belonging to a given class*

| | I | | II | | III | |
|---|---|---|---|---|---|---|
| | Coding | Non-coding | Coding | Non-coding | Coding | Non-coding |
| SS1 | 5·01 | 2·74 | 4·89 | 2·43 | 3·60 | 3·18 |
| SS2 | 4·46 | 2·57 | 4·63 | 2·88 | 3·42 | 2·98 |
| SS3 | 3·61 | 3·18 | 3·35 | 2·98 | 3·75 | 3·84 |

Using only those sequences comprising solely genes from a single class, the frequency of the most positively biased 7 tetranucleotides extracted from all CDSs present in each class (SS1, SS2 and SS3, respectively) was calculated. Comparison was performed independently between coding and non-coding regions in each class, a non-coding region class being defined by the class of the surrounding CDSs. In the Table the percentage of each tetranucleotide set is displayed, comparing coding to non-coding sequence in each class. For example, the 7 favoured tetranucleotides in class I (SS1) represent about 5% of all tetranucleotides in class I coding regions, corresponding to about twice (0·71%) the expected frequency (0·39%) for a tetranucleotide in the sequence. As visible in the Table, favoured tetranucleotides from classes I and II are overrepresented in coding as compared to non-coding regions of the same class, whereas those identified from class III tend to be present equally in coding as well as non-coding regions, independent of the class.

as positively biased tetranucleotides are much rarer in intergenic regions. This is very different in the case of class III where the same oligonucleotides are evenly distributed (Table 4).

### (c) *Is horizontal gene transfer involved in bacterial speciation?*

A final point must be emphasized. If class III, as strongly suggested, consists of genes inherited from horizontal transfer, gene products with a strong antimutator action (*mutD*, *mutT* and *mutH*) as members of this class, suggests a radical hypothesis: the usual state of micro-organisms in nature might correspond to a population of highly mutable individuals (Cox & Gibson, 1974; Painter, 1975). Some of them having discovered a stable environment would then be stabilized as a species, by acquisition of such genes. Obviously this would correspond to organisms that can be cultivated in the laboratory and account for the very interesting observation that single cells isolated from sea water or hydrothermal environments do not match cells extracted as cultivatable colonies from the same sources (Giovannoni *et al.*, 1990; Ward *et al.*, 1990). This is also well in line with the interpretation of population genetic data, recently reviewed by Maynard-Smith and colleagues (Maynard-Smith *et al.*, 1991), and provides independent support for a specific role of horizontal gene transfer in bacterial populations. Finally, this raises a more general question about partitioning a genome into well-defined classes of

genes. Is this derived from convergence caused by specific selective pressure, or is this the mark of the association of individual sets of genes present very early in evolution? Comparison between the organization of the genomes of widely separated organisms will help settle this puzzling issue.

We thank S. Brouillet, for her help in using the FCA software, and F. Michel and F. Lisacek for constructive comments on the manuscript. C.M. and T.R. were recipients of ORSAN fellowships.

## References

Bejar, S., Bouché, F. & Bouché, J. P. (1988). Cell division inhibition gene *dicB* is regulated by a locus similar to lambdoid bacteriophage immunity locus. *Mol. Gen. Genet.* **212**, 11–19.

Blake, R. D. & W. Hinds, P. (1984). Analysis of the codons bias in *E. coli* sequences. *J. Biomol. Struct. Dynam.* **2**, 593–606.

Cox, E. C. & Gibson, T. C. (1974). Selection of high mutation rates in chemostats. *Genetics*, **77**, 169–184.

Delorme, M. O. & Hénaut, A. (1988). Merging of distance matrices and classification by dynamic clustering. *Comput. Appl. Biosci.* **4**, 453–458.

Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. *Rev. Statist. Appl.* **19**, 19–33.

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature (London)*, **345**, 60–62.

Gouy, M. & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucl. Acids Res.* **10**, 7055–7074.

Hénaut, A., Limaiem, J. & Vigier, P. (1985). The origins of the strategy of codon use. *Biochimie*, **67**, 475–783.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Appl. Statist.* **23**, 340–353.

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.

Lebart, L., Morineau, A. & Warwick, K. A. (1984). *Multivariate Descriptive Statistical Analysis*, John Wiley and Sons, New York.

Lipman, D. & Wilbur, J. (1983). Contextual constraints on synonymous codon choice. *J. Mol. Biol.* **163**, 363–376.

Maynard-Smith, J., Dowson, C. & Spratt, B. (1991). Localized sex in bacteria. *Nature (London)*, **349**, 29–31.

Médigue, C., Viari, A., Hénaut, A. & Danchin, A. (1991). *Escherichia coli* molecular genetic map (1500 kb): update II. *Mol. Microbiol.* **5** (11), in the press.

Painter, P. R. (1975). Mutator genes and selection for the mutation rate in bacteria. *Genetics*, **79**, 649–660.

Ward, D. M., Weller, R. & Bateson, M. M. (1990). 16 S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature (London)*, **345**, 3–65.

*Edited by J. H. Miller*