Codon Bias Signatures, Organization of Microorganisms in Codon Space, and Lifestyle

A. Carbone, F. Képès,† and A. Zinovyev‡*

*Génomique Analytique, Université Pierre et Marie Curie, INSERM U511, 91, Bd de l'Hôpital, 75013 Paris, France;
†ATelier de Génomique Cognitive, CNRS ESA 8071/Genopole, 523, Terrasses de l'Agora, 91000 Evry, France;
‡Institut des Hautes Études Scientifiques, 35, route de Chartres, 91440 Bures-sur-Yvette, France

New and simple numerical criteria based on a codon adaptation index are applied to the complete genomic sequences of 80 Eubacteria and 16 Archaea, to infer weak and strong genome tendencies toward content bias, translational bias, and strand bias. These criteria can be applied to all microbial genomes, even those for which little biological information is known, and a codon bias signature, that is the collection of strong biases displayed by a genome, can be automatically derived. A codon bias space, where genomes are identified by their preferred codons, is proposed as a novel formal framework to interpret genomic relationships. Principal component analysis confirms that although *GC* content has a dominant effect on codon bias space, thermophilic and mesophilic species can be identified and separated by codon preferences. Two more examples concerning lifestyle are studied with linear discriminant analysis: suitable separating functions characterized by sets of preferred codons are provided to discriminate: translationally biased (hyper)thermophiles from mesophiles, and organisms with different respiratory characteristics, aerobic, anaerobic, facultative aerobic and facultative anaerobic. These results suggest that codon bias space might reflect the geometry of a prokaryotic "physiology space." Evolutionary perspectives are noted, numerical criteria and distances among organisms are validated on known cases, and various results and predictions are discussed both on methodological and biological grounds.

Introduction

Statistical analysis of DNA sequences and in particular of codon bias were performed from the moment that long chunks of DNA sequences were publicly available in the early eighties (Grantham et al. 1980; Wada et al. 1990), and the roots for these studies can be traced back to the sixties (Sueoka 1962; Zuckerkandl and Pauling 1965). However, with the increasing number of bacterial genome sequences from a broad diversity of species, this field of research has been revivified in the last 5 years (Koonin and Galperin 1997; Lin and Gerstein 2000; Radomski and Slonimski 2001; Knight, Freeland, and Landweber 2001; Sicheritz-Pontén and Andersson 2001; Daubin, Gouy, and Perrière 2002; Lin et al. 2002; Lobry and Chessel 2003; Sandberg et al. 2003). Pioneer work in inferring bacterial similarity relationships using large chunks of genomic sequences is attributable to Karlin. In a series of papers starting with (Karlin 1994; Karlin, Ladunga, and Blaisdell 1994), Karlin et al. showed how dinucleotide relative abundance values (profiles) of different DNA sequences samples of size $\geq$ 50 kb from the same organism are generally much more similar to each other than they are to profiles from other organisms, and that closely related organisms generally have more similar profiles than do distantly related ones.

The interest of comparing organisms leads to the problem of defining biologically meaningful spaces from which to extract new insight into organism similarities. Spaces rising from a direct statistical analysis of genomic sequences, based on dinucleotide frequencies (Karlin 1994; Karlin, Ladunga, and Blaisdell 1994; Karlin and Mrázek 1998), as well as codon usage, synonymous codon usage, and amino-acids usage (Kreil and Ouzounis 2001; Tekaia, Yeramian, and Dujon 2002) organize organisms roughly in a similar manner: relative distances among most phylogenetic genuses are preserved across spaces. There are pairs of organisms though, whose relative distance may vary considerably depending on the space one chooses (see later). The main motivation for this work was to define a space whose coordinates are mathematically well defined as well as justifiable by biological intuition, and to revisit organism distances within this framework. The mathematical rigor is a particularly important requirement for genome comparison; suitable biological properties can then be tested, validated, and possibly predicted across organisms.

An organism is defined through the set of preferred codons shaping its genome. The basic idea is simple and goes back to two main facts: first, the genetic code associates a set of sibling codons to the same amino acid, and some codons occur more frequently than others in gene sequences (Grantham et al. 1980; Wada et al. 1990); second is the hypothesis, formulated by Sharp (Sharp and Li 1987), that for each genome sequence *G*, there is a set of coding sequences *S*, constituting roughly the 1% of the genes in *G*, which is representative of the dominating codon bias in *G*. Many observations support this hypothesis: for bacteria and small eukaryotes like *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster* for instance, which are governed by translational bias, this set is constituted mainly by ribosomal, glycolytic, heat-shock proteins, and elongation factors; for *Pseudomonas aeruginosa*, the set contains the proteins with the highest *GC*3 content; for *Borrelia burgdorferi* the set is constituted solely by genes lying in leading strands (Carbone, Zinovyev, and Képès 2003). Combining the two facts together, one can define weights for codons on genes in *S*, which are representatives of codon preferences, as follows. Given an amino-acid *j*, its synonymous codons might have different frequencies in *S*; if $x_{i,j}$ is the number of times that the codon *i* for the amino acid *j* occurs in *S*,

Key words: Prokaryotes, Archaea, codon bias, codon space, microbial evolution, microbial lifestyle.

E-mail: carbone@ihes.fr.

Mol. Biol. Evol. 22(3):547–561. 2004
doi:10.1093/molbev/msi040
Advance Access publication November 10, 2004

then one associates to i a *weight* $w_{i,j}$ relative to its sibling of maximal frequency y_j in S

$$w_{i,j} = \frac{x_{i,j}}{y_j}. \quad (1)$$

Such weights have been successfully used by Sharp to correlate expression levels to translational codon bias (Sharp and Li 1987). (Notice that weights equal to 1 do not correspond to codons which are the most frequent over the entire genome: more than 10 amino acids in *Bacillus subtilis*, for instance, are preferentially coded with codons other than those which are the most frequent over the entire genome.) Weights calculated over S allow us to define the codon adaptation Index (CAI) (Sharp and Li 1987), which produces a rank of all genes in a genome agreeing with dominating codon bias: genes ranking the highest are the most biased and those ranking lowest are the less affected by selective bias. More generally, it has been shown that CAI correlates to *any* kind of dominating bias in genomes (like GC-content, preference for codons with G or C at the third nucleotide position, a leading strand richer in $G + T$ than a lagging strand), and not just to translational bias (Carbone, Zinovyev, and Képès 2003). Moreover, an algorithm for the automatic detection of S from the collection of all genes in a genome has been proposed in (Carbone, Zinovyev, and Képès 2003); since the algorithm is not based on any biological knowledge of the organism, it allows us to determine weights for those genomes for which not much biological information is available.

Weights are highly specific to a genome, they can be defined for any microorganism, they are good indicators of the evolutionary process under which the organism has gone, and they seem shaped by the metabolic constraints of the organism during evolution (Wagner 2000). For these reasons, we use codon weights to represent genomes (a genome becomes a [normalized] vector of 64 weights).

In the first part of this article, we present new and simple statistical criteria that correlate a bias of a given origin (content bias, translational bias, strand bias) to *CAI* values of genes. Each criterion, being bias specific, allows us to infer weak or strong tendencies of a genome toward the bias, and possibly provides a numerical evaluation of the strength. Suitable numerical thresholds are proposed, and they allow for an automatic detection of a *codon bias signature*, that is the collection of strong biases displayed by a genome. Two of the criteria allow us to determine whether an organism is affected by some (weak or strong) form of translational bias, and in this case to infer putative gene expression levels for the organism. This is done with no use of gene expression data (Jansen et al 2003) nor of gene classification and protein class comparison (Karlin and Mrázek 2000; Mrázek et al. 2001; Karlin et al. 2003). Our criteria can be applied to any genome for which no biological knowledge is yet available. All numerical criteria have been validated on previously established analysis of codon bias; contrary to what has been claimed by Anderson and Sharp (1996), a tendency toward translational bias has been detected for *Rickettsia prowazekii*. Predictions on newly sequenced genomes have been deduced.

In the second part of this article we describe a codon bias space where genomes are identified by their specific 64 codon weights, and we introduce a linear distance between genomes, which is new to comparative analysis. Distances among organisms are validated on known cases of strains, species, and established phylogenetic branches. Codon bias space is proposed as a novel formal framework to interpret genomic relationships and biologically important features including lifestyle and evolutionary trends. Principal component analysis is applied to codon bias space and confirms that although *GC* content has a dominant effect, optimal growth temperature explains the second principal component (Lynn, Singer, and Hickey 2002). As a consequence, thermophilic and mesophilic species can be identified and sharply separated by codon preferences. Using linear discriminant analysis, two more examples concerning lifestyle are studied, and suitable separating functions characterized by sets of preferred codons are provided to discriminate translationally biased (hyper)thermophiles from mesophiles, and organisms with different respiratory characteristics: aerobic, anaerobic, facultative aerobic, and facultative anaerobic. These results suggest that codon bias space might reflect the geometry of a prokaryotic "physiology space." Evolutionary perspectives are noted, and various results are discussed both on methodological and biological grounds.

Materials and Methods

Genomes and Replication Origins

Genomes, along with gene annotation, were retrieved from the Genomes directory of the GenBank FTP (see table 1). All coding sequences (CDS) were considered, including those annotated as hypothetical and those predicted by computational methods only. From each CDS, we excluded initiation and stop codons.

Information on the replication origin and terminus for 38 bacteria has been taken from the following web site, http://pbil.univ-lyon1.fr/emglib/emglib.html, where the prediction of these locations was based on the work of Lobry (1996). For most organisms in table 1, this information is still unknown.

Mesophilic, Thermophilic, and Hyperthermophilic Genomes

Mesophiles are organisms with an optimum growth temperature (OGT) near 37°C; thermophiles have OGT between 45 and 65°C, and hyperthermophiles have OGT $\geq$ 65°C, preferably around 80°C or higher (http://www.mblab.gla.ac.uk/dictionary).

Nucleotide Frequencies: Some Definitions

GC-content is the frequency of $G + C$ base pairs (bps); *GC3-content* is the frequency of $G + C$ bps at the codons third position (excluding Met, Trp, and termination codons); *XY-skew* is defined as $(N_X - N_Y)/(N_X + N_Y)$, where N_X, N_Y represent the frequencies of the nucleotides $X, Y \in \{A, T, G, C\}$, with $X \neq Y$.

Computation of CAI Values

The algorithm proposed by Carbone, Zinovyev, and Képès (2003) is used to detect a set S of genes which are

Table 1
Thermophilic, Hyperthermophilic, and Mesophilic Eubacteria and Archaea

	AQUIFICAE
1**H**	*Aquifex aeolicus*
	CYANOBACTERIA
2	*Nostoc sp*
3	*Synechocystis PCC6803*
4	*Thermosynechococcus elongatus*
	ACTINOBACTERIA
5	*Bifidobacterium longum*
6	*Corynebacterium efficiens YS-314*
7	*Corynebacterium glutamicum*
8	*Mycobacterium leprae*
9	*Mycobacterium tuberculosis CDC1551*
10	*Mycobacterium tuberculosis H37Rv*
11	*Streptomyces coelicolor*
	FIRMICUTES Bacillales
12	*Bacillus halodurans*
13	*Bacillus subtilis*
14	*Listeria innocua*
15	*Listeria monocytogenes*
16	*Oceanobacillus iheyensis*
17	*Staphylococcus aureus Mu50*
18	*Staphylococcus aureus MW2*
19	*Staphylococcus aureus N315*
	FIRMICUTES Clostridia
20	*Clostridium acetobutylicum*
21	*Clostridium perfringens*
22**H**	*Thermoanaerobacter tengcongensis*
	FIRMICUTES Lactobacillales
23	*Lactococcus lactis*
24	*Streptococcus agalactiae 2603*
25	*Streptococcus agalactiae NEM316*
26	*Streptococcus mutans*
27	*Streptococcus pneumoniae R6*
28	*Streptococcus pneumoniae TIGR4*
29	*Streptococcus pyogenes*
30	*Streptococcus pyogenes MGAS315*
31	*Streptococcus pyogenes MGAS8232*
	FIRMICUTES Mollicutes
32	*Mycoplasma genitalium*
33	*Mycoplasma pneumoniae*
34	*Mycoplasma pulmonis*
35	*Ureaplasma urealyticum*
	FUSOBACTERIALES
36	*Fusobacterium nucleatum*
	SPIROCHAETALES
37	*Borrelia burgdorferi*
38	*Treponema pallidum*
39	*Leptospira interrogans*
	CHLAMYDIALES
40	*Chlamydia muridarum*
41	*Chlamydia trachomatis*
42	*Chlamydophila pneumoniae AR39*
43	*Chlamydophila pneumoniae J138*
	PROTEOBACTERIA alpha
44	*Agrobacterium tumefaciens C58 Cereon*
45	*Agrobacterium tumefaciens C58 UWash*
46	*Brucella melitensis*
47	*Brucella suis 1330*
48	*Caulobacter crescentus*
49	*Mesorhizobium loti*
50	*Rickettsia conorii*
51	*Rickettsia prowazekii*
52	*Sinorhizobium meliloti*

Table 1. Continued

	PROTEOBACTERIA beta
53	*Neisseria meningitidis MC58*
54	*Neisseria meningitidis Z2491*
55	*Ralstonia solanacearum*
	PROTEOBACTERIA epsilon
56	*Campylobacter jejuni*
57	*Helicobacter pylori 26695*
58	*Helicobacter pylori J99*
	PROTEOBACTERIA gamma
59	*Buchnera aphidicola Sg*
60	*Buchnera sp*
61	*Escherichia coli K12*
62	*Escherichia coli O157H7*
63	*Escherichia coli O157H7 EDL933*
64	*Haemophilus influenzae*
65	*Pasteurella multocida*
66	*Pseudomonas aeruginosa*
67	*Salmonella thyphi*
68	*Salmonella thyphimurium LT2*
69	*Shewanella oneidensis*
70	*Shigella flexneri 2a*
71	*Wigglesworthia brevipalpis*
72	*Vibrio cholerae*
73	*Xanthomonas campestris*
74	*Xanthomonas citri*
75	*Xylella fastidiosa*
76	*Yersinia pestis CO92*
77	*Yersinia pestis KIM*
	CHLOROBIALES
78	*Chlorobium tepidum TLS*
	DEINOCOCCUS/THERMUS
79	*Deinococcus radiodurans*
	THERMOTOGALES
80**H**	*Thermotoga maritima*
	ARCHEOGLOBALES
81**H**	*Archaeoglobus fulgidus*
	METHANOBACTERIALES
82**H**	*Methanobacterium thermoautotrophicum*
	METHANOPYRALES
83**H**	*Methanopyrus kandleri*
	SULFOLOBALES
84**H**	*Sulfolobus solfataricus*
85**H**	*Sulfolobus tokodaii*
	THERMOPLASMALES
86**T**	*Thermoplasma acidophilum*
87**T**	*Thermoplasma volcanium*
	DESULFUROCOCCALES
88**H**	*Aeropyrum pernix*
	HALOBACTERIALES
89	*Halobacterium sp*
	METHANOCOCCALES
90**H**	*Methanococcus jannaschii*
	METHANOSARCINALES
91**T**	*Methanosarcina acetivorans*
92**T**	*Methanosarcina mazei*
	THERMOCOCCALES
93**H**	*Pyrococcus abyssi*
94**H**	*Pyrococcus furiosus*
95**H**	*Pyrococcus horikoshii*
	THERMOPROTEALES
96**H**	*Pyrobaculum aerophilum*

representative of the dominating codon bias in a given genome. This reference set S contains the 1% of the most biased genes of the genome (the size of S corresponds to the one suggested in Sharp's original work (Sharp and Li 1987)). From S, one computes weights $w_{i,j}$ for codon i and organism j as in equation (1). These weights $w_{i,j}$ are then used to compute the CAI for all genes, $CAI(g)=(\Pi_{k=1}^{L} w_k)^{1/L}$, where g is a gene, w_k is the weight of the kth codon in g, and L is the number of codons in g (Sharp and Li 1987).

Notice that the "preference" of a codon among synonymous ones is identifiable by codon weight equal to 1. Theoretically speaking, multiple synonymous codons (possibly all n synonymous codons of a n-fold degenerate amino acid) might take value 1, and one can think of those as being equally preferred. In practice, no equally preferred codons ever occurred in our analysis of 96 organisms. In particular, it should be noticed that equal codon preferences represent a *possible*, but merely theoretical, condition under which homogeneous codon composition—that is the absence of compositional bias, strand bias, and translational bias—can take place.

Codon weights, reference set S and CAI values are calculated with the program *CAIJava* written by the authors, which uses parsers of GenBank flat files from the *Biojava* (http://www.biojava.org) programming package. The idea of the algorithm is simple. It is an iterative algorithm that at iteration $i + 1$ computes codon weights based on a set S of genes selected at iteration i, then ranks all genes with respect to *CAI* value and selects a new set S, which has half the cardinality of the set determined at iteration i (if at the ith iteration the selected set is already constituted by the 1% of all genes, then the new set will also be constituted by 1% of genes) and whose genes score the highest. The process is repeated until 1% of genes have been selected and convergence is reached. At the start, S is the set of all genes. A description of the algorithm and a validation of the approach is reported in Carbone, Zinovyev, and Képès (2003). The program CAIJava is available at http://www.ihes.fr/~carbone/data.htm.

Plasmids

A chromosome is distinguished from a plasmid by assuming that it contains genes which are essential for metabolism under all growth conditions, i.e., housekeeping genes; plasmids generally provide gene product that can benefit the bacterium under certain conditions, such as resistance to antibiotics (Madigan, Martinko, and Parker 2000). Some prokaryotes contain more than one chromosome, such as *Methanococcus jannaskii* (3), *Vibrio cholerae* (2), members of the genus *Agrobacterium* (2) and *Brucella* (2). Of the organisms we considered, 22 contain plasmids, but for only 6 of them the ratio P/C, where P is the number of bps in the plasmids and C is the number of bps in the chromosome(s), is $>10\%$. In particular, *B. burgdorferi* which has a linear chromosome and 21 circular and linear plasmids, has $P/C = 66\%$.

The calculation of t-values, concerning leading and lagging strand bias, is done considering chromosomal CDSs only. This is particularly important to correctly evaluate those genomes like *B. burgdorferi* whose P/C is large.

Space of Organisms and Its Visualization

An organism is represented by a 64-dimensional vector, whose entries correspond to the 64 codon weights w_i of the organism computed for a set of genes S, which is representative of the dominating codon bias of the organism. (Stop codons *UAA*, *UAG*, *UGA*, and *UGG*, *AUG* with no synonymous codons could be disregarded. No substantial difference in the determination of the reference set S nor in the 3D visualization occurs.)

Hence, an organism is a point in the 64-dimensional space $[0\ldots1]^{64}$, where no special assumption is made on the space nor on the coordinate system. The set of points is visualized in 3 dimensions by using principal components analysis (PCA) (Hotelling 1933; Hand, Mannila, and Smyth 2001): first, every coordinate is normalized on unity standard deviation to take into account equally dominating as well as rare codons (following the standard procedure employed in PCA, the normalized weight w_i^{k*} for codon i in organism k is defined as $(w^k{}_i - \bar{w}_i)/\sigma_i$, where w_i^k is the weight of i in k, $\bar{w}_i$ is the average weight of i computed with respect to all organisms k, and σ_i is the standard deviation for the set of weights w_i^k, for all organisms k); then, three principal components for the cloud of points are calculated; finally, the cloud of points is projected orthogonally in the subspace of the three selected vectors and visualized by means of a 3D viewer. The projection of the points in the principal plane (defined by the first two principal axes) explains 58% of the variance (with the first component that explains 45% of the variance; this ensures that the PCA projection reflects well the total information embedded in the original data matrix). The three principal axes explain 65% of the variance.

Principal components analysis and the visualization of the space of organisms are done in *VidaExpert*, a tool developed by A.Z. A specialized 3D viewer is provided with *VidaExpert*. All software is available at http://www.ihes.fr/~carbone/data.htm. The interactive version of figure 1 (top) can be found at http://www.ihes.fr/~materials/organisms/htmlview.html.

Linear Analysis of the Space of Organisms

Linear discriminant analysis (LDA) (Fisher 1936) has been used to detect relevant patterns in the high-dimensional space of organisms: (1) hyperthermophiles, thermophiles, and mesophiles; (2) translationally biased (hyper)thermophiles and mesophiles; and (3) organisms with different respiratory characteristics. For each application, we construct a linear discriminant function $f(k) = \alpha_0 + \sum_{i=1}^{64} \alpha_i w_i^k$, where w_i^k is the weight of codon i in organism k, and where the separation coefficients $\alpha_i \in (-1, +1)$ are computed with the LDA algorithm. The purpose of LDA is to determine the coefficients α_i that discriminate best a set X from a set Y, that is optimize the ratio $(\textit{mean difference})^2/\textit{variance}$. For applications 1 and 3, we find a linear discriminant function f such that the set of k's with $f(k) > 0$ is exactly X (all true positives and no false negatives appear); for application 2, translationally biased mesophiles can be separated with 94 true positives and 2 false negatives, specificity $Sp = 97.78$, sensitivity

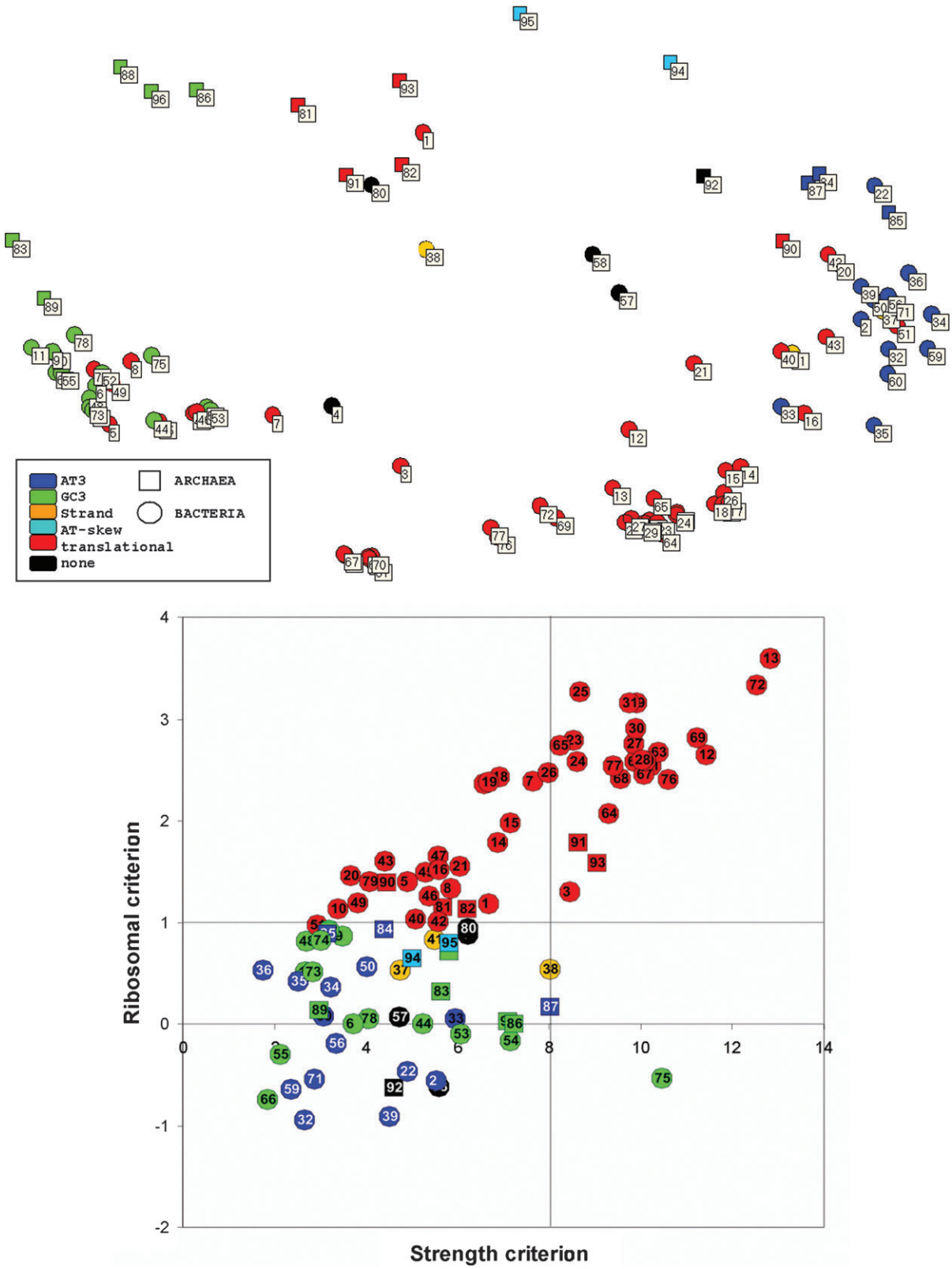


FIG. 1.—Top: Principal components analysis principal plane representation of the distribution of 96 organisms according to codon weights (numbers as in table 1). Archaea (squares) and Eubacteria (circles) are colored with a preferential order of codon biases: translational bias (red), GC3-bias (green), AT3-bias (blue), strand (orange), *AT*-skew bias (light blue), no bias (black). Bottom: Two-dimensional plot of the 96 organisms with the $\bar{z}_{Rib}$ scale on the *y*-axis and the $d(w^G, w^S)$ scale on the *x*-axis. The organisms with $d(w^G, w^S) > 8$ are all translationally biased by the ribosomal criterion with the exception of *X. fastidiosa* (75), a *GC*3 biased genome, with $d(w^G, w^S) = 10.43$; other two organisms, *T. volcanium* (87) and *T. pallidum* (38), which are *AT*3 and strand biased but not translationally biased, approximate closely the threshold $d(w^G, w^S) \approx 8$.

$Sn = 97.78$. Even if it is the full set of α_i values that defines f, the largest positive (smallest negative) values α_i defining f indicate codons that are preferentially used in X (Y). For 3, we trained (with leave-one-out cross-validation) the linear discriminant function f and tested prediction performance on the remaining data. We did this on organisms represented in 64 dimensions and on 2 dimensions after having applied PCA to the data set. We obtained 13.5% errors in the first case and 15.6% errors in the second. As expected, the number of variables required for optimal discrimination is greater in 64 dimensions (20) than in 2 (7). LDA was done in *VidaExpert*, and the training of the LDA function was done in *R*.

Distances

Let us consider two kinds of point sets representing a genome: the set of codon weights w_i, and its binarized form $\bar{w}_i$, where for each codon i, we approximate to 0 all weights $w_i \neq 1$, i.e., for those codons which are not preferred. Hence, this second set of points is constituted by values 0 and 1 only.

Distances between pairs of organisms are measured as "$\frac{1}{2}\ell_1$-distances" in codon space. Given two genomes G_1, G_2 and two collections $\bar{w}^1, \bar{w}^2$ of binarized weights $\bar{w}_i^1, \bar{w}_i^2$, we define the *binarized distance* between G_1 and G_2 to be

$$d_b(G_1, G_2) = \frac{\sum_{i=1}^{64} |\bar{w}_i^1 - \bar{w}_i^2|}{2} = \frac{1}{2}\ell_1(\bar{w}^1, \bar{w}^2) \qquad (2)$$

The coefficient $\frac{1}{2}$ in front of the usual ℓ_1-distance is considered because we want to count amino acids having different preferred codons exactly once. Intuitively, this distance represents the number of amino acids with different preferred codons. We speak about "binarized $\frac{1}{2}\ell_1$-distance." Similarly, we use $d(G_1,G_2) = \frac{1}{2}\ell_1(w^1,w^2)$ when collections of weights w_i^1, w_i^2 are considered. If not otherwise specified, with "$\frac{1}{2}\ell_1$-distance" we refer to $d(G_1, G_2)$.

A Tree Describing Distances

The tree of figure 2 has been constructed using the unweighted pair group method with arithmetic mean (UPGMA) as a distance method (with the program Neighbor, integrated in the PHYLIP package, and available at http://evolution.genetics.washington.edu). Figure 2 is used to illustrate an approximated distance between pairs of organisms (such a distance, read out of the tree by adding up the values along the shortest path that connects two leaves, is a priori neither an upper bound nor a lower bound to the effective distance). No fact is inferred from the tree in this article, besides the observation that it organizes organisms in three classes reflecting *AT*, *GC*, and translational bias. The same three groups of organisms are found by using distance methods as *NJ*, *BIO-NJ*, and *NNET* (from the SplitsTree 4.0 package).

Choice of Metrics and Over-Represented Families of Organisms

To choose a meaningful metrics is non-trivial. The $\frac{1}{2}\ell_1$-metrics and its binarized version are justified by simple intuition. Conclusions drawn using Euclidean metrics (less obvious to justify) remain compatible with our results. The three large families collecting *GC* rich, *AT* rich, and translationally biased organisms, for example, which are visualized in the tree of figure 2 for $\frac{1}{2}\ell_1$-distances, and in the Supplementary Material online for $\frac{1}{2}\ell_1$-binarized and Euclidean distances, are comparable. In particular, relative distances among organisms in these codon spaces remain *coherent*.

A few over-represented bacterial species, like γ-proteobacteria and firmicutes, bias the set of available sequenced genomes. Also, Archaea are relatively few compared to Eubacteria. As a consequence, a comparative analysis of species drawn on such a sample needs to be carefully evaluated. Namely, the clustering suggested by figure 2 might not reveal some features of the organisation due to over- and under- organism representation.

Prokaryotes Characteristics

For all references to the ecology, genetics, and physiology of prokaryotic organisms we follow closely the work of Balows et al. (1992) and Madigan, Martinko, and Parker (2000).

Criteria to Detect Codon Bias Signatures and Tendencies

It is commonly recognized that organisms might be subjected to codon biases of different origins. There are examples for which it is rather difficult to decide what is the most dominant codon bias, if it exists at all, as for *Helicobacter pylori* for instance, a rather homogeneous genome (Lafay, Atherton, and Sharp 2000) or for *Treponema pallidum*, which displays both a strong *GC*-skew bias (Lafay et al. 1999) and a strand bias. In fact, it seems more appropriate to think of biases in a "continuum" way instead of considering them as clear-cut properties, and to think that different biases might be present at the same time, with different strengths. Numerical criteria to detect the *tendency* of a genome toward a bias and the strength of this bias are desirable, and we shall provide a solution to this question.

The idea supporting our method is to correlate codon biases of different origins with a common measure, the *CAI* values of genes. The approach is justified by the fact that *CAI* is a *universal* measure to study codon bias and it has been proven to be highly correlated with dominant biases of different nature (Carbone, Zinovyev, and Képès 2003). For each genome and each kind of bias, we compute a correlation coefficient that expresses the strength of the bias for the genome. The numerical coefficients can be used to rank different genomes with respect to a given bias, and to detect whether a genome has a *tendency* for a bias (in this case, the correlation coefficient is expected to be rather high) or not.

For each criterion, we suggest a *threshold*, that is an indicator for *strong* bias; formally, if $T > 0$ is the threshold, then for all genome G and bias B, B is a strong bias for G if and only if the coefficient computed for the bias B is bounded by T. Thresholds allow us to *automatically* identify strong biases and define the *codon bias signature* of an organism to be the collection of its strong biases. The

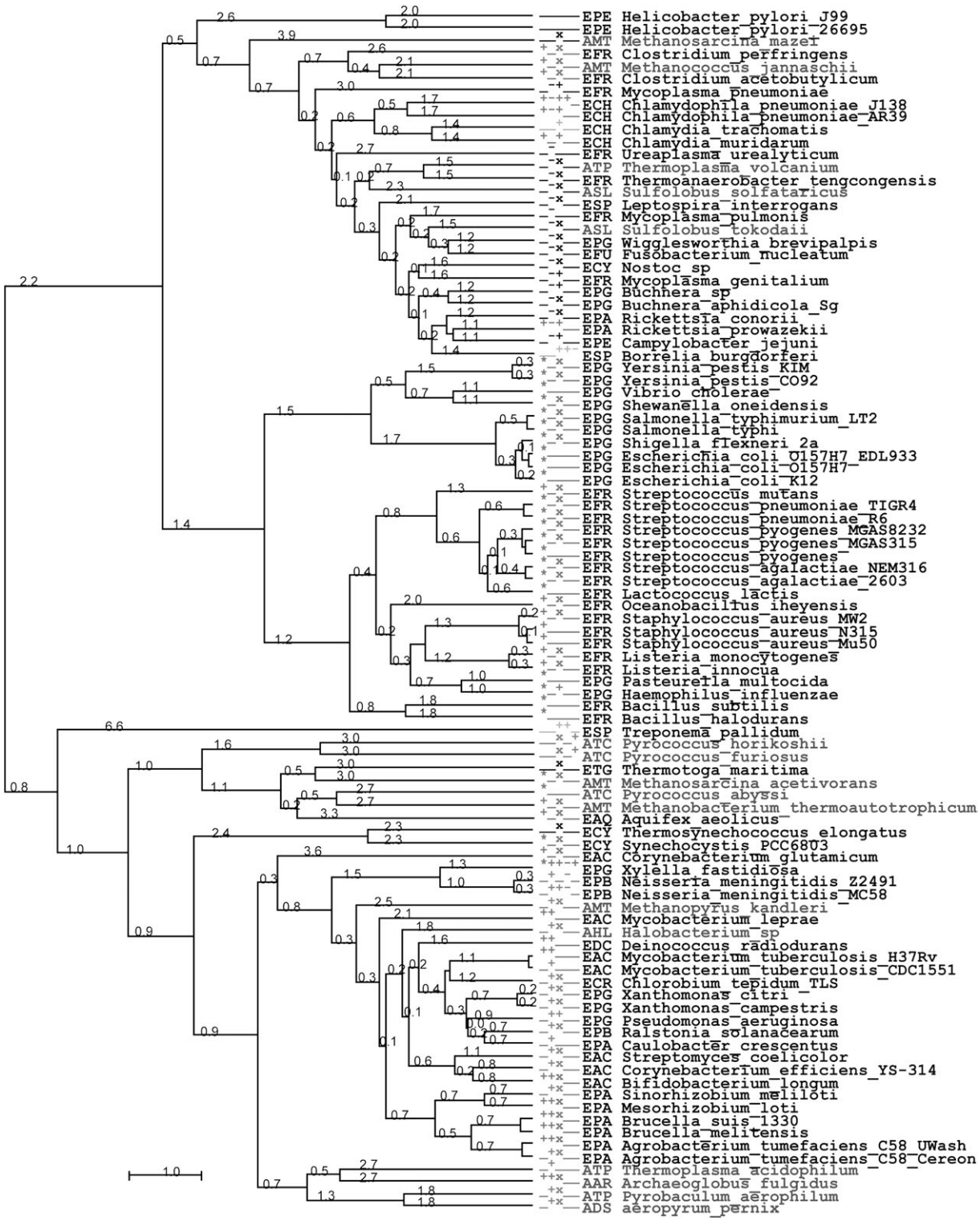


FIG. 2.—Tree constructed from the $\frac{1}{2}\ell_1$-distance matrix for the organisms in table 1. A three-letter code describes whether the organism is an Archaea (*A*, blue) or a Eubacteria (*E*, red), and its genus (shortened to two letters). Five consecutive positions occupied by the symbols +, *, –, _, × are interpreted as follows: 1. + translational bias by ribosomal criterion, * translational bias by both strength and ribosomal criteria; 2. + GC3-content, – *AT*3-content; 3. + strand bias, × no replication origin is known; 4. + *GC*-skew bias, – *CG*-skew bias; 5. + *AT*-skew bias, – *TA*-skew bias. The symbol _ indicates that no bias is present.

thresholds that we propose validate all strong biases that have been detected for organisms previously studied in the literature.

Even though a signature allows for an immediate "picture" of a genome, it is important to stress that signatures provide only a partial description, and that the most accurate one corresponds, in our view, to the entire collection of numerical values associated to different biases which highlights "tendencies." We say that a genome has a *strong tendency* toward a bias B if the coefficient computed for B is bounded by $T - \epsilon$, for some small $\epsilon \geq 0$. If $\epsilon = 0$ then we speak about a strong bias. Since the threshold T is defined for all genomes, it gives the possibility to compare genomes (and in particular to compare them through their signature).

If a genome presents no strong bias on nucleotide frequencies, we say that it has a *weak tendency* toward the content bias that presents the largest coefficient (in absolute value). This notion describes the nucleotide evolutionary pressure of a genome, and it is best used in the analysis of rather homogeneous genomes; for instance, it allows us to say that *H. pylori*, despite its empty signature, has a weak tendency toward *GC*-skew bias (in agreement with Grigoriev [2000]).

Ribosomal Criterion

This simple statistical test detects translational bias and it relies on the idea that for translationally biased genomes, the pool of ribosomal proteins has high *CAI* score compared to the average *CAI* value of all CDSs. In general, ribosomal proteins are not expected to be highly biased, and in particular, if bias exists, the interval within which *CAI* scores for ribosomal proteins vary might be rather different from genome to genome. We use this second observation to measure translational bias strength for a genome. More formally, we compute the *average* $\overline{CAI}$ and the standard deviation σ_{CAI} for *CAI* values of all CDSs, and define a z-score value for those CDS r annotated as ribosomal proteins; i.e., $(CAI(r) - \overline{CAI})/\sigma_{CAI}$. We call $\bar{z}_{Rib}$ the average of z-scores for ribosomal proteins and define the following criterion: an organism characterized by *translational bias* is expected to have high $\bar{z}_{Rib}$, i.e., >1.

Because ribosomal protein coding genes are highly conserved across species, they can easily be accessed by homology in organisms not yet well investigated, and this renders the criterion amenable.

Strength Criterion

This is an heuristic criterion for the detection of translational bias, which does not use any information coming from annotation of ribosomal proteins, and it consists solely of *statistical* analysis of CDSs. Let $w_i^k(G)$ be the weight calculated as in equation (1) over the whole set G of CDS for organism k, and let w_i^k be the weight calculated over the set of most biased genes S for k. Because of the existence of a particularly *strong* dominant codon bias in organisms affected by translational bias (that is, the frequency of a preferred codon compared to the frequencies of its synonymous codons is much higher in the set of most biased genes S than in the whole genome G), one expects the difference between $w_i^k(G)$ and w_i^k to be large, and to use this quantity as a criterion to detect translational bias. Thus, we use the $\frac{1}{2}\ell_1$-distance between $w_i^k(G)$ and w_i^k, defined as

$$d(w^G, w^S) = \frac{\sum_{i=1}^{64} |w_i^k(G) - w_i^k|}{2} = \frac{1}{2}\ell_1(w^G, w^S) \qquad (3)$$

to discriminate those organisms that likely are affected by translational bias by requiring $d(w^G, w^S) > 8$ (where $d(w^G, w^S) \in [1 \ldots 13]$ for the 96 organisms in table 1; see *Supplementary Material*). To explain the intuition behind this formula, let us consider its binarized version, namely the case where we set $\bar{w}_i^k(G) = 0$ if $w_i^k(G) \neq 1$, and $\bar{w}_i^k = 0$ if $w_{i,j} \neq 1$. In this simplified form, equation (3) counts the number of amino acids that have different preferred codons in the entire genome and in the set of most biased genes.

Such a numerical criterion, being based only on a statistical analysis of CDSs, is highly desirable but it does not provide a sufficient and necessary condition for translational bias. In fact, *not all* organisms satisfying translational bias are detected, and some extra organism, like *Xylella fastidiosa*, might be erroneously selected. We propose it though, because the combination of the two criteria for translational bias detection allows us to discriminate those genomes that are *strongly* translationally biased (that is those satisfying both criteria) from those that are *weakly* so (that is those that only satisfy the ribosomal criterion).

Content Criterion

*GC*3 bias is detected by comparing the *GC*3-content of each CDS with the corresponding *CAI* value, and asking the correlation coefficient (on all CDSs) be >0.7; correlation <-0.7 detects *AT*3-bias. *GC*-skew bias is detected with a correlation coefficient >0.5; correlation <-0.5 detects *CG*-skew bias. Thresholds 0.5 and -0.5 define *AT*-skew and *TA*-skew bias.

Strand Criterion

Strand bias says that most biased genes of a (circular or linear with bidirectional replication) genome are preferentially distributed in precisely one of its strands (typically the leading strand). This definition does not depend on gene function, and it allows us to detect strand bias for genomes whose strongest bias is of any origin. In particular, we make no hypothesis on high expressivity for most biased genes, and this is in concert with the finding of Rocha and Danchin (2003), who show that essential genes more than highly expressed are located on leading strands.

To detect strand bias we verify the statistical hypothesis on the two distributions of *CAI* values of genes in leading and lagging strands of chromosomes (see discussion on plasmids below). This has been done only for those genomes whose replication origin is known. We compute the t-value representative of the difference between the means of the two distributions and say that organisms with average t-value (taken as an absolute value) >0.25 have leading-lagging strand bias.

This criterion provides a way to check for strand bias which is independent of that based on the co-existence

between strand bias and *GC*-skew bias, proposed by Sueoka (1962) and McLean, Wolfe, and Devine (1998). (The use of this idea to detect replication sites is envisageable but out of the scope of this study.)

The Number of Codon Bias Signatures Is Limited

Translational bias is strongly correlated with *GC*3 content (in the sense that *GC*3 is the most prominent compositional content of a translationally biased organism), and most strand biased genomes in our collection are either *AT*3 or *GC*3. These observations justify the limited number of signatures we found, as it appears in figure 2. Also, it is worth mentioning that we detected three genomes with *GC*-skew bias, three with *CG*-skew bias, and three with *AT*-skew bias, but only 1 with a *TA*-skew bias.

Validation of Signatures and Tendencies

Tendencies and signatures obtained by applying the simple numerical criteria above (see figure 2 and *Supplementary Material* for the complete list of signatures and tendencies for genomes in table 1) are validated on known cases, and for some genomes, predictions have been drawn:

Pseudomonas aeruginosa is GC3 biased but also strand biased

Drawn from calculations of *CAI* values which were based on misleading manual selections of sets of most biased genes (Grocock and Sharp 2002; Gupta and Ghosh 2001; Kiewitz and Tümmler 2000), the dominating codon bias of *P. aeruginosa* gave origin to controversial opinions on the biology of this organism. This makes this genome a good test case for our criteria, which are also based on *CAI* analysis. In agreement with Grocock and Sharp (2002), we detect that *P. aeruginosa* has a very strong *GC*3-bias (see also Carbone, Zinovyev, and Képès [2003]), but also a strong tendency toward *GC*-skew, and a strong strand bias.

Genomes with Strand Bias and No GC-Skew Bias

Strand bias (0.95) is detected for *Haemophilus influenzae*, a genome with no *GC*-skew bias (0.05). Other organisms also display strand bias but no *GC*-skew bias: *Mycoplasma pneumoniae*, *Buchnera* sp., *M. genitalium*, and *Chlamydia trachomatis*. Besides *P. aeruginosa*, other organisms display strand bias and just a strong tendency toward *GC*-skew: *C. pneumoniae AR39*, *C. muridarum*, *C. jejuni*. Some others display both biases as strongly, like *B. burgdorferi* (with strand bias at 1.89 and *GC*-skew bias at 0.77) (Lafay et al. 1999; Carbone, Zinovyev, and Képès 2003).

Translationally Biased Genomes

In figure 1 (bottom), $\bar{z}_{Rib}$ and $d(w^G, w^S)$ values show that organisms known to be translationally biased are separable from all others with respect to suitable thresholds. Some of these organisms have been previously reported in the literature and validate our separation (Gouy and Gautier 1982; Sharp and Li 1987; Sharp et al. 1988; Médigue et al. 1991; Shields and Sharp 1987; Carbone, Zinovyev, and Képès 2003).

To validate the ribosomal criterion, we looked at the sets of most biased genes *S* determined by the evaluation of *CAI* values for each genome satisfying the ribosomal criterion, for which we claim a translational bias. We checked the annotation of the genes in the set of most biased genes, and we positively verified that the genes which typically are representative of translational bias, such as ribosomal, glycolytic, dehydrogenase, enolase, elongation factors, photo-system, heat-shock, and cold-shock proteins were consistently present in the set.

Weak Forms of Translational Bias—*Mycobacterium tuberculosis*

The coupled use of the two criteria detecting translational bias allows us to identify those genomes for which translational bias is *weakly* present. An example is *M. tuberculosis*, for which only one of the two strains *H37Rv* ($\bar{z}_{Rib} = 1.14$) and *CDC* 1551 ($\bar{z}_{Rib} = 0.87$) is characterized by translational bias, even though both strains have comparable codon preferences. Translational bias for this species cannot be detected by strength criterion, and this is an indicator for weak detection. This observation is compatible with the findings of de Miranda et al. (2000).

Tendencies Toward Translational Bias—*Rickettsia prowazekii*

For those organisms which only tend to the threshold $T = 1$, i.e., $\bar{z}_{Rib} = 1 - \epsilon$ for some small $\epsilon \geq 0$, one can check whether ribosomal proteins are present in the set of most biased genes or not. *R. prowazekii*, for instance, has $\bar{z}_{Rib} = 0.98$ and a set of most biased genes *S* whose 88% is made of ribosomal proteins. We conclude that it has a strong tendency toward translational bias, contrary to what has been claimed by Anderson and Sharp (1996), on the basis of a comparison of the amino acid composition patterns of 21 *R. prowazekii* proteins with that of a homologous set of proteins from *Escherichia coli*; there, it has been argued that translational selection has been ineffective in this species under the base that synonymous codon usage patterns are roughly similar in the 21 proteins, even though the data set includes genes expected to be expressed at very different levels. A finer analysis of the space of all *R. prowakezii* proteins indicates that the set of ribosomal proteins in *R. prowakezii* is separable from all other proteins by a linear discriminant function with no false positive nor true negatives (Sn = 100 and Sp = 100). This means that ribosomal proteins occupy a particular location in codon bias space and that there is a pressure on codon bias (especially on codons *aaa*, *aga*), even though the set of ribosomal proteins has unusually broad dispersion (compared to the typical case where translational bias is present).

Genomes with Empty Signature and Weak Tendencies

A few genomes display an empty codon bias signature, indicating the absence of strong biases of any

particular type, as for instance *Helicobacter pylori* (Lafay, Atherton, and Sharp 2000), but also *Thermosynechococcus elongatus*, *Thermotoga maritima*, and the Archaea *Methanosancina mazei*. *H. pylori* has a weak tendency toward *GC*-skew bias (Grigoriev 2000); *T. maritima* has a weak tendency toward *GC* and *GC*3 (Zavala et al. 2002); *T. elongatus* and *M. mazei* have weak tendencies toward *GC*3 and *AT*3 biases, respectively.

Microbial Codon Space and Lifestyle

A 2-dimensional projection of the 64-dimensional space of Eubacteria and Archaea organisms is illustrated in figure 1 (top), where the first principal PCA component (*x*-axis, explaining 45% of the variance) corresponds to *GC* content and the second principal PCA component (*y*-axis, explaining 13% of the variance) corresponds to optimal temperature growth. A non-linear shape in the distribution of points (as viewed best in 3D, not shown), roughly resembling a "horseshoe," splits the set of organisms into two well-defined subsets: the top half of the horseshoe is made by *hyperthermophiles* which lie "above" *thermophiles* (all Archaea in table 1 except the mesophilic *Halobacterium* sp., and the three hyperthermophilic bacteria *Aquifex aeolicus*, *T. maritima*, and *Thermoanaerobacter tencongensis*), and the bottom half by *mesophiles* (all Eubacteria in table 1 except the three hyperthermophilic species indicated above, and the mesophilic *Halobacterium* sp.). The division suggests a separation of the three lifestyle domains (hyperthermophiles, thermophiles, and mesophiles) based on codon bias in agreement with the division observed by Lynn, Singer, and Hickey (2002) for 40 organisms, and by Kreil and Ouzounis (2001) and Tekaia, Yeramian, and Dujon (2002) for 27 and 56 organisms, and based on amino-acids composition. (See also Torres de Farias and Manhães Bonato [2002] and Lobry and Chessel [2003].)

Codon Bias and Optimal Growth Temperature

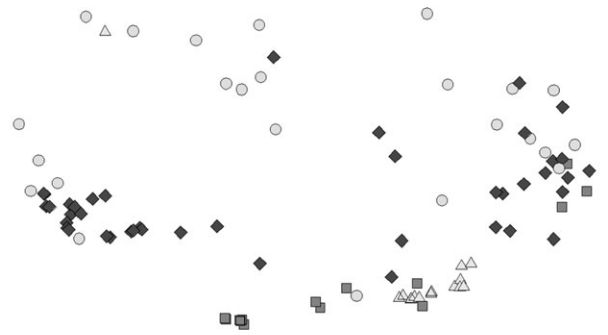
To study codon bias differences in (hyper)thermophilic and mesophilic genomes, we used LDA and determined that codons *cgt*, *cgc* are positive indicators for mesophiles, while *agg*, *ata*, *gga*, *cta*, *acg* are negative indicators; *agg* is a positive indicator for thermophiles; *cta*, *agt*, *ggg*, *agg*, *cca*, *ctc* are positive indicators for hyperthermophiles, while *cgc*, *cat*, *ggc*, *tcg* are negative indicators. These preferential codons code for Arg+Ile+Gly+Leu and separate mesophiles from (hyper)thermophiles; it is interesting to notice that distinguished preferred codons coding for Arg separate (hyper)thermophiles (*agg*) from mesophiles (*cgt*, *cgc*). The small number (4) of thermophiles is detected with preferred codon *agg* coding for Arg; hyperthermophiles are separated on preferential codons coding for Leu+Gly+Ser+Arg. There is no codon coding for Glu, Tyr or Val that is preferential only in mesophiles and thermophiles, or only in hyperthermophiles: the role played by these three amino acids (Kreil and Ouzounis 2001; Tekaia, Yeramian, and Dujon 2002) in hyperthermophilic proteins remains transparent at the nucleotide level. We conclude that while the division between (hyper)thermophiles and mesophiles is sharply determined by preferential codon bias (on Arg+Ile+Gly+Leu), the transition between hyperthermophiles and thermophiles is less clear and should be understood as gradual. Our set of 96 organisms confirms the hypothesis of gradual transition discussed in Tekaia, Yeramian, and Dujon (2002).

Translational Bias for Hyperthemophiles and Mesophiles

As expected, regions in codon space that collect the most *GC*3 and *AT*3 biased genomes, that is the two most extreme regions of the genomes distribution along the first principal PCA axis (interpreted by *GC*-content), contain (hyper)thermophiles and mesophiles. It is surprising though, to see that translationally biased organisms cluster in two groups localized in distinguished sites of codon space, one collecting (hyper)thermophiles and the other mesophiles. Knowing that preferred codons and isoacceptor tRNA content exhibit a strong positive correlation (Ikemura 1985; Bulmer 1987; Gouy and Gautier 1982), and that tRNA isoacceptor pools affect the rate of polypeptide chain elongation (Varenne et al. 1984; Buckingham and Grosjean 1986), this means that the set of preferred codons correlated with isoacceptors tRNA leading translational bias for (hyper)thermophiles is different than that for mesophiles. Applying LDA, we observe that a positive indicator for translationally biased (hyper)thermophiles genomes is *agg*, that positive indicators for translationally biased mesophiles are *gct*, *ctt*, *ttc*, *cag*, *act*, *cga*, *gtt*, *cgg*, *cat*, *tca*, *tat*, *cac*, *gtg*, *acc*, *aac* and that negative indicators are *acg*, *tcc*, *agc*, *aca*, *ccg*, *cca*, *agt*, *gca*, *aga*. If selection depended merely on some property of mRNAs that is important under conditions of high temperature (Lynn, Singer, and Hickey 2002), like increased mRNA stability at high temperature for instance, it is not clear whether translational efficiency could be effectively distinguished in hyperthermophiles. We showed that translational bias in hyperthermophiles can be clearly detected through codon analysis.

Aerobic and Anaerobic Respiration

Organisms sharing the same respiratory characteristics tend to group together in codon space as illustrated in figure 3. Linear discriminant analysis demonstrates that clusters in the figure are not an artifact of the 2-dimensional projection. Indeed, four groups are sharply characterized by distinguished sets of preferred codons with highly significant (positive and negative) separation coefficients: *tct*, *gtt*, *gcg* are positive indicators, and *ctt*, *tca*, *tcg*, *gtc*, *agt*, *aag* are negative indicators for facultative anaerobism; *tca*, *ctt*, *gac*, *tac*, *ttc*, *cac*, *aac* are positive indicators and *tct*, *tgc*, *aga* are negative indicators for facultative aerobism; *ccg*, *tta*, *gcg*, *cac*, *aaa*, *ctc*, *ctg*, *agt*, *ggg*, *gga*, *gtc*, *cca*, *ggc* are positive indicators and *cgt*, *tcc*, *ccc*, *cta*, *acc*, *gtg*, *tcg*, *cat*, *gaa* are negative indicators for anaerobism; *cgc*, *gta*, *gaa, caa*, *tgc*, *ccc*, *cct*, *gtg* are positive indicators and *aaa*, *ccg*, *ata*, *ggg* are negative indicators for aerobism. Within thermophiles, facultative aerobic are represented by *Pyrobaculum aerophilum*, and we



	Aer	An	FAer	FAn	Aer	An	FAer	FAn
Aer	39	7	0	3	36	3	0	2
An	2	17	1	0	4	21	1	0
FAer	1	0	14	0	2	0	14	0
FAn	0	1	0	11	0	1	0	12

FIG. 3.—Organisms and their respiration characteristics: aerobic (rhomboids), anaerobic (circles), facultative aerobic (triangles), facultative anaerobic (squares). Prediction tables after LDA training (leave-one-out cross-validation) on two dimensions after PCA (left, error rate 0.15) and on 64 dimensions (right, error rate 0.13).

expect new sequenced facultative aerobic thermophiles to be grouped in the same part of codon space.

The only facultative anaerobic organisms in table 1 are γ-proteobacteria; their position in space is not due to their phylogenetic closeness because all γ-proteobacteria which are aerobic are located in a sharply separated part of the space (not shown). The same holds for facultative aerobic represented solely by firmicutes; anaerobic firmicutes lie far apart in our codon space. It is important to stress that the transition between clusters should be considered as gradual rather than a clear-cut separation. In particular, aerobic and facultative aerobic organisms tend to be located closely as well as anaerobic and facultative anaerobic organisms.

Validation of Distances with Respect to Genomic Variability and Codon Bias

In our codon bias space, organisms distances vary considerably in a scale from 0 to 24. At distance <1 we typically find different strains of the same organism, and, usually, different species lie at distance >1. Distances reflect the genomic variability within the same species or within the same phylogenetic group, and they provide a numerical description of important differences in codon bias signatures among organisms within the same species or phylogenetic branch. Roughly speaking, one might estimate two organisms to be *close* in codon bias space, *weakly close, far* and *very far* if their distance lies in the intervals [0, 7), [7, 12), [12, 16), and [16, 24), respectively. Largest distances are detected between pairs of organisms which are *AT*3 and *GC*3 biased.

Organisms of Close Genomic Relationships: Some Examples

Consistently with what one expects, *Shigella flexneri 2a* lies at distance <0.5 from all strains of *E. coli*, while the two strains of *H. pylori* lie at distance ≈ 4, reflecting a certain degree of genomic and allelic diversity among the two strains (Wang, Humayun, and Taylor 1999), but not a large one as observed by Alm et al. (1999). Similarly, the three closely related *Mycoplasma*, known to have quite different genome composition, lie at distance ≈4–6.

Phylogenetic Groups Organization in Codon Space

It is instructive to analyze phylogenetic groups through codon bias differences which can be detected in codon space. For a first rough impression, one can look at the tree in figure 2, which represents $\frac{1}{2}\ell_1$-distances among genomes and groups—together *three* large families of organisms that turn out to be characterized by *GC* rich, *AT* rich, and translationally biased genomes. The sister tree collecting translationally biased genomes separates Firmicutes from γ-Proteobacteria, and the sister subtree corresponding to *GC*-rich genomes groups in different subtrees translationally biased Archaea, *GC*3-biased Eubacteria, and *GC*3-biased Archaea. A finer analysis leads to the observation that for those phylogenetic branches that present a variety of different signatures within the branch, organisms displaying the same signature are localized in the same region of codon space and are usually classified within a known subfamily of the phylogenetic branch.

γ-Proteobacteria

They split into 6 groups (see fig. 4), and the large distance among some of the groups (at times >15) is reflected in the signature: subgroups *G*2 and *G*6 (Enterobacteriales), *G*3 (Pasteurellales), *G*5 (Vibrionales and Alteromonadales) collect translationally biased genomes, and *G*1 (Enterobacteriales), *G*4 (Xanthomonadales) are *AT*3, *GC*3 biased.

α-, β- and ϵ-Proteobacteria

ϵ-Proteobacteria present a variety of codon biases; *Helicobacter* strains are rather homogeneous genomes and *Campylobacter jejuni* (at distance ≈8 from *Helicobacter* strains) displays *AT*3 bias and strand bias. Within the α-proteobacteria group, *Rickettsia* are *AT*3 biased whereas Caulobacterales and Rhizobiales are *GC*3 biased, and most of them display translational bias as well. β-proteobacteria are *GC*3 biased (see fig. 4).

Firmicutes and Actinobacteria

Mollicutes and Clostridia are *AT*3 biased or tend toward *AT*3 bias. Lactobacillales and Bacillales are all translationally biased, and Actinobacteria are either *GC*3 biased or tend toward *GC* bias.

Chlamydiales

Chlamydiales are at close distance and display a common codon bias signature: they are strand biased, and they all tend toward *AT*3 bias and translational bias.

Spirochaetales

Treponema pallidum and *B. burgdorferi* are characterized by strand and *CG*-skew bias, while *Leptospira interrogans* is *AT*3 biased. While *B. burgdorferi* and *L.*

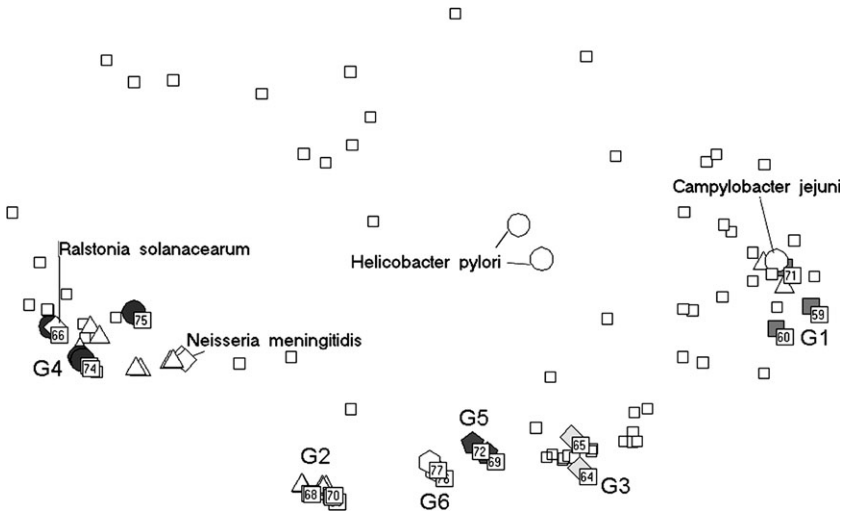


FIG. 4.—Proteobacteria α (white triangles), β (white rhomboids), ϵ (white circles). γ-proteobacteria (filled shapes) cluster into six sets (numbers as in table 1). *G*1: *B. aphidicola Sg*, *Buchnera sp*, *W. brevipalpis*; *G*2: *Salmonella*, *Escherichia*, and *S. flexneri 2a*; *G*3: *P. multocida*, *H. influenzae*; *G*4: *X. fastidiosa*, *X. campestris*, *X. citri*; *G*5: *V. colerae*, *S. oneidensis*; *G*6: *Yersinia*.

interrogans group in the same sister subtree (fig. 2), *T. pallidum* is far away from all species (see long branch in fig. 2).

Archaeal subgroups

Because of the restrained number of available complete archaeal genome sequences (16) we can only try to consider *Thermoplasmales*, *Sulfolobales*, *Thermococcales*, and *Methanosarcinales*. These species span a large range of codon compositions and biases, going from *GC*3 to *AT*3 along the first principal PCA component. *Sulfolobales* (*AT*3 biased) are located within a small relative distance ≈3; *Methanosarcinales* have a rather large relative distance ≈11, and they lie in two distinguished subtrees in figure 2: *M. mazei* tends toward *AT*3 bias, and *M. acetivorans* is translationally biased and tends toward *GC*3; *Thermococcales* are grouped in the same subtree in figure 2 because either they have a *AT*3 biased genome (*Pyrococcus furiosus*, *P. horikoshii*) or they tend to be *AT*3 biased (*P. abyssi*, which displays a translational bias). The *Thermoplasmales*, *Thermoplasma acidophilum* is *GC*3 biased and *T. volcanium* is *AT*3 biased.

Discussion

Detection of Strong and Weak Forms of Bias

The numerical criteria that we have introduced to detect codon bias allow us to treat genomes uniformly and faithfully compare species and strains. Our numerical methods allow for (1) a *quantitative* evaluation of whether an organism has a *strong* or *weak* form of bias (by computing the distance from the corresponding threshold) and (2) the detection of co-existing multiple biases (by using distinguished criteria). These features should be compared with the analysis demanded by methods like PCA and correspondence analysis, where principal components might be far from being unambiguously interpretable giving origin to misleading conclusions as discussed by Perrière and Thioulouse (2002). In particular, the interpretation of more than the first two or three principal components usually becomes quite difficult.

Thresholds are indicators of high bias, and their values confirm all previous studies; however, one expects formal statistical approaches to be employed for further tuning once larger sets of organisms become available. Also, our numerical approach provides, for each bias, quantitative values ranging within a continuous interval. Based on these values, we defined strong, weak, and absent forms of bias, but finer classifications are envisageable and can be introduced with the help of new appropriate definitions.

Codon Weights versus Codon Usage and Comparison of Spaces

"Preferred" codons, defined by high codon weights, should not be confused with "most frequent codons," defined by high codon usage. This is shown by Carbone, Zinovyev, and Képès (2003) (fig. 2) through an analysis of codon preferences for *H. pylori*, *E. coli*, and *C. elegans*. In *H. pylori*, a rather homogeneous genome, preferred codons are the most frequent codons, but for *E. coli* and *C. elegans*, preferred codons calculated on the set of most biased genes *S* are not the same as preferred codons computed over the whole genome, that is, the most frequent codons. We used codon weights to represent an organism and to suitably define a space of organisms; codon usage instead is not a good measure to accomplish this task. To verify this, we constructed a distance tree (based on $1/2\ell_1$ distance metric) among organisms represented as 64-dimensional vectors of codon usage (*CU*)—i.e., frequencies calculated over the whole genome, and of codon usage calculated over the set of most biased genes *S* (CU_S) (see trees in the Supplementary Material online). The same rough division among AT-rich, GC-rich, and translationally biased genomes seen to be true for the tree based on codon weights (fig. 2), holds true for the tree constructed with CU_S, but it is not satisfied by the tree based on *CU*. In

particular, for this latter tree, closely related phylogenetic groups sharing the same codon bias, like the three γ-proteobacteria Xantomonadales, *X. fastidiosa*, *X. campestris*, and *X. citri*, are not grouped together, contrary to what happens in our codon space (see *G*4 in figure 4). These observations make it inappropriate to employ codon usage for organism comparison. (Notice that the distinction between preferred codon and most frequent codon was also exploited in our analysis to define the strength criterion.)

GC Bias and Translational Bias

Our codon space demonstrates that translational bias is *independent* of GC bias. There are organisms for which this is not the case, as for *Drosophila* (Kliman and Hey 1994) for instance, where *GC* content is uniformly higher at silent sites in coding regions than in putatively neutrally evolving introns. Figure 1 (top) shows a wide distribution of translationally biased genomes (red) extending from the GC-rich region (left) toward the *AT*-rich region (right) of the space.

Codon Bias Space, Physiology and Habitat

Reasons supporting an evolutionary convergence of codon bias for organisms sharing similar physiology *and* living in similar habitats, might include (1) the need for a successful exchange of genes by lateral transfer, (2) the sharing of physical parameters such as temperature (preferred amino acids and codons related to thermal adaptation), (3) the sharing of chemical parameters such as nutrient supply (that would differentially affect pyrimidine and purine production), and (4) the sharing of biological parameters such as, for pathogens, the management of genetic variability through codon usage (to escape the immune system for instance).

Examples supporting these reasons are several. The bacterium *Aquifex aeolicus*, for instance, occupies the hyperthermophilic niche otherwise dominated by Archaea. After genome analysis, it seems likely that the archaeal genes in *Aquifex* have been introduced by horizontal gene transfer, on top of a typical bacterial gene repertoire, and have been retained owning to the specific selective advantage they provided by enabling the bacterium to thrive in high-temperature habitat (Aravind et al. 1998). A similar gene transfer has been observed for another hyperthermophilic bacterium, *Thermotoga maritima*. This transfer heavily influenced the codon bias of the two bacterial genomes, *Archaeoglobus fulgidus* and *Methanobacterium thermoautotrophicum*, which are also close to extreme thermophilic Archaea.

Other important examples are pathogens or symbionts. Most Eubacteria collected in the sister subtree of *AT*-rich genomes (*Buchnera*, *Chlamydiae*, *Spirochaetes*, *Mycoplasma*, *Rickettsia* genera, ϵ-proteobacteria; see figure 2, and also Rocha and Danchin [2002]), rely on their host for survival. A few other obligatory pathogens, such as *M. tuberculosis* and *M. leprae* are *GC*3 biased, or translationally biased, as *Shigella flexneri*, *Haemophilus influenzae*, and *Pasteurella multocida*. For these latter species, it might be that symbiosis genes are located in "islands" of lower $G + C$ content, as it is the case for *Mesorhizobium loti*, a *GC*3-biased pathogen hosted by *Lotus japonicus*. The higher energy cost and limited availability of *G* and *C* over *A* and *T*/*U* could be a basis for the understanding of the differences of free-living bacteria and obligatory pathogens (Rocha and Danchin 2002).

Sharing merely the habitat is not enough to be close in codon space. In this respect, *Staphylococcus aureus* shares common ecology but neither physiology nor genetics with *Neisseria meningitidis*. These two commensal bacteria have very different ways to survive outside a host, to colonize it, and be toxic. We observe them to be located rather far in our space ($\approx$ 13). On the other hand, *C. jejuni*, an extracellular pathogen of the digestive tract, and *Rickettsia*, an obligate intracellular parasite, would be distantly related in this space if only habitat were to matter, while they lie only at distance $\approx$2.

These examples suggest that it makes sense to investigate the connection between codon bias and environmental and physiological conditions, but that the task is far from being simple. A rigorous mathematical analysis that could consolidate this intuition would require the definition of a set of parameters to describe the physiology and ecology of Eubacteria and Archaea. This may include a description of the biotopes encountered by the bacteria, the doubling time, the genome size, the number of ribosomal operons, and so on. This characterization would allow us to define a "physiology space" and a suitable distance within it. Such a space could then be compared to the codon bias space defined in this article and the hypothesis could be tested. Notice that, if the intuition were confirmed, the detection of codon bias signatures for upcoming genome sequences could become a very important tool with which to infer valuable information on the physiology, ecology and possibly, the ecological conditions under which bacterial organisms evolved. For some of these organisms, this information cannot be otherwise obtained. (See also Wagner [2000].) In particular, new biological questions could arise, even on distantly related organisms like *Thermoplasma volcanium*, which is known to resemble bacterial mycoplasmas in that it lacks a cell wall and which turns out to be close to the mycoplasmas *M. genitalium*, *M. pneumoniae*, and *M. pulmonis* ($\approx$6) in our codon space.

Phylogenies and Codon Bias

Controversial phylogenies have been proposed several times, and reasons for these misinterpretations are several (Gribaldo and Philippe 2002). Many of the misleading examples are due to codon bias which, at times, depends on lateral gene transfer among phylogenetically unrelated taxa thriving in the same ecological niches (Ruepp et al. 2000). This is the case, for instance, for the unexpected relationship among *Thermoplasmales* and *Crenarchaeota* (Korbel, Huynen, and Bork 2002), which we find are indeed close in codon space. Our analysis is transversal to phylogenetic classifications and can help to refine the analysis of phylogenetic branches. One example are γ-proteobacteria, which we have seen divided into six distinct groups in codon space. Another example is

Deinococcus radiodurans, positive to the Gram coloration but deprived of the external membrane, unlike Gram-positive organisms. It is located close to the Gram-positive location in our space (with both a GC and a translational bias), as well as in many phylogenetic reconstructions (Daubin, Gouy, and Perrière 2002), while it is expected to have a basal position among bacteria (Woese 1987). It is possible that these controversial phylogenetic positions are simply due to the high GC content of this genome, but it might be also possible that two independent losses of the external membrane have occurred in high-GC-content and low-GC-content Gram-positive bacteria as argued by Daubin, Gouy, and Perrière (2002). If codon space could provide the opportunity to detect and study, in a systematic way, those genomes that live in the same ecological environment, which are susceptible to have similar physiology and to have successfully exchanged genes by lateral transfer, bacteria like *D. radiodurans* might be able to finally find the right place within phylogenetic classifications.

Supplementary Material

The file wdatcodonspace.xls contains basic statistics and codon bias analysis for all organisms, wdistcodonspace.xls contains ℓ_1-distances between organisms, LDA-separationfunctions.xls contains all LDA separation coefficients, Supplementary material.doc contains distance trees and accession numbers. These files are also available at http://www.ihes.fr/~carbone/data.htm.

Acknowledgments

We acknowledge the suggestion of M. Gromov to compare organisms through codon bias and thank two anonymous referees for helpful remarks. A.C. is grateful to S. Gribaldo for a stimulating discussion. F.K. was supported by CNRS, Genopole, and Conseil Régional d'Ile-de-France. A.C. is supported by a grant from the Fondation pour la Recherche Médicale. We also acknowledge the hospitality of the Institut des Hautes Etudes Scientifiques while this work was being accomplished.

Literature Cited

Alm, R. A., L. S. Ling, D. T. Moir et al. (20 co-authors). 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature **397**: 176–180.

Andersson, S. G., and P. M. Sharp. 1996. Codon usage and base composition in *Rickettsia prowazekii*. J. Mol. Evol. **42**:525–536.

Aravind, L., R. L. Tatusov, Y. I. Wolf, R. Walker, and E. V. Koonin. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet. **14**:442–443.

Balows, A., H. G. Truper, M. Dworkin, W. Harder, and K. H. Schleifer. (eds.) 1992. The Prokaryotes. Springer-Verlag, New York.

Buckingham, R. H., and H. Grosjean. 1986. The accuracy of mRNA-tRNA recognition. Pp. 83–126 *in* T. B. L. Kirkwood, R. Rosenberger, and D. J. Galas, eds. Accuracy in Molecular Processes: Its Control and Relevance to Living Systems, Chapman & Hall Publishers, London.

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature **325**:728–730.

Carbone, A., A. Zinovyev, and F. Képès. 2003. Codon Adaptation Index as a measure of dominating codon bias. Bioinformatics **19**:2005–2015.

Daubin, V., M. Gouy, and G. Perrière. 2002. A phylogenetic approach to bacterial phylogeny: evidence of a core of genes sharing a comon history. Genome Res. **12**:1080–1090.

de Miranda, A. B., F. Alvarez-Valin, K. Jabbari, W. M. Degrave, and G. Bernardi. 2000. Gene expression, amino-acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J. Mol. Evol. **50**:45–55.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**:179–188.

Gouy, M., and Ch. Gautier. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**:7055–7070.

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.

Gribaldo, S., and H. Philippe. 2002. Ancient phylogenetic relationships. Theoret. Popul. Biol. **61**:391–408.

Grigoriev, A. 2000. Graphical genome comparison rearrangements and replication origin of *Helicobacter pylori*. Trends Genet. **16**:376–378.

Grocock, R. J., and P. M. Sharp. 2002. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. Gene **289**:131–139.

Gupta, S. K., and T. C. Ghosh. 2001. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene **272**:63–70.

Hand, D., H. Mannila, and P. Smyth. 2001. *Principles of Data Mining*. A Bradford Book, MIT Press, Cambridge, Mass.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**:417–441,498–520.

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

Jansen, R., J. Harmen, H. J. Bussemaker, and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res. **31**:2242–2251.

Karlin, S. 1994. Statistical studies of biomolecular sequences: score based methods. Phylos. Trans. R. Soc. Lond. Ser. B **344**:391–401.

Karlin, S., I. Ladunga, and B. E. Blaisdell. 1994. Heterogeneity of genomes: measures and values. Proceedings of the National Academy of Sciences **91**:12837–12843.

Karlin, S., and J. Mrázek. 1998. Prokaryotic genome-wide comparisons and evolutionary implications. In *Bacterial genomes, physical structure and analysis*, edited by F. J. de Brujin, J.R. Lupski and G.M. Weinstock, Kluwer Academic Publishers, Boston.

———. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J. Bacteriol. **182**:5238–5250.

Karlin, S., M. J. Barnett, A. M. Campbell, R. F. Fisher, and J. Mrázek. 2003. Predicting gene expression levels from codon biases in α-proteobacterial genomes. Proc. Nat. Acad. Sci. USA **100**:7313–7318.

Kiewitz, C., and B. Tümmler. 2000. Sequence diversity of *Pseudomonas aeruginosa*: impact on population structure and genome evolution. J. Bacteriol. **182**:3125–3135.

Kliman, R. M., and J. Hey. 1994. The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. Genetics **137**:1049–1056.

Knight, R. D., S. J. Freeland, and L. F. Landweber. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. **2**: at http://genomebiology.com/2001/2/4/research/0010.

Koonin, E. V., and M. Y. Galperin. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr. Opin. Genet. Dev. **7**:757–763.

Korbel, J. O., M. A. Huynen, and P. Bork. 2002. SHOT: a Web server for the construction of genome phytogenies. Trends Genet. **18**:158–162.

Kreil, P. D., and C. A. Ouzounis. 2001. Identification of thermophilic species by the amino-acids composition deduced from their genomes. Nucleic Acids Res. **29**:1608–1615.

Lafay, B., A. T. Lloyd, M. J. McLean, K. M. Devine, P. M. Sharp, and K. H. Wolfe. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. **27**: 1642–1649.

Lafay, B., J. C. Atherton, and P. M. Sharp. 2000. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology **146**:851–860.

Lin, J., and M. Gerstein. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. **10**: 808–818.

Lin, J., D. Qian, P. Bertone, R. Das, N. Echols, A. Senes, B. Stenger, and M. Gerstein. 2002. GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. Nucleic Acids Res. **30**:4574–4582.

Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13**:660–665.

Lobry, J. R., and D. Chessel. 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J. Appl. Genet. **44**:235–261.

Lynn, D. J., G. A. Singer, and D. A. Hickey. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res. **30**:4272–4277.

Madigan, M. T., J. M. Martinko, and J. Parker. 2000. Brock Biology of Microorganisms, 9th edition. Prentice-Hall Inc., Englewood Cliffs, N. J.

McLean, M. J., K. H. Wolfe, and K. M. Devine. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryotic genomes. J. Mol. Evol. **47**:691–696.

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

Mrázek, J., D. Bhaya, A. R. Grossman, and S. Karlin. 2001. Highly expressed and alien genes of the *Synechocystis* genome. Nucleic Acids Res. **29**:1590–1601.

Perrière, G., and J. Thioulouse. 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res. **30**:4548–4555.

Radomski, J. P., and P. P. Slonimski. 2001. Genomic style of proteins: concepts, methods and analysis of ribosomal proteins from 16 microbial species. FEMS Microbiol. Rev. **25**:425–435.

Rocha, E. P., and A. Danchin. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. **18**:291–294.

Rocha, E. P., and A. Danchin. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nature Genet. **34**:377–378.

Ruepp, A., W. Graml, M. L. Santos-Martinez, K. K. Koretke, C. Volker, H. W. Mewes, D. Frishman, S. Stocker, A. N. Lupas, and W. Baumeister. 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature **407**:508–513.

Sandberg, R., C. I. Bränden, I. Ernberg, and J. Cöster. 2003. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino-acids usage and G+C content. Gene **311**:35–42.

Sharp, P. M., and W-H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acid Res. **15**:1281–1295.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. **16**:8207–8211.

Shields, D. C., and P. M. Sharp. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both traditional selection and mutational biases. Nucleic Acids Res. **15**:8023–8040.

Sicheritz-Pontén, T., and S. G. Andersson. 2001. A phylogenomic approach to microbial evolution. Nucleic Acids Res. **29**:545–552.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48**:582–592.

Tekaia, F., E. Yeramian, and B. Dujon. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene **297**:5160.

Torres de Farias, S., and M. C. Manhães Bonato. 2002. Preferred codons and amino-acds coupled in Hyperthermophiles. Genome Biol. **3**:preprint0006.1–0006.18.

Varenne, S., J. Buc, R. Lloubès, and C. Lazdunski. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. **180**:549–576.

Wada, K. S., R. Aota, F. Tsuchiya, T. Ishibashi, T. Gojobori, and T. Ikemura. 1990. Codon usage tabulated from GenBank genetic sequence data. Nucleic Acids Res. **18**(Suppl):2367–2411.

Wagner, A. 2000. Inferring lifestyle from gene expression patterns. Letter to the editor, Mol. Biol. Evol. **17**:1985–1987.

Wang, G., M. Z. Humayun, and D. E. Taylor. 1999. Mutation as an origin of genetic variability in *Helicobacter pylori*. Trends Microbiol. **7**:488–493.

Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**:221–271.

Zavala, A., H. Naya, H. Romero, and H. Musto. 2002. Trends in codon and amino acid usage in *Thermotoga maritima*. J. Mol. Evol. **54**:563–568.

Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. J Theor. Biol. **8**:357–366.

Brian Golding, Associate Editor

Accepted November 2, 2004