BIOINFORMATICS

Vol. 19 no. 16 2003, pages 2005–2015
DOI: 10.1093/bioinformatics/btg272

Codon adaptation index as a measure of dominating codon bias

A. Carbone[1,*], A. Zinovyev[1] and F. Képès[2]

[1]Institut des Hautes Études Scientifiques, 35, route de Chartres, 91440 Bures-sur-Yvette, France and [2]ATelier de Génomique Cognitive, CNRS ESA 8071/Genopole, 523 Terrasses de l'Agora, 91000 Evry, France

Received on February 6, 2003; revised on April 15, 2003; accepted on April 25, 2003

ABSTRACT

We propose a simple algorithm to detect dominating synonymous codon usage bias in genomes. The algorithm is based on a precise mathematical formulation of the problem that lead us to use the Codon Adaptation Index (CAI) as a 'universal' measure of codon bias. This measure has been previously employed in the specific context of translational bias. With the set of coding sequences as a sole source of biological information, the algorithm provides a reference set of genes which is highly representative of the bias. This set can be used to compute the CAI of genes of prokaryotic and eukaryotic organisms, including those whose functional annotation is *not* yet available. An important application concerns the detection of a reference set characterizing translational bias which is known to correlate to expression levels; in this case, the algorithm becomes a key tool to predict gene expression levels, to guide regulatory circuit reconstruction, and to compare species. The algorithm detects also leading–lagging strands bias, GC-content bias, GC3 bias, and horizontal gene transfer. The approach is validated on 12 slow-growing and fast-growing bacteria, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*.

Availability: http://www.ihes.fr/~materials.

Contact: carbone@ihes.fr

INTRODUCTION

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). Biased codon usage may result from a diversity of factors: GC-content, preference for codons with G or C at the third nucleotide position (Lafay *et al*., 1999), a leading strand richer in G+T than a lagging strand (Lafay *et al*., 1999), horizontal gene transfer which induces chromosome segments of unusual base composition (Moszer *et al*., 1999), and in particular, translational bias which has been frequently noticed in fast growing prokaryotes and eukaryotes (Sharp and Li, 1987; Sharp *et al*., 1986; Médigue *et al*., 1991; Shields and Sharp, 1987; Sharp *et al*., 1988; Stenico *et al*., 1994). Three main facts support the idea of 'translational impact': highly expressed genes tend to use only a limited number of codons and display a high codon bias (Grantham *et al*., 1980; Sharp and Li, 1987), preferred codons and iso-acceptor tRNA content exhibit a strong positive correlation (Ikemura, 1985; Bennetzen and Hall, 1982; Bulmer, 1987; Gouy and Gautier, 1982), and tRNA isoacceptor pools affect the rate of polypeptide chain elongation (Varenne *et al*., 1984).

To study the effect of translational bias on gene expression, Sharp and Li (Sharp and Li, 1987) proposed to associate to each gene of a given genome a numerical value, called codon adaptation index (CAI), which expresses its synonymous codon bias. The idea is to compute a *weight* (representing relative adaptiveness) for each codon from its frequency within a chosen small pool of highly expressed genes S, and combine these weights to define the CAI(g) value of each gene g in the genome. For Sharp *et al.*, the hypothesis driving the choice of S is that, for certain organisms, highly expressed genes in the cell have highest codon bias, and these genes, made out of frequent codons, are representative for the bias. Based on this rationale, one can select a pool of ribosomal proteins, elongation factors, proteins involved in glycolysis, possibly histone proteins (in eukaryotes) and outer membrane proteins (in prokaryotes) or other selections from known highly expressed genes, to form the representative set S. Then, CAI values are computed and are checked to be compatible with genes known to be highly or lowly expressed in the cell. If this is the case, then predictions are drawn with some confidence on expression levels for genes and open reading frames, even with no known homologues. Even if conceptually clear, this framework has been misused several times in the literature and incorrect biological consequences have been derived for gene expression levels of organisms which do *not* display a dominant translational bias, as discussed in (Grocock and Sharp, 2002). This confusion motivated us to search for a methodology based on a precise mathematical formulation of the problem to detect the existence of translational bias.

*To whom correspondence should be addressed.

But the main motivation for us came from the recognition that an increasing number of genome sequences will be available for organisms for which biological knowledge consists merely of a sketched morphological and ecological description. For these organisms, it might not be evident how to define the reference set *S*, or how to identify a reliable testing set which can ensure that predictions meet a satisfiable confidence level. Still, one would like to detect if translational bias holds for these genomes and if so, to predict their gene expression levels. If not, one would like to know the origin of their dominating bias.

We propose an algorithm that uses the notion of CAI as a *measure* to detect the most dominant codon bias in the genome, *regardless* of whether this bias is translational or not. The algorithm screens all genes of an organism and it *selects* a reference set *S* which scores the highest values in the CAI scale. A screening of the genes in the set allows to identify the kind of synonymous codon usage bias which drives the genome under examination. If *S* contains proteins involved in translation and glycolysis, then one derives that the bias is translational, and that CAI values can be safely correlated to gene expression levels. If no translational bias is present, it is then possible to successfully correlate CAI values to GC content, GC3 bias, GC skew bias, leading–lagging strand bias, and so on. We discuss some examples later.

The algorithm is based on no biological assumption, in particular concerning the biological functions of the organism. The key point is that dominant codon bias in a set of coding sequences is a notion which is *independent* of biological knowledge. It can be precisely formalized in purely mathematical terms and used to detect a representative set of sequences which lead the dominating bias. It is important to stress though, that a biological evaluation of the reference set is a crucial step to use it appropriately.

One novel technical aspect of our analysis is a *revised* definition of CAI that allows the automatic detection of dominant codon bias for both prokaryotic and *eukaryotic* genomes. For these latter, it has been noticed (Duret and Mouchiroud, 1999) that gene length in the genomes of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana* displays a strong negative correlation with codon usage bias. This means that a careless selection along the iteration of the algorithm of short and long genes might yield a set *S* with a *heterogeneous* codon distribution (where rare codons appear with high frequency in certain long genes), and consequently, to the calculation of a codon adaptation index which is undesirably affected. For eukaryotic genomes, whose long genes make up about half of their coding part, the selection of genes in *S* needs to be guided by such a criteria. For this, we introduce a correcting factor in the original formula proposed by Sharp and Li (Sharp and Li, 1987), compute the global codon adaptation index, (gCAI) and determine a set *S* of coding sequences with high gCAI values. Ultimately, CAI values can be computed on codon weights calculated on *S*.

Table 1. The full set of genomes considered in this paper with their number of CDSs

Mycoplasma pulmonis	480	SG
Mycobacterium tuberculosis	4187	SG
Treponema pallidum	1031	SG
Helicobacter pylori	1566	SG
Pseudomonas aeruginosa	5567	SG
Borrelia burgdorferi	1638	SG
Haemophilus influenzae	1709	FG
Salmonella enterica	4600	FG
Staphylococcus aureus	2714	FG
Lactococcus lactis	2266	FG
Bacillus subtilis	4100	FG
Escherichia coli	4289	FG
Saccharomyces cerevisiae	6305	LE
Caenorhabditis elegans	17 078	HE
Drosophila melanogaster	14 146	HE

SG, slow-growing bacteria; FG, fast-growing bacteria; LE, lower eukaryotes; HE, higher eukaryotes.

MATERIALS AND METHODS

Sequence data

The whole genomes along with gene annotation were retrieved from the Genomes directory of GeneBank FTP (Table 1 and supplementary material). All sequences marked as CDS were considered, including those annotated as hypothetical and those predicted by computational methods only. From each coding sequence, we excluded initiation and stop codons.

Transcriptional data for *Saccharomyces cerevisiae* are taken from the study reported in (Holstege *et al*., 1998) and based on high-density oligonucleotide arrays technology (downloaded from http://www.wi.mit.edu/young/expression.html in 1999, now available at http://www.ihes.fr/~materials). They concern a set of 4849 genes.

Space of coding sequences and visualization

A coding sequence is represented by a 64-dimensional vector, whose entries correspond to the 64 relative codon frequencies in the sequence. (The entries for codons UAA, UAG, UGA, UGG, AUG could be discharged: UGG, AUG have no synonymous codons and UAA, UAG, UGA are stop codons and they do not code for amino acids. Considering 59 dimensions instead of 64 would not make any substantial difference neither in the determination of the reference set nor in the visualization of the coding sequences in three-dimensions.) Recall that the *frequency* of a codon i in a sequence g is the number of occurrences of i in g (where g is intended to be split in consecutive non-overlapping triplets corresponding to amino-acid decomposition), and that the *relative frequency* of i in g is the frequency of i in g divided by the number of codons in g. Notice that for each vector representing a coding sequence, the sum of its entries must equal 1.

Hence, a coding sequence is a point in the 64-dimensional space $[0 \ldots 1]^{64}$, where no special assumption is made on the space nor on the coordinate system. The set of points can be visualized in three dimensions by using principal components analysis (PCA) (Hotelling, 1933; Hand *et al*., 2001): first, every coordinate is normalized on unity standard deviation to take into account equally dominating as well as rare codons [following the standard definition employed in PCA, the normalized value of relative frequency $x'_{i,j}$ for codon i in gene j is defined as $(x_{i,j} - \bar{x}_i)/\sigma_i$, where $x_{i,j}$ is the relative frequency of i in j, $\bar{x}_i$ is the average relative frequency of i computed with respect to all coding sequences, and σ_i is the standard deviation for the set of frequency values $x_{i,j}$, for all j]; then, three principal components for the cloud of points are calculated using the Euclidean distance measure; finally, the cloud of points is projected orthogonally in the subspace of the three selected vectors and visualized by means of a specialized three-dimensional viewer (see below).

Other methods of multivariate analysis, as Correspondence Analysis and Principal Factorial Analysis, have been used, much more frequently than PCA, to investigate codon usage (Médigue *et al*., 1991; Perrière and Thioulouse, 2002).

Reference sets, codon frequency tables, tables of gCAI and CAI values were 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 algorithmic behaviour in the space of codon usage was visualized in VidaExpert, a tool developed by A.Z. A specialized three-dimensional viewer is provided with VidaExpert. All software is available at http://www.ihes.fr/~materials.

Codon composition and codon bias

For each genome sequence G and some set of coding sequences S in G, *codon bias* is measured with respect to its synonymous codon usage. 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, then one associates to i a *weight* $w_{i,j}$ relative to its sibling of maximal frequency y_j in S

$$w_{i,j} = \frac{x_{i,j}}{y_j}.$$

A codon with maximal frequency in S is called *preferred* among its sibling codons. To each gene g in G, Sharp and Li (Sharp and Li, 1987) associated a value in $[0, 1]$, called CAI defined as

$$\mathrm{CAI}(g) = \left(\prod_{k=1}^{L} w_k \right)^{1/L}$$

where L is the number of codons in the gene, and w_k is the weight of the kth codon in the gene sequence. Genes with CAI value close to 1 are made by highly frequent codons.

Codon bias and length of genes in eukaryotes: the need of a revised statistics

We introduce a new definition of *weight*

$$\bar{w}_{i,j} = \frac{|S^i|}{|S|} \cdot \frac{x_{i,j}}{y_j}$$

where S^i is the set of coding sequences in S that contain at least one occurrence of codon i, and $|S^i|, |S|$ denote the number of coding sequences in S^i and S. The factor $|S^i|/|S|$ denotes the probability that a codon i appears in a gene of S. If the set S has a *highly homogeneous* codon distribution, i.e. if genes are made out of the same pool of codons, then the factor is expected to take almost always value 1, with no dramatic effect on the weight. In general, the factor discriminates against those codons that happen to appear in a few number of genes in S, even if their occurrence is very pronounced within those genes. Such a situation might arise, for instance, when some long gene belongs to S, since rare codons become likely to appear there (possibly in a high absolute frequency). [Alternatively, one could define the weights $\bar{w}_{i,j}$ as $(|S^i|/|S|) \cdot (x_{i,j}/\Sigma x_{l,j})$, where $x_{l,j}$ denotes the frequency of codon l synonymous to i. The value $x_{i,j}/\Sigma x_{l,j}$ is in $[0, 1]$ and denotes the probability that i is selected within all synonymous codons for the amino-acid j.]

A value in $[0, 1]$, called gCAI, is associated to each gene g, and it is defined as

$$\mathrm{gCAI}(g) = \left(\prod_{k=1}^{L} \bar{w}_k \right)^{1/L}$$

where L is the number of codons in the gene, and $\bar{w}_k$ is the weight of the kth codon in the gene sequence. Genes with gCAI value close to 1 are made by highly frequent codons.

Some measures for codon usage

GC-*content* is defined to be the frequency of G + C basepairs, and GC3-*content* is the frequency of G + C basepairs at the third coding position (excluding Met and Trp, and termination codons). *XY-skew*, where $X, Y \in \{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\}$ and $X \neq Y$, is defined as $(X - Y)/(X + Y)$, that is the relative distance between X-frequency and Y-frequency; its value is positive and high when the sequences are made by many Xs and a few Ys. Finally, for circular genomes, genes might happen to be located on *leading* and *lagging strands*; their codon usage is influenced accordingly. To measure the connection between strands and bias we use the standard t-value for calculating the difference between mean CAI values in leading and lagging strands.

A strategy to search for the most biased reference set

We propose an automatic procedure to search for a set S which is representative of the codon usage in the genome. Precisely,

one wants to find a set S which contains about 1% of predicted coding sequences ($\approx$ 50–150 sequences) and which allows to compute weights $\bar{w}_{i,j}$ that *maximize* the gCAI values of the genes in S, i.e. the gCAI values of coding sequences in G/S (that is, those that are not in G) are smaller than all gCAI values of sequences in S

$$\mathrm{gCAI}(G/S) \leq \mathrm{gCAI}(S) \tag{1}$$

where gCAI values are computed on S. In other words, the highest gCAI values are obtained on the selected set S. Condition (1) expresses a sort of *self-consistency principle* for S.

Among all sets that satisfy (1), one wants to choose the set S which is representative of the family of codons that appear in most genes with the highest frequency (in within the genome). In formal terms this means that if $c_1, \ldots, c_{20}$ are preferred codons for S, and $d_1, \ldots, d_{20}$ are preferred codons for the entire genome G, then we look for the set S that minimizes

$$\sum_{i=1}^{20} \chi(c_i, d_i) \tag{2}$$

where $\chi(c_i, d_i) = 1$ if $d_i \neq c_i$ and $= 0$ otherwise. Condition (2) expresses the meaning of *dominating* codon bias in G.

An exhaustive search for the best reference set satisfying (1) and (2) asks for too much computational time. In fact, one should search through $\binom{X}{Y}$ sets, where the binomial coefficient $\binom{X}{Y} = X!/[Y!(X-Y)!]$, X is the number of coding sequences of the genome and Y is $X/100$. This means that for a genome of 6000 coding sequences, like the genome of *S.cerevisiae* for instance, the number of sets to be checked would be more than 2^{360}. The algorithm that we propose, is based on the belief that for all genomes there exists a pool of coding sequences that contain few and very frequent codons, and that the bias induced by such codons can guide the search. Maximum values are expected when codon distribution in S is the most homogeneous, and when long genes containing rare codons do not belong to S. These two properties are controlled by the factor $|S^i|/|S|$.

It is important to stress that the added complexity of $|S^i|/|S|$ is meaningful in the process of automated delineation of highly biased reference sets in *eukaryotes*. In the last section, we discuss, based on the concrete case of *C.elegans*, the malfunctioning of the algorithm when this factor is not taken into account. For bacterial genomes, this factor does not significantly contribute and gCAI values are highly correlated (correlation coefficient >0.98) with CAI values.

THE ALGORITHM

The algorithm is iterative. At each iteration step k, it computes the gCAI values of the coding sequences in G from codon weights that are calculated with respect to a selected set S_k. Codons that do not appear in S_k take weight 0.01 by default. At step 1, let S_1 be the set of all coding sequences in G; at step $k+1$, define S_{k+1} to be the x_{k+1}% of the genes with highest gCAI value at step k, where $x_{k+1} = x_k/2$, and $x_1 = 100\%$. In particular, by dividing at each step the number of genes by 2, the procedure soon ($\approx$8–15 steps for prokaryotes, and 15–25 steps for eukaryotes) converges to some small set S_k. If the number of coding sequences in S_k is smaller than the 1% of all coding sequences, then step $k+1$ is applied to S_k containing the 1% of coding sequences with highest gCAI_k value on the previous iteration. The algorithm terminates when it converges to a small set S_k such that $\mathrm{gCAI}_{k+1}(g) = \mathrm{gCAI}_k(g)$ for all g in S_k, i.e. $S_{k+1} = S_k$, where gCAI_k represents the gCAI values which are obtained at step k. It might happen that a finite sequence of sets $S_k, S_{k+1}, \ldots, S_{k+r}$ (i.e. $S_{k+r+1} = S_k$) is found instead. In this case, we say that the procedure *oscillates* between $r+1$ sets. To detect a unique convergent set S_k, we take away from S_{k+1} the gene with smallest gCAI_k value and re-iterate, that is we look for a reference set of size possibly smaller than 1%. We do so (by decrementing a set by one) until a unique convergent set is found.

The choice of fixing the smallest size of S_k sets at 1% of the genome size corresponds roughly to the size of the reference sets proposed by Sharp *et al*. in (Sharp and Li, 1987; Sharp *et al*., 1986; Shields and Sharp, 1987; Sharp *et al*., 1988; Stenico *et al*., 1994; Andersson and Sharp, 1996).

Once the convergent reference set S_k is computed, weights $w_{i,j}$ and CAI(g) values, for all genes g, can be calculated.

Rationale and a localized version

The rationale of the algorithm is based on the belief that for all genomes there exists a small pool of coding sequences that contain a small number of very frequent codons. From the very first iterations, such codons lead to determine the small set of highly biased coding sequences which are representative of the codon bias dominating the entire genome.

This intuition is supported by the following numerical analysis: on the algorithm above, we fixed S_1 to be some *randomly* selected set made of 1% of all coding sequences, and we fixed $x_{k+1} = 1\%$, for all iterations k. This 'randomized' version of the algorithm has been applied several times to the genomes in Table 1, and always it produced the reference set S obtained with the original version (data not shown). Numerical simulations were based on coding sequences chosen randomly, but most of them happened to be also 'uniform' on the space of coding sequences: this space is usually shaped into a few clusters (see later discussion), and the number of representative coding sequences in the random set coming from each cluster was proportional to the size of the cluster. For non-uniform random selections, i.e. localized in space, the algorithmic convergence to S cannot be ensured. In this case, the algorithm detects *local* codon biases (see later discussion).

Convergence and uniqueness of the solution

The algorithm converges to some finite set S of coding sequences with high gCAI value because, at each iteration, the set S_{k+1} is chosen to contain coding sequences with stronger bias than in S_k. Notice that all genomes analysed in this paper converged immediately to a unique set S_k of size 1%, that is no oscillation was detected.

The discussion at the end of the previous section points out that the reference set S, for genome sequences, is detectable by the randomized version of the algorithm. This implies that S does *not* depend on x_i values, and we say that the gCAI(g) [CAI(g)] value is *unique*, for each gene g.

To conclude, let us observe that the self-consistency principle (1), saying that S is a fixed point for the algorithm, is satisfied by our definition of stable set. Also, condition (2) is satisfied with high probability at each iteration step, that is with high probability $\sum_{i=1}^{20} \chi(d_i^{k+1}, d_i^k)$ is minimized by the selected set S_{k+1}, at each iteration step k. (The formal argument justifying this fact is based on the observation that synonymous codons with weights $\ll 1$ calculated on S_k tend to preserve the same frequency distribution in S_{k+1}.)

APPLICATION TO VARIOUS GENOMES

The algorithm has been applied to 15 prokaryotic and eukaryotic genomes. For all of them, a convergent set was found after $K = 15$ iterations, except in the case of *C.elegans* where $K = 20$. By plotting gCAI_k values for all genes (see Fig. 1), we observe that the trend associated to the convergent iteration K (bottom plot), can be very different for different species. Coding sequences of high gCAI_1 values (on the right of the top curve) tend to conserve their high values in iteration K (bottom plot), while coding sequences with low gCAI_1 value typically take smaller values in successive iterations. The latter fact is due to the codon biases of the sets S_j, which are expected to associate lower weights to a larger number of codons than S_k, for $k < j$.

The differences between first and last plots, i.e. $|\mathrm{gCAI}_1(g) - \mathrm{gCAI}_K(g)|$ for each gene g, reveal a homogeneous distribution of codons in *H.pylori*, the existence of a pool of very biased genes detectable from the first iteration in *E.coli*, and a biased set which is not detectable from the first iteration in *C.elegans* (notice the high number of peaks along the plots at iteration K). Figure 2 illustrates the change of codon frequency in the first and the last iterations of the algorithm for these three representative organisms.

DYNAMICS OF THE ALGORITHM

The dynamics of the algorithm can be studied by looking at the way the clouds S_k move in the 64-dimensional space of sequences, or, analogously, by following the trajectory of average codon usage for the sets S_k. We first say some words about the space.

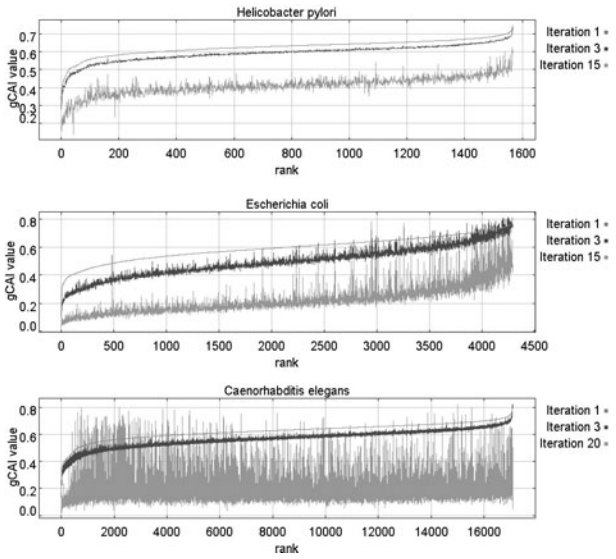

Fig. 1. Plot of the gCAI_k values of coding sequences in *H.pylori*, *E.coli*, *C.elegans* for $k = 1, 3, K$, where K corresponds to convergent iterations. The x-axis ranks coding sequences, and its maximum value is the number of coding sequences in the organism. The rank follows increasing gCAI_1 values (seen in the smooth increasing curve on the top). In steps 3 and K, we see how values adapt to the codon bias that has been detected.

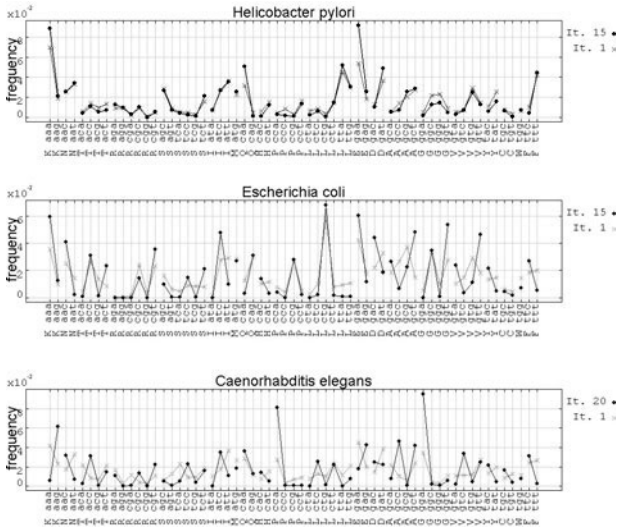

Fig. 2. Frequency distribution of codons in the sets S_k for $k = 1, K$, in *H.pylori*, *E.coli*, *C.elegans*. Recall that S_1 is the set of all coding sequences. The frequencies of synonymous codons are connected by consecutive lines for an easier visualization.

Structure of the 64-dimensional spaces of coding sequences

Recall that coding sequences are points in the 64-dimensional space $[0 \ldots 1]^{64}$, and that the cloud they form can be visualized

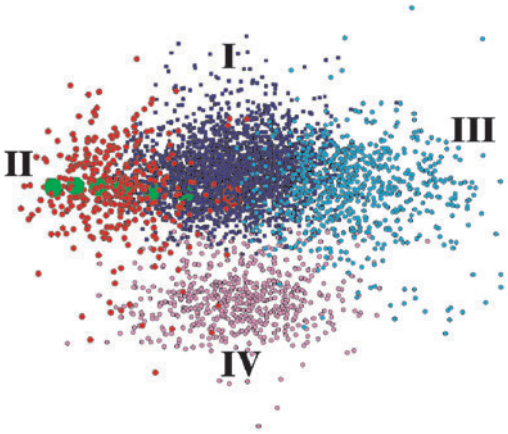

Fig. 3. *E.coli*: four clusters are shown in blue (class I), pink (class IV), red (class II) and green (class III). Left: the *rabbit head*; this view has been realized after rotating slightly the three-dimensional-projection along the first and second principal components. Right: two-dimensional-projection along the second and third principal components; notice the pink cluster 'below' the rabbit head. The clustering was done with the *K-Means* method. The trajectory of convergence of the algorithm applied to this genome is described by a sequence of green balls (representing codon usage of successive sets S_k) starting with a red one (computed for S_1; the red ball is barely visible). The trajectory starts in the blue cluster and ends-up in the red one which contains genes coding for ribosomal proteins.

in three dimensions by normalizing the points (i.e. each vector coordinate), calculating three principal components and performing an orthogonal projection in the subspace of the three principal vectors. The shape of the cloud of points, for *fast growing* organisms, reminds a 'rabbit head' and it was observed in (Médigue *et al*., 1991) for *E.coli* (see Fig. 3). It looks as formed by a dense cluster of coding sequences (class I), with two much sparser sets of points protruding from it (the 'ears' of the rabbit, class II and III). Class I comprises genes that maintain a low or intermediary level of expression, but that at times can be expressed at a very high level; in contrast class II contains genes that are constitutively expressed at a high level, most of them are involved in translation, in protein folding, in transcription, in DNA binding; class III contains integration host factors, insertion sequences, genes behaving as mutators when inactive, but also genes controlling cell division, outer membrane proteins, catabolic operons.

Below class I one finds another dense set of points, which we call class IV, rather well separated by the main one (see Fig. 3). It mainly includes proteins encoded with hydrophobic amino-acids, as membrane proteins. Figure 3 shows this cluster structure for *E.coli*. It is detected in 64 dimensions by PCA and then projected along the three principal components. [Class IV was not detected in (Médigue *et al*., 1991) because Factorial Analysis and RSCU data tables for representing codons were used, and RSCU does not distinguish amino-acid composition.] This distinguished positional organisation of coding sequences in the 64-dimensional space has been observed several times for specific genomes (Perrière and Thioulouse, 2002).

For *slow growing* organisms, the cloud of points is constituted, in some cases, by class I and class IV, where class II and III are not distinct from class I. Other times, for instance for spirochaetes, the two dense clouds have a less clear-cut shape (Lafay *et al*., 1999).

Clusters and the dynamics of the algorithm

The *codon usage* of the whole set of coding sequences, i.e. the vector whose ith coordinate represents the average usage of codon i in the set (formally, each coordinate of the vector is defined as x_i/N, where x_i is the frequency of codon i in S_1 and N is the number of codons in S_1), sits in the most dense cluster (see red ball in Fig. 3). Codon usages of the sets of coding sequences S_k lie on the trajectory depicted by green balls in Fig. 3.

In Fig. 4, the clouds S_k are shown for the projected sequence space of *B.subtilis* at consecutive iterations: the codon usage of S_1 is located in class I (the large black ball is hidden in it) and it gradually moves towards class II, along a trajectory indicated by consecutive medium sized black balls, which represent the codon usage of the sets S_k, for increasing k. The algorithm displays the same regular behaviour for all fast-growing bacteria and eukaryotic organisms analysed in this paper: the trajectory starts at class I and terminates in class II.

For slow-growing bacteria, the algorithmic trajectory starts and ends in class I. Typically such genomes display dominant leading strand bias, GC3 bias, GC or AT skew bias, or a homogeneous composition (Lafay *et al*., 1999, 2000; Grocock and Sharp, 2002).

VALIDATION

Detection of translational bias in fast growers

For fast-growing micro-organisms, reference sets detected by the algorithm contain genes which encode almost exclusively proteins involved in translation, protein folding and glycolysis. This composition has been obtained for *S.enterica*, *S.aureus* and *L.lactis* for which no table of weights was previously compiled and no reference set was proposed (see supplementary material), as well as for organisms for which translational bias has been previously investigated as *H.influenzae* (Perrière and Thioulouse, 1996), *B.subtilis* (Shields and Sharp, 1987), *E.coli* (Sharp and Li, 1987), *S.cerevisiae* (Sharp *et al*., 1986), *D.melanogaster* (Sharp *et al*., 1988), *C.elegans* (Stenico *et al*., 1994). For *E.coli* and *H.influenzae*, two Gram-negative bacteria with an outer membrane, major outer-membrane proteins are additionally included by the algorithm in the reference set.

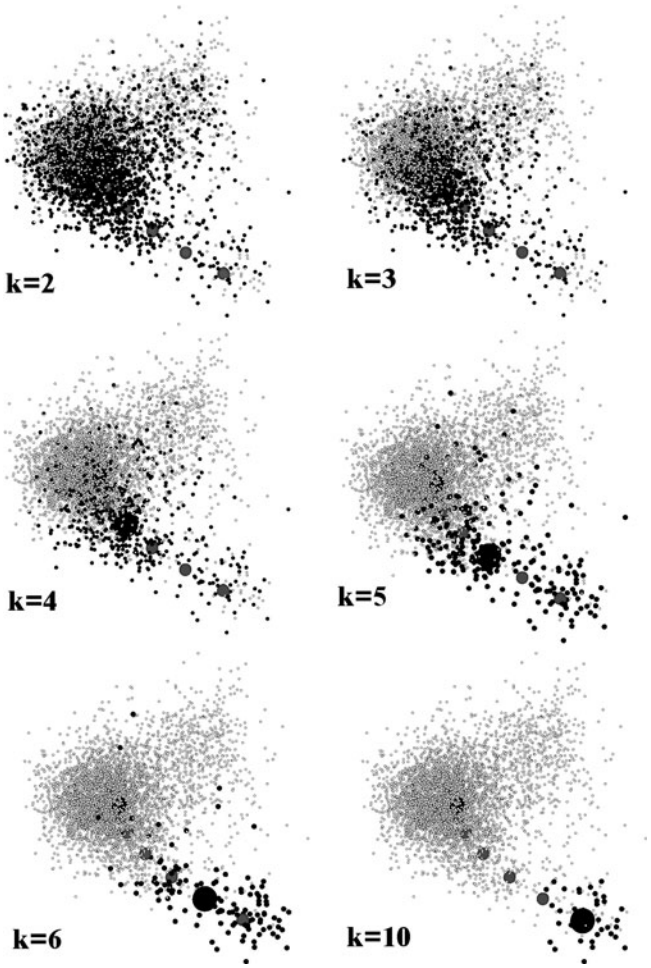

Fig. 4. *B.subtilis*: the dynamics of the algorithm from iteration 2 (top left) to iteration 10 (down right) visualized on the rabbit head. Coding sequences are represented by dots; black dots correspond to sequences in S_k, for $k = 2, \ldots, 10$. The black cloud and its large black ball (corresponding to average codon bias in S_k) progressively move from (sequences in) class I towards class II, for increasing values of k. The algorithm stabilizes around sets S_k, for $k \geq 7$, whose codon usage are localized in the same small neighbourhood.

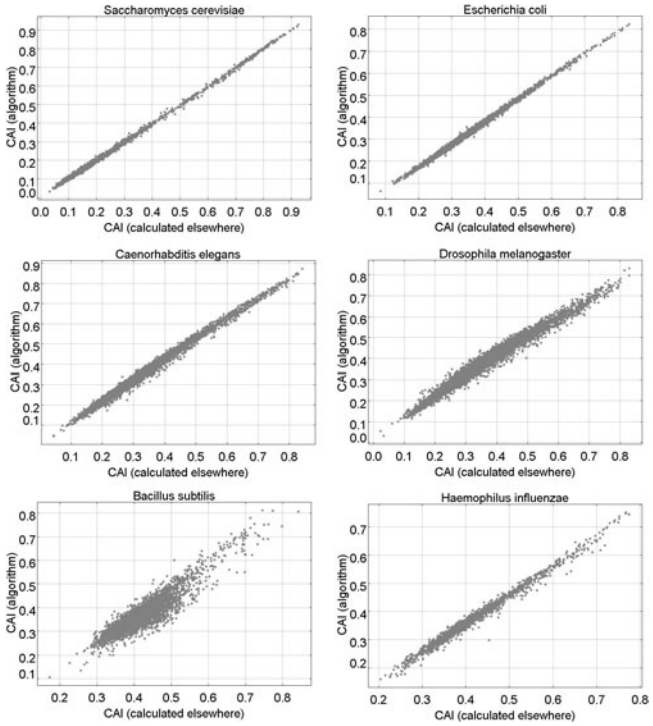

Fig. 5. CAI values computed on the reference set found by our algorithm (y-axis) are plotted with CAI values computed by Sharp *et al.* (x-axis) for *S.cerevisiae*, *E.coli*, *C.elegans*, *D.melanogaster*, *B.subtilis* and *H.influenzae*.

Besides proteins involved in translation and glycolysis, the reference sets for higher eukaryotes comprise proteins involved in ATP production (by mitochondria) and in the cytoskeleton. For *C.elegans*, the reference set additionally contains histone and collagen proteins.

The lengths of genes vary considerably within each reference set, from $\approx$100aa for cold-shock proteins, to $\approx$700aa for heat-shock and glycolytic proteins, with the vast majority $\approx$ 200–300aa long. There are also proteins which are 1000aa long such as myosin proteins, up to 1600aa long as vitellogenin in *C.elegans* (see supplementary material also).

When comparing our CAI values with those computed on reference sets defined by manually selecting highly expressed genes in the cell, we obtain a remarkably good correlation between the values as illustrated in Fig. 5. For *E.coli*, *S.cerevisiae* and *H.influenzae* (Sharp and Li, 1987; Sharp *et al.*, 1986; Perrière and Thioulouse, 1996) all points are well distributed along the diagonal. For *C.elegans*, the correlation is done with data in (Stenico *et al.*, 1994); our reference set contains 172 proteins and it is much larger than the one computed for micro-organisms. The same is true for *D.melanogaster* [with data in (Sharp *et al.*, 1988)], where our reference set contains 129 coding sequences. The CAI values of *B.subtilis* have been defined in (Shields and Sharp, 1987) on a set which comprised only seven genes and given its small size, it is surprising that the correlation is so high.

For *S.cerevisiae*, the bias captured by our algorithm is correlated with transcript steady-state levels and reflects the transcriptional load of mRNA present in the cell (see Fig. 6, left). This correlation supports the intuition that the cost of cellular macromolecular synthesis would be increased by producing an abundant transcript encoding lowly expressed proteins, or vice versa, by producing a poor transcript encoding highly expressed proteins. It also suggests that the bias detected by the analysis of codon usage is intrinsically related to both transcription and translation [as argued in (Sharp and Li, 1987; Sharp *et al.*, 1986)].

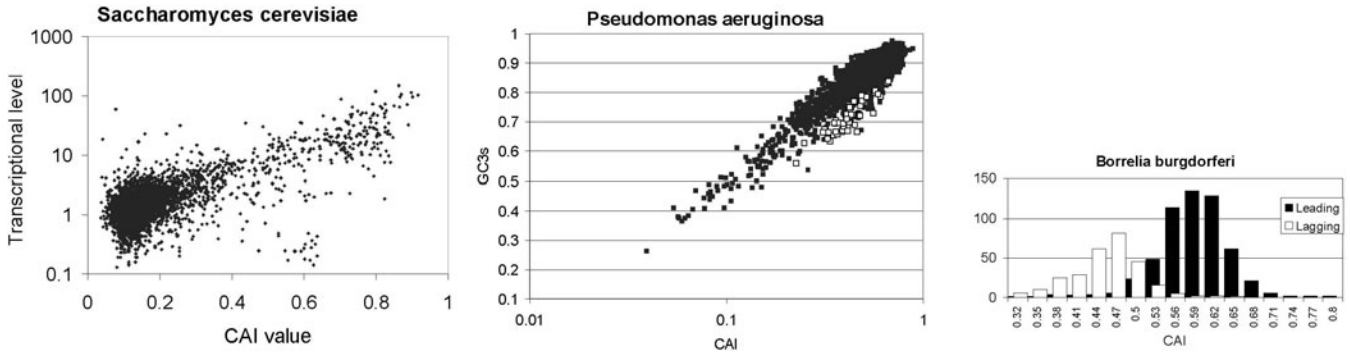

Fig. 6. Different biases detected by the algorithm. Transcriptional bias, left: transcription levels of *S.cerevisiae* genes are plotted (in log scale) with CAI values. GC3s bias, centre: GC3s values of *P.aeruginosa* are plotted with CAI values (in log scale); the open squares correspond to ribosomal genes. Leading–lagging strand bias, right: number of genes in leading and lagging strands of *B.burgdorferi* are plotted with CAI values.

(For all fast growing organisms considered in the paper, high gCAI values are about 10–15% smaller than CAI values, and low gCAI values are at most 30% smaller than CAI values. gCAI values are highly correlated with CAI values, with a correlation coefficient of 0.98–0.99 in the worst case.)

Detection of translationally optimal codons in slow growers: *M.pulmonis* and *M.tuberculosis*

M.pulmonis is a slow growing pathogen with a small number of genes, distributed within a four cluster structure in the 64-dimensional space of coding sequences (data not shown; the space reminds the one illustrated in Fig. 3 for *E.coli*); ribosomal proteins are grouped in one of these clusters. Our algorithm converges towards ribosomal proteins first, to deviate afterwards towards a group of lipoproteins, outliers for the distribution and highly GC-rich. As a result, the reference set (constituted by seven genes) includes three ribosomal proteins, three lipoproteins and one protein of unknown function. This suggests that the primary source of codon variation in *M.pulmonis* is in the use of a subset of codons which are translationally optimal.

An analogous description holds for *M.tuberculosis*. This genome is highly GC3 biased and the correlation coefficient between GC3 and CAI values is $r = 0.93$. We find a reference set of 41 genes that includes three ribosomal proteins, two translation factors, and many properties involved in glycolysis suggesting a selection for optimal codons being more effective on genes expressed at high levels. This was also noticed in (Andersson and Sharp, 1996).

Detection of GC3 bias: *Pseudomonas aeruginosa PA01*

On *P.aeruginosa* genome, the algorithm detects a reference set S which contains neither ribosomal proteins nor elongation factors. This suggests that translational bias is not dominating, and that CAI values computed from S are not representative of gene expression levels. Notice, for instance, that the highly expressed gene *oprI* turns out to have CAI value 0.26.

CAI values are well correlated with GC3-content though, with a correlation coefficient $r = 0.83$; the correlation coefficient between the logarithmic value of CAI and GC3 is $r = 0.90$ (see Fig. 6, middle). It is interesting to see that if the algorithm is run from an initial set S' containing genes constituting the core of the gene expression machinery (i.e. genes coding for ribosomal proteins, elongation factors, and so on), then, nevertheless, it detects S as a reference set. This 'proves' that GC3 bias is *much* more dominant than translational bias.

The dominating codon bias in *P.aeruginosa* gave origin to controversial opinions on the biology of this organism. This was due to calculations of CAI values which were based on misleading manual selections of reference sets (Grocock and Sharp, 2002; Gupta and Ghosh, 2001; Kiewitz and Tümmler, 2000).

Detection of GC skew bias: *Treponema pallidum*

A genome where GC skew content is known to be the dominating codon bias, is *T.pallidum* (Lafay *et al*., 1999). Our CAI values meet the highest correlation coefficient with GC skew values, with $r = 0.659$.

Homogeneous genomes: *Helicobacter pylori*

On *H.pylori*, the algorithm returns a reference set S which is constituted essentially by coding sequences with 'unknown' function. CAI values have a low correlation with all forms of bias (notice that the strongest one is a mild correlation with GC skew values, with $r = 0.358$), and the gCAI_k values of genes along different iterations vary very little (see Fig. 1). All these observations support the hypothesis of *homogeneity* of the codon distribution on this slow growing micro-organism, and a lack of translational bias as a dominating bias (Lafay *et al*., 2000).

Detection of leading strand bias: *Borrelia burgdorferi*

For the spirochaete *B.burgdorferi*, it was shown that the main factor shaping codon usage is the strand-specific mutational pressure (Lafay *et al*., 1999). [Some translational selection was shown to exist (Perrière and Thioulouse, 2002) but it does not constitute the dominating bias.] The leading strand of replication is G+T-rich, and therefore genes placed on that strand (565 genes) display a strong bias towards those basis at the silent sites, while the opposite biases are found in genes placed on the lagging strand (286 genes) (McInerney, 1998; Lafay *et al*., 1999). Figure 6 (right) shows that genes with the highest CAI values lie in the leading strand. The t-value calculated for the difference between mean CAI values in leading and lagging strands is $t = 1.8565$, which corresponds to the confidence level of 93.8%, suggesting that the leading and lagging strands determine the dominating bias in agreement with (McInerney, 1998; Lafay *et al*., 1999).

Local codon bias: detection of a bias on the lagging strand of *B.burgdorferi*

Given a point x in the 64-dimensional space of coding sequences, let us consider the 1% of the sequences S_x which are closest to x with respect to some distance metric, for instance the Euclidean metric. By applying the algorithm to the reference set S_x, it might happen that S_x satisfies the self-consistency principle (1). This is possible even when S_x is not representative of the dominating bias, and simply means that besides the *global* bias, which is dominating, there is a *local* bias which is represented by S_x.

For *B.burgdorferi*, if the algorithm is run on a random set of genes selected among those on the lagging strand, the convergent set also lies on the lagging strand. The local bias for the lagging strand is shaped mainly by GC3-content and very mildly (negatively) related to GC-skew.

Local codon bias: detection of horizontal gene transfer

On all random selections of 1% of genes from sequences in class I and in class II of *B.subtilis* (see Fig. 4), the algorithm converged to the *same* reference set located in class II and containing genes coding for the translation machinery. When random sets were selected from class III, the algorithm converged to a reference set lying in class III. This suggests that coding sequences in class III follow a different codon bias from the rest of the genome. Class III [made of coding sequences which are A+T-rich (Nicolas *et al*., 2002)] contains horizontally transferred genes (Médigue *et al*., 1991): transposons, insertion sequences and proteins involved in phage-related functions, in adaptation to atypical conditions, and in detoxification.

We repeated the numerical test on the genome of *E.coli* whose shape in the projected 64-dimensional space is again made of classes I, II, III and IV (Fig. 3). When random sets were selected in class III (as well as in class I and class IV), the algorithm always converged to the same reference set in class II, showing that codon usage is less biased in *B.subtilis* than in *E.coli*, in agreement with (Shields and Sharp, 1987). A local fixed point exists though for genes selected on the most extreme AT-rich region in class III: the algorithm does not leave this set. The same behaviour was observed for *S.enterica*. In *C.elegans*, the algorithm always escaped chosen sets to converge towards the most translationally biased set in class II. No other fixed point, other than the dominating one, was found. [On codon bias, base composition and gene transfer see (Syvanen, 1994; Guidon and Perrière, 2001; Koski *et al*., 2001).]

CODON BIAS AND THE LONG GENE EFFECT: *S.CEREVISIAE* AND *C.ELEGANS*

We give a numerical justification for introducing the notion of 'global Codon Adaptation Index'. We had argued that an algorithmic analysis of codon bias should take into explicit account gene length because of a combination of two facts: half genetic information (defined as the number of base pairs) sits in long genes (>2000 bp long), and a strong negative correlation between codon usage and protein length was observed for Eukaryotes (Duret and Mouchiroud, 1999). In particular, one expects long coding sequences to undesirably influence the behaviour of the algorithm in successive steps: rare codons that appear in long coding sequences which are consistently selected at successive steps of the procedure, augment their relative weight with k, and bias the codon usage accordingly, possibly deviating the algorithmic behaviour. The factor $|S_k^i|/|S_k|$ included in the gCAI formula takes care of these situations.

A mild form of the problem appears with *S.cerevisiae*. This genome is the only one among unicellular organisms that displays an oscillatory behavior among two reference sets S_1, S_2 differing for exactly one gene, if weights $w_{i,j}$ are used instead of $\bar{w}_{i,j}$, i.e. if we compute CAI_k values instead of gCAI_k values at each iteration. This is due to two genes which flip in and out the reference set, RPL9B (576 bp long) and YEF3 (3135 bp long). YEF3 is by far the longest gene in the set and it spoils the statistics when CAI values are considered. With the gCAI formula no such effect appears.

A much stronger form of long gene effect is displayed in *C.elegans*. Here, the algorithm converges but it 'fails' to determine a set S, representative of the translational bias, if we compute CAI_k values instead of gCAI_k values. By applying the algorithm to the coding sequences of *C.elegans*, one observes that the algorithmic trajectory starts and ends in Class I because of the presence, in the sets S_k, of long genes as lrp-1 of length $= 14\,262$ bp and converging CAI_K value $= 0.7$. Even though coding sequences in Classes II and III are the most biased, the trajectory does not escape Class I

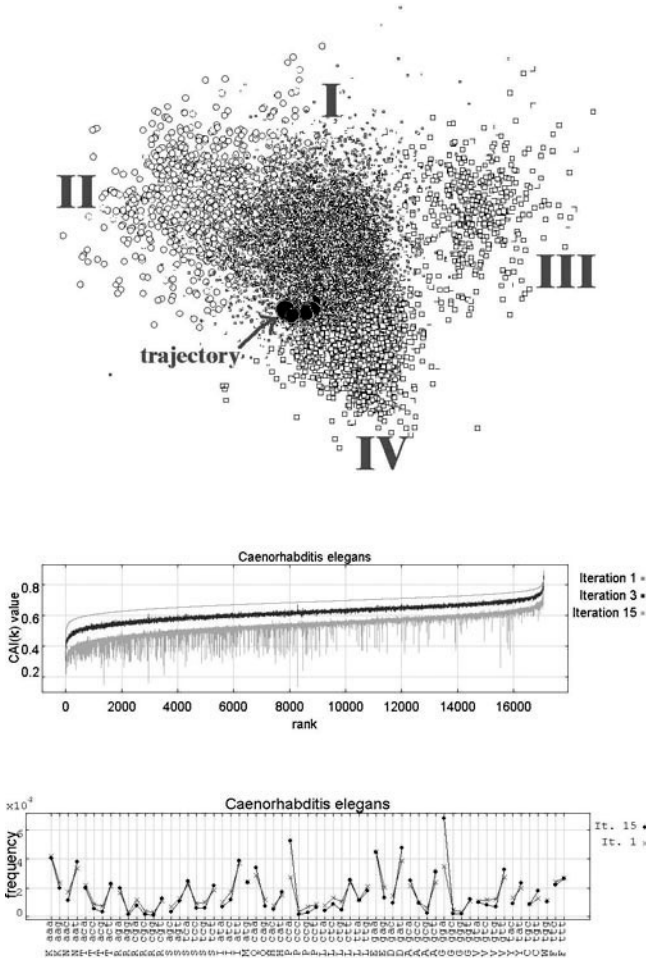

Fig. 7. *C.elegans*. Top: a view of the projected 64-dimensional space of codon frequencies. Notice the shape of the 'rabbit head' (classes I and IV correspond to the dense cloud, and classes II and III are the 'ears of the rabbit'). The trajectory of the algorithm is indicated by large black balls and it is visibly trapped in cluster C1 (compare to Fig. 3). The algorithm computes CAI_k values instead of gCAI_k values at each iteration, and converges in 15 steps. Below: CAI_k values of each coding sequence at iteration $k = 1, 3, 15$; plot of codon frequencies in S_1 and S_{15}. Compare with Figs 1 and 2.

(see top of Fig. 7). The analysis (based on CAI_k) restricted to single chromosome sequences of *C.elegans* shows that chromosomes II, IV, V and X display a trajectory that ends in Class II as one would expect, but that this is not the case for chromosomes I and III (data not shown). In Chromosome I, the long genes that influence the behaviour of the algorithm encode the membrane protein lrp-1, but also the DNA replication protein C44E4.1a of length $= 11\,595$ bp and $\mathrm{CAI}_k = 0.60$, and the dynein protein F18C12.1 of length $= 10\,630$ bp and $\mathrm{CAI}_k = 0.63$. In chromosome III the long gene that influences the behaviour of the algorithm is the hypothetical protein K07E12.1 of length $= 39\,168$ bp and $\mathrm{CAI}_k = 0.599$.

DISCUSSION

We introduced a method to study dominating codon biases in genomes and we validated it over several known unicellular organisms which have been previously investigated with Correspondence Analysis (CA) (Perrière and Thioulouse, 2002). A few new organisms, as *D.melanogaster* and *C.elegans*, whose genomes are much larger in size and provide a computational obstacle to CA, are also considered. The biological impacts of this new approach will be discussed elsewhere; they comprise the definition of new quantitative measures for comparing species, and the reconstruction and validation of regulatory circuits and metabolic pathways on known and less known organisms.

ACKNOWLEDGEMENTS

We would like to thank Tamara Front for implementing a preliminary version of the algorithm. A.C. and A.Z. are grateful to Misha Gromov, Magda Konarska and Michael Savageau for stimulating discussions. F.K. was supported by CNRS, genopole and Conseil Régional d'Ile-de-France.

REFERENCES

Andersson,S.G.E. and Sharp,P.M. (1996) Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology*, **142**, 915–925.

Bennetzen,J.L. and Hall,B.D. (1982) Codon selection in Yeast. *J. Biol. Chem.*, **257**, 3026–3031.

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

Duret,L. and Mouchiroud,D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **96**, 4482–4487.

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

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

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

Guindon,S. and Perrière,G. (2001) Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol. Biol. Evol.*, **18**, 1838–1840.

Gupta,S.K. and Ghosh,T.C. (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., Mannila,H. and Smyth,P. (2001) *Principles of Data Mining*. *A Bradford Book*. MIT Press, Cambridge, MA.

Holstege,F.C.P., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S. and Young,R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

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.

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

Koski,L.B., Morton,R.A. and Golding,G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.*, **18**, 404–412.

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

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

McInerney,J.O. (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA*, **95**, 10698–10703.

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

Moszer,I., Rocha,E.P.C. and Danchin,A. (1999) Codon usage and lateral gene transfer in *Bacillus Subtilis*. *Curr. Opin. Microbiol.*, **2**, 524–528.

Nicolas,P., Bize,L., Muri,F., Hoebeke,M., Rodolphe,F., Dusko Ehrlich,S., Prum,B. and Bessières,Ph. (2002) Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acid Res.*, **30**, 1418–1426.

Perrière,G. and Thioulouse,J. (1996) On-line tools for sequence retrieval and multivariate statistics in molecular biology. *Comput. Appl. Biosci.*, **12**, 63–69.

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

Sharp,P.M., Tuohy,T.M.F. and Mosurski,K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiate highly and lowly expressed genes. *Nucleic Acids Res.*, **14**, 8207–8211.

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

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

Stenico,M., Loyd,A.T. and Sharp,P.M. (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.*, **22**, 2437–2446.

Syvanen,M. (1994) Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.*, **28**, 237–261.

Varenne,S., Buc,J., Lloubès,R. and Lazdunski,C. (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., Aota,R., Tsuchiya,F., Ishibashi,T., Gojobori,T. and Ikemura,T. (1990) Codon usage tabulated from GenBank genetic sequence data. *Nucleic Acids Res.*, **18** (Suppl), 2367–2411.