Persistent biases in the amino acid composition of prokaryotic proteins

Géraldine Pascal,[1,2] Claudine Médigue,[1] and Antoine Danchin[2]*

Summary

Correspondence analysis of 28 proteomes selected to span the entire realm of prokaryotes revealed universal biases in the proteins' amino acid distribution. Integral Inner Membrane Proteins always form an individual cluster, which can then be used to predict protein localisation in unknown proteomes, independently of the organism's biotope or kingdom. Orphan proteins are consistently rich in aromatic residues. Another bias is also ubiquitous: the amino acid composition is driven by the G + C content of the first codon position. An unexpected bias is driven, in many proteomes, by the AAN box of the genetic code, suggesting some functional biochemical relationship between asparagine and lysine. Less-significant biases are driven by the rare amino acids, cysteine and tryptophan. Some allow identification of species-specific functions or localisation such as surface or exported proteins. Errors in genome annotations are also revealed by correspondence analysis, making it useful for quality control and correction.
***BioEssays* 28:726–738, 2006.**

Introduction

Natural selection drives adaptation of living organisms through subtle variations that integrate over many generations. As the main effectors of the cell's life and architecture, the sequences of proteins integrate features that are directly related to their function and indirectly result from a great diversity of processes contributing to the organism's survival. Each amino acid displays specific structural and physico-chemical properties that combine into the final outcome (Table 1). At each position of each protein sequence, the electric charge, the metabolic cost and codon availability, to choose a few properties, have integrated all kinds of selections, shufflings or drifts during the course of evolution, only a few of which are required (typically less than 10% of all positions) for the function of the catalytic centres. Genome studies analyse the bulk of the proteins of an organism, its proteome, as the conceptual translation of protein-coding DNA sequences.[1–8] With the development of rapid sequencing methods, the genome sequence of prokaryotic organisms, Bacteria and Archaea, have become widely available. Several analyses of the codon-usage bias or the amino acid usage in reference proteomes have been published, such as those of *Thermotoga maritima*,[9] *Pseudomonas aeruginosa*,[10] Buchnera species,[11] or the three model prokaryotes *Escherichia coli, Bacillus subtilis* and *Methanococcus jannaschii*.[12] These studies revealed remarkable and possibly universal rules that we attempted to relate to the biological function and evolution of the proteins of the organisms of interest. To substantiate or disprove the observations derived from this very limited set, we analysed a representative sample of the proteome of the prokaryote world (we chose 28 organisms, excluding closely related organisms) (Fig. 1) to look for rules that would reflect the constraints of evolution, of the environment, or of both. For the sake of generality, we analysed the proteome of one organism from each phylogenetic branch for which a complete genome was available (Fig. 1). Sometimes, several organisms belonging to the same branch were included in the analysis because they displayed interesting biological features that may interfere with the amino acid composition of their proteome (Table 2). Superimposed on the rules that previous work had already partially suggested and that the present work show are indeed universal,[12] we find variations that tell much about the history and environment of each organism, allowing us, by analysing a protein sequence within a proteome, to make inferences about its function and structure. This will help in the discovery of new functions.

[1]Genoscope/CNRS UMR 8030, Atelier de Génomique Comparative, Evry, France.
[2]Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, Paris, France.
*Correspondence to: Antoine Danchin, Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, Paris, France.
E-mail: adanchin@pasteur.fr
DOI 10.1002/bies.20431
Published online in Wiley InterScience (www.interscience.wiley.com).

Overview of the analysis

The choice of each amino acid residue in a protein results from superposition of a wide range of selection pressures, some indirect (such as the metabolic cost to obtain each residue[13]), or the availability of pathways for the synthesis of nucleotides,[14] with a limited contribution from the nature of the protein function. Global features, such as the genomic G + C content (from 28.6% to 72.1%), the optimal growth temperature (from 10°C to 103°C), the ecological niche (living in animal or plant tissues, in soil, in marine, alkaline or acidic

Table 1. Physico-biochemical features of amino acids

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Hydrophobicity																				
Hydrophilic			1	1			1		1			1		1	1					
Hydrophobic	1	1			1			1		1	1							1	1	1
Volume																				
Big					1		1								1				1	1
Medium			1									1	1				1			
Small	1	1				1										1				
Atoms																				
Nitrogen							2		1			1		1	3				1	
Oxygen																1	1			1
Sulphur		1									1									
Carbon	3	3	4	5	9	2	6	6	6	6	5	4	5	5	6	3	4	5	11	9
Side chain																				
Aromatic					2		1												2	2
Basic							1		2						3					
Acid			1	1																
Metabolic cost																				
ATP	0	3	1	1	4	1	6	3	2	3	3	2	2	2	6	1	3	3	5	4
NAD(P)H	0	3	0	−1	1	0	−2	1	2	0	6	0	1	−1	0	−1	2	1	1	0
Dayhoff classes																				
astpg	1					1							1			1	1			
mliv								1		1	1							1		
denq			1	1								1		1						
rkh							1		1						1					
fyw					1														1	1
Post-translational modifications																				
Phosphorylation			1				1									1	1			1
Carbamylation									1											
Glycosylation												1								
Methylation			1				1		1											
Others																				
Number of codons	4	2	2	2	2	4	2	3	2	6	1	2	4	2	6	6	4	4	1	2
G + C-content	0.83	0.50	0.50	0.50	0.17	0.83	0.50	0.11	0.17	0.39	0.33	0.17	0.83	0.50	0.72	0.50	0.50	0.50	0.67	0.17
Role in aging			1									2								
metabolic steps from pyruvate	1	8	4	6	11	7	20	12	13	8	11	5	10	7	12	6	9	4	14	11

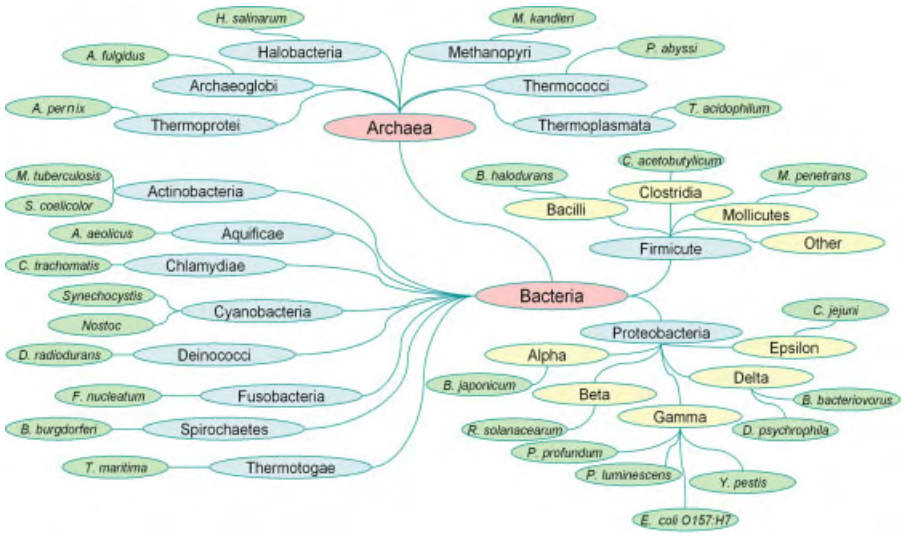

Figure 1. Phylogenetic tree of the 28 organisms of this study.

environments, etc), the extent of phylogenetic divergence (Fig. 1) and other constraints due to differences in growth rate or pathogenicity, all contribute to the final outcome. The methods used in the study are summarised in Box 1. Because selection pressure acts on individual proteins, we analysed the impact of these various constraints on each protein (excluding the shortest ones) of each proteome of the representative organisms, and not on the concatenated proteomes, as proposed in a previous study[15] (Table 2). For this purpose, we used correspondence analysis (CA) (see Box 2). It sometimes happens that the difference between the amino acid compositions of particular groups of proteins is so important that several well-defined clouds are formed. Furthermore, the more a cloud is extended, the more its extremities are interesting to study, because they reveal large divergence in composition and, frequently, are correlated with the proteins' function. Finally, we observed an intermediate level of complexity in the proteomes of interest, with rules that apply not to the whole domain of prokaryotes, but to some domains or to some general environmental constraints, such as temperature. These are described and placed in relation to the niche of the organism.

A universal rule: integral inner membrane proteins cluster together

The amino acids that comprise proteins are roughly split into two major classes, depending on their interaction with water. In the cytoplasm, they contribute to protein folding, with the hydrophobic amino acids usually clustered within the inside of the protein and the hydrophilic to the outside. In contrast, membrane-associated proteins interact with a highly hydrophobic lipid bilayer, which they usually perform through sequences of exposed hydrophobic residues. In particular, hydrophobic alpha helices made of 19–22 residues span the lipid bilayer.[16] Proteins that are embedded in the membrane, with limited outside stretches (Integral Inner Membrane Proteins, IIMPs), are rich in hydrophobic residues[17,18] including a significant amount of Phe, Leu and Ile, while they have only a few charged residues, Asp, Arg, Glu and Lys, mostly located outside of the lipid core of the membrane.[12,17,19] These proteins, atypical in their amino acid composition, differ from the proteins from other cellular compartments, including those of the outer membrane, when it exists.[20,21]

While this feature seemed a fairly general property, it was of interest to explore whether those proteins would group together as a specific cloud of proteins in each organism of interest. Remarkably, in all the organisms analysed, a well-separated cloud was observed. This cloud is distinguished by one single factor, hydrophobicity of proteins, often brought about by leucine and phenylalanine, versus charged residues (See Supplementary Figs 1, 2 and 3 for this article on the BioEssays website (http://www.interscience.wiley.com/jpages/0265-9247/suppmat/2006/28/v28.xxx.html)). In a previous study based on model prokaryotes, this homogeneous class was shown to comprise exclusively IIMPs.[12] Most of the proteomes studied here have not been experimentally characterized. While we can be confident that the isolated cluster driven by the hydrophobic versus charged residues is made of IIMPs, we tentatively named the corresponding proteins Probable Inner Integral Membrane Proteins (PIIMPs).

The presence of a single consistent class of proteins in such a large diversity of organisms is particularly surprising as the study includes both Bacteria and Archaea, and organisms

Table 2. First three columns: description of studied organisms (A: Archaea, B: Bacteria)

Organisms	Classification	Features	GC bias (Axes)	Lys/Asn Bias vs. Opposites (Axes)	Cys & Trp bias (Axes)
A. pernix	A:Thermoprotei		G1 (2), G2 (3)	No bias	Cys (3)
A. fulgidus	A:Archaeoglobi		G1, GC1 (3)	No bias	Cys (2 & 4)
H. salinarum	A:Halobacteria		GC1 (2)	No bias	No bias
M. kandleri	A:Methanopyri		G1(3)	No bias	Cys (4)
P. abyssi	A:Thermococci		No bias	No bias	Cys (3)
T. acidophilum	A:Thermoplasmata		G1 (3)	Asn A2 vs. T2 (2)	Cys (4)
M. tuberculosis	B:Actinobacteria	pathogen	No bias	Gly Asn vs. Arg Glu (1), Asn vs. Ala (2)	No bias
*M. tuberculosis**				Lys Glu Asn vs. Ala (1)	
S. coelicolor	B:Actinobacteria	G + C high	No bias	Lys Asn Ile vs. Arg Ala (1)	Trp (4)
A. aeolicus	B:Aquificae		GC2, GC, C2 (2)	No bias	Cys (3)
C. trachomatis	B:Chlamydiae		GC2, C2 (3)	No bias	No bias
Synechocystis	B:Cyanobacteria: Chroococcales		No bias	Asn Gln vs. Met Val (4)	No bias
Nostoc	B:Cyanobacteria: Nostocales		GC1, G1, GC (3)	Aromaticity Asn vs. GC1 G1 GC (3)	No bias
D. radiodurans	B:Deinococci		No bias	Lys Ile Asn vs. Leu Ala (1)	Cys & Trp (4)
B. halodurans	B:Firmicutes:Bacilli	halophile	G1 (2)	No bias	Cys (4)
C. acetobutylicum	B:Firmicutes:Clostridia		G1, GC1 (2)	Asn Ser A2 vs. T2 (3)	No bias
M. penetrans	B:Firmicutes:Mollicutes		GC1, GC, G1 (2)	Asn T1 vs. GC1 GC (2)	No bias
F. nucleatum	B:Fusobacteria	G + C low	GC2, C2, GC (1) G1, GC1 (2)	A2 Lys Asn vs. GC2 C2 GC (1)	Trp (4)
B. japonicum	B:Proteobacteria: Alpha		C1, GC1 (2)	A1 A2 Lys vs. C1 GC1 (2)	No bias
R. solanacearum	B:Proteobacteria: Beta		No bias	Lys Asn vs. Ala (1)	No bias
B. bacteriovorus	B:Proteobacteria: Delta		GC2 (1)	A2 Lys vs. T2 GRAVY AROMO (2)	No bias
D. psychrophila	B:Proteobacteria: Delta	psychrophile	GC1, G1, GC (3)	A1 Aromaticity Asn vs. GC1 G1 GC (3)	No bias
C. jejuni	B:Proteobacteria: Epsilon		GC1, G1 (2)	A2 Lys vs. GC2 Gly GC (1) Asn Ser vs. Cys (3)	Cys (3)
E. coli O157:H7	B:Proteobacteria: Gamma	pathogen	C1, GC1 (2) G1 (3)	A1 Asn vs. C1 GC1 (2)	No bias
P. luminescens	B:Proteobacteria: Gamma	pathogen	GC, GC1 (1) G1 (3)	A1 Lys Asn vs. GC GC1 (1)	No bias
P. profundum	B:Proteobacteria: Gamma	psychrophile	G1, GC1 (2)	Tyr Asn Ser vs. GC1 (3)	No bias
Y. pestis	B:Proteobacteria: Gamma	pathogen	C1 (2) GC1 (3)	A1 Asn Lys vs. C1 Leu GC (2)	No bias
B. burgdorferi	B:Spirochaetes	G + C low	GC2, GC (1) GC1, GC, G1 (2)	A2 A1 Lys Asn vs. GC2 Gly GC Val (1)	Trp (4)
T. maritima	B:Thermotogae		G1, GC2, GC1 (2)	No bias	Trp (3)

The fourth column gives which axis is built by genome G + C content bias per each organism. The fifth column gives which axis where there is a bias including asparagine (Asn) and/or lysine (Lys) and their opposite biases. The sixth column gives which axis is built by cysteine (Cys) or tryptophan (Trp) bias.
*Proteome of *M. tuberculosis* without PE, PPE and PE-PGRS proteins.

that live in extremely different environments. The case of the industrial bacterium *Clostridium acetobutylicum* is of particular interest (Fig. 2A), as these cells produce a mixture of acetone and butanol. Hydrophobic interactions, which are crucial to membrane organisation, are dramatically modified in solvents such as acetone. The cellular membrane is made up of lipids and it is therefore remarkable that the amino acid composition of *C. acetobutylicum* membrane proteins looks similar to that of the membrane proteins of organisms living in more typical conditions. In the same way, it is worth noticing that PIIMPs constitute a consistent class in Archaea, which have a membrane bilayer formed of lipids completely different from those of Bacteria (ethers instead of esters, in particular).[22] As a consequence, this feature of the amino acid distribution in the proteins that form the proteome of prokaryotes would be useful for valid annotation of the corresponding class of genes in genome projects.

Highly expressed ancestral proteins display common biases

The majority of the expressed proteins in fast-growing organisms constitutes the translation machinery. These core proteins are considered ancestral, and are generally used as markers for phylogenetic analyses, in parallel with studies involving ribosomal RNA.[23,24] They have been used as a reference to compare the proteomes of all organisms, including those that grow poorly or slowly. Generally, highly expressed proteins have a biased amino acid composition,

frequently linked to the way that they use the genetic code, their Codon Adaptation Index (CAI). The highly expressed genes use a subset of optimal codons as a result of selection for efficient translation of their mRNAs.[1,6,8,25–27] Furthermore, ribosomal proteins are characterised by an enrichment in basic amino acids (Lys and Arg) and small, hydrophobic residues (Ala, Val and Gly).[28,29]

The cloud of proteins forming a proteome in Correspondence Analysis can be separated into further clusters (see coloured clusters in Fig. 2 and Supplementary Figs 1, 2 and 3 for this article on the BioEssays website (http://www.interscience.wiley.com/jpages/0265-9247/suppmat/2006/28/v28.xxx.html)).[30] To analyse the content of the different clusters in terms of protein function, we first determined how the proteins conserved in most genomes (''persistent'' proteins) distribute among the clusters.[31] This analysis demonstrated that proteins in the clusters indeed differ in terms of functional annotations. As an example, in the representatives of Firmicutes (*Bacillus halodurans*) and Gamma-proteobacteria (*E. coli* O157:H7), these persistent proteins were mostly found in a single cluster (52% of the persistent proteins make the cluster labelled with yellow circles for *B. halodurans* and 62% of the persistent proteins are in the orange triangle cluster for *E. coli*). To further understand the functional constraints that might drive this clustering, we analysed ribosomal proteins. In the majority of the proteomes (22 out of 28), the ribosomal proteins belonged to one

Box 1: Material and methods

Correspondence analysis[65] was used to analyse the distribution of amino acids in the proteins of the proteomes (see Box 2). Clustering into consistent classes used the dynamic clouds method[66] that automatically clusters the proteins located close to one another.

All complete proteomes of interest were extracted from the Genome Review databank (http://www.ebi.ac.uk/genomes/). In order to avoid constraints linked to the molecular processes of initiation and termination of translation, all proteins used in our study were truncated by 10 amino acids from their N-terminal end, and 5 amino acids from their C-terminal end (there is an over-representation of hydrophilic residues near both termini of proteins[67]). To reduce the influence of stochastic variations that may occur in small proteins, only proteins longer than 100 residues (after truncating) were retained.

Box 2: Correspondence Analysis

Correspondence analysis (CA) is a multivariate method that belongs to the Principal Component Analysis (PCA) class of multivariate methods meant to extract information from large data tables (contingency tables) associating objects and properties (in our case proteins and their amino acid residues). The goal of this class of methods, which all rest on the assumption that the statistical behaviour of the dataset of interest follows Laplace-Gauss ('normal'') statistics is to explore the links between objects in the data table as consistent function of properties. To this aim, a distance is calculated between objects, using a particular measure. In PCA, it is usually based on providing the same weight to each object of interest. This has many unwanted consequences when analysing heterogeneous datasets. In contrast to PCA, CA therefore uses a more sophisticated measure and considers the chi-squared metrics[65] to calculate the distance between objects, a common statistical way to provide a similar weight to objects the type of which can vary widely in a collection to be analysed. One important feature of CA is that it allows representation simultaneously of the lines and the columns of the contingency table (here the proteins in the space of amino acids or the amino acids in the space of proteins are represented jointly). Correspondence analysis (CA)[65] was used to identify the major factors that shape variation in amino acid usage among proteins of the organism of interest. The analyses were based on absolute frequencies (i.e. actual residues counts in each protein) in order to avoid introducing some unwanted biases.[68] Correspondence analysis was applied on the data table including all proteins of an organism as described by their amino acid usage, to determine an orthogonal space, or factorial space, with dimension 19. The axes (called factors) are constructed and ordered in such a way to represent the space with decreasing order of importance (each axis therefore contributes to the information on the structure of the CA space). The axes are represented in a decreasing order of importance as quantified by their corresponding 'inertia''[69] (axis 1 is the most informative, followed by axis 2 and so on). Sequences that have a similar amino acid composition appear as neighbours. Following CA, an automatic clustering method, the dynamic clouds method,[30,70] was subsequently used to interpret the graphical representation in terms of clusters with common properties.

or two clusters. For example, despite separation of the *Campylobacter jejuni* proteome into four classes (Fig. 2B), all 66 identified proteins cluster together. This systematic gathering into one or two clusters shows once again that ribosomal proteins, and more generally persistent proteins (data not shown), are remarkably conserved in prokaryotes in terms of their amino acid composition, despite enormous phylogenetic divergence or variety of growth environments.

Rare amino acids create specific clusters

Not all amino acids are equal in terms of their frequency in proteins. Some are systematically rare, while others are frequent. Among the former, cysteine (Cys) and tryptophan (Trp) are particularly important as their rarity (each represents less than 1% of the amino acids of the proteomes) often results in a strong bias in the proteome cloud's shape. Cys contains the highly reactive sulphhydryl group, and takes part in numerous active sites. It also plays an important role in structural stabilisation of exported proteins in forming disulphide bridges. Trp is an aromatic amino acid with a voluminous side chain. Its specific role is not well established: besides its aromaticity, it is mildly hydrophobic and often contributes to the stabilisation of protein structures; it has a positive impact on the folding of proteins, because indole can accept hydrogen bonds under certain circumstances.[32,33] Trp is not, however, used very frequently, perhaps because it is very costly in terms of metabolism and quite reactive towards reactive oxygen species (also, it is usually coded by only one codon, TGG, or two in many mycoplasma, including TGA). These two residues, Cys and Trp, are so rare in proteomes that proteins having several of these amino acids are atypical and therefore worth investigation.

As shown in Table 2, a consequence of this scarcity is that one of the first four axes that best describes the proteome is often led by the frequency of the rare residues Cys and Trp. Analysis of the extremities of cysteine-biased axes shows that most Cys-rich proteins systematically belong to the class of metal-binding proteins (iron- or zinc-binding in general), presumably via Cys residues (data not shown). In some organisms (e.g. *Thermoplasma acidophilum*), many of these proteins are annotated as 'of unknown function'; inference from annotation of the other genome sequences suggests that the corresponding proteins belong to similar classes. As expected perhaps, organisms atypical in their G + C content, such as *Borrelia burgdorferi* or *Fusobacterium nucleatum* (A + T-rich) and *Streptomyces coelicolor* (G + C-rich), show a Trp distribution bias in one of their factorial first four axes. After analysis of the proteins located at the extremities of Trp bias axes, we observed that this was due to a variety of independent causes and not a single common one (data not shown). In *B. burgdorferi*, the bias mostly affects proteins involved in the translation machinery. This is likely due to the fact that the corresponding genes are located in the leading strand of the DNA, which is biased in G + T.[34] Furthermore, biochemical experiments suggest involvement of Trp residues in RNA binding.[35,36] In *F. nucleatum*, proteins are mostly involved in the construction of the envelope, with a significant proportion of proteins predicted to be on the surface of the cell. Interestingly, this seems also to be the case of *S. coelicolor*, where an excess of Trp is expected because of the high G + C content of the cells: in this case, mostly in proteins of the cell envelope and in particular heme-binding proteins.[37]

The G + C content of the coding DNA sequences creates an unexpected bias in amino acid composition

The amino acids present in proteins are constrained by the nucleotide composition of the cognate genes. Because there is often a strong bias in the composition of the leading and lagging strands of chromosomes, distinguished from each other by an enrichment in G + T and C + A, respectively, proteins coded from the leading strand are enriched in valine relative to those coded from the lagging strand, which are enriched in isoleucine and threonine.[34,38–41] Moreover, there is a general bias in GNN codons in CDSs, possibly acting as a ratchet-like mechanism during the translation elongation process that could influence the overall amino acid composition of the proteomes.[42] In the same way, the overall G + C base composition of genomes influences strongly the choice of amino acids that constitute the corresponding proteomes.[12,43,44] The G + C content will drive the codon usage bias and not the reverse.[45] As a consequence, because it is adapted to the optimal growth temperature of the organism,[46–48] the resulting constraints on the proteome create the first discriminant factor for thermophilic organisms.[15,49]

The role of nucleotides in codons is not symmetrical (the third position of codons often results in synonymous substitutions in proteins) and this may impact on the amino acid content of a proteome depending on the G + C content of the genome. It has long been noticed that the second codon position shows the highest correlation with the specific nature of the amino acid, with T associated with hydrophobic residues and A with hydrophilic residues. When present, the bias due to the second position of the codons' nucleotides was highly correlated with a bias in amino acids (or amino acids similar to each other according to Dayhoff's classification). For example, the *Aquifex aeolicus* G + C bias at the second position of codons (Table 2) overlaps with the alanine, glycine and proline biases (GCN, GGN and CCN respective codons) and impacts on many of the cellular metabolism proteins. Remarkably, however, we observed that the G + C genes' content influenced the amino acid composition of the proteins in an unexpected way, frequently following a G + C gradient at the *first* codon position. We analysed phylogenetically consistent groups of Bacteria and Archaea to see whether this resulted from a separation between different classes of protein

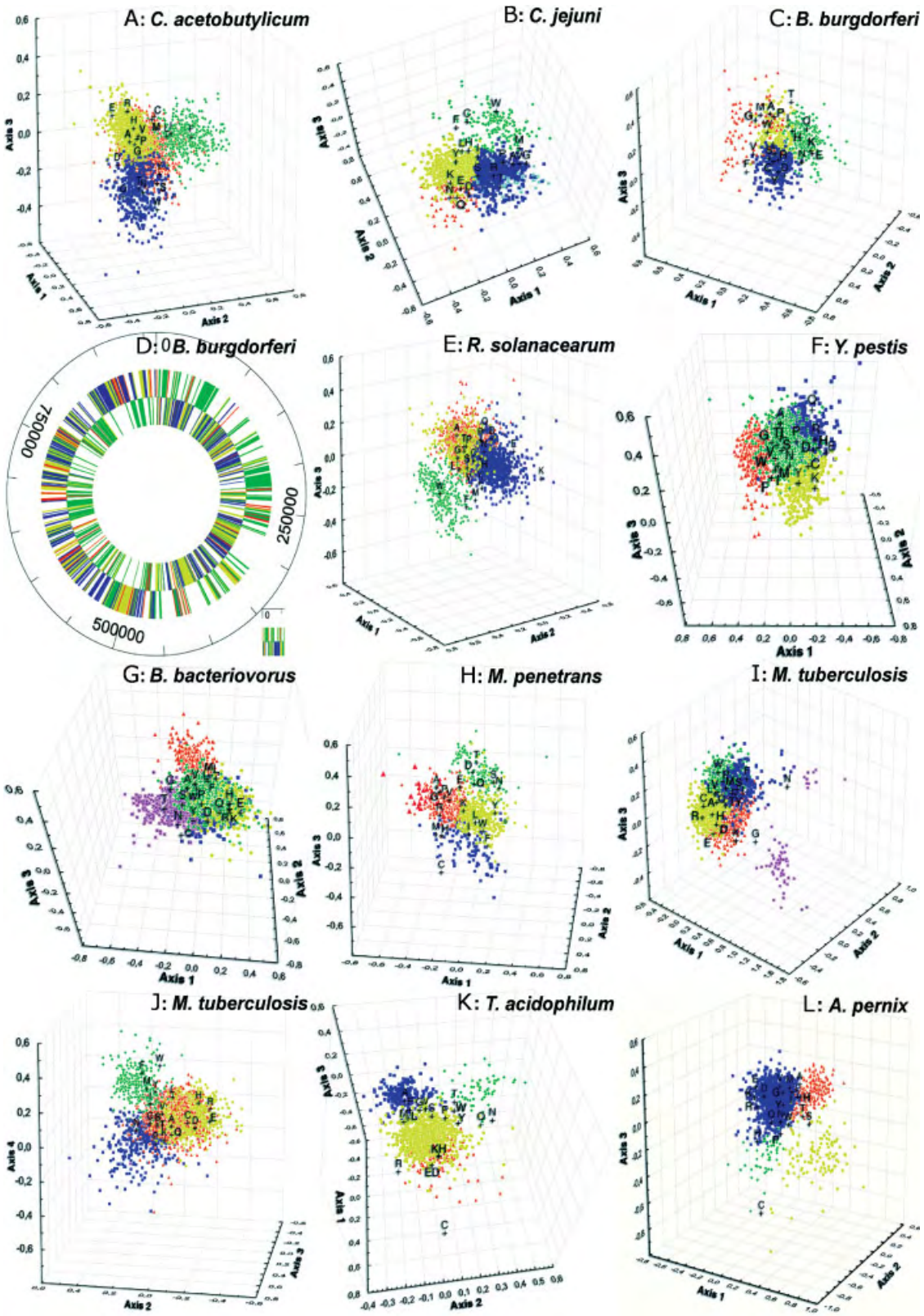
A: C. acetobutylicum
B: C. jejuni
C: B. burgdorferi
D: B. burgdorferi
E: R. solanacearum
F: Y. pestis
G: B. bacteriovorus
H: M. penetrans
I: M. tuberculosis
J: M. tuberculosis
K: T. acidophilum
L: A. pernix
Axis 1
Axis 2
Axis 3
0
250000
500000
750000

Figure 2.

functions (Table 2). This study demonstrated that a bias driven by the G + C content at the first codon position was probably not associated with specific biological functions.

While this G + C-content-driven bias seems fairly ubiquitous, its raison-d'être in terms of functionality in proteins is not obvious. This bias creates a discrimination between Asn, Cys, Ile, Lys, Met, Phe, Ser, Thr, Trp, Tyr on the one hand and Ala, Asp, Gln, Glu, Gly, His, Pro, Val on the other hand. Each class has some common properties (aromaticity, sulfur content, hydroxyl group in the first class; negative electric charge, relatively smaller size in the second class) but nothing really compelling (Table 1). This bias might be the consequence of a remnant of the origin of translation associated with some optimisation of the translation machinery, which seems to prefer GNN codons.[42] However, in the case of the G + C-poor genome of *B. burgdorferi* (Fig. 2C), where the G + C bias at the first position of codons is correlated with the overall G + C content of the organism, a remarkable separation of proteins into four clusters was observed. Two of those (green diamond and blue square clusters, Fig. 2C) (separated along axis 2 and correlated with the G + C content at the first position of codons) contained almost exclusively proteins encoded by the leading strand or the lagging strand of the chromosome (Fig. 2D). This particular split between genes has previously been noticed in several studies.[34,50]

Aromatic amino acids tag orphan proteins

As new genome sequences continue to be deciphered, a particular class of proteins of unknown function is becoming prominent. While most of the predicted proteins are similar to counterparts in other genomes, approximately 10% of the proteins do not look like anything known in other genomes, except in those that are highly related. These proteins are usually named orphan proteins. Interestingly, they are enriched in aromatic amino acids.[12,51] Aromatic amino acids are very costly in terms of metabolic requirements, supporting the hypothesis that they are newly created proteins that did not yet have time to be optimised in terms of cost versus benefit. As a consequence, they are markers of the ''self'' of a given species. One of their possible functions is that of stabilising agents for multimeric complexes. Those with this type of hypothetical function have been named ''gluons''.[12] An aromatic amino acid bias was rarely observed in Archaea but frequently in Bacteria. This is shown in Table 3, where the content of proteins of unknown function (hypothetical or putative proteins) has been computed for the 10% located in the CA cloud at the extremity of the axis driven by aromaticity. This value has been subsequently compared to the percentage of unknown proteins in the whole proteomes. Two interesting features stand out: (1) most aromatic proteins show a significant proportion of unknown proteins (1.1–2.5 times the average, depending on the proteome), and (2) this gradient is more pronounced when the axis driven by aromaticity appears early in the order of the axes organising the CA space. This is most likely due to the contribution of orphan proteins, as observed in the proteome of model organisms,[12] substantiating our previous hypotheses about the involvement of aromatic residues in the creation of new functions, such as protein complex stabilisation.

A persistent bias is generated by AAN codons

As we go down the list of importance of the CA axes, we find features that become more and more specific for a given species. However, we still uncovered a remarkable bias that persists in many genomes: numerous organisms (19 of the selected organisms) present a clear bias due to a gradient in lysine (Lys) and/or asparagine (Asn) content along one of the CA factorial axes (Table 2). While this bias exists in *B. subtilis* (41% G + C), it is almost absent in *B. halodurans* (44% G + C), a halophilic Bacillus. It is present in a single Archaeon living in an acidic biotope, *T. acidophilum* (46% G + C).

Interestingly, in *Deinococcus radiodurans*, the AAN bias is the first factor of the CA cloud. The proteins responsible for the bias are clearly linked to the protein biosynthesis machinery, with Lys contributing most to the bias. In the case of *C. acetobutylicum* (Asn bias on axis 3) the cluster of Asn-rich proteins (428 proteins) was dominated by enzymes involved in polysaccharide biosynthesis and turnover, and often present in the cellulosome.[52] *C. acetobutylicum* is known to have an abundance of polymer degradation systems.[53] In line with a specific relationship between Asn enrichment and the surface of the cell, we also observed that many proteins linked to

Figure 2. First CA space (axes 1, 2 and 3) with the PIIMPs group indicated in parentheses, except when indicated otherwise **A:** *Clostridium acetobutylicum* (green diamonds), **B:** *Campylobacter jejuni* (green diamonds), cyan stars represent proteins tagged ''ribosomal protein'', **C:** *Borrelia burgdorferi* (orange triangles), **D:** distribution along leading and lagging strands of chromosome of genes of CA clustered proteins of *B. burgdorferi*. **E:** *Ralstonia solanacearum* (green diamonds), **F:** *Yersinia pestis* (orange triangles), **G:** *Bdellovibrio bacteriovorus* (orange triangles), pink stars represent G + C rich gene proteins, **H:** *Mycoplasma penetrans* (blue squares), green diamonds represent Ser- and Thr-rich proteins, large red triangles and the large khaki circle represent proteins tagged ''ribosomal protein'', **I:** *Mycobacterium tuberculosis*, PE/PPE/PE-PGRS pink stars, **J:** axis 2, 3 and 4 of *M. tuberculosis* without PE, PPE and PE-PGRS proteins (green diamonds), **K:** *Thermoplasma acidophilum* (blue squares), orange triangles represent cysteine rich proteins, yellow circles represent cytoplasmic proteins, green diamonds represent extracellular proteins, **L:** *Aeropyrum pernix* (orange triangles), blue squares and green diamonds represent proteins which may be false-positive.

Table 3. This table shows which organisms have a bias due to aromatic amino acids

Organism	Axis	% unknown in 10% aromatic rich	%unknown in proteome
A. pernix	no bias		
A. fulgidus	no bias		
H. salinarum	3	49	33
M. kandleri	no bias		
P. abyssi	no bias		
T. acidophilum	no bias		
M. tuberculosis	no bias		
S. coelicolor	no bias		
A. aeolicus	no bias		
C. trachomatis	3	44	23
Synechocystis	no bias		
Nostoc	3	62	35
D. radiodurans	no bias		
B. halodurans	2	44	22
C. acetobutylicum	4	25	21
M. penetrans	no bias		
F. nucleatum	4	40	22
B. japonicum	3	37	30
R. solanacearum	no bias		
B. bacteriovorus	no bias		
D. psychrophila	2	65	26
C. jejuni	4	26	23
E. coli O157:H7	3	46	23
P. luminescens	3	41	30
P. profundum	no bias		
Y. pestis	3	39	21
B. burgdorferi	2	53	24
T. maritima	2	54	22

Moreover, it displays the percentage of unknown proteins content among 10% of proteins situated at extremity of CA axis built by aromaticity bias (aromatic-rich extremity). The percentage of unknown proteins in the global proteome is also indicated.

flagella and the cell wall belonged to this same cluster. Likewise, a Lys-dominated bias was observed on axis 2 of *Bdellovibrio bacteriovorus* and an Asn-dominated bias on axis 3 of *C. jejuni*. In *B. bacteriovorus*, the group formed on axis 2 is mostly composed of hypothetical proteins (two-thirds), while most of the rest are linked to proteins involved in the outer surface of the cell: outer membrane proteins, flagella, cell wall, secreted proteases or other extracellular activities. A similar situation holds for *C. jejuni*, with proteins linked to flagella, outer membrane, chemotaxis and proteolysis forming the bulk of the Asn-rich proteins. *Ralstonia solanacearum* represents a situation where the bias involves both Lys and Asn. The cluster defined by this bias (the yellow circles shown in Fig. 2E) comprises mostly outer membrane proteins, porins, siderophore-iron transporter activity and calcium ion binding proteins. In the same way, the *Yersinia pestis* proteome also has a bias in Asn and Lys, on axis 2 (Fig. 2F). Proteins situated on the Asn/Lys-enriched extremity are often putatively exported proteins or proteins located in the outer membrane. The most remarkable feature of this particular bias is that it behaves as if Asn and Lys had some common physico-chemical feature that would account for them being coded by the same box (AAN) of the genetic code. Both are hydrophilic but, in general, Lys is positively charged. However, when appropriately screened from the environment, the amino-terminal end of Lys can be a doublet electron-donating group. In known metabolism, the most similar amino acid would be ornithine. This amino acid, however, is not among those present in proteins, the underlying biochemical reason being that, upon activation as an amino acid adenylate and loading onto tRNA, it would become cyclised.[54] One may therefore conjecture that asparagine might use its end amide group in the same way as lysine does with its distal amine group, with a shorter chain, as a substitute for ornithine, thereby accounting for the selection pressure that has coded both amino acids from a common genetic code box.

Zooming in on some interesting cases

Most biases described in this study were persistent among prokaryotes, indicative of common trends of selection, probably associated with inevitable physico-chemical constraints. Sometimes however, a bias showing up in one of the CA axes, belonged to only one organism. Three of the corresponding features, which are usually highly revealing about the lifestyle of an organism, are presented in the next paragraphs.

G+C content bias of Bdellovibrio bacteriovorus

B. bacteriovorus, a medium G + C content bacterial parasite, displays a strong bias mediated by the G + C content of its genes at the *second* position of its codons. This bias builds a well-separated cluster of proteins in the CA space (pink stars, axis 1, Fig. 2G). Remarkably, approximately 30% of the proteins of *B. bacteriovorus* are annotated as "protein precursors" (i.e. as proteins in the form that they assume before post-translational modifications and processing, usually for export). This was so unexpected that we verified that all these proteins were authentic proteins, using the MiCheck tool.[55] The distribution of these proteins into five clusters, determined using dynamic clustering,[30] is not uniform. Indeed, whereas four clusters contained between 20 and 30% of "protein precursors", the pink star protein cluster displayed in Fig. 2G is composed of approximately 70% of this kind of protein. In addition, these proteins stem from genes that are G + C rich. Interestingly, the genome analysis did not identify particular regions that would deviate from the average G + C content (except for four regions generally composed of A + T rich genes).[56] This shows that the method could be used as a complement to detect properties that are not identified by more standard genome analysis approaches.

Atypical PIIMPs of Mycoplasma penetrans

Although the *M. penetrans* CA presents a PIIMPs cluster similar to that of all other prokaryotes, this cluster is not driven by an opposition between Phe and Leu and charged amino acids but by an opposition between charged (Lys, Glu) and uncharged hydrophilic residues (Thr, Ser). At one extremity of the axis, a cluster of Ser+Thr rich proteins is formed (green diamond cluster, Fig. 2H), with about half of them annotated as hypothetical, and about one-third as membrane-associated proteins (blue square cluster content: PIIMPs). Yánez and co-workers have described the original structure of the membrane of this Mollicute.[57] *M. penetrans* has a typical elongated, flask-shaped morphology, with two internal compartments that permit the cells to adhere, then to penetrate into human cells. Ser+Thr rich proteins are often involved in adhesion to membranes[58–60] and it is not unexpected that the small proteome of this pathogen could be biased by one specific category of proteins. The proteome of *M. penetrans* displayed a second unusual feature: the cluster formed on axis 2 of the CA, built by opposition of Asn to the codons G + C content, contained almost all the ribosomal proteins of the organism (Fig. 2H), except for an interesting exception. Unexpectedly, protein MYPE1290, annotated as a "ribosomal protein" is located completely outside the ribosomal protein cluster, in another cluster comprising many enzymes (large khaki circle, Fig. 2H). This prompted us to check its annotation. In the metabolic pathway database MetaCyc,[61] MYPE1290 is annotated as a enzyme (alanine acetyltransferase, putative, EC number 2.3.1.128). Using its sequence to browse the proteome of Firmicutes, we uncovered that it is most probably an enzyme modifying a ribosomal protein (e.g. YdiD in the SubtiList database). Combining this description and the protein localisation in the CA space, we can therefore be confident that this protein is not a structural constituent of the ribosome, but rather a protein involved in post-translational modification. This further illustrates that the methodology presented here is a powerful complement to other methods used for proteome annotation.

Pathogenicity proteins of Mycobacterium tuberculosis

The CA cloud of *M. tuberculosis* is appreciably atypical, as shown in Fig. 2I. Axis 1 allows the formation of a first cluster (pink stars in Fig. 2I), opposing Gly and Asn to Arg. This cluster mostly contains the PPE, PE and PE-PGRS family proteins (some of these are also located in the blue cluster, at the positive extremity of axis 1). These proteins are rich in Gly and certain members of these families could be located in the mycobacterial cell wall.[62] This cluster is well-separated on axis 2 due to opposition of Asn and Phe versus Ala. Remarkably, this opposition differentiates PPE proteins (positive extremity of axis 2) from PE-PGRS proteins (negative extremity of axis 2). The bias is so intense that it may hide other biases. Indeed, the first CA axis represents approximately 35% of the total information, while the average information of axis 1 in the CA of all others organisms is between 15% and 25%. To overcome the contribution of this unwanted bias, we computed a new CA omitting the proteins of the PPE, PE and PE-PGRS families (approximately 140 proteins). The most interesting outcome of the new analysis is the opposition between Pro and Leu, Phe on axis 4, separating two clusters of proteins (Fig. 2J). All the identified functions of the proteins in one of the clusters correspond to IIMPs; we can therefore reasonably predict that all proteins of this cluster belong to that category. The second one (blue square cluster) is very homogeneous and when annotated its members are all somehow involved in pathogenicity, leading to the hypothesis that the whole cluster is composed of proteins involved in pathogenic processes. This comprises: (1) hydrolases, all associated with activity at the cell surface or to proteins anchored in the membrane (proteases, peptidases, murein hydrolases, complex carbohydrate hydrolases, lipases), (2) weak complexity proteins, (3) protein kinase-like regulators and (4) other proteins present at the membrane surface (integral membrane proteins, oxido-reductases etc.). In addition to providing us with a remarkable way to make inferences about actors of pathogenicity, these observations lead to the idea that CA shows several levels of complexity and that by suppressing some clusters, we could observe novel

characteristics associated with specific amino acid usage in proteins.

Perspective: using CA as a tool for functional annotation.

In addition to the clear characterisation of PIIMPs, CA could be used as a versatile tool for protein functional annotation to refine the annotation of already annotated proteins, and to propose new functional categories for unannotated proteins. Furthermore, as demonstrated in the case of *M. penetrans*, CA can help to identify probable annotation errors and, in general, to annotate new genome sequences. This can be illustrated by two final examples. The CA of *T. acidophilum* is quite unusual. Indeed, CA separates the *T. acidophilum* proteome into four distinct clusters (Fig. 2K). Three clusters are associated with a specific cellular compartment, while the fourth one, driven by cysteine, is composed of iron-binding proteins, mostly proteins containing iron–sulfur clusters. Not surprisingly, the PIIMP cluster is composed of very hydrophobic proteins. In contrast, the yellow circle cluster (Fig. 2K) groups together cytoplasmic proteins, while extracellular proteins (green diamond cluster) are clustered according to their asparagine content.

To investigate further whether CA could be used to improve genome sequence annotation, we chose to work on the proteome of *Aeropyrum pernix*, as revised by Natale et al.[63] As in the case of *T. acidophilum*, the CA that we obtained is atypical. Four well-separated protein clusters are formed and the correlations between the amino acids constructing the axes differ from those other organisms, except for the hydrophobicity-driven axis 1. The orange triangle cluster, shaped on this axis, indeed contains membrane proteins (Fig. 2L). After exploration of the functional annotations of proteins of the three other clusters, two were found to be composed (green diamond and blue square clusters) almost exclusively of unknown proteins (98% and 90%, respectively), whereas the last one (yellow circle cluster) contains the rest of the proteome (with 36% unknown proteins). This may indicate that, in spite of reannotation efforts, the annotated sequence of this genome still contains many erroneous proteins.

Conclusion

The availability of large datasets derived from genome studies makes them amenable to refined statistical analyses. It is remarkable and perhaps surprising that analysis of the collective behaviour of such simple objects as individual codons[64] or amino acids (this work) allow us to uncover unexpectedly high correlations between the function and/or the structure of corresponding biological entities. Proteins with similar amino acid composition play similar role (e.g. pathogenicity) or belong to common structures (e.g. the cytoplasmic membrane or the cell's envelope). While we have defined broad clusters with common amino acid properties, it appears that a finer grain analysis reveals that neighbouring proteins often have common properties. This is the case of outer membrane proteins of gamma-proteobacteria, for example. However, despite our care in trying to span as much as possible of the tree of prokaryotic life, our choice of organisms for analysis, which reflected faithfully the availability of genomes in databases, could not escape being biased by the available complete proteomes (predicted from complete genomes). Some of the rules that we have uncovered might be qualified by analysis of new proteomes, and we expect that new functional relationships, associated to new specific biases, will appear as the full proteomes from organisms living in usual or difficult niches become known. This work must be understood as a first step in deeper knowledge of what life is.

Acknowledgments

This work was supported by the European Union Network of Excellence BioSapiens, grant LSHG CT-2003-503265, the French Ministry of Research ACI IMPBio Blastsets and MicroScope. We wish to express our gratitude to David Vallenet and Zoé Rouy for their contribution on data exploration. We thank Aurélie Lajus for her help on running MiCheck and Susan Cure for her help in writing the manuscript.

References

1. Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res 22:3174–3180.
2. Blake RD, Hinds PW. 1984. Analysis of the codon bias in E. coli sequences. J Biomol Struct Dyn 2:593–606.
3. Grantham R, Gautier C, Gouy M, Mercier R, et al. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62.
4. Grantham R, Gautier C, Gouy M, Jacobzone M, et al. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74.
5. Grantham R, Gautier C, Gouy M. 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res 8:1893–1912.
6. Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074.
7. Karlin S, Mrazek J, Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899–3913.
8. Médigue C, Rouxel T, Vigier P, Hénaut A, et al. 1991. Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol 222: 851–856.
9. Zavala A, Naya H, Romero H, Musto H. 2002. Trends in codon and amino acid usage in Thermotoga maritima. J Mol Evol 54:563–568.
10. Gupta SK, Ghosh TC. 2001. Gene expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa. Gene 273:63–70.
11. Palacios C, Wernegreen JJ. 2002. A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes. Mol Biol Evol 19:1575–1584.
12. Pascal G, Médigue C, Danchin A. 2005. Universal biases in protein composition of model prokaryotes. Proteins 60:27–35.

13. Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci USA 99:3695–3700.
14. Rocha EP, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet 18:291–294.
15. Tekaia F, Yeramian E, Dujon B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297:51–60.
16. Tie JK, Nicchitta C, von Heijne G, Stafford DW. 2005. Membrane topology mapping of vitamin K epoxide reductase by in vitro translation/cotranslocation. J Biol Chem 280:16410–16416.
17. Ulmschneider MB, Sansom MS. 2001. Amino acid distributions in integral membrane protein structures. Biochim Biophys Acta 1512:1–14.
18. van Geest M, Lolkema JS. 2000. Membrane topology and insertion of membrane proteins: search for topogenic signals. Microbiol Mol Biol Rev 64:13–33.
19. Wallin E, von Heijne G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 7:1029–1038.
20. Perriere G, Lobry JR, Thioulouse J. 1996. Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. Comput Appl Biosci 12:519–524.
21. Guerdoux-Jamet P, Henaut A, Nitschke P, Risler JL, et al. 1997. Using codon usage to predict genes origin: is the Escherichia coli outer membrane a patchwork of products from different genomes? DNA Res 4: 257–265.
22. Pereto J, Lopez-Garcia P, Moreira D. 2004. Ancestral lipid biosynthesis and early membrane evolution. Trends Biochem Sci 29:469–477.
23. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579.
24. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090.
25. Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295.
26. Karlin S, Mrazek J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol 182:5238–5250.
27. Pan A, Dutta C, Das J. 1998. Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: translational selection versus mutational bias. Gene 215:405–413.
28. Karlin S, Mrazek J, Campbell AM. 1998. Codon usages in different gene classes of the Escherichia coli genome. Mol Microbiol 29:1341–1355.
29. Lin K, Kuang Y, Joseph JS, Kolatkar PR. 2002. Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics. Nucleic Acids Res 30:2599–2607.
30. Diday E. 1971. Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. Rev Stat Appliquée 19:19–33.
31. Fang G, Rocha E, Danchin A. 2005. How essential are nonessential genes? Mol Biol Evol 22:2147–2156.
32. Clark EH, East JM, Lee AG. 2003. The role of tryptophan residues in an integral membrane protein: diacylglycerol kinase. Biochemistry 42: 11065–11073.
33. Zhu K, Jutila A, Tuominen EK, Patkar SA, et al. 2001. Impact of the tryptophan residues of Humicola lanuginosa lipase on its thermal stability. Biochim Biophys Acta 1547:329–338.
34. Rocha EP, Danchin A, Viari A. 1999. Universal replication biases in bacteria. Mol Microbiol 32:11–16.
35. Dresios J, Chan YL, Wool IG. 2002. The role of the zinc finger motif and of the residues at the amino terminus in the function of yeast ribosomal protein YL37a. J Mol Biol 316:475–488.
36. Skinner LM, Jackson MP. 1997. Investigation of ribosome binding by the Shiga toxin A1 subunit, using competition and site-directed mutagenesis. J Bacteriol 179:1368–1374.
37. David PS, Dutt PS, Wathen B, Jia Z, et al. 2000. Characterization of a structural model of membrane bound cytochrome c-550 from Bacillus subtilis. Arch Biochem Biophys 377:22–30.
38. Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. Trends Genet 13:240–245.
39. Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665.
40. Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238:65–77.
41. Lobry JR, Louarn JM. 2003. Polarisation of prokaryotic chromosomes. Curr Opin Microbiol 6:101–108.
42. Brooks DJ, Fresco JR. 2003. Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins. Gene 303:177–185.
43. Lobry JR. 1997. Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205:309–316.
44. Wilquet V, Van de Casteele M. 1999. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res Microbiol 150:21–32.
45. Knight RD, Freeland SJ, Landweber LF. 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:research0010.0011–0010.0013.
46. Musto H, Naya H, Zavala A, Romero H, et al. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett 573:73–77.
47. Lynn DJ, Singer GA, Hickey DA. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 30:4272–4277.
48. Naya H, Zavala A, Romero H, Rodriguez-Maseda H, et al. 2004. Correspondence analysis of amino acid usage within the family Bacillaceae. Biochem Biophys Res Commun 325:1252–1257.
49. Kreil DP, Ouzounis CA. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 29:1608–1615.
50. Lafay B, Lloyd AT, McLean MJ, Devine KM, et al. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642–1649.
51. Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14:1036–1042.
52. Perret S, Belaich A, Fierobe HP, Belaich JP, et al. 2004. Towards designer cellulosomes in Clostridia: mannanase enrichment of the cellulosomes produced by Clostridium cellulolyticum. J Bacteriol 186:6544–6552.
53. Nolling J, Breton G, Omelchenko MV, Makarova KS, et al. 2001. Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. J Bacteriol 183:4823–4838.
54. Jakubowski H. 1999. Misacylation of tRNALys with noncognate amino acids by lysyl-tRNA synthetase. Biochemistry 38:8088–8093.
55. Cruveiller S, Le Saux J, Vallenet D, Lajus A, et al. 2005. MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. Nucleic Acids Res 33:W471–479.
56. Rendulic S, Jagtap P, Rosinus A, Eppinger M, et al. 2004. A predator unmasked: life cycle of Bdellovibrio bacteriovorus from a genomic perspective. Science 303:689–692.
57. Antonio Yáñez LC, Olivier Neyrolles, Encarnación Alonso, Marie-Christine Prévost, et al. 1996 Mycoplasma penetrans Bacteremia and Primary Antiphospholipid Syndrome. July 14–19; Orlando, FL, USA.
58. Siboo IR, Chambers HF, Sullam PM. 2005. Role of SraP, a Serine-Rich Surface Protein of Staphylococcus aureus, in binding to human platelets. Infect Immun 73:2273–2280.
59. Handley PS, Correia FF, Russell K, Rosan B, et al. 2005. Association of a novel high molecular weight, serine-rich protein (SrpA) with fibril-mediated adhesion of the oral biofilm bacterium Streptococcus cristatus. Oral Microbiol Immunol 20:131–140.

60. Plummer C, Wu H, Kerrigan SW, Meade G, et al. 2005. A serine-rich glycoprotein of Streptococcus sanguis mediates adhesion to platelets via GPIb. Br J Haematol 129:101–109.
61. Krieger CJ, Zhang P, Mueller LA, Wang A, et al. 2004. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32:D438–D442.
62. Brennan MJ, Delogu G. 2002. The PE multigene family: a 'molecular mantra' for mycobacteria. Trends Microbiol 10:246–249.
63. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, et al. 2000. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol 1:RESEARCH0009.
64. Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, et al. 2006. Codon usage domains over bacterial chromosomes. PLOS computational biology 2:e37.
65. Benzecri JP. 1973; L'analyse des données, L'Analyse des Correspondances. Paris, France: Dunod Edition.
66. Delorme MO, Henaut A. 1988. Merging of distance matrices and classification by dynamic clustering. Comput Appl Biosci 4:453–458.
67. Rocha EP, Danchin A, Viari A. 1999. Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res 27:3567–3576.
68. Perrière G, Thioulouse J. 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 30:4548–4555.
69. Lebart T, Morineau A, Warwick K. 1984. Multivariate Descriptive Statistical Analysis. Hoboken, NJ: John Wiley and Sons.
70. Delorme MO, Hénaut A. 1988. Merging of distance matrices and classification by dynamic clustering. Comput Appl Biosci 4:453–458.