Universal Biases in Protein Composition of Model Prokaryotes

Géraldine Pascal,[1,2]* Claudine Médigue,[1] and Antoine Danchin[2]
[1]*Genoscope/CNRS UMR 8030, Atelier de Génomique Comparative, Evry, France*
[2]*Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, Paris, France*

ABSTRACT **The levels of cellular organization in living organisms are the results of a variety of selection pressures. We have investigated here the final outcome of this integrated selective process in proteins of the best known microbial models *Escherichia coli*, *Bacillus subtilis*, and *Methanococcus jannaschii*, supposed to have undergone separate evolution for more than 1 billion years. Using multivariate analysis methods, including correspondence analysis, we studied the overall amino acid composition of all proteins making a proteome. Starting from and further developing previous results that had pointed out some general forces driving the amino acid composition of the proteomes of these model bacteria, we explored the correlations existing between the structure and functions of the proteins forming a proteome and their amino acid composition. The electric charge of amino acids measured against hydrophobicity creates a highly homogeneous cluster, made exclusively of proteins that are core components of the cytoplasmic membrane of the cell (integral inner membrane proteins). A second bias is imposed by the G+C content of the genome, indicating that protein functions are so robust with respect to amino acid changes that they can accommodate a large shift in the nucleotide content of the genome. A remarkable role of aromatic amino acids was uncovered. Expressed orphan proteins are enriched in these residues, suggesting that they might participate in a process of gain of function during evolution. Proteins 2005; 60:27–35.** © **2005 Wiley-Liss, Inc.**

Key words: amino acids; hydrophobicity; GC content; aromaticity; orphans; multivariate analysis

INTRODUCTION

Living organisms are subjected to a variety of selection pressures that act not only at the level of the global phenotype but at each level of the cell's organization. It is usually assumed that proteins, for example, would be mainly subjected to selection pressures associated with their function. This is indeed the case, but if one compares proteins with identical function in organisms widely dispersed throughout evolution, one discovers that apart from a few 10's of amino acid residues, almost every residue of the protein can be replaced by any other amino acid, without dramatic change in the protein's function.[1,2] As a case in point in bacteria, proteins with similar functions will change in amino acid composition when their gene moves from the leading replicated strand to the lagging strand.[3] This not only indicates that proteins are extremely robust structures but also demonstrates that the chemical identity of each residue at most sites might integrate subtle cues derived from selection pressure operating at other places in the cell.[4] An amino acid in a protein integrates a preference for a certain DNA base composition, a preference for a codon, and a preference for that particular amino acid. This results from constraints at the level of DNA and RNA structure, as well as constraints at the level of nucleotide biosynthesis, or amino acid biosynthesis. Akashi and Gojobori have investigated the latter in a study suggesting that, indeed, the energy cost required to synthesize a given amino acid has some bearing on the overall composition of proteins.[5] Furthermore, Lobry has shown that differences in mutational bias can explain some variations in amino acid composition in bacterial species.[6] Starting from these studies that predated genome studies, we undertook a thorough analysis of the various factors that may affect amino acid composition in reference proteomes as a prerequisite to creating new ways to compare proteins from widely distant organisms, a notoriously difficult task. As a first step we analyzed the distribution of amino acids in proteins of model bacteria, where their function is best known, in contrast to that of the vast majority of proteins predicted from genome sequences.

An average protein of *Escherichia coli* K-12 is 300 residues long. If all things were kept equal, it should therefore have about 15 residues of each type of amino acid

Grant sponsor: Innovation and Technology Fund of the government of the SAR Hong Kong, China (program BIOSUPPORT), granted to A. Danchin for the creation of the HKU-Pasteur Research Centre. Grant sponsor: BioSapiens EU; Grant number: LSHG-CT-2003-503265. Grant sponsor: French Ministry of Foreign Affairs. Grant sponsor: French Centre National de la Recherche Scientifique; Grant numbers: CNRS-UMR 8030 and URA 2171. Grant sponsor: Institut Pasteur (Paris, France).

*Correspondence to: Géraldine Pascal, Genoscope/UMR 8030, Atelier de Genomique Comparative, 2 rue Gaston Cremieux, 91006 Evry Cedex, France, or to Antoine Danchin, Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France. E-mail: gpascal@genoscope.cns.fr or adanchin@pasteur.fr

Received 4 August 2004; Accepted 14 January 2005

Published online 22 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20475

(of course, this also depends on the number of available codons, the metabolic cost of amino acid synthesis, chemical reactivity, etc.). Furthermore, if the amino acid content is normalized to the length of the protein and the number of proteins with a given density of a given residue is plotted, one would expect, all things being equal, to witness a normal distribution, with the exception of a small number of proteins with no residues of a given amino acid (the numbers considered are finite, and small). Whereas this is true for some amino acids (A, G, L, P, R, V), this is not true for others. In particular, the behavior of glutamate and aspartate is surprising: One observes a clear biphasic distribution, suggesting that proteins of *E. coli* K-12 form at least 2 classes. In addition to the presence of shoulders in some distributions, others are hyper-Gaussian (H, N, S, T, Y) (data not shown). This prompted us to study the amino acid composition in proteins using multivariate analysis methods.

Correspondence analysis (CA) has long been used to analyze the codon usage and amino acids composition bias of sequences of many organisms, in particular models of procaryotes.[7,8] As early as 1994, Lobry and Gautier explored the *E. coli* K-12 available data using this method, well before the complete sequence and other essential information on this organism was available.[9] Further studies substantiated that during replication, complementary DNA strands were subjected to different mutation constraints, which might be reflected at the level of amino acid composition in proteins.[10] More recently, global studies used this approach to extend investigation to a number of organisms without entering the details of the proteomes.[11–13] Taken together, these studies suggested that exploration of the detailed knowledge of genomes might allow significant integration of the statistical data with functional information about genes and genomes. This prompted us to analyze the *E. coli* K-12, *Bacillus subtilis*, and *Methanococcus jannaschii* proteomes to establish whether rules of amino acid composition might exist while placing them in a functional genomics perspective. In the present work, we correlate explicitly, using knowledge derived from genome programs, the rules that can be extracted from the statistical analysis with the function of gene and gene products as they are integrated in a coherent view of the cell. Rather than draw conclusions from uncertain annotation data, we have chosen to restrict our analysis to these 3 proteomes, because they represent well-known bacterial models of Gram-negative organisms, Gram-positive organisms, and the best known archaebacterium. Moreover, these 3 organisms are very distant in the phylogenetic tree; *E. coli* K-12 and *B. subtilis* are separated by more than 1.5 billion years. These 3 proteomes have been studied extensively for many years; consequently, the annotations in data banks are the most complete and the most correct. In choosing these 3 proteomes, we are avoiding the pitfall of basing our analysis on subcellular functions and localizations that are erroneous due to the increasingly rapid sequencing of genomes and the lack of experimental verification of information from these sequences.[14]

MATERIALS AND METHODS

Correspondence Analysis, Statistics, and Data Clustering

We used correspondence analysis[15] to identify the major factors that shape variation in amino acid usage among proteins of the organism of interest. These analyses were based on absolute frequencies in order to avoid introducing other biases.[16] 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 according to the information they represent. They are presented in a decreasing order of importance as quantified by their corresponding "inertia."[17] Proteins and amino acids can be represented jointly in the obtained factorial space. Sequences that have a similar amino acid composition appear as neighbors.

As an additional tool, the CODONW software (http://www.molbiol.ox.ac.uk/cu/) was used for each of the 3 complete sequences, to help interpret the results. It computes the hydrophobicity levels (GRAVY score),[18] the G+C content of genes for specific proteins, and the aromaticity level (the relative frequency of aromatic amino acids) of each sequence, which are correlated with position on the main discriminating axis.

Data Sets

The complete proteome of *E. coli* K-12 is from the EcoGene17 database (http://bmb.med.miami.edu/ecogene/ecoweb/), which is known to be the most complete and reliable database for this organism. The data for *B. subtilis* and *M. jannaschii* used in this study are from the latest versions of the complete genomes in EMBL format, release 76 of 7 July 2003 and release 72 of 18 July 2002 respectively. They are available on the FTP site of the International Nucleotide Sequence Database (GenBank/EMBL/DDBJ) at 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[19]). In order to reduce influence of stochastic variations that may occur in small proteins, only proteins longer than 100 residues (after truncating) were retained. After formatting, 3652 proteins from *E. coli* K-12, 3465 proteins from *B. subtilis*, and 1460 proteins from *M. jannaschii* were analyzed.

RESULTS

In this study, we used CA to explore the nature of the links that exist between the proteins forming the proteome of prokaryotes and the amino acid residues they contain, starting with the suggested biases created by hydrophobicity and by the local G+C content, as described in the literature.[9,11,12] New relations between amino acid composition and features of proteins were uncovered, such as a

bias created by the aromaticity of proteins. The proteomes of *E. coli* K12, *B. subtilis*, and *M. jannaschii* have been analyzed in order to identify the common rules that drive amino acid composition of Gram-negative and Gram-positive prokaryotes and Archaea, and the differences that are typical of their metabolism or their structures. The data presented are analyzed using the best annotated models. Their generality has been substantiated by a similar analysis of the proteomes of a variety of other organisms (data not shown).

Inertia of CA

Inertia of CA can serve as a guide in determining the relative importance of a given factor. Regarding the *E. coli* K-12 data, about 41% of the inertia was distributed into the 3 first factors, and 43% for the *B. subtilis* data and for the *M. jannaschii* data. To measure their importance, these figures should be compared with an expected average inertia per axis of approximately 5% (100% of information would share equivalently in 19 axes). For this set of organisms, more than 75% of the information was present in the first 10 factors. CA was used to summarize and simplify the data. Analysis of the information carried in the CA was limited in the present study to the first 3 axes, which represent the most significant part of the whole information.

Hydrophobicity, a Discriminant Factor of Proteins

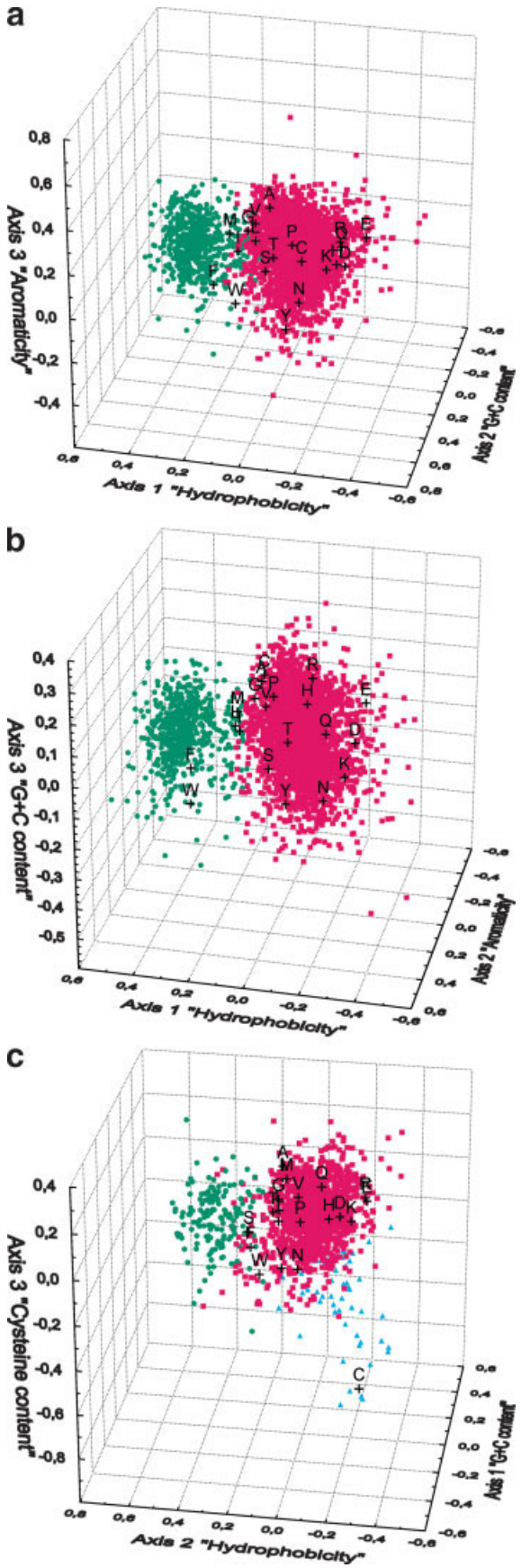
The distribution of proteins on the factorial plane made of the 2 first axes displayed 2 well-separated groups of proteins that stood out prominently (groups A and B), shown as individual clusters in all 3 model prokaryote proteomes. In *M. jannaschii*, a third small group was further revealed [Fig. 1(a–c)]. The contraposition of charged residues (E, D, and K) and of large hydrophobic amino acids (F and L) determined the clear separation of group A from group B. This discriminating factor strongly correlates with GRAVY score (*M. jannaschii*, $r = 0.87, p < 10^{-4}$; *E. coli* K-12, $r = 0.95$, $p < 10^{-4}$; *B. subtilis*, $r = 0.96$, $p < 10^{-4}$). Following the approach of Lobry and Gautier,[9] we identified the subcellular location of each known protein of *E. coli* K-12 and *B. subtilis* group A in Swiss-Prot databank (http://www.expasy.org) and GenProtEC database (http://genprotec.mbl.edu/). Thus, all group A proteins have an integral inner membrane location. Moreover, these integral inner membrane proteins (IIMPs) possess a transmembrane portion that makes at least 30% of the total length of the protein. Remarkably, all outer membrane proteins were found in group B. This indicated that proteins in group A are selected out of a very stringent property of their amino acid composition. After pooling the 3 proteomes together (Fig. 2), we could observe the strong conservation of group A in spite of the considerable difference in the envelope structure among the 3 organisms. These results substantiate and extend the previous observation of an hydrophobicity bias in the bacterial proteome[9,11,12] by providing an unambiguous link between the subcellular location of the proteins and their amino acid composition.

G+C Content Bias

As noticed in previous works, in addition to the bias driven by the hydrophobicity of the proteins, the overall G+C content of genomes of organisms appears to drive a second bias in the amino acid composition of the proteome. This feature is common to the 3 model prokaryotes studied in the present work. Contrary to expectation, however, this bias is not due to a bias in the second position of the codons, which is driving most of the physicochemical nature of the amino acids for which they code. Remarkably, the effect of the G+C content was apparently correlated to the first codon position (*M. jannaschii*, $r = 0.86$, *E. coli* K-12, $r = -0.59$, *B. subtilis*, $r = 0.71$; $p < 10^{-4}$ for each value). This bias seems to be the major bias for *M. jannaschii* (axis 1 of CA) and appears as the second and the third one for, respectively, *E. coli* K-12 and *B. subtilis* (axes 2 and 3). For *E. coli* K-12, this CA factor is driven by the opposition between by K and N (A+T rich codons) on one side and by L (neither A+T rich nor G+C rich codons) and R (G+C rich codon) on the other side. This probably accounts for the weaker G+C content correlation on that axis. Moreover, we find a correlation between discriminating CA axes and A and C at the first nucleotide position of codons (respectively for A and C: *E. coli*: $r = 0.75, p < 10^{-4}$, $r = -0.81, p < 10^{-4}$; *B. subtilis*: $r = -0.59, p < 10^{-4}$, $r = 0.57, p < 10^{-4}$).

Aromaticity of Proteins

As we go from one axis to the next one, the weight of characters retained through evolution is progressively less prominent and influenced by the specific nature of the organism. Indeed, the biases reflected in the third axis differ in the 3 model organisms, splitting into 2 types of behavior: (1) the model Bacteria present a bias based on the aromaticity of proteins, found in axis 2 for *B. subtilis* and in axis 3 for *E. coli* K-12, and (2) the model Archaeon forms a third group of proteins, which is very clearly correlated with the percentage of cysteine in proteins along CA axis 3 ($r = -0.91$, $p < 10^{-4}$) [Fig. 1(c)]. A bias driven by the aromaticity of the proteins is contained in a further axis, with a rather weak contribution to inertia. This third group of *M. jannaschii* proteins was found to be constituted of proteins rich in cysteine residues (< 5% of proteome). To put aside this extreme bias, cysteine was removed from the CA. This resulted in the construction of a cloud of points where groups identical to the A and B groups found in Bacteria, were obtained (Figure 3). Interestingly however, the third CA axis of this new plot, driven on one side by residues F, L and on the other side by residues N, T, did not correlate significantly with any parameters investigated and will deserve further investigation, using other proteomes of Archaea. The fourth axis, having an inertia below 7%, was correlated with the aromaticity of proteins ($r = -0.48$, $p < 10^{-4}$). The low dispersion observed in these and subsequent axes did not warrant further consideration. The bacteria *B. subtilis* and *E. coli* K-12 have a much more pronounced correlation in favor of aromaticity of proteins due to opposition between residues A and G versus Y, W, and F (*B. subtilis*:



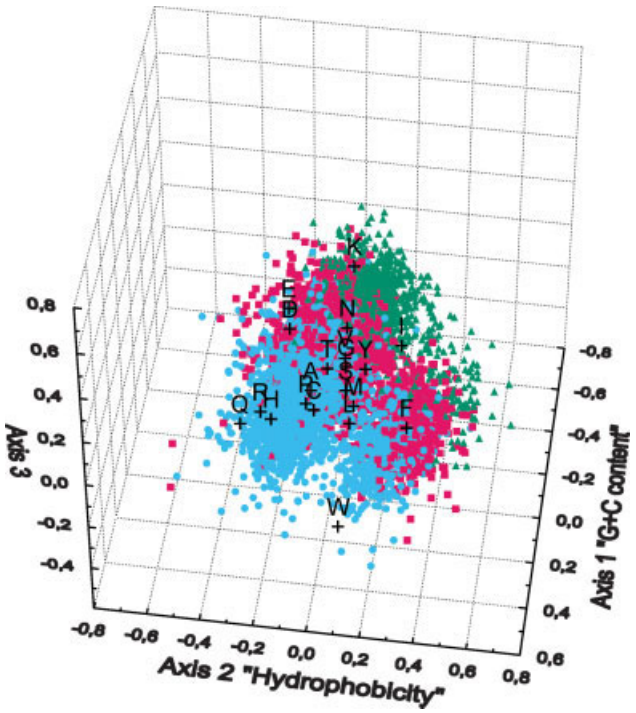


Fig. 2. The 3 first axes of CA of the 3 pooled proteomes: *E. coli* K-12 in blue circles; *B. subtilis* in pink squares; and *M. jannaschii* in green triangles.

$r = -0.68, p < 10^{-4}$; *E. coli* K-12: $r = -0.72, p < 10^{-4}$). In order to analyze the correlation between the bias in aromatic amino acids and protein function, the proteins situated at the extremities of the axis that discriminated aromaticity of proteins ($2 \times 10\%$ of the total number of proteins) were extracted and compared with each other. We distinguished 2 classes of proteins, termed "high aromatic" and "low aromatic" classes. A large number of proteins of unknown function constituted the group of proteins rich in aromatic amino acids, whereas a number of proteins with housekeeping functions constituted the group of proteins in which aromatic residues are scarce. Especially noteworthy was the presence of ribosomal proteins in the latter group. Two classes of proteins with unknown function exist. They are either shared between a variety of related or less related genomes, or completely original to the genome of interest. In each of the extreme protein groups, the number of orphan proteins (proteins with no resemblance to any other protein in the databases) was studied by sequence alignment comparisons using BLASTP.[20] An orphan protein was defined as a protein with a sequence displaying a similarity score with other

Fig. 1. Distribution of the protein sequences on the CA factorial space determined by the 3 first factors. Green circles represent proteins in group A; pink squares represent proteins in group B; and cyan triangles represent proteins in the third group of *M. jannaschii*. Amino acids are represented by black crosses. (**a**) For *E. coli* K-12, group A represents 18.8% of the analyzed proteome. (**b**) For *B. subtilis*, group A represents 19.8% of the analyzed proteome. (**c**) For *M. jannaschii*, group A represent 10.6% of the analyzed proteome.

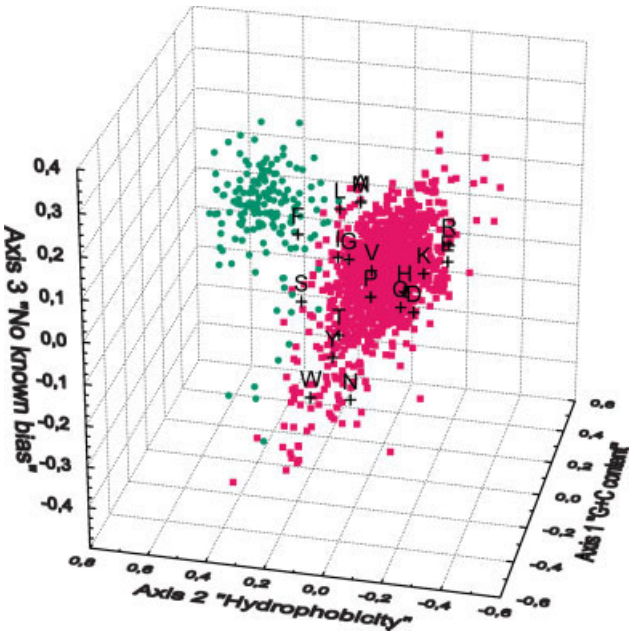


Fig. 3. The 3 first axes of CA of *M. jannaschii* when cysteine is omitted. Green circles represent proteins in group A (12% of analyzed proteome). Pink squares represent proteins in group B.

proteins lower than an *e*-value of less than 1^{e-3} in Swiss-Prot or in TrEMBL. In the case of *B. subtilis* and *E. coli*, this identifies 12.3% and 6.3% of "high aromatic" proteins as orphans, respectively, contrasting with, respectively, 2.02% and 0.55% of orphans in the "low aromatic" proteins. Using the transcriptome data of *B. subtilis* available from our previous work,[21] we checked whether orphan protein genes were expressed. Except for 4 proteins corresponding to genes that had not been included in the transcriptome data set, all of the orphan proteins of *B. subtilis* were found to be expressed (data not shown). These 2 classes were compared using a test of comparison of proportions based on observed proportions. This test revealed that "high aromatic" proteins were enriched in orphan proteins. It was also noted that, for *B. subtilis* and *E. coli* K-12 respectively, 38% and 44% of the orphan proteins were found in the 10% of "high aromatic" proteins retained for the test. A Wilcoxon test allowed us to show that the set of orphan proteins of *B. subtilis* and *E. coli* K-12 was richer in aromatic amino acids than the ensemble of other proteins (non-orphans) (*E. coli* K-12: $p < 5 \cdot 10^{-5}$; *B. subtilis*: $p < 2 \cdot 10^{-9}$).

DISCUSSION

CA With a Very Significant Inertia

In this study, meant to substantiate and extend previous work to correlate amino acid composition of model proteomes to structural and functional properties of their proteins, we restricted most of our analysis to the information given by the 3 first axes of CA. Furthermore, in order to refine exploration, when a property was clearly driven by a single amino acid, we extended the study by performing CA with the set of amino acids lacking the biasing one. Fortunately, the main characteristics of proteomes appeared factor by factor, unfolding in a consistent manner. A first factor linked protein function and hydrophobicity. A second one emphasized the proteomic relationship between the G+C content of the genome and the related genes. And a third factor shifted focus to the aromaticity of proteins in association to the origins of their functions.

Hydrophobicity Versus Electric Charge of Proteins Characteristic of Their Subcellular Location

Previous works either classifying proteomes as bulk entities[13] or analyzing partial or global proteomes[9,22] suggested that the proteome was subject to a bias driven by hydrophobicity. We show here that CA of the amino acid content of the proteome discriminated prokaryotic proteins according to hydropathy in a way that was consistent with the subcellular localization of proteins. Proteins that were known to be integral components of the cytoplasmic membrane all clustered into a uniquely defined group. Remarkably, this group was consistently observed in the 3 model organisms we analyzed despite their enormous phylogenetic divergence. It is known that the envelopes of Gram-positive and Gram-negative bacteria differ mainly in their structure. Gram-positive bacteria usually have a thick peptidoglycan multilayer on their surface, at the exterior of the cytoplasmic membrane, whereas Gram-negative bacteria have a more complex structure. The latter have 2 membranes, of which 1 is at the surface of the peptidoglycan layer, with the following order (from interior to exterior): plasma (inner) membrane, peptidoglycan, and external (outer) membrane. The membrane structure is contradictory; it plays the role of a filter and of a transporter at the same time. It is semipermeable, letting small molecules such as sugars or salts enter, but prevents the entry of large molecules such as proteins. Membranes of Archaea differ from those of Bacteria because they lack peptidoglycan as a constituent of their membranes. Furthermore, their plasma membrane presents a lipid composition which is different from that of all other organisms.[23] Despite these considerable structural differences, we observed that proteins imbedded in the plasma membrane of the 3 model prokaryotes were characterized by proteins that were quite different from the other proteins of the organism in their amino acid composition.[24] This distinction is seen in the composition of groups A and B of the CA, and it is so clear cut that CA followed by dynamic clustering could be suggested as a straightforward tool for the identification of IIMP. As substantiation we analyzed the behavior of 83 proteomes from a variety of genomes and found the exact same splitting of the proteome into at least 2 groups, A and B (data not shown). In group A, the scattering observed in the amino acid composition space shows a very clear separation: hydrophobic residues versus charged residues. This observation substantiated other studies describing the structure and amino acid composition of membrane proteins[25]: Hydrophobic residues (F, L, and I) show a clear preference for the transbilayer region, whereas charged residues (E, K, D, and R) are located on the extracellular side of the membrane (basic residues

preferably inside because of the negative electric potential on the inside of the cell).

Generally, the solubility in water of the lateral chains of the amino acids increases with the electric charge that they possess, and their capacity to establish hydrogen bonds. The hydrophobic transmembrane segments of the polypeptide chains of the IIMPs are essentially composed of amino acid residues with non-polar lateral chains; the peptide bonds between these amino acids are, however, polar themselves, and when they are immersed in the lipid bilayer (away from water), they tend to establish hydrogen bonds between each other, which causes refolding of the segment as an α-helix (20–30 amino acid residues). α-helix IIMPs make 4 main classes of functions: energy transduction, ion channeling without adenosine triphosphate (ATP) consumption, water diffusion (aquaporin), and active transport of molecules (requiring ATP). This universality in amino acid composition of α-helical IIMPs could be explained by the huge selective constraints imposed by the essential role played by these proteins in the cell, by their lipidic environment, and by the large electrostatic potential (usually 100,000 volts/cm) to which they are submitted.[26] In contrast to the first group, group B is not a homogeneous class, and it is sometimes split into other classes. It contains also some membrane-associated proteins, as the outer membrane proteins in Gram-negative bacteria and some proteins at least partially imbedded in the inner membrane, most of which are sensor and receptor proteins. These latter proteins possess either a large cytoplasmic or external domain, or both, and this accounts for their being excluded from the narrow class of integral inner membrane proteins. The proteins with large transmembrane domains (> 30% of the total length of the protein), which are in group B and not in group A, were either integral outer membrane proteins (OmpF, OmpG, OmpX, OmpT, LamB) with β-barrel structures, or unknown proteins (YcdP: probably a permease component for hemin transport, defective in *E. coli* K12; YibH: could belong to an efflux pump). They differ from IIMPs by their code, such as the non-random frequency of the tripeptide motif: aromatic–random–aromatic of porins with β-barrel structures.[27] In summary, the present study brings about a novel feature of a particular consistent class of IIMPs that have key cellular roles, while they can be distinguished from all others membrane protein types. This particular feature will be useful for functional annotation of new genomes, by restricting the domain of hypotheses about putative functions of unknown proteins to a narrower set.

Genome G+C Content Biasing Amino Acid Proteome Composition

The genome's G+C content is known to bias the amino acid proteome composition[28] (i.e., a significant part of the choice of preferred amino acids is not determined by the selection pressure on the proteins, but rather on the genome). Indeed, we observed a corresponding correlation on the discriminant axes for each of the 3 organisms, although the selection pressure, which would have been expected to be mostly driven by the second codon position (the most discriminating one), was found to come from the first position instead. Previous work has documented that, in the majority of thermophiles, the genome's G+C% is the first factor that influences the amino acid composition, as reported, for example, by Lynn et al. in 2002 based on an analysis of usage correspondences of synonymous codons.[29] This may explain why the codon usage bias driven by the genome's G+C% determining a specific amino acid composition of proteins in *M. jannaschii* is greater than the variability due to selection pressure on the subcellular localization of proteins.

The thermophilic biotope of the organism is not the sole cause of this bias, as seen with axis 2 of the CA of *E. coli* K-12 proteome and axis 3 of the CA of *B. subtilis* that split their proteome according to the G+C content of the codons. In previous studies based on incomplete proteomes, Palacios and Wernegreen[7] and Lobry and Gautier[9] identified a somewhat similar axis 2 (conjugated to axis 3 in Palacios and Wernegreen) in the *E. coli* K-12 and interpreted it as the result of the selection pressure acting on the proteins' gene expression. The present results do not substantiate this interpretation. Indeed, the bias due to adenine and cytosine at the first nucleotide position of codons of genes, that would give proteins rich in L, P, H, and Q and poor in I, M, T, N, K, and S amino acids or, inversely, does not appear to have a straightforward interpretation. In any event, this bias does not involve the level of gene expression, as demonstrated by the weak correlation of the Codon Adaptation Index (CAI) with discriminant axes ($r = 0.22$, $p < 10^{-4}$ for axis 2 of *E. coli* K-12 and $r = 0.16$, $p < 10^{-4}$ for axis 3 of *B. subtilis*). The CAI score measures empirically the synonymous codon usage bias. It is positively correlated with the expressivity of a gene.[30]

What could be the impact of a bias of the first nucleotide in the codon for an amino acid in a proteome? Enrichment in A+T at the first position of codons indicates a specific selection pressure that is not simply driven by the hydrophilicity/hydrophobicity controlled by the second codon position. It separates between V, A, D, E, G, P, H, and Q, and F, S, Y, C, W, I, M, T, N, and K, with L and R not discriminating. This partially matches a shift from amino acids with a small volume to those with a large volume, from those with no sulfur to those containing sulfur, from non-aromatic to aromatic and to the acquisition of serine.[31] This trend is at a cost: The richer an organism is in A+T at the first position of codons, the higher the metabolic cost of amino acid production. However, this can also be an advantage. Indeed, it is well established that pathogens, symbionts, as well as phage and plasmid sequences and insertion sequences, are richer in A+T.[32] And, although the high G+C content of genomes is associated with low carbon content in proteins, pathogens, and bacterial symbionts, even those that are A+T-rich, are less subject to limitations of carbon in the environment, because they are linked to nutritional resources furnished by their host[33] and thus stay competitive in terms of metabolic resources. A+T enrichment may also derive from other selective constraints: Hyperthermophiles such as *M. jannaschii* are

often A+T-rich despite the greater heat stability of the GC pair. This bias makes these organisms more resistant to cytosine deamination (C → U), which is activated at high temperature,[34] and they are thus more likely to retain their original sequence. It has also been demonstrated in numerous studies that enrichment in A+T is evidence of recent evolution. Based on results of a study by Sorimachi,[35] serine, considered to be one of the important residues in biological evolution, exhibits increased concentration as evolution proceeds in all organisms studied. It should also be noted that serine contributes to the formation of new protein functions during evolution.[35] This argument is further supported by the work of McDonald et al.,[36] which compares several proteins of mesophiles and thermophiles of *Methanococcus* and *Bacillus* species. This study demonstrates that, in the case of *Methanococcus*, enrichment in A+T is not necessarily due to changes in the temperature of the environment, but rather to the evolution of species.

In summary, the present study uncovers a remarkable impact of the genome G+C content on the corresponding proteome that does not display any significant association either with the optimal growth temperature of the organism or its gene expressivity.

Aromaticity, Source of Novel Proteins

The two first factors explaining the amino acid composition of the prokaryote proteomes were universal. In contrast, the third discriminant factor, at first sight, was not: The Archaeon *M. jannaschii* exhibits a different behavior compared to the other two model Bacteria. Indeed, the third axis of CA of the *M. jannaschii* proteome presented a strong bias driven by the cysteine composition of the proteins. It has been reported that CXXC clusters are more frequent at high temperature, which would allow thermophiles like *M. jannaschii* to somehow protect these residues, as isolated cysteines would otherwise have a tendency to decrease with increasing temperature.[37] It is therefore possible that enrichment in cysteines is due to formation of clusters (perhaps stabilized by metals or formation of sulfur bonds). This would perhaps explain the special bias created by the strong presence of cysteine in a certain number of *M. jannaschii* proteins. Because of this effect, which is particular to this organism, we explored whether CA would once again lead to some kind of universal constraint on the proteomes, omitting cysteine from the analysis. When the cysteine bias was suppressed, we observed that the third axis of the *M. jannaschii* proteome was formed by the opposition of the L and F amino acids to the N and T amino acids. Interestingly, these oppositions suggested a separation between proteins integrated into the membrane (biased in L and F) and excreted proteins. Indeed, Perrière and Thioulouse surmised that in Gram-negative bacteria, the periplasmic proteins were characterized by N, P, Q, and T residues,[16] which are known to slow the folding of proteins, a phenomenon required for proteins that have to be exported. Further exploration is needed to see whether this observation could be used to help predicting excreted proteins in *M. jannaschii*. The next CA axis in *M. jannaschii* correlated with the aromaticity of proteins ($r = 0.48, p < 10^{-4}$), suggesting a particular role of aromatic amino acids. This role was much stronger in *E. coli* K-12 and in *B. subtilis*, where an amino acid composition bias opposed aromatic amino acids to A and G residues allowing separation of housekeeping proteins from orphan proteins. It is worth remarking that aromatic amino acids' defective proteins do not constitute a functionally random class of proteins, since ribosomal proteins make the bulk of this class. Ribosomes must belong to the very first organelles that made the first cells, and it is most likely that ribosomal proteins were present in the ancestral proteomes, providing them with a long time for evolution. In this context, it is interesting that ribosomal proteins from Archaea differ significantly in amino acid content from those of Bacteria, because some are common with those of Eukaryotes while they are absent from Bacteria.[38] This might account for the observation that the separation between proteins rich in aromatic amino acids and ribosomal proteins did not stand out prominently in *M. jannaschii*. Ribosomal proteins are essential to the cell's life, and because protein synthesis is at the core of macromolecule metabolism, they must be expressed at a high level. These proteins are enriched in basic amino acids (K, R: they interact with RNA) and in small hydrophobic residues (A, V, G). The latter have a low metabolic cost. This goes in the direction of the selection pressure for high expression of proteins.[39] Furthermore residues A and G are often considered as belonging to the first amino acids present at the origin of the first cells, and this is consistent with the presumably primitive nature of ribosomal proteins. This contrasts with enrichment in aromatic amino acids, which are considered newcomers that progressively invaded the genetic code.[40] Because only a few codons code for aromatic amino acids, and because of their large metabolic cost, they are usually rare in proteins, especially in highly expressed proteins.[5,7] They have, however, extremely interesting physicochemical properties that are witnessed frequently in the 2D and 3D structure of proteins.[9,41,42] Indeed, aromatic residues often interact with one another, in particular in α-helices,[24] usually in an orthogonal interaction (edge-to-face).[43] They have been recruited for a variety of roles in the interaction between protein subunits and with nucleic acids. In particular, it has been observed that transmembrane α-helices are richer in aromatic amino acid residues than cytoplasmic α-helices,[24] strongly suggesting that aromatic amino acids play a particularly important stabilizing role in an hydrophobic context. These residues would play a major role in the structure and function of membrane proteins by allowing interaction between the exposed faces of the α-helices. They would also allow stable anchoring of proteins to the membrane.[24] Membranes must be flexible and must adjust rapidly to variations in the environment, and aromatic amino acids might provide them with the necessary flexibility while conserving stable interactions. Remarkably, we observed not only that small hydrophobic residues were opposed to aromatic amino acids in the overall pattern of protein composition organi-

zation, but also that ancestral housekeeping proteins were opposed to orphan proteins. The latter can either be considered as old, but evolving extremely fast, or recent acquisitions or creations. The opposition between A, G, V, and Y, F and W, as well as the amino acid composition of housekeeping proteins, would strongly argue in favor of the latter hypothesis with orphan proteins as recent acquisitions. This is indeed the general consensus about their origin (some think that they might also be coded by pseudogenes, but pseudogenes, deriving from ancestral genes, should not differ in amino acid composition from the bulk of the genes).[44] Orphan proteins are usually small (~150 residues) and their genes are AT-rich in the third codon position.[45] They are often present in bacteriophages. Furthermore, through the processes of recombination/excision, they might generate sequences from bits and pieces that would not have counterparts anywhere else. Genomes, like all living processes, are subject to a process of selective stabilization, and those proteins, when recruited for a function increasing somehow the fitness of their host (despite their metabolic cost), would stay there. The very fact that they are "orphans" creates a discrimination between the self and the non-self of the species. One is therefore compelled to uncover a function that would have some mark of "self" for an organism. Proteins are certainly not isolated in the cytoplasm of cells: They must both interact with a precise set of subunits or other factors and avoid interacting with unrelated proteins. We propose that many of those orphan proteins could be factors promoting this stabilization process (i.e., being non-catalytic subunits of complexes). The versatility of aromatic amino acids residues that are present in these "gluons" would provide them with a quick adaptation process that would allow them to be recruited frequently, thus accounting for the apparent ubiquity of orphans in genomes. Once recruited, they would slowly evolve to less costly material by losing their amino acid residues as time elapses, thereby forming classes of well-adapted proteins that would now be transmitted by horizontal transfer from organism to organism, while losing their status of orphan would be helped by intervention of phages, which constitute the major vector of introduction of proteins in genomes.[32,46]

CONCLUSIONS

Despite their essential role in catalysis and protein folding, most amino acid residues in proteins are not subject to such dramatic selection pressure that would link them to the function of the proteins they encode. As a matter of fact, usually less than 10% of the residues are submitted to strict functional constraints. Analyzing the 3 model proteomes of Gram-positive and Gram-negative Bacteria, as well as the model for Archaea, we uncovered two universal biases that affect proteome amino acid composition and a third one that is Bacteria-specific. These results were also observed on a large set of 80 proteomes (data not shown). Proteins integrated in the inner membrane can easily be identified from the bulk; a particular constraint, of still unknown nature, drives the amino acid content of the proteome, as a function of the G+C content of the first position of their codon; and, finally, orphan proteins are unusually rich in aromatic amino acid residues. Preliminary observations suggest that these rules are ubiquitous. Deeper analysis will help us understand the nature of the selection pressure driving these universal biases.

ACKNOWLEDGMENTS

We thank Stephane Cruveiller, Eduardo Rocha, and Cédric Cabau for their critical comments and suggestions, and Susan Cure for her help in writing the manuscript.

This work was initiated as a core genomics program at the HKU-Pasteur Research Centre in Hong Kong.

REFERENCES

1. Beyer A. Sequence analysis of the AAA protein family. Protein Sci 1997;6:2043–2058.
2. Ma B, Wolfson HJ, Nussinov R. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. Curr Opin Struct Biol 2001;11:364–369.
3. Rocha EP, Danchin A. Ongoing evolution of strand composition in bacterial genomes. Mol Biol Evol 2001;18:1789–1799.
4. Dean AM. Selection and neutrality in lactose operons of *Escherichia coli*. Genetics 1989;123:441–454.
5. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 2002;99:3695–3700.
6. Lobry JR. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 1997;205:309–316.
7. Palacios C, Wernegreen JJ. A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high-expression genes. Mol Biol Evol 2002;19:1575–1584.
8. Guerdoux-Jamet P, Henaut A, Nitschke P, Risler JL, Danchin A. Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? DNA Res 1997;4:257–265.
9. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 1994;22:3174–3180.
10. Lobry JR, Sueoka N. Asymmetric directional mutation pressures in bacteria. Genome Biol 2002;3(10).
11. Dumontier M, Michalickova K, Hogue CW. Species-specific protein sequence and fold optimizations. BMC Bioinformatics 2002;3.
12. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29:1608–1615.
13. Tekaia F, Yeramian E, Dujon B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 2002;297:51–60.
14. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 2002;18:1641–1649.
15. Benzecri J-P. L'analyse des données, L'Analyse des Correspondances. Paris, France: Dunod Edition; 1973.
16. Perrière G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 2002;30:4548–4555.
17. Lebart T, Morineau A, Warwick KA. Multivariate descriptive statistical analysis. New York: Wiley; 1984.
18. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
19. Rocha EP, Danchin A, Viari A. Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res 1999;27:3567–3576.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
21. Sekowska A, Robin S, Daudin JJ, Henaut A, Danchin A. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. Genome Biol 2001;2(6).

22. Rispe C, Delmotte F, van Ham RC, Moya A. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res 2004;14:44–53.
23. Cavalier-Smith T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 2002;52:7–76.
24. Koshi JM, Bruno WJ. Major structural determinants of transmembrane proteins identified by principal component analysis. Proteins 1999;34:333–340.
25. Ulmschneider MB, Sansom MS. Amino acid distributions in integral membrane protein structures. Biochim Biophys Acta 2001;1512:1–14.
26. Nilsson I, Johnson AE, von Heijne G. How hydrophobic is alanine? J Biol Chem 2003;278:29389–29393.
27. Wimley WC. Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci 2002;11:301–312.
28. Sandberg R, Branden CI, Ernberg I, Coster J. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene 2003;311:35–42.
29. Lynn DJ, Singer GA, Hickey DA. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 2002;30:4272–4277.
30. Sharp PM, Li WH. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15:1281–1295.
31. Wilquet V, Van de Casteele M. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res Microbiol 1999;150:21–32.
32. Rocha EP, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet 2002;18:291–294.
33. Baudouin-Cornu P, Schuerer K, Marliere P, Thomas D. Intimate evolution of proteins: proteome atomic content correlates with genome base composition. J Biol Chem 2004;279:5421–5428.
34. Frederico LA, Kunkel TA, Shaw BR. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 1990;29:2532–2537.
35. Sorimachi K. Evolutionary changes reflected by the cellular amino acid composition. Amino Acids 1999;17:207–226.
36. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol Biol Evol 1999;16:1785–1790.
37. Rosato V, Pucello N, Giuliano G. Evidence for cysteine clustering in thermophilic proteomes. Trends Genet 2002;18:278–281.
38. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 2002;30:5382–5390.
39. Lin K, Kuang Y, Joseph JS, Kolatkar PR. 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 2002;30:2599–2607.
40. Brooks DJ, Fresco JR. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol Cell Proteomics 2002;1:125–131.
41. Thomas A, Meurisse R, Brasseur R. Aromatic side-chain interactions in proteins: II. Near- and far-sequence Phe-X pairs. Proteins 2002;48:635–644.
42. Thomas A, Meurisse R, Charloteaux B, Brasseur R. Aromatic side-chain interactions in proteins: I. Main structural features. Proteins 2002;48:628–634.
43. Hunter CA, Singh J, Thornton JM. Pi–pi interactions: the geometry and energetics of phenylalanine–phenylalanine interactions in proteins. J Mol Biol 1991;218:837–846.
44. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res 2003;13:2213–2219.
45. Daubin V, Ochman H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 2004;14:1036–1042.
46. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF. Origins of highly mosaic mycobacteriophage genomes. Cell 2003;113:171–182.