

# Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis

Fredj Tekai<sup>a,\*</sup>, Edouard Yeramian<sup>b</sup>, Bernard Dujon<sup>a</sup>

<sup>a</sup>Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR927, Univ. P. M. Curie, Paris), Institut Pasteur, 25, Rue du Dr Roux, F-75724 Paris Cedex 15, France

<sup>b</sup>Centre de Bioinformatique, Génopole, Institut Pasteur, 25, Rue du Dr Roux, F-75724 Paris Cedex 15, France

Received 15 April 2002; received in revised form 18 July 2002; accepted 29 July 2002

Received by S. Salzberg

## Abstract

Can we infer the lifestyle of an organism from the characteristic properties of its genome? More precisely, what are the relations between easily quantifiable properties from genomic sequences, such as amino-acid compositions, and more subtle characteristics concerning for example lifestyles or evolutionary trends? Here, we seek a global picture for such properties, based on a large number (56) of complete genomes, including significant numbers of representatives from the three domains of life. We consider the amino acid compositions of the predicted proteomes, and we use correspondence analysis, as a multivariate method to extract the relevant information from the large-scale data. From these analyses we derive a series of conclusions, concerning lifestyles, as well as physico-chemical and evolutionary trends: (1) correspondence analysis of the amino acid compositions permits discrimination between the three known lifestyles (mesophily/thermophily/hyperthermophily). (2) For various organisms, amino-acid composition properties are essentially driven by GC content, and to a significantly lesser extent by growth temperatures associated with lifestyles. Roughly speaking, the respective contributions of these two components are 57 and 20%. It is notable that these proportions are essentially unchanged with respect to a previous analysis (Nature 393 (1998) 537), which involved only 15 genomes, available at the time. (3) In terms of amino acid compositional biases, two specific ‘signatures’ for thermophily (in a broad sense, including hyperthermophily) can be detected. First, thermophilic species display a relative abundance in glutamic acid (Glu), concomitantly with the depletion in glutamine. Second, in thermophilic species, the relative abundance in Glu (negative charge) is significantly correlated (Pearson correlation coefficient  $r = 0.83$  with  $P < 0.0001$ ), with the increase in the lumped ‘pool’ lysine + arginine (positive charges). This correlation (absent in mesophiles) could be interpreted on a physico-chemical basis, relevant to the thermostability of proteins. (4) Statistically significant differences are observed between the average lengths of the genes in the surveyed species, which follow their distribution between the three domains of life. Also a significant difference is observed between the average lengths of thermophilic ( $283.0 \pm 5.8$ ) versus mesophilic ( $340 \pm 9.4$ ) genes. It is thus possible that the ‘general’ shortening of the primary sequences in thermophilic proteins plays a role in thermostability. (5) Considering various combinations of conservation properties (genes conserved exclusively in eukaryotes, in archaea, in bacteria, in combinations of two domains, etc.) correspondence analysis reveals a trend towards thermophilic-hyperthermophilic profiles for the most conserved subset of genes (ancient genes). (6) When limited to the subset of species-specific genes, correspondence analysis leads to a different picture for the clustering of genomes following amino-acid compositions: for example, the ‘core’ specific part of a genome can bear lifestyle signatures different from those of the complete genome.

Various results are discussed both on methodological and biological grounds. The evolutionary perspectives opened by our analyses are noted. © 2002 Published by Elsevier Science B.V.

**Keywords:** Hyperthermophiles; Mesophiles; Thermostability; Amino acid composition; Evolution; Multivariate analyses

## 1. Introduction

One major aim of large-scale genomic projects is to reach a global understanding of the physiological functioning of living organisms. Such understanding must encompass the

puzzling discovery that certain organisms live in extreme conditions of temperature, pressure, and salinity, which were originally thought to be incompatible with life (for a recent review see Rothschild and Mancinelli, 2001, and references therein). With the genomic sequences of these organisms becoming available, it is rather surprising that no striking genomic counterparts seem to be associated with such extreme lifestyles. For example, at the DNA level, an

Abbreviations: GC, guanine–cytosine; LGF, lifestyle genomic flag

\* Corresponding author. Tel.: +33-1-4568-8509; fax: +33-1-4061-3456.

E-mail address: tekaia@pasteur.fr (F. Tekai).

intuitively appealing idea would associate thermophily-hyperthermophily with GC-richness (to avoid undue ‘melting’ of genomic DNA). Yet it appears that the organisms with the highest GC contents are neither hyperthermophiles, nor thermophiles. It can be noticed, nevertheless, that at the level of particular sequences (rRNA genes) a correlation was observed—in prokaryotes—between GC contents and optimal growth temperatures (Galtier et al., 1999). Similarly, at the level of proteins, various partial characterizations led to elusive results. In such studies (involving most often limited sets of sequence data), the composition of proteins was described in terms of individual contributions concerning the various amino-acids (for a recent example see Haney et al., 1999). Even for certain completely sequenced genomes, amino-acid compositions have been only partially characterized (see, for example, Deckert et al., 1998; Nelson et al., 1999). Overall, diverse trends in the amino-acid compositions—as associated with the different lifestyles—could be derived. Trends detected in certain studies (see, for an early example, Argos et al., 1979) were not confirmed by further determinations on larger sets of protein data. A detailed overview can be found in a recent revue on hypothermophilic enzymes (Vieille and Zeikus, 2001).

In order to address the question of the characterization of genomic properties, at a large-scale level, appropriate methodological approaches must be adopted to provide relevant profiles (‘signatures’) for each organism. Correspondence analysis (Benzecri, 1973) aims to allow such global treatments. Based on this formulation, a multivariate analysis is performed here, on the available predicted proteomes for the completely sequenced genomes (including a rather significant number of representatives from the three phylogenetic domains of life). The results of this study are presented in a synthetic picture. From this picture, a series of conclusions are derived which extend and complete previously obtained results. With this respect, notably, a first treatment based on correspondence analysis was performed for the characterization of the *Mycobacterium tuberculosis* genome (Cole et al., 1998), as compared to the 15 genomes available at the time.

Here, beyond a significant increase in the number of genomes considered (56), a series of new questions are addressed concerning the detailed properties of amino acid compositions. These questions are put in perspective with a series of other works, relevant or not to multivariate formulations. On methodological grounds, two works—with similar interests than those here—relied on principal component analysis (Thompson and Eisenberg, 1999; Kreil and Ouzounis, 2001). Accordingly, the respective advantages in the usage of the different approaches are mentioned. Some of the conclusions obtained here overlap with results discussed in a rather large series of disparate works. The unified picture here should thus provide a better understanding of the relative importance of the various components involved in the intricate relations connecting amino acid compositions with other biologically important features including lifestyles and evolutionary trends. Several evolutionary-

oriented questions addressed here are entirely new with respect to previous works.

## 2. Materials and methods

Correspondence analysis (Benzecri, 1973; Greenacre, 1984) is a powerful method for the multivariate exploration of large-scale data. It has been applied in various research areas, including genomic analyses [for example, McInerney (1998) and Tekaia et al. (1999)]. To extract relevant informations from raw data, correspondence analysis relies on the projection of high-dimensional information onto low-dimensional spaces. Such projections, onto a plane, allow direct visual inspection of significant trends, which are often difficult to grasp in high-dimensional spaces. The dimensions of the considered spaces are of course relevant to the number of variables and observations involved in the study (such as, here, the variables associated with the different amino-acids, and the observations associated with the different genomes). In this multivariate method—as applied to numerical data matrices—we can construct an orthogonal system called factorial axes, corresponding to the low-order projections on planes called factorial planes. An important virtue of this construction is that the characteristic properties of the observations and the variables are displayed simultaneously on the factorial planes. A transition formula allows the coordinates of a given observation (respectively variable) to be calculated as a function of the variables (respectively observation) coordinates. The method is called after the ‘correspondence’ between the analysis of observations and that of variables. In this analysis, each factorial axis represents a fraction of the whole information contained in the analysed matrix. The statistical significance of this fraction determines the relative confidence attached to the displayed observations and/or variables, on the corresponding axis. The orthogonality of the factorial axes allows the summation of their corresponding information fractions. For example, the fraction of total information included in the first factorial plane is obtained by summing the fractions corresponding to the first ( $F_1$ ) and to the second ( $F_2$ ) factorial axes.

Correspondence analysis also allows consideration of dummy variables and observations (called ‘illustrative variables’, respectively ‘illustrative observations’), as additional variables (respectively observations) which do not contribute to the construction of the factorial space, but can be displayed on this factorial space. With such a representation it is possible to determine the proximity between observations and variables and the illustrative variables and observations.

### 2.1. Amino-acid relative composition matrix

The amino-acid compositions of 56 completely sequenced genomes were calculated using the *freqaa.pl* script (<http://www-alt.pasteur.fr/~tekaia/HYG/scripts.html>). These genomes included seven eukaryotes, 14 archaeal and

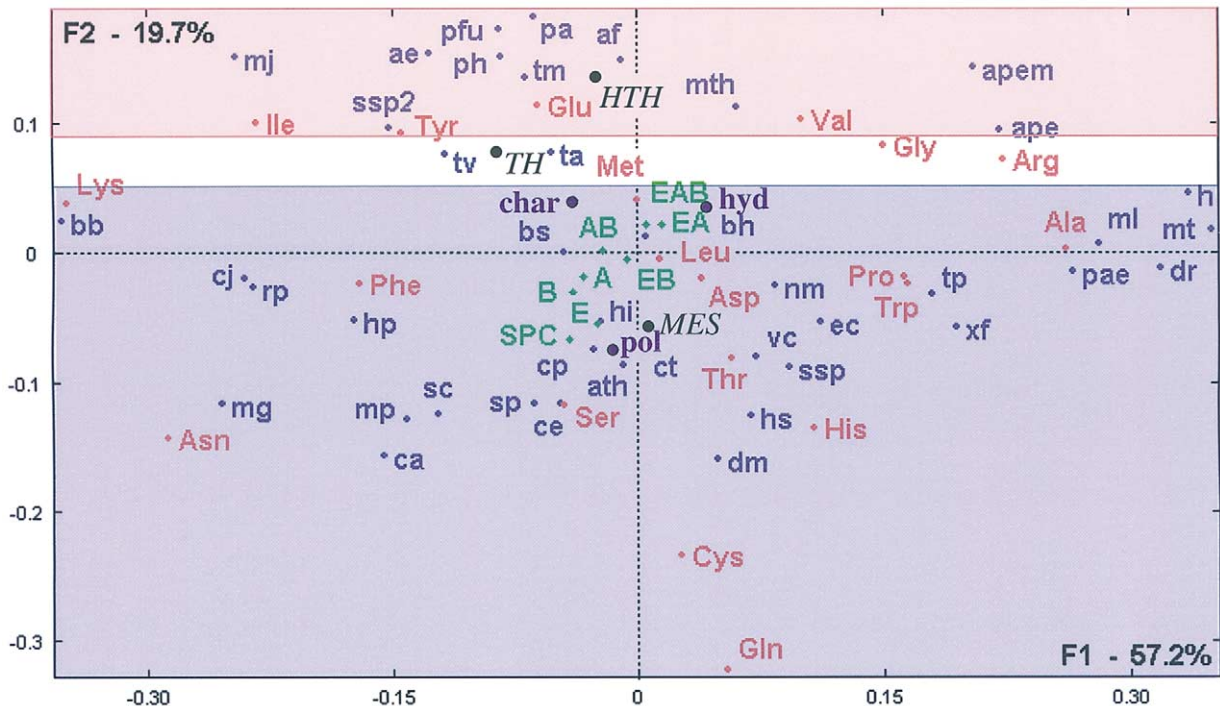


Fig. 1. Factorial plane representation for the distribution of species according to amino-acid compositions. Distribution of the surveyed species according to their relative amino-acid composition (see Section 2), as obtained on the first factorial space ( $F_1$  and  $F_2$ ) by correspondence analysis. This plane represents 76.9% of the total information embedded in the original data matrix. The relative amino acid compositions were calculated from the published predicted proteomes of the following species (each genome is indicated on the graph by an abbreviation<sup>+</sup> (references for the original publications for the various genomes can be found at <http://wit.integratedgenomics.com/GOLD/Genome.Refs.html>). Eukaryota (seven genomes): *Arabidopsis thaliana* (ath), *Candida albicans* (cacd)\*, *Caenorhabditis elegans* (ce), *Drosophila melanogaster* (dm), *Homo sapiens* (hs)\*\*, *Saccharomyces cerevisiae* (sc), *Schizosaccharomyces pombe* (sp). Archaea (14 genomes): *Aeropyrum pernix* K1 (ape), (a modified version [Natale et al., 2000] is also used and is denoted apem), *Archaeoglobus fulgidus* (af), *Halobacterium* sp. NRC-1 (h), *Methanobacterium thermoautotrophicum* (mth), *Methanococcus jannaschii* (mj), *Pyrococcus abyssi* (pa), *Pyrobaculum aerophilum* (pyae), *Pyrococcus horikoshii* (ph), *Pyrococcus furiosus* (pfu), *Sulfolobus solfataricus* P2 (ssp2), *Sulfolobus tokodaii* (sto), *Thermoplasma acidophilum* (ta), *Thermoplasma volcanium* (tv). Bacteria (35 genomes): *Agrobacterium tumefaciens* (agrt), *Aquifex aeolicus* (ae), *Bacillus subtilis* (bs), *Bacillus halodurans* (bh), *Borrelia burgdorferi* (bb), *Buchnera* sp. APS (b), *Campylobacter jejuni* (cj), *Chlamydia pneumoniae* (cp), *Chlamydia trachomatis* (ct), *Deinococcus radiodurans* (dr), *Escherichia coli* (ec), *Haemophilus influenzae* (hi), *Helicobacter pylori* (hp), *Listeria innocua* (lin), *Listeria monocytogenes* (lmo), *Mesorhizobium loti* (mm), *Mycobacterium tuberculosis* H37R (mt), *M. tuberculosis* CDC 1551 (mtc), *Mycoplasma genitalium* (mg), *Mycobacterium leprae* (ml), *Mycoplasma pneumoniae* (mp), *Neisseria meningitidis* (nm), *Pseudomonas aeruginosa* (pae), *Rickettsia prowazekii* (rp), *Salmonella typhi* (sty), *Sinorhizobium meliloti* (sm), *Staphylococcus aureus* Mu50 (samu50), *Staphylococcus aureus* N315 (san315), *Streptococcus pyogenes* M1 (spy), *Synechocystis* sp. (ssp), *Thermotoga maritima* (tm), *Treponema pallidum* (tp), *Vibrio cholerae* (vc), *Xylella fastidiosa* (xf), *Yersinia pestis* (yp). Charged amino-acids ('char'), polar/uncharged amino-acids ('pol') and hydrophobic amino-acids ('hyd') are considered in the analysis as illustrative variables (see Section 2.2). On the same plane, the hyperthermophilic (HTH), thermophilic (TH) and mesophilic (MES) profiles are reported as illustrative observations, along with the profiles associated with genes conserved in the various eukaryal (E), archael (A) and bacterial (B) species (and the corresponding intersections EA, EB, AB and EAB; see Section 2.3). <sup>+</sup>: Species abbreviations used in Figs. 1 and 2. \*: Sequence data were obtained from CandidaDB genome database <http://www.genolist.pasteur.fr/CandidaDB/>, constructed from data available at the Stanford Genome Technology Center website at: <http://www-sequence.stanford.edu/group/candida>. \*\*: 21,724 protein sequences downloaded from the ncbi ftp server on February 23, 2001 ([ftp://ncbi.nlm.nih.gov/genomes/H\\_sapiens/protein/](ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/protein/)).

35 bacterial species (see legend of Fig. 1). The raw amino-acid counts were transformed into relative compositions (or percentages)  $T_{ij}$  [with  $T_{ij}$  defined as the number of occurrences of the amino-acid  $j$  in the species  $i$ , divided by the total number of amino acids in  $i$ , multiplied by 100]. For the *Aeropyrum pernix* genome (ape), two distinct entries were considered [following the original (ape) annotation by Kawarabayasi et al. (1999) and a modified (apem) annotation by Natale et al. (2000)]. Altogether, a matrix of 56 observations (species, with the two entries ape and apem) and 20 variables (amino-acids) was obtained ([\[alt.pasteur.fr/~tekaia/aafreq.html\]\(http://alt.pasteur.fr/~tekaia/aafreq.html\)\). The hyperthermophilic and thermophilic species considered are listed in Table 1, with their respective optimal growth temperatures \(as reported in the literature\).](http://www-</a></p>
</div>
<div data-bbox=)

## 2.2. Illustrative variables

Three supplementary variables were considered in the analyses, as illustrative variables: 'char' for charged amino acids (Asp (D), Glu (E), Lys (K), Arg (R) and His (H)), 'pol' for polar/uncharged amino acids (Gly (G), Ser

Table 1  
Thermophilic and hyperthermophilic species<sup>a</sup>

Species	Growth temperature (°C)	Code	Prot. avg size	Gln	Glu	Asp	Arg + Lys
<i>Archaea</i>							
<i>P. furiosus</i> (pfu)	113	hth	266.5	1.8	8.9	4.4	13.4
<i>P. aerophilum</i> (pyae)	100	hth	251.8	2.1	7.0	4.3	12.2
<i>P. horikoshii</i> (ph)	98	hth	275.9	1.6	8.3	4.3	13.2
<i>P. abyssi</i> (pa)	96	hth	303.6	1.7	8.8	4.6	13.5
<i>A. permix</i> K1 (ape)	95	hth	237.1	1.9	6.6	3.9	11.2
<i>A. permix</i> K1 (apem*)	95	hth	279.0	1.8	7.3	4.2	11.7
<i>M. jannaschii</i> (mj)	85	hth	287.0	1.5	8.7	5.5	14.3
<i>A. fulgidus</i> (af)	83	hth	275.4	1.8	8.9	4.9	12.7
<i>S. solfataricus</i> P2 (ssp2)	80	hth	282.3	2.1	6.8	4.7	12.4
<i>S. tokodaii</i> (sto)	80	hth	268.4	2.1	7.0	4.6	12.2
<i>M. thermoautotrophicum</i> (mth)	65	hth	281.4	1.9	8.1	5.9	11.4
<i>T. volcanium</i> (tv)	60	th	297.1	2.1	6.4	5.5	11.6
<i>T. acidophilum</i> (ta)	59	th	306.6	2.2	6.0	5.7	11.1
<i>Bacteria</i>							
<i>A. aeolicus</i> (ae)	85	hth	317.0	2.0	9.6	4.3	14.3
<i>T. maritima</i> (tm)	80	hth	315.3	2.0	8.9	5.0	13.1

<sup>a</sup> For thermophilic (th) and hyperthermophilic (hth) species, the optimal growth temperatures are tabulated, along with the mean values for the protein lengths, glutamine (Gln), glutamic acid (Glu), aspartic acid (Asp) and arginine + lysine (Arg + Lys) amino acid compositions. The definitions of 'hyperthermophile' and of 'thermophile' are taken from the 'Dictionary of Cell and Molecular Biology' (<http://www.mblab.gla.ac.uk/dictionary/graphic.html>). In hyperthermophiles (excluding ape) the mean proportions (and the corresponding standard errors) for Glu and Gln are respectively  $8.19 \pm 0.27$  and  $1.87 \pm 0.06$ . \*Corresponds to a modified annotation of *A. permix* (Natale et al., 2000).

(S), Thr (T), Asn (N), Gln (Q), Tyr (Y) and Cys (C)) and 'hyd' for hydrophobic amino-acids (Leu (L), Met (M), Ile (I), Val (V), Trp (W), Pro (P), Ala (A) and Phe (F)). The amino-acid composition values attributed to the supplementary variables were obtained by summing the respective contributions of the corresponding amino acids, in the various species (for example, for the variable char, the contributions of the amino acids Asp, Glu, Lys, Arg and His, are summed). The classification of amino-acids follows Deckert et al. (1998).

### 2.3. Illustrative observations

For various specific analyses, a series of illustrative observations were considered. Mesophilic (MES), thermophilic (TH) and the hyperthermophilic (HTH) profiles were constructed, by taking the mean values for the amino acid compositions in the corresponding species, together with the three illustrative variables above. The three profiles MES, TH and HTH were handled as illustrative observations.

A systematic proteome comparison of the considered species was performed to detect gene conservation, in connection with the amino acid distribution properties. Such a comparison permits determination of the set of genes that are exclusively conserved in one of the three domains of life, in the various combinations of two domains, or at the intersection of the three domains (see <http://www-alt.pasteur.fr/~tekaia/domspec.html>). The methodology behind this treatment has previously been described in detail (Tekaia et al., 1999; Tekaia and Dujon, 1999). Here we recall briefly the methodology used. The proteome of each species

was compared to that of each other surveyed species, using the blastp program (Altschul et al., 1997), with the pam250 substitution matrix and the seg filter (Wootton and Federhen, 1993). For such comparisons, a limit of significance for the blastp probability scores was first determined for each of the genomes considered (Tekaia and Dujon, 1999). For each such genome, a set of random sequences (simulated proteic sequences) was generated, with size and amino-acid compositions set to the mean values for the coding sequences in the actual proteome. Each such random sequence was compared with the proteic database of the cognate organism and the best probability scores were recorded, as for actual sequences. For each organism, when used as target, the limit of significance for the comparisons was set as the highest blastp probability score with less than 5% of pseudosignificant matches. For example, probability score limits were set at  $10^{-9}$  for *Saccharomyces cerevisiae* [for details see Tekaia et al. (2000)].

Based on the above procedures, the following subsets of genes were determined:

1. genes specific to each species [SPC: genes with no matches outside the genomes of the corresponding species];
2. genes conserved exclusively in one domain of life [E: in eukaryal species, A: in archaeal species and B: in bacterial species];
3. genes conserved in a combination of two domains [EA: eukaryal and archaeal, EB: eukaryal and bacterial and AB: archaeal and bacterial], or at the intersection of the three domains [EAB: eukaryal, archaeal and bacterial].

Some cautionary remarks should be made on the definition of species-specific genes, those without detectable matches outside their own genome. Of course, with such a definition, the notion of ‘specific’ depends on the set of species considered. On theoretical grounds, it should be almost impossible to guarantee that no match will be ever found for a given gene, outside the genome considered. Also, by considering progressively closer genomes, at some point a match will be found for almost every gene. On practical grounds, following these observations, it is possible to rely on a reasonable definition for specific genes, as developed in a previous work for *S. cerevisiae* (Malpertuy et al., 2000). Following this work, it appears that the comparison of genes from a given species with sets of genes from species belonging to different phyla permits detection of asymptotic limits. Such limits do not change, essentially, when new comparisons are performed with genes belonging to yet other phyla. Of course, when new genes are considered belonging to the same phylum especially in the extreme case of very close genomes, then additional matches will be obtained. Accordingly, our definition of species specific-genes follows this phylum-dictated evaluation (see Malpertuy et al., 2000, for further discussions and details). Because of the close phylogenetic relationships, and in some cases the small sizes of the corresponding genomes, certain species considered here display only very few species-specific genes [for example, *Mycoplasma genitalium* (mg), *Buchnera* sp. APS (b), *Staphylococcus aureus* Mu50 (samu50) or *S. aureus* N315 (san315)]. As a further illustration the inclusion of *M. tuberculosis* CDC 1551 (mtc) in the analysis, results in a reduction of the number of specific genes in *M. tuberculosis* H37R (mt) from 707 to 61. With these cautionary remarks, and in order to facilitate the interpretations below, an exhaustive table is provided (<http://www-alt.pasteur.fr/~tekaia/domspec.html>) with the percentages of specific genes in the various genomes.

For each of the above subsets of genes, specific amino-acid composition matrices were calculated, with the same basic structure: 23 columns (20 entries for the various amino acids and three entries for the illustrative variables char, pol and hyd) and 56 lines (corresponding to the considered species). Following the subsets above, the matrices were built with variable number of blank lines [corresponding to species with no specific genes in the considered subset: for example, in the matrix associated with E, the lines associated with mg (*M. genitalium*) are left blank, as no mg gene is exclusively conserved in the E domain]. For each such matrix, mean proportions were calculated for each amino-acid, and each illustrative variable (char, pol and hyd), leading to eight illustrative observations: SPC, E, A, B, EA, EB, AB and EAB (with the same notations as for the gene subsets above, for simplicity). Finally, all the data were gathered in a single matrix of size 67 lines by 23 columns. The 23 columns correspond to the 20 types of amino acids, along with the three illustrative variables (char, pol and hyd), and the 67 lines correspond to the 56 genomes, along with the 11 illustrative observations (MES, TH, HTH, SPC, E, A, B, EA, EB, AB and

EAB). We call this matrix (see <http://www-alt.pasteur.fr/~tekaia/genomesaa.html>) the ‘amino-acid composition genomic matrix’.

### 3. Results

Correspondence analysis was applied to the amino-acid composition genomic matrix (see Section 2). The resulting distribution of amino acids, along with the surveyed species, the illustrative variables (char, pol, hyd) and the illustrative observations (MES, TH, HTH, SPC, E, A, ..., EAB) is shown in Fig. 1 [with the representation of the first factorial plane (see Section 2)]. From this synthetic output the following series of conclusions can be drawn.

#### 3.1. Correspondence analysis discriminates the three lifestyles

The first, most striking, conclusion is a clearcut discrimination between the known lifestyles of the various organisms considered. This discrimination is highlighted by the three coloured regions which slice the first factorial plane into three disjoint domains. The region in red gathers all the hyperthermophiles [either archaea or bacteria]: *Methanococcus jannaschii* (mj), *Archaeoglobus fulgidus* (af), *Pyrococcus horikoshii* OT3 (ph), *Pyrococcus abyssi* (pa), *Pyrococcus furiosus* (pfu), *Aquifex aeolicus* (ae), *Thermotoga maritima* (tm), *Methanobacterium thermoautotrophicum* (mth), *Sulfolobus solfataricus* P2 (ssp2) and *Aeropyrum pernix* K1 (apem, as annotated by Natale et al., 2000). It is interesting to note a shift in the position of the *A. pernix* genome on the factorial plane, which follows the method of annotation [with ape corresponding to the original annotation by Kawarabayasi et al. (1999)]. This observation could provide an (indirect) confirmation of the analysis of Cambilau and Claverie (2000), which reported discrepancies in the original annotation of the *A. pernix* genome. A specific region (in white) can be attributed to the thermophiles *Thermoplasma acidophilum* (ta) and *Thermoplasma volcanium* (tv). Finally, the region in blue is the territory for the mesophiles considered. We shall call this 3-coloured representation the ‘lifestyle genomic flag’ (LGF), as associated with the discrimination between the three major lifestyles.

The discrimination should be considered, of course, as reflecting a ‘continuum’ rather than a clear-cut separation between the three regions. In fact, the borders between the three coloured regions were set graphically, and somewhat arbitrarily for their exact positions. It is nevertheless striking that the second factorial axis reflects rather accurately the scale in optimal growth temperatures of the various species (from bottom to top in increasing temperatures). In this respect, it is important to point out that the distribution of the species following the vertical axis (allowing the discrimination between the lifestyles), is obtained independently of the tabulated know temperatures in Table 1. The discrimination obtained with correspondence analysis is solely

based on amino acid-compositions. Following this observation, the second factorial axis in the LGF representation could thus be called a ‘temperature-axis’. On the other hand, as expected from their definitions, the three illustrative observations (HTH, TH, MES) appear on the LGF as barycenters for the species belonging respectively to the hyperthermophiles, thermophiles and mesophiles regions.

The discrimination between lifestyles, obtained here with the LGF representation, was already apparent in the initial analysis of 15 genomes (Cole et al., 1998). With a different methodology, concerning the multivariate analyses, similar conclusions were reached by Thompson and Eisenberg (1999, 20 genomes) and Kreil and Ouzounis (2001, 27 genomes). As the additional genomes considered here included notably two thermophiles (which are not classified as hyperthermophiles), the LGF introduces a further sharp distinction between the lifestyles [mesophiles/thermophiles/hyperthermophiles, as compared to the discrimination mesophile/thermophiles]. As already mentioned, this enhanced discrimination (as well as the definition of the lifestyles themselves), should be considered in the perspective of a ‘continuum’.

### 3.2. GC content versus growth temperature

In addition to the vertical axis above, we consider the distribution of the species following the first ‘horizontal’ factorial axis. Following this axis, the distribution reflects essentially the GC contents of the species (as calculated for the coding regions of the corresponding DNA genomic sequences. This distribution is in increasing order, from left to right: from *Buchnera* sp. APS (b, 27.6%) and *Borrelia burgdorferi* (bb, 28.8%), at the most left side, to *M. tuberculosis* (mt, 65.9%), *Deinococcus radiodurans* (dr, 67.2%) and *Halobacterium* sp. (h, 68.5%). On the other hand, species which are scattered parallel to the vertical axis, correspond to roughly the same nucleotide content. For example, this is the case for *P. abyssi* (pa, 45.3%), *T. maritima* (tm, 46.4%), *T. acidophilum* (ta, 47.2%), *Bacillus subtilis* (bs, 44.2%), *Haemophilus influenzae* (hi, 38.8%) and *Chlamydia pneumoniae* (cp, 41.3%). As such, we can describe the first factorial axis as a ‘GC-axis’.

Since the two factorial axes are orthogonal, we can deduce that the properties on GC-content and lifestyle ‘relevant to thermophily, and reflected in the temperature properties’ are essentially not correlated. The results of correspondence analysis also suggest that GC content is the most important determinant of amino-acid composition properties, with the second most important being temperature (as related to lifestyle). This conclusion derives from the proportions of informations projected following the two factorial axes:  $F_1$  (as ‘GC-axis’) representing 60.9% of the total information, and  $F_2$  (as ‘temperature-axis’), representing 16.6% of the total information. It is interesting to note that the figures are essentially similar to those in Cole et al. (1998): 55.2% for  $F_1$  and 22.6% for  $F_2$ . Accordingly, the

proportions obtained here could be considered as asymptotically stable values. These figures provide a quantitative basis for the recurring conclusion that global amino-acid composition is essentially under the influence of GC-pressure (see, for example, Singer and Hickey, 2000).

### 3.3. Global physico-chemical trends

An interesting virtue of correspondence analysis is that illustrative parameters may be used (see Sections 2.2 and 2.3). With such parameters, we can derive visually the global trends associated with the usage of the various families of amino acids, as classified by their main physico-chemical properties (see Section 2.2). As seen in Fig. 1, we observe that the global trend for the usage of charged or hydrophobic amino acids points towards thermophily, whereas the global trend for the usage of polar/uncharged amino acids is in the opposite direction (higher representation in eukaryotes and some bacterial species including *Mycoplasma pneumoniae* (mp), *Chlamydia trachomatis* (ct) and *C. pneumoniae* (cp)). This observation may be expressed in the following terms: the trends pol- > char or pol- > hyd are essentially temperature-driven, whereas the trend char- > hyd appears to be connected essentially to the GC-content property (the line joining the positions ‘char’ and ‘hyd’ on the factorial plane is parallel to the second axis).

### 3.4. Amino acid composition: refined characterization of thermophily

Beyond the global amino acid distribution properties displayed by the LGF, it is useful to consider in detail some extreme positions (‘cardinal points’) on the LGF. We observe that, roughly parallel to the temperature-axis, the bottom-top positions are occupied respectively by glutamine (Gln) and glutamic acid (Glu). As for the left-right positions, roughly parallel to the GC-axis, we shall consider the lysine (Lys)-arginine (Arg) couple. Strictly speaking the most extreme, at the right, cardinal amino acid would be alanine (Ala). But the couple Lys-Arg is considered because of their common physico-chemical property: both amino acids bear positive charges. We next consider the properties associated with these cardinal amino-acids.

(a) Gln/Glu acid signature for hyperthermophiles: we observe that hyperthermophiles and thermophiles alike are reduced in Gln, whereas only hyperthermophilic species are increased in Glu acid (see Table 1; and also <http://www-alt.pasteur.fr/~tekaia/aafreq.html>). This rule holds for all the species considered, including the modified version of *A. permix* (apem) but with the exception of the original annotated version of *A. permix* (ape) (Kawarabayasi et al., 1999). It is also interesting to consider the codon usage associated with the apparent importance of Glu acid in hyperthermophiles. Indeed the uniformly high score associated with this amino acid, in the various hyperthermophiles, hides very different proportions of the two corresponding codons

[namely GAA and GAG, with the ratio between them varying from 0.47, such as in *A. fulgidus* (af), to 0.97, such as in *P. horikoshii* OT3 (ph), and up to 1.87, such as in *A. aeolicus* (ae)].

(b) Glu/Lys + Arg signature for hyperthermophiles: considering now the left-right Lys-Arg cardinal amino acids, an interesting observation can be made, which may have a functional basis. Hyperthermophiles, are characterized by a relative increase in the lumped sum Lys + Arg. The increase in Lys + Arg is nevertheless not entirely specific to hyperthermophiles, as similar increases are observed in some other non-hyperthermophilic species. However, focussing on the hyperthermophilic species (the raw data are displayed in Table 1, along with the growth temperatures), we can observe that Lys + Arg is (very) significantly correlated with Glu (Pearson correlation coefficient  $r = 0.83$  with  $P < 0.001$ ). A weaker but still significant correlation is also observed ( $r = 0.55$  with  $P < 0.04$ ) between this relative increase in Lys + Arg and the optimal growth temperatures, as reported in the literature and shown in Table 1. More importantly, there is no such correlation between Glu and the lumped sum Lys + Arg in the mesophilic species ( $r = 0.29$ ).

An attractive and plausible physico-chemical basis may exist, for this result. It is becoming increasingly clear that higher-order oligomerization of proteins in thermophilic proteins (as compared to their mesophilic counterparts) may represent one of their most important stabilizing mechanisms (for a general review see, for example, Vieille and Zeikus, 2001): ‘an ever-increasing number of hyperthermophilic proteins are known that have a higher oligomerization state than their mesophilic homologues’. Potentially the observed correlation may reflect, at least in part, the participation of the negatively charged Glu residues, in ionic bonds with the positively charged Lys and Arg residues, which act to stabilize such higher-order multimers. It is interesting to note that the remaining charged amino acid, aspartic acid (Asp), appears not to participate significantly in this compensatory negative/positive (charged) correlation. In fact, the distribution of Asp appears to be quite homogeneous throughout the species (this is shown in detail for the thermophilic/hyperthermophilic species in Table 1). At a global level, this feature is highlighted on the LGF representation of Fig. 1, with Asp being close to the origin of the factorial axes (in sharp contrast with the cardinal positions of the three other charged amino acids). Pearson correlation coefficient ( $r = -0.18$ ,  $P = 0.52$ ) confirms that for the thermophilic/hyperthermophilic species, there is no significant correlation between Asp and the pool Arg + Lys. Interestingly, with the relatively homogeneous distribution of Asp, the correlation between the negative pool (Asp + Glu) and the positive pool (Arg + Lys) (Pearson correlation coefficient  $r = 0.6810$ ,  $P < 0.005$ ), is weakened with respect to the correlation concerning Glu/Arg + Lys. Since Asp possesses the shortest side chain (1 CH<sub>2</sub>) of the charged amino acids, it seems possible that this amino acid

makes ionic bonds less easily, being more deeply located at protein surfaces than Glu residues.

### 3.5. Gene lengths and lifestyle

While the issue is somehow separate from the previous analysis, it is interesting, in an evolutionary context, to revisit yet another property that may be related to phylogeny and/or lifestyle: is there a relation between the lengths of coding regions and thermophilic stability of the corresponding proteins? Considering mean lengths for the proteins of the 56 species, one way analysis of variance revealed a significant difference ( $F = 185.3$ ,  $P < 0.0001$ ) for such means following the three domains of life: eukarya ( $462.4 \pm 9.7$ ), bacteria ( $316 \pm 3.6$ ) and archaea ( $278.4 \pm 5$ ) (mean values with the corresponding SEMs). All three *t*-tests comparing the three pairs of mean lengths, showed significant differences ( $P < 0.0001$ ). This result is in full agreement with a recent paper, devoted to the subject (Zhang, 2000). Interestingly the obtained values should now correspond to asymptotically stable values, being very close to those obtained by Zhang (2000), despite the fact that they were obtained on the basis of only 22 genomes [with the values 449 for eukarya (two species), 330 for bacteria, and 270 for archaea]. Also, the significance of the reported differences, in the mean protein lengths, appears to be robust with respect to different annotation methods. For example, the mean value 278.4 for archaea was obtained with ape, whereas with apem the value of the mean is 281.5.

Concerning lifestyle, also a significant difference ( $t = 3.6$ ,  $df = 54$ ,  $P < 0.0007$ ) is observed between mean protein lengths in (hyper)thermophiles ( $283.0 \pm 5.8$ ) and in mesophiles ( $340 \pm 9.4$ ). Accordingly, the shortening of coding sequences could be yet another important determinant of the thermal stability of thermophilic proteins.

### 3.6. Ancient conserved genes, thermophily and evolutionary trends

We consider the observations on mean compositions of specified subsets of genes (E, A, B, EA, EB, AB, EAB and SPC, see Section 2.3), reported in the LGF (Fig. 1). The distribution of the corresponding points on the LGF appears to follow a general trend. If an arrow is defined connecting SPC (species specific genes) to EAB (genes conserved in the three domains of life), we observe a ‘progressive’ trend between these two extreme values: following the direction of the arrow, we find first the three points specific to the three domains of life taken separately (E for eukaryotes, B for bacteria and A for archaea), followed by the three points corresponding to the intersections between two different domains (EB, AB and EA, very close to EAB). Now, the direction of the arrow, from SPC to EAB tends towards increased temperatures (thermophily), and increased GC-contents.

Since the set of most conserved genes can be assimilated to the most ancient ones, this result suggests that the ances-



#### 4. Conclusions and discussions

Correspondence analysis is a powerful and efficient exploratory method to extract significant properties, and trends, from large-scale genomic data. These features are illustrated here with an analysis of the amino acid composition characteristics of the genomes. The work has both methodological and biological implications.

Methodologically, the use of correspondence analysis to analyse various large-scale genomic problems possesses several significant advantages including notably simplicity and straightforwardness. For example, obtaining the factorial plane, which corresponds here to the LGF representation, from the raw data did not involve any manipulations to the basic equations of correspondence analysis. In contrast, other multivariate methods may require a priori hypotheses on the choice of various free-parameters. Correspondence analysis also represents a powerful technique to integrate several types of auxiliary informations including illustrative variables and observations within a unified picture. A recent publication on the application of correspondence analysis to microarray data (Fellenberg et al., 2001) provides a good illustration of the strengths of this methodology.

The LGF representation provides a clear discrimination between lifestyles, solely on the basis of amino-acid composition. Even though we consider this discrimination to reflect a ‘continuum’, it is interesting to point out the refinement in the discrimination, with the two thermophiles separated from the hyperthermophiles. As an increasing number of thermophilic species (not classified as hyperthermophiles) becomes available, it will be interesting to see the extent to which this refinement is confirmed. This picture lacks, nevertheless, representatives from cold-loving species (or psychrophiles). It will be of great interest to include in this analysis genomes such as that of *colwellia* sp. 34H (sequenced at the TIGR), when they become available (updated analyses will be made available at: <http://www-alt.pasteur.fr/~tekaia/ooaimg.html>). Preliminary remarks reported recently (Deming, 2002), suggest that the amino acid composition signature of such species may provide important new information to refine the global picture of lifestyles and genomes.

The LGF provides a quantitative picture for the relative importance of the two main components shaping the amino-acid composition: GC-pressure (with about 60% of the factorial information following this axis) and temperature (as related with lifestyle, with about 15% of the factorial information following this axis). These estimates for the relative contributions of the two components may confidently be considered as asymptotically stable (with the number of genomes), since they are little changed from the original analysis in Cole et al. (1998).

Beyond the global picture of the LGF, the results of correspondence analysis also point towards more specific properties, concerning certain amino-acids. It seems possible that the thermostability may be correlated with certain biases in

the content of particular amino-acids. However, in most cases, the increasing amount of available genomic data, has tended to contradict correlations reported from analysis of smaller sets of data (for a recent reviews see, for example, Vieille and Zeikus, 2001; Sterner and Liebl, 2001). A rather striking illustration of the need to consider large data sets consisting of many genomes, is provided by the extreme ‘cardinal points’ for the positions of the amino acids on the LGF. Some of the first statistical analyses on relations between amino-acid composition and thermophily, concluded that substitutional trends such as Gly- > Ala and Lys- > Arg are characteristic of thermophiles, as compared to mesophiles (see Vieille and Zeikus, 2001, for review). However, the LGF clearly shows that Lys- > Arg is, in fact not such a good predictor of thermophily, since the two amino acids lie at essentially the same horizontal level in the factorial plane.

Taking these restrictions into account, it is all the more interesting to extract some specific trends from the LGF, which may be more elaborate than simple biases concerning certain amino acids. The interesting extreme ‘cardinal position’ amino acids appear to be the couples Gln-Glu acid and Lys-Arg. The observation that hyperthermophiles and thermophiles (only two species) are both reduced in Gln, whereas only hyperthermophilic species are increased in Glu acid will need to be further assessed, when more genomes from thermophilic species become available. Jaenicke and Bohm (2001) already noted that Gln, seems to be significantly discriminated against in hyperthermophiles and suggested that an increased rate of deamination of this residue at high temperatures might explain this observation. However the same discrimination is not observed for asparagine, as noted by these authors and fully confirmed by the LGF. Most interestingly, the significant correlation observed in all hyperthermophilic species between the relative increase in the negatively charged Glu acid and the lumped sum of positively charged Lys + Arg, is totally absent from the mesophilic species. The possible association with essential physico-chemical properties – such as an increased number of oligomerizations in thermostable proteins – would render this correlation hypothesis still more significant.

The evolutionary considerations underlying the LGF provide novel and distinct pictures of the complex genomic events associated with shifts between lifestyles.

With the illustrative observations E, A, B, EA, EB, AB, EAB (describing genes specifically found in eukarya, archaea, etc., see Sections 2.3 and 3.6), a clear trend appears on the LGF, with increased universality (pool of genes common to all species) tending towards the hyperthermophilic lifestyle. Since the EAB probably contains the most ancient genes, following the conservation criterion, based on this representation, it is tempting to associate ancient genes with thermophilic (and/or hyperthermophilic) lifestyle. This observation would be coherent with relatively well-accepted hypotheses that life emerged in high-thermal

niches (for discussions see, for example, Miller and Lazcano, 1995; Forterre, 1996). It is interesting to point out that on the LGF all the speciation trends that lead from the common pool to eukarya, as well as bacteria and archaea, also show the tendency towards ‘lower temperatures’.

## Acknowledgements

This work is dedicated to the memory of Alexis Harrington. We thank Michael Nilges, Horst Feldmann, Annie Kolb and Richard Miles for careful reading of the manuscript, and helpful comments. A series of insights and comments from the referees helped in the reshaping of the manuscript, with the introduction of important clarifications. This work was generously supported by funding from the Pasteur Institute (notably through a DVPI contract). EY thanks support from the CNRS and the ‘Ministère de la Recherche’ (Programme Bioinformatique 2000 and Action Concertée Incitative Physicochimie de la Matière Complexe). B.D is Professor of Molecular Genetics at Univ. P. M. Curie and a member of the Institut Universitaire de France.

## References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Argos, P., Rossmann, M.G., Grau, U.M., Zuber, H., Frank, G., Tratschin, J.D., 1979. Structural comparisons of heme binding proteins. *Biochemistry* 18, 5698–5703.
- Benzecri, J.-P., 1973. L’analyse des données, L’Analyse des Correspondances, 2, Dunod, Paris, France.
- Cambillau, C., Claverie, J.M., 2000. Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.* 275, 32383–32386.
- Cole, S.T., et al., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544.
- Deckert, G., et al., 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 392, 353–358.
- Deming, J.W., 2002. Psychrophiles and polar regions. *Curr. Opin. Microbiol.* 5, 301–309.
- Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., Vingron, M., 2001. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. USA* 98, 10781–10786.
- Forterre, P., 1996. A hot topic: the origin of hyperthermophiles. *Cell* 85, 475–480.
- Galtier, N., Tourasse, N., Gouy, M., 1999. A non-hyperthermophilic common ancestor to extant life forms. *Science* 283, 155–157.
- Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. 1st Edition. Academic press, London, p. 223.
- Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I., Woese, C.R., Olsen, G.J., 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus species*. *Proc. Natl. Acad. Sci. USA* 96, 3578–3583.
- Jaenicke, R., Bohm, G., 2001. Thermostability of proteins from *Thermotoga maritima*. *Methods Enzymol.* 334, 438–469.
- Kawarabayasi, Y., et al., 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* 6, 83–101.
- Kreil, D.P., Ouzounis, C.A., 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* 468, 109–114.
- Malpertuy, A., et al., 2000. Genomic exploration of the hemiascomycetous yeasts: ascomycetes-specific genes. *FEBS Lett.* 487, 113–121.
- McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA* 95, 10698–10703.
- Miller, S.L., Lazcano, A., 1995. The origin of life – did it occur at high temperatures? *J. Mol. Evol.* 41, 689–692.
- Natale, D.A., Shankavaram, U.T., Galperin, M.Y., Wolf, Y.I., Aravind, L., Koonin, E.V., 2000. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.* 1, 1–19.
- Nelson, K.E., et al., 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329.
- Rothschild, L.J., Mancinelli, R.L., 2001. Life in extreme environments. *Nature* 409, 1092–1101.
- Singer, G.A., Hickey, D.A., 2000. Nucleotide bias causes a genome wide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* 17, 1581–1588.
- Sternier, R., Liebl, W., 2001. Thermophilic adaptation of proteins. *Crit. Rev. Biochem. Mol. Biol.* 36, 39–106.
- Tekaia, F., Dujon, B., 1999. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J. Mol. Evol.* 49, 591–600.
- Tekaia, F., Lazcano, A., Dujon, B., 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* 9, 550–557.
- Tekaia, F., et al., 2000. Genomic exploration of the hemiascomycetous yeasts: 3. Methods and strategies used for sequence analysis and annotation. *FEBS Lett.* 487, 17–30.
- Thompson, M.J., Eisenberg, D., 1999. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J. Mol. Biol.* 290, 595–604.
- Vieille, C., Zeikus, G.J., 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1–43.
- Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149–163.
- Zhang, J., 2000. Protein-length distributions for the three domains of life. *Trends Genet.* 16, 107–109.