1608–1615 *Nucleic Acids Research, 2001, Vol. 29, No. 7*

Identification of thermophilic species by the amino acid compositions deduced from their genomes

David P. Kreil[1,2,*] and Christos A. Ouzounis[2]

[1]University of Cambridge and [2]European Bioinformatics Institute, Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Received November 7, 2000; Revised and Accepted February 12, 2001

ABSTRACT

The global amino acid compositions as deduced from the complete genomic sequences of six thermophilic archaea, two thermophilic bacteria, 17 mesophilic bacteria and two eukaryotic species were analysed by hierarchical clustering and principal components analysis. Both methods showed an influence of several factors on amino acid composition. Although GC content has a dominant effect, thermophilic species can be identified by their global amino acid compositions alone. This study presents a careful statistical analysis of factors that affect amino acid composition and also yielded specific features of the average amino acid composition of thermophilic species. Moreover, we introduce the first example of a 'compositional tree' of species that takes into account not only homologous proteins, but also proteins unique to particular species. We expect this simple yet novel approach to be a useful additional tool for the study of phylogeny at the genome level.

INTRODUCTION

The properties of thermophilic proteins have been examined extensively in the past two decades. In particular, it has been investigated whether thermophily can be detected at the amino acid level, e.g., in the form of preferences towards particular residue types or other structural features. One approach to address this issue has been to analyse the amino acid compositions of thermophilic proteins for comparison with their mesophilic counterparts. Such studies have detected some preferences of thermophilic proteins for particular amino acids (1,2), but general rules have not yet emerged (3). Factors that have been reported to increase thermal stability are: a higher number of salt bridges on the surface [first proposed by Perutz more than 20 years ago (4)], tighter internal packing of hydrophobic residues, additional hydrogen bonds, and others. Apparently, all these factors contribute towards the greater stability of thermophilic proteins (3).

The rapid progress in the sequencing of genomes has opened many new avenues of research. In particular, a comprehensive comparison of global amino acid compositions as deduced from the genomes of different organisms is now possible. Using clustering and statistical methods, we have conducted such an extensive analysis to examine possible correlations of amino acid composition with the most prominent phenotypes of the species under consideration. Here we present the results for the genomes of six archaea, 19 bacteria, and the eukaryotic organisms yeast and *Caenorhabditis elegans*.

As we have shown in the present study, using two different approaches, several factors that determine amino acid composition can be deduced. Although the GC content of the coding sequences is the dominant influence on amino acid composition, it is possible to identify thermophilic species on the basis of their amino acid compositions alone.

MATERIALS AND METHODS

Data sources and tools

For each of the 27 fully sequenced genomes listed in Table 1, all deduced proteins, together with their corresponding coding sequences, were obtained from public databases available at the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk) and the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). Data-extraction and scripting were performed using SRS, a system that provides integrated access to multiple molecular biology databases (5). Amino acid compositions were determined using aacomp (6). Non-redundant sequence sets were created using nrdb90 (7). EPCLUST, a tool developed for expression profile analysis (8) was used for hierarchical clustering. Principal components analysis (PCA), examination of distributions and all post-hoc significance tests were performed using the statistics package SPSS-9 (9).

Exploratory data analysis

For all organisms, we determined global amino acid compositions, yielding a matrix where the rows represent the data sources listed in Table 1, and the 20 columns correspond to the respective percentage amino acid content. Each column of the matrix was then standardised. Standardised scores (*Z*-scores) were used throughout the analysis, i.e., all 20 amino acids were treated with equal weight. The 25 bacterial and archaeal species were chosen as a well-defined reference group for this normalisation; the two eukaryotic species were excluded, because they may represent an unbalanced sample of the eukaryotic world.

*To whom correspondence should be addressed. Tel: +44 1223 494 663; Fax: +44 1223 494 468; Email: kreil@ebi.ac.uk

Table 1. Abbreviations for data set entries

ID	Source
afa	*Aquifex aeolicus*
agf	*Archaeoglobus fulgidus*
appk1	*Aeropyrum pernix* K1
bbc	*Borrelia burgdorferi*, chromosome
bs	*Bacillus subtilis*
ce	*Caenorhabditis elegans*, chromosomes I–V and X
cj	*Campylobacter jejuni* NCTC11168
cp	*Chlamydia pneumoniae*
ct	*Chlamydia trachomatis*
dcrdr1c1	*Deinococcus radiodurans* R1, chromosome 1
dcrdr1c2	*Deinococcus radiodurans* R1, chromosome 2
ec	*Escherichia coli* K-12 MG1655
hbp26695	*Helicobacter pylori* 26695
hbpj99	*Helicobacter pylori* J99
hbpnrc100	*Halobacterium sp. NRC-1* plasmid pNRC100
hi	*Haemophilus influenzae*
mbt	*Mycobacterium tuberculosis* H37Rv
mbtat	*Methanobacterium thermoautotrophicum*
mcjc	*Methanococcus jannaschii*, chromosome
mpg	*Mycoplasma genitalium*
mpp	*Mycoplasma pneumoniae*
pca	*Pyrococcus abyssi*
pch	*Pyrococcus horikoshii* OT3
rp	*Rickettsia prowazekii* Madrid E
scer	*Saccharomyces cerevisiae*, chromosomes 1–16
scpcc6803	*Synechocystis* PCC6803
tp	*Treponema pallidum*
ttm	*Thermotoga maritima*
uu	*Ureaplasma urealyticum*

The set includes six thermophilic archaea, 17 mesophilic bacteria, two thermophilic bacteria and two eukaryotes. The ID column lists the labels used in this paper, while the Source column describes the origin of the corresponding data sets and includes the species and strain names where applicable.

A hierarchical clustering tool (http://www.ebi.ac.uk/microarray) was employed to group species according to similar amino acid composition. The unweighted pair group method with arithmetic mean (UPGMA) and a Euclidean distance measure were used for construction of the hierarchical tree.

To aid interpretation of the grouping, the data were subjected to PCA (10), see http://www.spss.com/tech/stat/Algorithms.htm for algorithmic details. Respective results were compared after: (i) oblique rotation; (ii) and (iii) two types of orthogonal rotation; and (iv) no rotation of factor axes. Step (iv) served as a test for possible correlations among factors. The assignment of principal factors was supported through introduction of two further variables: the GC ratio (GC counts versus AT counts) and a binary variable, *therm*, indicating thermophily. The normalisation step for the GC ratios obtained from the coding sequences for all studied proteins was identical to the procedure for amino acid compositions described above. The binary variable, *therm*, was set to zero or one for the mesophilic and thermophilic species, respectively. PCA was repeated after adding either or both of the two additional variables to the amino acid composition data. SPSS syntax scripts were prepared using SRS. In addition, the magnitudes of the raw correlations of all variables to the GC ratio and *therm* have been calculated.

Sensitivity analysis, sampling adequacy and significance

The sensitivity of the grouping achieved by hierarchical clustering was examined for various choices of protein sequence sets, normalisation reference and clustering method. Protein sequence sets were optionally made non-redundant. Normalisation was either relative to all bacteria and archaea, or relative to archaea only. Miscellaneous clustering methods were tried, such as UPGMA (average linkage), the maximum distance method (complete linkage), minimum distance (single linkage), or the weighted pair group method with arithmetic mean (WPGMA).

Similarly, the sensitivity of PCA was tested for various choices of data sets, normalisation reference and rotation of factor axes. A range of statistical criteria for the suitability of PCA were employed. In particular, the issue of sampling adequacy was addressed (see Appendix).

The distributions of the amino acid compositions found in the given groups of thermophiles and mesophiles were assessed with post-hoc significance tests for differences, as appropriate. Assumptions of these tests were always verified, and all tests were repeated with weighted cases for similar effective group sizes (a critical requirement for the strict validity of the tests). The two chromosomes of *Deinococcus radiodurans* were merged into one data set, and the two strains of *Helicobacter pylori* were weighted down to ensure a balanced representation within the group of mesophiles. The PCA was repeated to verify that this weighting did not affect any conclusions of the PCA.

For each of the raw correlations to GC ratio and *therm*, the two-tailed significance (P value) has also been determined.

RESULTS

From the genomes of six archaea, 19 bacteria and two eukarya, the respective amino acid compositions were determined. These were examined after hierarchical clustering and subjected to PCA. Our results have proven to be remarkably robust to variations of data sets and choice of algorithms. Addition of the GC ratios produced a further stabilising effect. All of our conclusions, however, are drawn from results derived from amino acid composition alone.

Hierarchical clustering by amino acid composition

The grouping that can be obtained from hierarchical clustering by standardised amino acid composition is shown in Figure 1. The coloured blocks show which amino acids are over- (red) or under-represented (green). The scale for the dendrogram shows Euclidean distance. Even though no phylogenetic information in terms of genomic or protein sequence has been

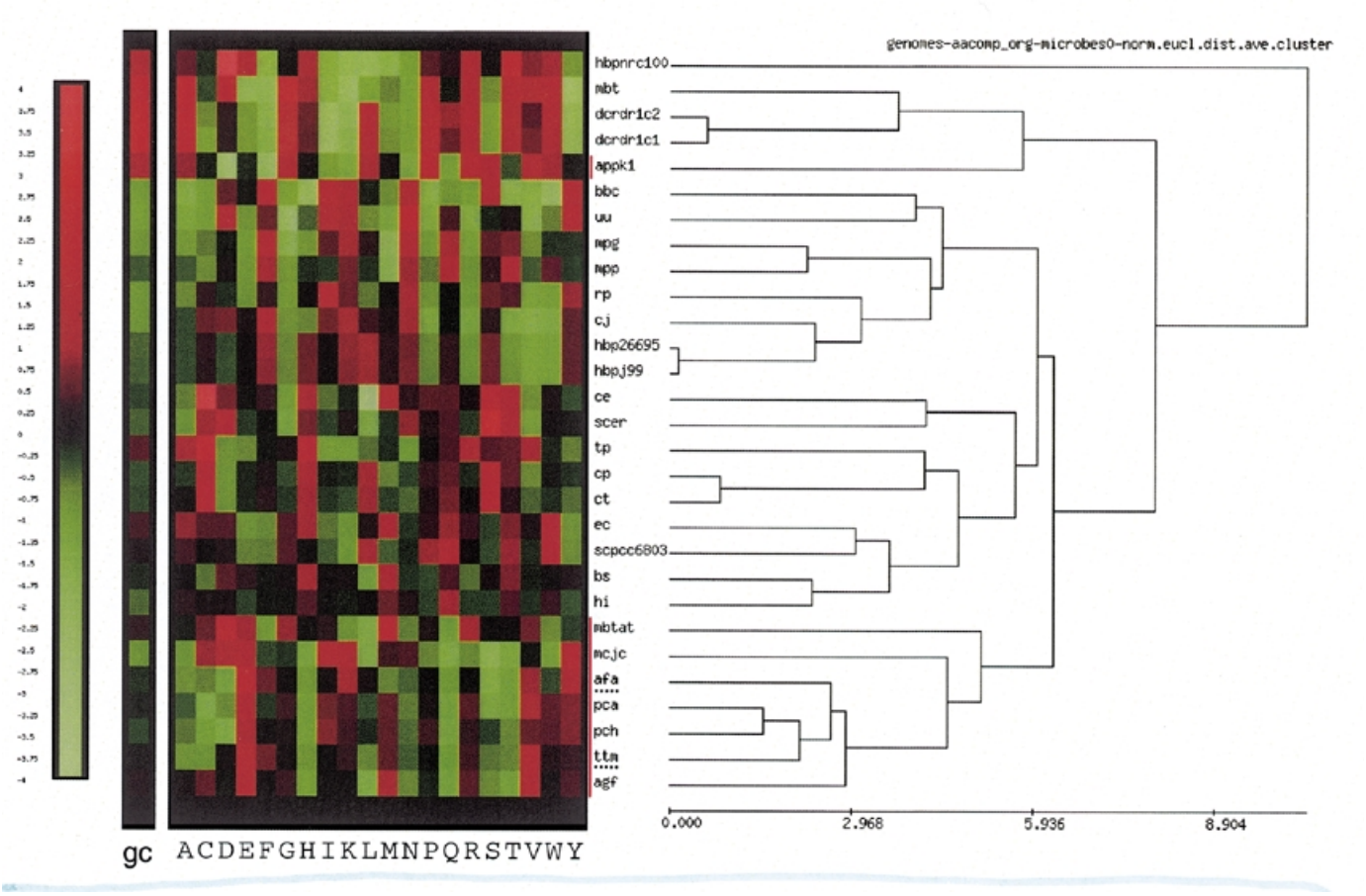

Figure 1. Standardised amino acid composition data of completely sequenced organisms grouped by hierarchical clustering. The GC ratios are shown for reference but were not used for the clustering process. Amino acids are abbreviated by the standard one letter code. The labels indicating the data sets for each row are explained in Table 1. In this figure, labels for thermophiles are marked with a red vertical bar, the thermophilic bacteria are highlighted by a dotted underline. The coloured blocks show normalised values as seen from the colour bar at the left. Red and green mean more and less than average, respectively. The scale for the dendrogram represents Eucledian distance. See Materials and Methods for details.

incorporated in tree construction, phylogenetically-related organisms were found as proximate neighbours. The closest pair was formed by the two strains 26695 and J99 of *H.pylori* with a distance of 0.2, followed by *Chlamydia pneumoniae* and *Chlamydia trachomatis* at 0.8, and *Pyrococcus abyssi* and *Pyrococcus horikoshii* at 1.5. Further, the two chromosomes of *D.radiodurans* have a distance of 0.6, which is larger than the distance between the two *H.pylori* strains (Fig. 1). This shows that there is remarkably large variation of long-range averaged amino acid composition even within an organism's genome. The two eukaryotes also have amino acid compositions that are more similar to each other than to any of the other organisms included in this study. The tree in Figure 1 successfully detects very similar patterns of amino acid composition as reflected in the detection of closely related species (see Discussion).

The clustering segmented the organisms into three distinct groups. One group contained organisms with an unusually high GC ratio. It was formed by the chromosomes of *D.radiodurans*, *Mycobacterium tuberculosis* and also the thermophilic archaeon *Aeropyrum pernix*. The range of the raw GC content observed in this group is 57–67%, in contrast to the average (standard deviation) of 47% (13%) computed over all bacteria and archaea. Only the amino acid compositions were used as input for the clustering process, not the GC ratios.

The other organisms split into two major groups. One group contains only thermophiles, namely all the remaining archaea (*Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Aquifex aeolicus*, *P.abyssi* and *P.horikoshii*) and the two bacteria *Thermotoga maritima* and *Archaeoglobus fulgidus*. This group thus consists of all the thermophiles except *A.pernix*, the clustering position of which is strongly influenced by its unusually high GC ratio.

While the exact topology of the trees did depend on the choice of normalisation and the algorithm for their construction, the resulting grouping was not changed significantly (see Appendix). Elimination of redundancy in protein sequence sets had no observable effect at all.

PCA of amino acid compositions

PCA was performed to identify underlying factors in the grouping found from hierarchical clustering. Plotting all organisms in reduced dimensions (Fig. 2) produces a clear separation of thermophiles and mesophiles along the second principal axis. In contrast, both thermophiles and mesophiles are found over a range of similar positions along the first

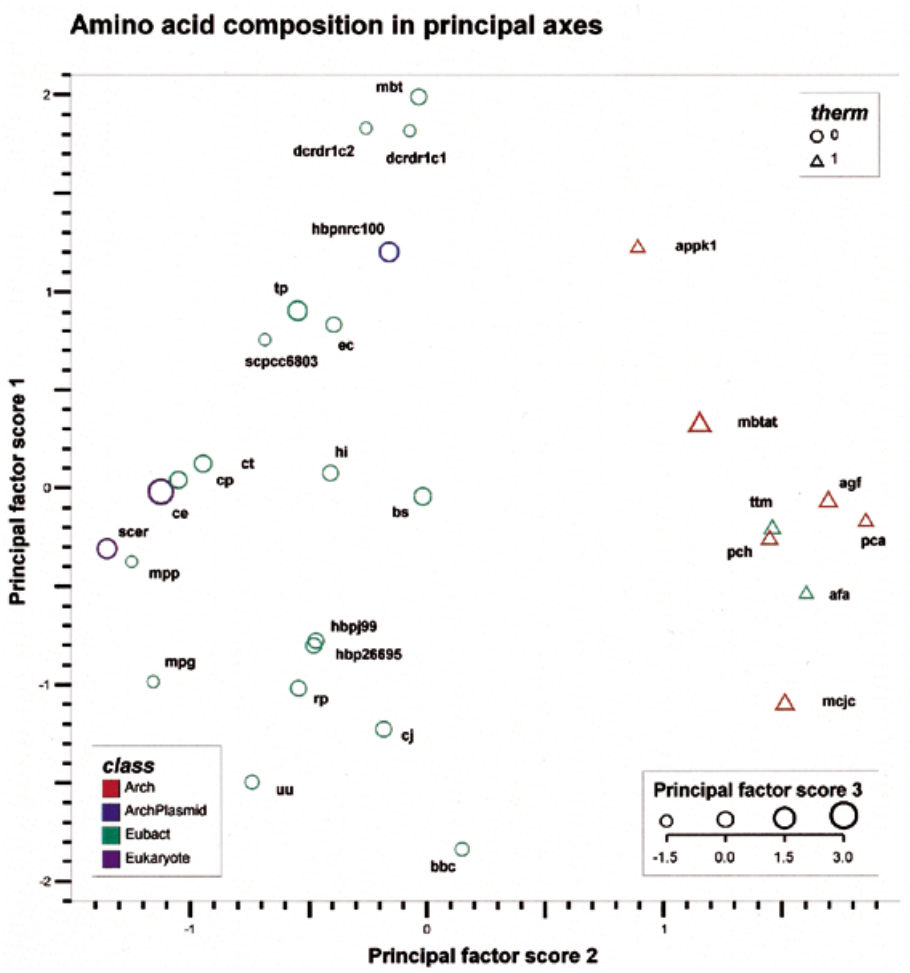

Figure 2. Reduced dimensionality plot showing the main principal components of the global amino acid compositions. The first principal axis (vertical) corresponds to GC ratio (see text). The second principal axis (horizontal) shows a clear separation of thermophiles and mesophiles, denoted by triangles and circles, respectively. The third principal component is depicted by symbol size (see insert for scale). Colour groups the sources into archea (red), bacteria (green) and eukaryotes (purple). The plasmid (the outgroup for hierarchical clustering, Fig. 1) is shown in blue. The graph is a projection, and distances are therefore not directly comparable to the distances observed in Figure 1. See text for discussion. For an explanation of data set labels see Table 1.

principal axis. Within the cluster of thermophiles, *A.pernix* is the organism most distant from its centre, although *M.thermoautotrophicum* and *M.jannaschii* are also slightly set apart from the relatively compact core formed by the others.

The contribution of the original variables to the principal components is shown in Figure 3. The component loadings can be interpreted as correlation coefficients (10). Absolute component loadings ≥0.6 are commonly considered as 'high loading', i.e., strong evidence. The plot also displays the correlations to the observed GC ratio and *therm*, the binary variable indicating thermophily (see Materials and Methods). These are shown for reference but have not been used as PCA input.

Clearly, component 1 strongly correlates with the GC ratio (94%, Fig. 3). This suggests that this factor can be seen as the influence of GC pressure on global amino acid composition. All of the individual amino acid biases firmly support such a conclusion. From the proportion of G and C bases in the respective codons for each amino acid, more Ala, Gly, Pro, Arg and Trp can be expected in organisms with a high GC ratio. This is in agreement with the very high positive component 1 factor loadings that were found for these amino acids. Similarly, fewer Ile, Phe, Lys, Asn and Tyr are expected to be seen at higher GC ratios. Indeed, very high negative factor loadings were observed for these amino acids. For the remaining amino acids, no bias is expected from their respective codons, a fact reflected in only low-to-moderate factor loadings for these amino acids (Fig. 3).

The obtained correlations for Met and Val at first seem surprising. (i) Met, encoded by ATG, would be expected less frequently in organisms of high GC ratio, but this trend is lacking from our analysis (Fig. 3). (ii) More Val is found in organisms with high GC ratio, although this bias would not be expected from the codons for Val (GTX). From observed substitutions, however, it is known that the amino acids Ile, Leu, Met and Val form a closed group of conservatively interchangeable residues [see BLOSUM62 matrices (11), for example]. The many conservative replacements of Ile (encoded by AT[A/T/C]) forced by positive GC pressure could explain the higher-than-expected observed counts of Val and Met.

As seen from Figure 2, component 2 clearly reflects membership in the group of thermophiles. This is firmly

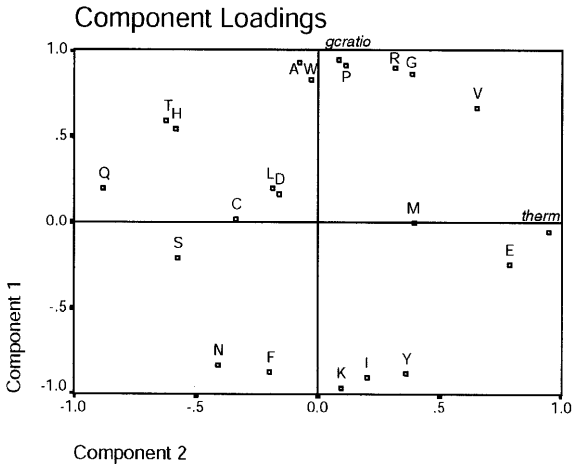

Figure 3. Component loadings for the main principal components. Component loadings can be interpreted as correlation coefficients (10). This plot shows to what degree the original variables contribute to the principal components. The figure further displays the correlations to the observed GC ratio and *therm*, the binary variable indicating thermophily (see Materials and Methods). These are shown for reference but have not been used as PCA input. Component loadings with an absolute value of ≥0.6 are commonly considered as high.

supported by the very high correlation (94%, see Fig. 3) with the binary variable *therm*. Oblique rotation showed that intercorrelation of the principal factors 1 and 2 is small (~14%). The component loadings (Fig. 3) suggested that the coupling might be mediated by Val, Thr and His, which were the only amino acids with moderate (or, in the case of Val, high) component loadings for both factors.

The strongest contributions to component 2 come from Gln, Glu and Val. Gln shows high negative correlation, whereas Glu and Val exhibit high positive correlations to the component. An apparent tendency for less Thr was also observed. These biases are already seen in the hierarchical clustering diagram (Fig. 1). Further inferences that can be made from only moderate component 2 factor loadings are biases for less His, Ser and Asn.

The two identified factors, to which we assigned GC pressure and thermophily, explain at least 65% of the total variance in amino acid compositions (see Appendix). Our analysis to a large degree also confirms their independence.

Different methods of weighting, factor axes rotation and exclusion of individual variables did change the appearance of reduced dimensionality plots, yet the separation of thermophiles was always clearly maintained (see Appendix for details). Factor loadings were very stable with only a few exceptions (e.g., Ser, see Appendix). A limitation to non-redundant sequence sets was of almost no consequence.

Statistical evidence and specific features of the thermophilic species

Starting from the distinct groups of thermophiles and mesophiles as obtained by PCA, analysis of factor loadings and the raw correlations were combined with post-hoc tests, where necessary, to determine significant differences between the two detected groups.

The component loadings for factor 1 are remarkably clear with very high factor loadings (>80%) for Ala, Gly, Pro, Arg, Trp, Ile, Phe, Lys, Asn and Tyr, and a high factor loading for Val. The conclusions derived from these are further supported by the strong absolute raw correlations (~70–90%) at high significance (~10^{-4}–10^{-12}).

The analysis of factor 2, however, warrants a closer look, as only Gln and Glu have very high component loadings. Indeed, these represent the strongest cases. Table 2 summarises the results and most of the statistical evidence that will be discussed here.

(i), (ii) Gln and Glu have very high factor loadings, and highly significant very strong raw correlations with the binary variable *therm*. The thermophiles have significantly less Gln and more Glu than the mesophiles.

(iii) Val has a moderate-to-high factor loading, and a significant strong raw correlation with *therm*. The thermophiles tend to have more Val than the mesophiles.

Table 2. Statistical evidence sorted by strength

Amino acid	PCA factor loading[a]	Raw correlation[a]	Significance	Raw Δ (S.D.)	Δ (S.D.)	Significance	Statistic
Gln (Q)	–90%	–80%	~10^{-8}	–2.18 (0.31)	–1.76 (0.25)	~10^{-4}	*t*-test
Glu (E)	80%	80%	~10^{-6}	2.27 (0.40)	1.73 (0.31)	~10^{-4}	*t*-test
Val (V)	50 to 65%	60%	~10^{-3}	1.57 (0.42)	1.40 (0.38)	~2×10^{-3}	*t*-test
Thr (T)	–65%	60%	~10^{-3}	–0.84 (0.25)	–1.31 (0.39)	~5×10^{-3}	*t*-test[b]
His (H)	–40 to –60%	–60%	~10^{-3}	–0.44 (0.15)	–1.22 (0.42)	1%[b]	*t*-test[b]
Ser (S)	–30 to –60%[b]	–40%	1%	–1.11 (0.51)	–1.18 (0.54)	5%	*t*-test
Asn (N)	–30 to –40%	–35%	3%	–1.94 (n/a)[b]	–1.05 (n/a)[b]	<2%	Median/Mann–Whitney[b]
Arg (R)	20 to 30%[b]	25%	>5%	>0 (n/a)	>0 (n/a)	1%	Mann–Whitney

For each amino acid, the range of PCA factor loadings for component 2, the raw correlation to the binary variable *therm* and its significance are displayed. The Raw Δ column shows the average difference between thermophiles and mesophiles in raw percentage points; Δ gives the equivalent in standardised scores. The last columns report the significance of the observed difference and the test statistics that have been used.

[a]Approximate figures.

[b]See Appendix for discussion.

n/a, not applicable.

(iv) Thr also has high factor loading, and a significant strong raw correlation with *therm*. The *t*-test indicates that the thermophiles have less Thr than the mesophiles with $P \approx 5 \times 10^{-3}$. While this *t*-test *P* value itself is doubtful, the test can be considered significant, e.g., at the 5% level (see Appendix).

There is moderate evidence for the roles of His, Ser and Asn:

(i) His has a moderate-to-high factor loading and a significant strong raw correlation with *therm*. The *t*-test indicates that the thermophiles have less His than the mesophiles with $P = 1\%$. While this *t*-test *P* value itself is doubtful, the test can be considered significant, e.g., at the 5% level (see Appendix).

(ii) Ser has a low-to-high factor loading, and a significant moderate correlation with *therm*. The higher, rotated factor loadings are doubtful considering its low measure of sampling adequacy (see Appendix). The *t*-test indicates that the thermophiles have less Ser than the mesophiles.

(iii) Asn has a low-to-moderate factor loading and a significant moderate raw correlation with *therm*. The thermophiles seem to have less Asn than the mesophiles (Westenberg–Mood Median test, see Appendix).

Arg had low factor loading and no significant correlation with *therm*. Thermophiles appear to have more Arg than mesophiles at $P = 2\%$ (Mann–Whitney *U*-test). Considering the low factor loading for component 2, this could be due to some other, at present unidentified, factor.

Trends for other amino acids had no backing from post-hoc tests.

DISCUSSION

We have analysed the global amino acid compositions deduced from the completely sequenced genomes of 27 species (Table 1), employing different methods of exploratory data analysis. Our results discern several underlying factors that influence amino acid composition. In this study, the two most prominent observations were the dominant effect of GC pressure and the clear identification of the thermophilic species.

The 'compositional tree' (Fig. 1) is, to our knowledge, the first of its kind, and its structure is affected by a variety of factors. Phylogeny is clearly reflected in the tree, exemplified by the close proximity of the two strains of *H.pylori*, and other related species. A comparison of the amino acid compositions of two sets of homologous proteins can be expected to yield such a result. Indeed, in each pairwise comparison of species, a certain number of homologues and paralogues are included, and similarities in their amino acid compositions can be expected to be higher between closely related species. The tree shown in Figure 1, however, also reflects the influence of numerous genes that are unique to a given species and that, therefore, uniquely contribute to its global amino acid composition.

Besides these general considerations, several specific factors can be distinguished. Hierarchical clustering first grouped organisms with high GC ratio together. Moreover, PCA found GC ratio to be the most important factor. The pronounced effect of GC content on the amino acid composition of encoded proteins has long been recognised (12) and thoroughly studied (13). High GC content will favour certain codons and, thus, the corresponding amino acids. This is not only true for individual proteins or groups of proteins but, as shown here, also for entire genomes. Indeed, GC content appears to be the dominant factor, with a stronger influence on amino acid composition than either phylogeny or adaptation to extreme environments (e.g., thermophily). Again, results from earlier studies (14–18) are corroborated by the analysis reported here.

In this context, we suggest an explanation for the well known positive correlation of Val content with GC ratio, and the lack of a negative correlation for Met, neither of which can be deduced from their codons (Val = GTX, Met = ATG): conservative replacements of Ile (encoded by AT[A/T/C]) forced by positive GC pressure would raise the observed counts of both Val and Met, and could account for the detected correlations (Fig. 3). Moreover, Met is the principal initiation codon, which has to be conserved to maintain the activity of the genes. Thus, ~10% of all Met residues are invariant.

Hierarchical clustering suggested thermophily as the next most important factor. All thermophilic species were clustered together with the exception of *A.pernix*, which has an unusually high GC ratio. PCA has further yielded a clear distinction of all the thermophilic species based on amino acid composition, independent of the effects of GC pressure. Hence, thermophily seems to be the next strongest determining factor for the global amino acid composition of an organism. While only a few specific indications of thermophily are apparent from the genomic sequences themselves (19), amino acid composition clearly groups both thermophilic bacteria and archaea together. This suggests that evolution in a hot environment has strongly selected for a certain average amino acid distribution in thermophiles.

Thermophiles are very divergent, both in terms of their phylogeny and their physiological properties (20). Within the group of thermophiles, one therefore expects to see further differentiation. Indeed we can distinguish by: (i) phylogeny, even after subtracting the effect of high GC ratio (see Fig. 2), the crenarchaeon *A.pernix* is found at a little distance from the other thermophiles; and (ii) metabolism, the methanogens *M.jannaschii* and *M.thermoautotrophicum* are both also excluded from the central core of the cluster (see Fig. 1). Other environmental adaptations would also be expected to play a role (e.g., halophily). A detailed analysis of these further factors and their effects on amino acid composition still requires a higher number of completely sequenced genomes.

The two thermophilic bacteria can be considered to be part of ancient lineages (i.e., to have branched early in evolution) and are generally believed to be closer to archaea. For example, one finds 47% homology between *T.maritima* and archaea versus only 17% between typical bacteria and archaea (21). Their close grouping together in our analysis would support such a deep branching. Therefore, it will be interesting to extend our analysis once completely sequenced non-thermophilic archaea become available.

Performing this analysis only for species with completely sequenced genomes minimises the problem of sampling bias. Along with the small size of data sets, sampling bias has been a major cause of the varying results of studies that compared individual families (22). More recent studies have drawn advantage from comparing individual proteins from one organism to their homologues in a fully sequenced close relative. Two such studies have also examined the total amino acid composition of a set of proteins from *M.jannaschii* and their homologues from mesophilic *Methanococcus* species (1,2). While most of their findings are in agreement, there also are several differences that may be due to incomplete data sets.

An extensive survey of amino acid composition of thermophiles covering *Methanococcus* and *Bacillus* species found significant genus-specific differences (2). Only studies with data covering a broad phylogenic range will be able to discern the many contributing factors to separate common and specific influences on the observed amino acid compositions.

In conclusion, we have analysed the global amino acid compositions deduced from completely sequenced genomes. From this first study it appears that this can be a useful additional tool to search for correlations resulting not only from common descent but moreover from other factors such as thermophily. It will be interesting to extend our analysis as the complete genomic sequences of organisms from different phyla are determined.

APPENDIX

Post-hoc test details

The *t*-test indicated that the thermophiles contained less Thr than the mesophiles with $P \approx 5 \times 10^{-3}$. The group of mesophiles, however, failed Shapiro–Wilk's test for normality at $P = 2\%$. While the *t*-test *P* value itself is therefore doubtful, the test can be considered significant (e.g., at the 5% level) due to the extreme robustness of the *t*-test to violations of the normality assumption. The less powerful Mann–Whitney *U*-test failed to give a significant result.

The *t*-test indicated that the thermophiles had less His than the mesophiles with $P = 1\%$. The group of thermophiles, however, failed both the Kolmogorov–Smirnov and Shapiro–Wilk tests for normality, at $P = 2$ and 3%, respectively. While the *t*-test *P* value itself is therefore doubtful, the test could be considered significant (e.g., at the 5% level) due to the extreme robustness of the *t*-test to violations of the normality assumption. The less powerful Westenberg–Mood median test failed to give a significant result.

Since both the power of the tests and the robustness of the *t*-test to violations of normality improve with sample size, corroboration of the above results for Thr is expected as more completely sequenced genomes become available. The results for His are also likely to be confirmed.

With respective median Asn contents of 3.4 and 5.4%, the thermophiles seemed to have less Asn than the mesophiles at $P = 2\%$ (Westenberg–Mood). The Mann–Whitney *U*-test gave an even smaller *P* at <1%. Looking to establish a difference of means, the *t*-test was not significant ($P > 5\%$), and the group of thermophiles failed the Kolmogorov–Smirnov test for normality at $P = 2\%$. Consequently, Table 2 shows a difference of medians rather than means for Asn. After Bonferroni correction, the tests are significant at $P < 5\%$.

Stability and robustness of hierarchical clustering

Unless specified otherwise, the same results were obtained with or without inclusion of GC ratios as input to clustering.

Using amino acid compositions normalised relative to the 25 reference species (see Materials and Methods) as input, almost all clustering methods yielded the complete group of high GC ratio. (The only exception was single linkage clustering without the GC ratio variable.)

All methods grouped 6–7 thermophiles together, where *A.pernix* was apparently excluded because of its high GC ratio.

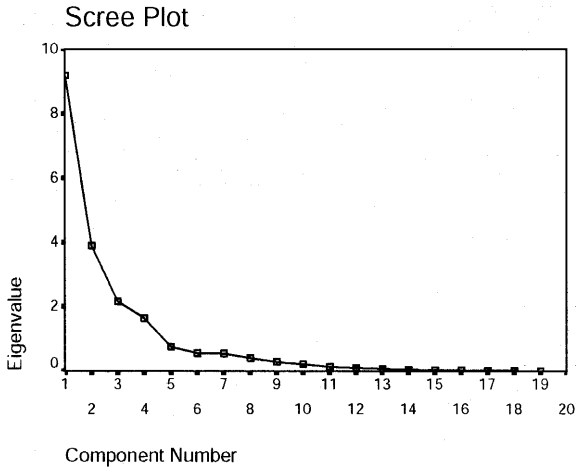

Figure 4. Scree plot of extracted eigenvalues. The eigenvalue, or characteristic root, for a given factor reflects the variance in all the original variables that is accounted for by that factor (10). The first two factors already account for >65% of the variation in the original 20 variables (see Appendix).

Some methods also failed to include either *M.thermoautotrophicum* or *M.jannaschii*. These two species contain more Asp and less Trp than the other thermophiles. The large plasmid pNRC100 from *Halobacterium sp. NRC-1* was sorted as outgroup by all methods except complete linkage.

With amino acid compositions normalised relative to the group of archaea, all methods first grouped all eight thermophiles together. Three of the eight examined methods also grouped *Borrelia burgdorferi* as the most distant entry with the thermophiles. This is the only mesophile with low Gln content. Together with its low His and Thr contents, this can explain the observed grouping. The other five methods gave a single cluster of thermophiles only.

Almost all clustering methods also yielded the complete remaining group of high GC ratio. (The only exception was single linkage clustering without the GC ratio as a variable.) The large plasmid pNRC100 from *Halobacterium sp. NRC-1* was sorted as outgroup by all methods except single linkage.

Stability and robustness of PCA

Great care was taken to ensure and verify the reliability of obtained PCA results. PCA of all 20 variables extracted four components when following the Kaiser–Guttman criterion (23) of retaining only factors with an eigenvalue >1 (Fig. 4). Factors 1 and 2 already explained 65% of the total variance in the original variables. Adding components 3 and 4 (for which we have no conceptual assignment at present) increased this coverage to 84%. Bartlett's test of sphericity is highly significant with $P \approx 10^{-102}$ (statistic for 190 degrees of freedom $\sim 10^3$), and residual correlations are sufficiently small. Despite the vastly grown available sequence data, however, the number of completely sequenced organisms is still rather low for a PCA of 20 variables. It is therefore not surprising to find a low Kaiser–Meyer–Olkin measure of sampling adequacy.

Examination of the measure of sampling adequacy for individual variables detected a small number with poor quality: Ser, Asp, Cys and Met. Excluding these seems to be a reasonable option given that they all had low unrotated factor loadings.

The high rotated factor loadings of Ser appear doubtful considering its low measure of sampling adequacy. Discarding only one of these variables already sufficed to make the overall measure of sampling adequacy acceptable (e.g., when discarding Ser: 58%). Subsequent removal of the others increased it to a respectable 72% (24), with Bartlett's test still being highly significant at $P < 10^{-73}$ (statistic for 120 degrees of freedom >500), and residual correlations being sufficiently small. Consequently, the anti-image matrix also improved, and the coverage of total variance by components 1 and 2 was raised to 78%.

Despite smaller changes to the component matrix the overall picture is robust, i.e., based on the new values one arrives at the same conclusions as before. We therefore opted to deal with the complete, unaltered set of variables.

We found only insignificant correlations between factors after rotation (<20%), and the main factors of interest (the two factors with the highest eigenvalues) only show a correlation of 14%. It is thus reasonably justified to assume orthogonality of factors and use the unrotated results. Two rotation methods preserving orthogonality were also explored. They slightly changed the separation of groups in reduced dimensionality plots, but did not simplify the interpretation of factor loadings for our data. Factor loadings changed only for some variables under rotation. In the discussion of results above we have always reported the observed range of factor loadings.

ACKNOWLEDGEMENTS

We are grateful to Dr G.Kreil for many interesting discussions and ideas, and to J.Vilo for EPCLUST and his friendly help. We thankfully acknowledge assistance from the Computing Services of the University of Cambridge through their facilities. This work was supported by the European Molecular Biology Laboratory. C.A.O. acknowledges support for his laboratory from the European Commission (DGXII—Science, Research and Development), the Medical Research Council (UK) and IBM Research.

REFERENCES

1. Haney,P.J., Badger,J.H., Buldak,G.L., Reich,C.I., Woese,C.R. and 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.
2. McDonald,J.H., Grasso,A.M. and Rejto,L.K. (1999) Patterns of temperature adaption in proteins from *Methanococcus* and *Bacillus*. *Mol. Biol. Evol.*, **16**, 1785–1790.
3. Jaenicke,R. and Böhm,G. (1998) The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.*, **8**, 738–748.
4. Perutz,M.F. (1978) Electrostatic effects in proteins. *Science*, **201**, 1187–1191.
5. Kreil,D.P. and Etzold,T.M. (2000) SRS—access to molecular biological databanks and integrated data analysis tools. In Higgins,D. and Taylor,W. (eds), *Bioinformatics—A Practical Approach.* Oxford University Press, Oxford, UK, pp. 215–241.
6. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol.*, **183**, 63–98.
7. Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.
8. Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.
9. SPSS Inc. (1998) SPSS version 9.0 for Windows. Chicago, USA.
10. Jackson,J.E. (1991) *A User's Guide to Principal Components*. John Wiley & Sons, New York.
11. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
12. Sueoka,N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl Acad. Sci. USA*, **47**, 1141–1149.
13. Lobry,J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**, 309–316.
14. Gu,X., Hewett-Emmet,D. and Li,W.H. (1998) Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica*, **103**, 383–391.
15. Filipski,J. (1990) Evolution of DNA sequence. Contributions of mutational bias and selection to the origin of chromosomal compartments. *Advances in Mutagenesis Research*, Vol. 2. Springer-Verlag, Berlin, pp. 1–54.
16. Benachenhou-Lahfa,N., Labedan,B. and Forterre,P. (1994) PCR-mediated cloning and sequencing of the gene encoding glutamate dehydrogenase from the archaeon *Sulfolobus shibatae*: identification of putative amino-acid signatures for extremophilic adaption. *Gene*, **140**, 17–24.
17. Taupin,C.M.J. and Leberman,R. (1999) Archaeabacterial seryl-tRNA synthetases: Adaption to extreme environments and evolutionary analysis. *J. Mol. Evol.*, **48**, 408–420.
18. Wilquet,V. and Van de Casteele,M. (1999) The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Res. Microbiol.*, **150**, 21–32.
19. Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M. *et al.* (1998) The complete genome of the hyperthermophile bacterium *Aquifex aeolicus*. *Nature*, **392**, 353–358.
20. Stetter,K.O. (1999) Extremophiles and their adaptation to hot environments. *FEBS Lett.*, **452**, 22–25.
21. Nelson,K.E., Clayon,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A. *et al.* (1999) Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.
22. Böhm,G. and Jaenicke,R. (1994) On the relevance of sequence statistics for the properties of extremophilic proteins. *Int. J. Pept. Protein Res.*, **43**, 97–106.
23. Guttman,L. (1953) Image theory for the structure of quantitative variables. *Psychometrica*, **18**, 277–296.
24. Kaiser,H.F. and Rice,J. (1974) Little jiffy, mark IV. *Educ. Psychol. Meas.*, **34**, 111–117.