Université d'Ottawa · University of Ottawa

A Comparative Analysis of Synonymous Codon Usage Patterns in Forty Completely Sequenced Bacterial Genomes.

By

David J. Lynn

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

University of Ottawa

In partial fulfillment of the requirements for the

M.Sc. degree in the

Ottawa-Carleton Institute of Biology

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A3 – Adenine content at the third codon positions.

T3 – Thymine content at the third codon positions.

G3 – Guanine content at the third codon positions.

C3 – Cytosine content at the third codon positions.

GC3 – Guanine + Cytosine content at the third codon positions.

GT3 – Guanine + Thymine content at the third codon positions.

CT3 – Cytosine + Thymine content at the third codon positions.

bp – Base pair

kb – Kilo base

LDA – Linear Discriminant Analysis

NCBI – National Centre for Biotechnology Information

NS – Non-significant

ORF – Open Reading Frame

RSCU – Relative Synonymous Codon Usage

Acknowledgements

I would like to acknowledge the contribution that Greg Singer and Somayeh Nourian have made to this study. I would also like to thank Dr. Andrew Lloyd, Trinity College Dublin, for all his helpful comments on this study and Dr. Cliona O'Farrelly of the Educational and Research Centre, St. Vincent's University Hospital, Dublin for her support. Thanks to my parents without whom none of this would have been possible and also to my girlfriend Jenny for sticking with me while I was in Canada. Last but not least I would like to thank my supervisor Dr. Donal Hickey for all his help, above and beyond the call of duty.

Abstract

The complete genomic sequences of a large number of bacteria which are now publicly available presented the opportunity to investigate the major factors shaping synonymous codon usage variation within and among these species. To do this, we analysed synonymous codon usage in forty completely sequenced bacterial genomes, both inter- and intra-genomically, using correspondence analysis. Within each genome, we investigated the importance of various mutational biases, including compositional bias and strand bias, in shaping differences in codon usage between genes within a particular species. We found that translational selection plays a crucial role in determining synonymous codon usage in almost all the genomes investigated. The overall picture that is emerging from this study is that no one factor but, rather, a combination of factors influence codon usage in a genome to produce a pattern that is unique to that species.

Analysis of synonymous codon usage in a combined dataset of all genes from the forty different species also resulted in some interesting findings. We found that much of the variation among species is related to genomic G+C content. We observed a difference in synonymous codon usage between thermophilic bacteria and non-thermophiles, which is not related to G+C content. This variation is primarily due to a differing usage of arginine and isoleucine codons in the two groups.

Résumé

Les séquences complètes du génome d'un grand nombre de bactéries qui sont présentement disponibles offrent l'opportunité d'étudier le fonctionnement des principaux facteurs qui déterminent la variation des codons synonymes au sein et parmi ces espèces. Nous avons eu recours à l'analyse de correspondance pour comprendre le fonctionnement des codons synonymes de quarante séquences de génomes bactériens, à la fois au sein de chaque génome et à travers l'ensemble des génomes étudiés. Au sein de chaque génome, nous avons déterminé l'importance des différents biais mutationnels, dont les biais par composition et les biais de brins, en identifiant les rôles des codons des différents gènes d'une espèce donnée. Nous avons montré que la sélection traductrice joue un rôle crucial dans la détermination des rôles des codons synonymes chez presque tous les génomes étudiés. L'image d'ensemble qui se dégage de cette étude est que ce n'est pas un facteur unique, mais une combinaison de facteurs qui amènent un codon à s'exprimer selon un modèle unique à cette espèce.

L'analyse du rôle des codons synonymes de l'ensemble des gènes des quarante espèces étudiées apporte également des données intéressantes. Nous avons montré que la plupart des variations d'une espèce à l'autre est une fonction de la teneur en G + C du génome. Nous avons aussi observé une différence entre les bactéries thermophiles et non-thermophiles, qui n'est pas fonction de la teneur en G + C du génome. Cette variation est avant tout

attribuable à l'usage différent des codons de l'arginine et de l'isoleucine dans les deux groupes.

Introduction

1.1 Background

It has been long established that usage of synonymous codons between different organisms is non-random and that genes within a particular genome have a broadly similar pattern of codon usage (Grantham *et al.*, 1980). However, significant within-species variation is also evident. For instance, in *Escherichia coli*, the first species to have its codon usage extensively studied, synonymous codon usage was discovered to be primarily governed by translational selection for those 'optimal' codons, which most efficiently recognize the most abundant tRNA species (Ikemura, 1981). Since these early studies it has become widely accepted that highly expressed genes have relatively high frequencies of the optimal codons, while other genes, expressed at lower levels, have codon usage patterns that are primarily determined by the mutational bias (G+C bias) of that particular organism (Bulmer, 1991). This pattern of codon usage has been reported for a number of prokaryotic genomes, such as *Bacillus subtilis* (Shields and Sharp, 1987), *Haemophilus influenzae* (McInerney, 1997) and *Mycobacterium tuberculosis* (Andersson and Sharp, 1996a) and also in the eukaryotes, *Saccharomyces cerevisiae* (Ikemura, 1982), *Drosophila melanogaster* (Shields *et al.*, 1988) and *Caenorhabditis elegans* (Stenico *et al.*, 1994), although the specific optimal codons used differ depending on the organism.

Codon usage in organisms with highly biased base contents, such as *Streptomyces* (Wright and Bibb, 1992), *Mycoplasma capricolum* (Ohkubo *et al.*, 1987) and *Micrococcus luteus* (Ohama *et al.*, 1990), has previously, been believed to be determined by mutational bias, with little or no translational selection. A Similar situation has been reported in the A+T rich genome of *Rickettsia prowazekii* (Andersson and Sharp, 1996b). In these species the mutational bias is so strong that any affect due to translational selection appears to be swamped.

Until recently, translational selection and mutational bias were believed to be the most important forces influencing codon usage in prokaryotic genomes, with factors such as gene length (Eyre-Walker, 1996; Powell and Moriyama, 1997), hydrophobicity (de Miranda *et al.*, 2000) and amino acid conservation (de Miranda *et al.*, 2000) playing less significant roles. The role of other factors in shaping codon usage is now becoming evident. In a number of bacterial genomes, the leading strand of replication has been found to be G+T rich in comparison to the lagging strand (Perriere *et al.*, 1996; Francino and Ochman, 1997; McLean *et al.*, 1998). This strand specific base composition asymmetry is commonly referred to as strand bias. Although the mechanisms that create strand bias are not fully understood, two major hypotheses have been proposed. The first focuses on replication associated asymmetries (Francino and Ochman, 1997) while the second hypothesis argues that strand mutational biases are generated primarily during transcription and transcription-coupled repair (Mrazek and Karlin, 1998).

Multivariate statistical analysis of codon usage in *Borrelia burgdorferi* has revealed that the genes can be divided into two distinct groups, genes located on the lagging strand and genes located on the leading strand. Chi-square analysis revealed that there were significant differences in codon usage between these two groups, such that leading strand genes used significantly more codons ending in G or U and lagging strand genes used more codons ending in A or C leading to the conclusion that strand bias is the major cause of codon usage variation in this genome (McInerney, 1998). Similar results have been reported for *Treponema pallidum* (Lafay *et al.*, 1999) and *Chlamydia trachomatis* (Romero *et al.*, 2000)

Previous studies of codon usage have tended to focus on single genomes or even a small subset of genes from one species. The large number of completely sequenced bacterial genomes now available presented the opportunity to investigate synonymous codon usage variation in a large number different species simultaneously. Although our investigations were primarily data driven, the objective of the study was to provide an overview of synonymous codon usage variation not only among the genes in each genome but also among the genomes themselves, by carrying out a multivariate statistical analysis of all the genes from each of the genomes as one data set. Due to the large size of our data set we hoped to be able to tease out trends in the data that may have been near impossible to find on a smaller scale or by comparing one genome at a time. We wanted to uncover the affects that strand bias, compositional bias and translational selection have on determining synonymous

codon usage variation within and among each of the forty genomes. We were also interested in determining what characteristics were responsible for different bacteria sharing similar codon usage patterns.

1.2 Natural Selection on Codon Usage:

1.2.1 Codon Usage Bias and tRNA abundance

In 1980, Grantham *et al* discovered that codon choices among genes of the same genome are non-random. This finding led them to propose the Genome Hypothesis that "each gene in a genome tends to conform to its species' usage of the codon catalog" (Grantham *et al.*, 1980). Early studies of codon usage focused on the *E.coli* and *S.cerevisiae* genomes. In these species a bias towards a subset of codons for each amino acid has been found, although the subset of biased codons is different in the two genomes. A strong correlation between the occurrence of these codons and the most abundant tRNA species was uncovered. This pattern of non-random codon usage suggested that selection at the level of translation was influencing codon choice (Ikemura, 1981; Ikemura, 1982). Since these 'optimal' codons were recognized by the tRNA species present at the highest concentrations, this would improve the efficiency and speed of translation.

Since these early studies correlations between codon usage and tRNA abundance have been reported in numerous genomes. In the prokaryotes,

Salmonella typhimurium and *M.capricolum* tRNA abundance was quantified by two-dimensional gel electrophoresis. Synonymous codon choice was correlated with the relative amount of the isoacceptor tRNAs in these species (Ikemura, 1985; Yamao *et al.*, 1991). Analysis of codon usage in the complete genome of the bacteriophage T7 revealed that codon usage was influenced by host tRNA abundance (Sharp *et al.*, 1984).

More recently, cellular levels of individual tRNAs have been quantified in the completely sequenced genome of *B.subtilis* and again an obvious relationship between tRNA abundance and synonymous codon choice was found (Kanaya *et al.*, 1999). Interestingly, tRNA levels were found to be proportional to the copy number of the respective tRNA genes, demonstrating a gene-dosage effect on the levels of tRNA. This allowed for the investigation of codon usage in seventeen other completely sequenced genomes, for which tRNA levels had not been quantified. Codon usage bias in these organisms was determined to be related to the level of optimal codon use, predicted by tRNA gene copy number (Kanaya *et al.*, 1999). In the eukaryotes, *D.melanogaster* and *C.elegans* similar results have also been found (Moriyama and Powell, 1997; Duret, 2000).

1.2.2 Codon Usage Bias and Expression Level

Following the Genome Hypothesis it quickly became apparent that there was also considerable variation in codon usage among genes from the same

genome. Multivariate analysis of thirteen highly and sixteen lowly expressed genes in *E.coli* revealed notable variation in codon choice between the two groups (Grantham *et al.*, 1981). A later study of 83 *E.coli* genes uncovered a strong correlation between codon composition and mRNA expressivity, such that the frequency of optimal codons was higher in genes expressed at high levels (Gouy and Gaultier, 1982). Another investigation of codon usage in 165 *E.coli* genes confirmed that there was "a consistent trend of increasing bias with increasing gene expression level" (Sharp and Li, 1986). Implementation of factorial correspondence analysis on 780 genes from the *E.coli* genome revealed that the genes clustered into three classes (Médigue *et al.*, 1991). Class one consisted of genes that were lowly expressed, class two consisted of highly expressed genes, while class three were mostly genes of foreign origin. Again the highly expressed genes had the highest frequencies of optimal codons.

An early study of codon usage in 56 *B.subtilis* genes using correspondence analysis identified certain genes with a high codon bias. These genes were the very highly expressed genes (Shields and Sharp, 1987). Since then the complete genome sequence has become available and correspondence analysis of codon usage has revealed similar results to *E.coli*. There are three categories of genes which correspond to the same classes as identified in *E.coli*, lowly expressed genes with low codon bias, highly expressed genes with high bias, and genes of foreign origin (Kunst *et al.*, 1997; Moszer, 1998). A similar relationship between codon bias and expression level has been

reported for a number of other prokaryotes, such as *Corynebacteria* (Malumbres *et al.*, 1993), *Lactobacilli* (Pouwels and Leunissen, 1994), *M. tuberculosis* (Andersson and Sharp, 1996; Pan *et al.*, 1998), *H.influenzae* (Pan *et al.*, 1998) and *Mycobacterium leprae* (de Miranda *et al.*, 2000).

The relationship between codon bias and gene expression has also been found in fungi. Cluster analysis of relative synonymous codon usage in 110 *S.cerevisiae* genes revealed two distinct groups of genes, one of which had the most extreme codon bias and were highly expressed. Correspondence analysis of synonymous codon usage in *Aspergillus nidulans* (45 genes), *Candida albicans* (28 genes), and *Kluyveromyces lactis* (47 genes) identified a single major trend in each of the datasets (Lloyd and Sharp, 1991; Lloyd and Sharp, 1992; Lloyd and Sharp, 1993; respectively). At one end of this trend were the lowly expressed genes and at the other were the highly expressed genes that were highly biased towards usage of the optimal codons. Similar analyses in the eukaryotes, *Dictyostelium discoideum* (Sharp and Devine, 1989), *C.elegans* (Stenico *et al.*, 1994; Duret, 2000), *D.melanogaster* (Shields *et al.*, 1988), *Plasmodium falciparum* (Musto *et al.*, 1999), *Giardia lamblia* (Lafay and Sharp, 1999) and *Entamoeba histolytica* (Romero *et al.*, 2000) has also revealed a correlation between the frequency of optimal codons and gene expression level.

1.2.3 Translation Efficiency

The Translational Efficiency Hypothesis proposes that natural selection favors a codon usage pattern that increases the rate of protein synthesis and thus a maximal growth rate (Xia, 1998). This model is supported by a number of different lines of evidence. The correlation between the frequency of optimal codons and tRNA abundance and the fact that highly expressed genes tend to have a higher bias towards using the optimal codons are both evidence of selection for translational efficiency (see above). Other experimental evidence has come from a variety of sources.

It was concluded, from investigation of highly expressed coding regions in the bacteriophage MS2 and in E.coli, that the optimal codons were those that optimized the codon-anticodon interaction energy and that this was part of a strategy to maximize the efficiency of translation (Grosjean and Fiers, 1982). Similarly, it was found that codons that were used very frequently in highly expressed genes in *E.coli* select aminoacyl-tRNAs more rapidly than do rarely used codons (Curran and Yarus, 1989). This suggested that the speed of tRNA selection was a determining factor in biasing synonymous codon usage.

By inserting synthetic oligonucleotides into a highly expressed gene in *E.coli* it was shcwn that the maximum level of translation could be reduced by unfavorable codon usage (Robinson *et al.*, 1984). A similar experiment in the lacZ gene of *E.coli* revealed a reduction in the rate of translation that was approximately six-fold between the wildtype gene and a gene that was altered to

use infrequent codons (Sorensen *et al.*, 1989). It has also been determined that tRNA species that recognize the optimal codons increase in *E.coli* as growth rate increases. This is evidence that codon bias towards optimal codons is a strategy to support a maximal growth rate through translational efficiency (Emilsson and Kurland, 1990; Dong *et al.*, 1996; Berg and Kurland, 1997).

1.2.4 Translational Accuracy

If selection biases codon usage to enhance translational accuracy then selection should be more evident at codons encoding amino acids that are functionally important within a protein (Akashi *et al.*, 1998). In a study of 38 *Drosophila* genes compared between three species, the frequency of optimal codons was found to be significantly higher at codons conserved for amino acids than at nonconserved codons. Furthermore, optimal codons were more frequent in the conserved zinc-finger and homeodomain regions than in the rest of 28 transcription factor genes (Akashi, 1994). A comparison between 548 genes in *C.elegans* and *Homo sapiens* revealed similar results (Marais and Duret, 2001). However, in *E.coli* this pattern of codon usage has not been found (Hartl *et al.*, 1994).

A positive correlation between synonymous codon usage bias and gene length in *E.coli* has been shown in genes of similar expression levels. It has been proposed that since the cost of producing a protein is proportional to its length, selection for codons that improve accuracy should be stronger in longer

genes (Eyre-Walker, 1996). In the eukaryotes, however, the opposite effect has been found casting doubt on this theory (Duret and Mouchiroud, 1999; Moriyama and Powell, 1998).

1.3 Mutational Bias on Codon usage

The G+C content of bacterial genomes is widely known to vary across taxa. It has been proposed that this compositional bias is due to mutation rates (G+C to A+T) and (A+T to G+C) that are not equal and are species-dependent (Sueoka, 1988). G+C content can also vary between genes within the same genome. This may be due to unequal mutation rates driven by variable of nucleotide pools during DNA synthesis (Wolfe, 1991) differences in the rate of DNA damage across species or differences in the rate and efficiency of DNA repair (Martin, 1995).

Proposals that selection may be responsible for composition bias (Bernardi and Bernardi, 1986), such that an increase in G+C content would allow for survival at higher temperatures are no longer widely accepted, since organisms that live at similar temperatures may have very different G+C contents (Martin, 1995)

1.3.1 Mutation Selection Drift

As discussed above, highly expressed genes in many species tend to have high frequencies of optimal codons and thus a highly biased codon usage pattern. But what of the lowly expressed genes? Lowly expressed genes in *B.subtilis* were found to have nucleotide frequencies that were similar among different codon positions and on complementary strands. It was proposed that in these genes where translational selection was relaxed, the codon usage pattern

largely reflected the mutational bias present in the genome (Shields and Sharp, 1987). In a study of 145 yeast genes and 339 *E.coli* genes, weakly expressed genes also showed the effects of mutational bias (Bulmer, 1990). To account for this pattern of codon usage the Selection-Mutation-Drift Theory of synonymous codon usage was proposed (Bulmer, 1991). This theory posits that codon usage patterns in unicellular organisms are due to a "balance in a finite population between selection favoring an optimal codon for each amino acid and mutation together with drift allowing the persistence of non-optimal codons". Selection is likely to be stronger in highly expressed genes because these are translated more often (Bulmer, 1991). Patterns of synonymous codon usage consistent with this theory have been described in the prokaryotes *M.tuberculosis* (Andersson and Sharp, 1996a; Pan *et al.*, 1998), *H.influenzae* (Pan *et al.*, 1998) and *M.leprae* (de Miranda *et al.*, 2000); the fungi *A.nidulans* (Lloyd and Sharp, 1991), *C.albicans* (Lloyd and Sharp, 1992) and *K.lactis* (Lloyd and Sharp, 1993) and in the other eukaryotes, *C.elegans* (Stenico *et al.*, 1994), *D.melanogaster* (Shields *et al.*, 1988), *P.falciparum* (Musto *et al.*, 1999), *G.lamblia* (Lafay and Sharp, 1999) and *E.histolytica* (Romero *et al.*, 2000). The situation in vertebrate genomes is very different. In these genomes there are large regions of different base composition, termed isochores and as a result codon usage between genes is highly variable (For review see Bernardi, 2000).

1.3.2 Codon Usage in Highly Biased Genomes

In organisms with highly biased base contents, such as the G+C rich *Streptomyces* (Wright and Bibb, 1992) or *Mycoplasma capricolum* (Ohkubo, *et al.*, 1987) codon usage is proposed to be predominantly determined by mutational bias, with little or no translational selection. *M.capricolum* has a genomic G+C content of only 25% and this bias is so strong that 93% of the codons end in A or U. In the other extreme *Micrococcus luteus* has a genomic G+C content of 74% and 95% of codons end in G or C (Ohama *et al.*, 1990). A Similar situation has been reported in the A+T rich genome of *Rickettsia prowazekii* (Andersson and Sharp, 1996b). In these species the compositional bias is such that any affect due to translational selection appears to be swamped. It should be noted that these results are based on the small number of gene sequences available at the time.

1.3.3 Strand-Specific Mutational Bias

In 1995 the first complete prokaryotic genome sequence of *Haemophilus influenzae* was published (Fleishmann *et al.*, 1995). Previous reports of asymmetric mutational pressure, such that the G+T and A+C contents of one strand were not equal to 50%, in the SV40 virus, polyomavirus and in the mitochondrial DNA from a number of eukaryotes, prompted investigation of the available prokaryotic sequences. In 1996, G+C skew and A+T skew analysis of the complete *H.influenzae* genome and the partial genome sequences of *E.coli*

and *B.subtilis* were published (Lobry, 1996a). G+C skew, the quantity (G – C)/(G + C), and AT skew, the quantity (A – T)/(A + T), was measured around the three genomes using a sliding window. All three genomes were found to have significant switches in G+C and A+T skews at the origin of replication, although the G+C skew was observed to be considerably stronger than the A+T skew. The direction of the skew was such that the leading strand of replication was richer in G+T than the lagging strand. When only the intergenic regions were considered, where selection pressures should be at a minimum, the intensities of both the G+C and A+T skews were seen to increase, indicating that the asymmetric substitution was due to mutation. When coding sequences were analysed, the intensity of the skews was lower, but were highest at the first and third codon positions, again suggestive of a mutational bias. The complete genome sequence of *Mycoplasma genitalium* became available soon after the *H.influenzae* sequence was released and was demonstrated to exhibit a strong G+C skew and a weaker A+T skew (Lobry, 1996b) This analysis was used to confirm that the origin of replication was located, as had been predicted, between dnaA and dnaN (Fraser *et al.*, 1995)

The complete genome sequences of *E.coli* and *B.subtilis* were published the year following Lobry's analysis (Blattner *et al.*, 1997) (Kunst *et al.*, 1997). Examination of the complete *E.coli* genome revealed that the leading strand was significantly richer in G (26.22%) than C (24.58%) and there was slightly more T (24.69%) than A (24.52%). A sharp change in the sign of the G+C skew at the origin and terminus of replication was observed, confirming Lobry's analysis of

the partial genome sequence. The publishers of the *B.subtilis* genome sequence also reported a similar nucleotide compositional asymmetry between the leading and lagging strands, and found that the GC skew inverted at the origin of replication.

G+C skew analysis was also performed on the genomic sequence of the spirochaete *Borrelia burgdorferi*, the causative agent of Lyme disease (Fraser, 1997). It was demonstrated that the G+C skew is uniformly negative from 0 to 450kb and uniformly positive from 450kb to the end of the chromosome, the switch in skew located at the putative origin of replication. Recently, nascent DNA strand analysis was used to physically map the *B.burgdorferi* origin to a 240bp sequence between dnaA and dnaN, where the switch in G+C skew occurs (Picardeau *et al.*, 1999).

To determine how general the asymmetrical strand biases were, 12 complete prokaryotic genomes were analysed, using consistent methods for each genome to allow for comparisons to be made between them (McLean *et al.*, 1998). Previous to this, there was little consistency in the G+C skew analysis of the different genomes. Nine eubacteria, *B.subtilis, B.burgdorferi, E.coli, H.influenzae, Mycoplasma pneumoniae, M.genitalium, T.pallidum, Helicobacter pylori* and *Synechocystis sp.* and three archaea, *Archeoglobus fulgidus, Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii*, were analysed. The G+C skew in each genome was evaluated using a large window size of 300,000 nucleotides and concentrated on the third codon positions, which are more likely to show mutational influences. A strong G+C skew was

demonstrated in eight eubacteria, only *Synechocystis* and the three archaebacteria did not exhibit one. A strong A+T skew was found in six of the eubacteria. *H.pylori* and *H.influenzae* exhibited a weak A+T skew and *Synechocystis* showed none. In all genomes with strong skews the skews switched sign at the probable origin and terminus of replication, such that the leading strand was G+T rich in most eubacteria. The only exceptions to this were *M.genitalium* and *M.pneumoniae* which were C+T rich. It was proposed that the pressure which creates the skew was independent of the pressures that determine the G+C content, since the two spirochaetes, *B.burgdorferi* and *T.pallidum*, have similar skew patterns, yet have the lowest and highest G+C contents, respectively, of all the genomes analysed (McLean *et al.*, 1998).

Around the time that the paper above was published, similar analysis was carried out on ten complete prokaryotic sequences, ten complete herpesvirus genomes and on other large viral and phage genomes (Mrazek and Karlin, 1998). G+C skew was assessed using a 50kb sliding window, and they reported similar results to McLean *et al*, as regards the bacterial genomes. The *E.coli, B.subtilis* and *M.genitalium* genomes all exhibited a strong GC skew, the *H.influenzae*, *H.pylori* and *M.pneumoniae* genomes all showed a marginal G+C skew, while *Synechocystis sp., A.fulgidus, M.thermoautotrophicum* and *M.jannashii* were observed to have an irregularly fluctuating G+C skew around the genome. Of the ten complete herpesvirus genomes, only three, HHV6, HHV7 and HCMV, had a strand compositional asymmetry, which changed at the

lytic origin of replication (*ori*L). A number of other long viral genomes analysed did not reveal clear strand asymmetry over large portions of the genomes.

Recently, fifteen complete prokaryotic genomes have been analysed for compositional asymmetry in genes at the level of nucleotides, codons and amino acids. The same twelve genomes analysed previously (McLean *et al.*, 1998), and the complete genomes of *Aquifex aeolicus*, *C.trachomatis*, and *M.tuberculosis* were investigated through the use of a classical statistical tool, linear discriminant analysis (LDA) (Rocha *et al.*, 1999). A bias acting at the level of nucleotides, codons and amino acids was demonstrated in nine of the species. *M.jannaschii, M.genitalium* and *M.pneumoniae* revealed ambiguous plots, while *A.aeolicus, A.fulgidus* and *Synechocystis sp.* showed no significant bias at all. G+C skew has also been reported for the *Xyllela fastidiosa* genome (Gautier, 2000).

A new method of detecting G+C skew has been proposed, known as cumulative G+C skew, which is the sum of the G+C skew in adjacent windows from an arbitrary start to a given point in a sequence (Grigoriev, 1998). This method has been demonstrated to reveal polarity switches in G+C skew which have previously been difficult to detect with the sliding window method, such as in the *M.pneumoniae* genome. Analysis of fourteen microbial genomes by this method, demonstrates a leading strand G+C skew in twelve of the genomes, with characteristic V-shaped cumulative G+C skew diagrams, indicative of bi-directional replication between a singular *ori* and *ter*. In all twelve cases the minimum of these diagrams is located at the origin, while the maximum

coincides with the terminus in eleven cases, the exception being *B.subtilis*. It is proposed that local extremities in the cumulative G+C skew diagrams may represent recent sequence inversions or the integration of foreign DNA into the chromosome (Grigoriev, 1998). In the archaea conventional G+C skew analyses have failed to identify strand asymmetry. By applying the cumulative technique G+C skew has been demonstrated in *M.thermoautotrophicum* (Grigoriev, 1998; Lopez *et al.*, 1999), *Pyrococcus horikoshii* (Lopez *et al.*, 1999), *Pyrococcus abyssi* and *Pyrococcus furiosus* (Myllykallio *et al.*, 2000).

Strand asymmetry has also been reported in organelle genomes. The two strands of the chloroplast genome of the green alga *Euglena gracilis*, have been shown to exhibit a strand asymmetry that switches at the origin of replication and at a location halfway around the genome, such that the leading strand is G+T rich (Morton, 1999). Asymmetry has also been demonstrated in 25 complete mammalian mitochondrial genomes, such that the transcribed heavy strand is G rich compared to the non-transcribed light strand.

1.3.4 Causes of Strand Compositional Asymmetry

Although the mechanisms that create strand compositional asymmetry are a long way from being completely understood, two papers have reviewed the most plausible hypotheses (Mrazek and Karlin, 1998) (Francino and Ochman, 1997). Two principle mechanisms have been proposed to explain strand compositional asymmetry. The first mechanism focuses on replication associated asymmetries

while the second hypothesis argues that strand mutational biases are generated primarily during transcription and transcription-coupled repair.

Mrazek and Karlin (1998) have put forward a number of replication related theories such as different mutational rates between leading and lagging strands, enzymological asymmetry and replication fork asymmetry. Replication itself is an asymmetrical process, the leading strand being replicated continuously, while the lagging strand is replicated via short Okazaki fragments (Frank and Lobry, 1999). At least one additional enzyme is required to synthesize the lagging strand – DNA primase, which is required to synthesis the RNA primers necessary for the synthesis of the Okazaki fragments. The use of different enzymes, to synthesize leading and lagging strands, has been proposed to allow for variation in error rates between strands (Mrazek and Karlin, 1998). In addition to this, the replication fork is structurally asymmetrical, such that leading strand replication proceeds by unwinding very short templates, while lagging strand replication involves the exposure of long single-stranded regions, which are susceptible to mutation and may facilitate primer-template misalignments (Francino and Ochman, 1997).

The second major hypothesis is that transcriptional effects can account for DNA strand asymmetry (Francino and Ochman, 1997). There are two transcription-dependent processes that could result in different mutation rates on the transcribed and non-transcribed strands, transcription–coupled repair and deamination. Transcription-coupled repair is a process that corrects lesions in the transcribed strand of expressed genes and has been characterised in both

prokaryotic and eukaryotic systems (Frank and Lobry, 1999). Bulky lesions, such as pyrimidine dimers, cause the RNA polymerase to stall on the template strand. These stalled polymerases are recognized by a transcription-repair coupling factor, which promotes the activity of nucleotide-excision-repair enzymes to remove the lesion. This only occurs on the template strand, which therefore results in an asymmetry between the two strands. In addition to transcription-coupled repair, another transcription related mechanism, deamination, can also contribute to the strand asymmetry. While RNA is synthesized on the transcribed strand, a portion of the other strand is single-stranded and is therefore prone to deamination. C deaminates to U over 100 times faster in single-stranded DNA, contributing to the strand compositional asymmetry (Francino and Ochman, 1997). In summary, it seems most likely that both replicational and transcriptional associated mutational pressures lead to the strand compositional asymmetries that have been observed in the genomes previously discussed.

If genes were orientated randomly between the leading and lagging strands, then the asymmetric effects on the coding and non-coding strands would cancel each other out. It has been observed, however, that the majority of genes in a genome are located on the leading strand (Brewer, *et al.*, 1988; McLean *et al.*, 1998). It has been proposed that there is a selective advantage for the transposition of genes to the leading strand from the lagging strand, at the level of replication, since head-on collisions between the RNA and DNA polymerases results in a slower replication rate in a gene that is transcribed in

the opposite direction to replication, i.e. genes on the lagging strand (Brewer *et al.*, 1988; French, 1992). The other selective advantage, at the level of transcription, is to maintain most of the highly expressed genes on the leading strand (McLean *et al.*, 1998; McInerney, 1998). Replication forks can proceed passively behind the transcription complex in genes located in the direction of replication but are disrupted on the lagging strand. Thus, genes that are transcribed more often have a selective advantage to be located on the leading strand.

1.3.5 Strand Compositional Asymmetry and Codon Usage Bias

Until recently, the combined influences of G+C base compositional bias and the effects of translational selection have been considered to be the most important factors to affect codon usage variation. The first indication that the selection-mutation model might not be able to always explain the variation in codon usage in prokaryotes came with the publication of the *M.genitalium* genome. The G+C base composition along the genome of *M.genitalium* varies from an average of about 17% G+C at one end of the genome to about 34% G+C at the other end, with an accompanying variation in codon usage (Kerr *et al.*, 1997). It was suggested this phenomenon was linked to replication.

The first time that strand asymmetry was demonstrated to be the most important cause of codon usage variation in an organism was in the *B.burgdorferi* genome (McInerney, 1998). Correspondence analysis (see

materials and methods section) was performed on the relative synonymous codon usage (RSCU) values of all the potential and known ORFs in the *B.burgdorferi* genome. A plot of the two most important axes after correspondence analysis resulted in the separation of the genes into two clusters along axis 1, the axis which explains the most amount of variation in the dataset. The two clusters divided the genes into two groups, genes located on the lagging strand and genes located on the leading strand. Chi-square analysis revealed that there were significant differences in codon usage between genes located on the lagging strand and genes located on the leading strand, such that leading strand genes used significantly more codons ending in G or U and lagging strand genes used more codons ending in A or C (McInerney, 1998).

A similar analysis was carried out on the genomes of *B.burgdorferi* and *T.pallidum* and it was determined again that the primary influence on codon usage in these genomes is whether a gene is transcribed in the same direction as replication, or not (Lafay *et al.*, 1999). Correspondence analysis of RSCU values for 1881 genes pooled from both species revealed two major trends in codon usage. Firstly, there was a difference in codon usage between species, but, more interestingly, the second most significant trend in the dataset was to separate the genes into two clusters, leading strand genes and lagging strand genes, in each species. The separation was less pronounced in the *T.pallidum* genome, but obvious nonetheless. Chi-square analysis revealed that leading strand genes utilize significantly more codons ending in G or U, while lagging

strand genes use more codons ending in A or C, in accordance with the strand compositional asymmetry.

Linear discriminant analysis has recently been used to demonstrate that in nine out of twelve complete prokaryotic genomes, the leading strand is biased towards codons ending in G or U, the most extreme cases being *B.burgdorferi, T.pallidum and C.trachomatis* (Rocha *et al.*, 1999). There is also some evidence of strand asymmetry affecting codon usage in the chloroplast genome of *E.gracilis* (Morton, 1999).

2.0 Materials and Methods

Nucleotide sequences and annotation tables of forty complete bacterial genomes were downloaded from the National Center for Biotechnology Information (NCBI) ftp site (ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria). The genomic sequences of eight archaeabacteria and thirty-two eubacteria were used in the analyses: *Aeropyrum pernix* (Kawarabayasi *et al.*, 1999), *Aquifex aeolicus* (Deckert *et al.*, 1998), *Archaeoglobus fulgidus* (Klenk *et al.*, 1997), *Bacillus halodurans* (Takami *et al.*, 2000), *Bacillus subtilis* (Kunst *et al.*, 1997), *Borrelia burgdorferi* (Fraser *et al.*, 1997), *Buchnera sp. APS* (Shigenobu *et al.*, 2000), *Campylobacter jejuni* (Parkhill *et al.*, 2000a), *Chlamydophila pneumoniae AR39* (Read *et al.*, 2000) *Chlamydophila pneumoniae CWL029* (Kalman *et al.*, 1999), *Chlamydophila pneumoniae J138* (Shirai *et al.*, 2000), *Chlamydia trachomatis* (Stephens *et al.*, 1998), *Chlamydia muridarum* (Read *et al.*, 2000), *Deinococcus radiodurans* (White *et al.*, 1999), *Escherichia coli K-12* (Blattner *et al.*, 1997), *Escherichia coli 0157:H7* (Perna *et al.*, 2001), *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Halobacterium sp.* (Ng *et al.*, 2000), *Helicobacter pylori 26695* (Tomb *et al.*, 1997), *Helicobacter pylori J99* (Alm *et al.*, 1999), *Lactococcus lactis* (Bolotin *et al.*, 2001), *Methanobacterium thermoautotrophicum* (Smith *et al.*, 1997), *Methanococcus jannaschii* (Bult *et al.*, 1996), *Mycobacterium tuberculosis* (Cole *et al.*, 1998), *Mycoplasma genitalium* (Fraser *et al.*, 1995), *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), *Neisseria meningitidis MC58* (Tettelin *et al.*, 2000), *Neisseria meningitidis*

Z2491(Parkhill *et al.*, 2000b), *Pasteurella multocida* (May *et al.*, 2001), *Pseudomonas aeruginosa* (Stover *et al.*, 2000), *Pyrococcus abyssi* (Heilig, unpublished), *Pyrococcus horikoshii* (Kawarabayasi *et al.*, 1998), *Rickettsia prowazekii* (Andersson *et al.*, 1998), *Synechocytis sp. PCC6803* (Kaneko *et al.*, 1996), *Thermoplasma acidophilum* (Ruepp *et al.*, 2000), *Thermotoga maritima* (Nelson *et al.*, 1999), *Treponema pallidum* (Fraser *et al.*, 1998), *Ureaplasma urealyticum* (Glass *et al.*, 2000), *Vibrio cholerae* (Heidelberg *et al.*, 2000) and *Xylella fastidiosa* (Simpson *et al.*, 2000). The GenBank accession numbers, G+C contents and optimal temperatures for the forty species used in the analyses are shown in Table 1 (see end of this section).

Codon bias was measured using the relative synonymous codon usage (RSCU) values, which were calculated using the codonW program written by John Peden, University of Nottingham, which is available at ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z. The RSCU value for a codon is a measure of its usage, relative to other codons for the same amino acid. An RSCU value of one indicates a uniform codon usage, while RSCU values greater than one represent codons that are used more frequently than the other codons for a particular amino acid (Sharp and Li, 1987). RSCU has been regarded as a better measure than raw codon statistics because it is less susceptible to bias in the data due to varying gene sizes.

Correspondence analysis (Greenacre, 1984), also implemented using the codonW program, was carried out on the RSCU values for each genome individually and for all the 84,162 genes from the forty genomes collected into a

single data set, to investigate trends in the data that would be difficult to determine from single-genome analyses. With such a massive dataset some sort of multivariate statistical method is required to elucidate trends and common features from the contingency data. Correspondence analysis is the most appropriate and regularly used method for this type of data, where the values are not independent (Shields and Sharp, 1987), (Sharp and Devine 1989), (Mèdigue *et al.*, 1991), (Wright and Bibb, 1992), (Andersson and Sharp, 1996), (McInerney, 1998), (Lafay *et al.*, 1999), (Lafay *et al.*, 2000), (Romero *et al.*, 2000). Correspondence analysis plots codon usage statistics for each gene in an n-dimensional hyperspace along n orthogonal axes, where n is, in this case, the number of synonymous codons used, i.e. 59. This results in a 'cloud' of points, which in the absence of any codon usage bias would not be skewed (Greenacre, 1984). The analysis identifies orthogonal axes through the data such that axis 1 explains the most variation in the dataset, with subsequent axes explaining a diminishing proportion of variation. The two axes which account for the greatest proportion of variation in the dataset, axis 1 and axis 2 were plotted for each genome and for the combined dataset. All correspondence analyses plots were graphed with Microsoft® Excel 2000. A Perl program was written to calculate the third codon position nucleotide content for each gene. Regression and correlation analyses between axis 1 and base composition at the third codon positions and axis 2 and base composition at the third codon positions were performed using the Microsoft® Excel 2000 data analysis tool.

Table 1. Phylogenetic breakdown, G+C contents, optimal growth temperatures and GenBank accession numbers for the forty completely sequenced bacteria used in the analysis. Optimal growth temperature data taken from The DSMZ database (http://www.dsmz.de/) and the American Type Culture Collection database (http://www.atcc.org/).

Organism	G+C	Optimal Growth Temp.	Accession No.
Archaea:			
Crenarchaeota			
Aeropyrum pernix	56.3%	90°C	NC_000854
Euryarchaeota			
Archaeoglobus fulgidus	48.5%	85°C	NC_000917
Halobacterium sp.	67.9%	37°C	NC_002607
Methanobacterium thermoautotrophicum	49.5%	65°C	NC_000916
Methanococcus jannaschii	31.4%	85°C	NC_000909
Pyrococcus abyssi	44.7%	97°C	NC_000868
Pyrococcus horikoshii	42%	95°C	NC_000961
Thermoplasma acidophilum	46%	59°C	NC_002578
Eubacteria:			
Aquificales			
Aquifex aeolicus	43.4%	95°C	NC_000918
Firmicutes			
Bacillus halodurans	43.7%	30°C	NC_002570
Bacillus subtilis	43.5%	30°C	NC_000964
Lactococcus lactis	35.4%	30 C	NC_002662
Mycoplasma genitalium	32%	37°C	NC_000908
Mycoplasma pneumoniae	40%	37°C	NC_000912
Urealplasma urealyticum	25.5%	37°C	NC_002162
Mycobacterium tuberculosis	65.6%	37°C	NC_000962
Spirochaetales			
Borrelia burgdorferi	28.6%	37°C	NC_001318
Treponema pallidum	52.8%	37°C	NC_000919
Thermotogales			
Thermotoga maritima	46%	80°C	NC_000853
Thermus/Deinococcus group			
Deinococcus radiodurans	67%	30°C	NC_001263/4
Chlamydia			
Chlamydia trachomatis	41.3%	37°C	NC_000117
Chlamydia muridarum	40.3%	37°C	NC_002182
Chlamydophila pneumoniae CWL029	40.6%	35°C	NC_000922
Chlamydophila pneumoniae AR39	40.6%	35°C	NC_002179
Chlamydophila pneumoniae J138	40.7%	35°C	NC_002491
Proteobacteria			
Rickettsia prowazekii	29.1%	35°C	NC_000963
Neisseria meningitidis Z2491	51.8%	37°C	NC_002203
Neisseria meningitidis MC58	51.5%	37°C	NC_002183
Buchnera sp.	26.3%	?	NC_002528
Escherichia coli K12	50.8%	37°C	NC_000913
Escherichia coli 0157	50.5%	37°C	NC_002655
Haemophilus influenzae	38%	37°C	NC_000907
Pasteurella multocida	41%	37°C	NC_002663
Pseudomonas aeruginosa	66.6%	37°C	NC_002516
Xylella fastidiosa	52.7%	26°C	NC_002488
Vibrio cholerae	47%	37°C	NC_002505/6
Campylobacter jejuni	30.6%	37°C	NC_002163
Helicobacter pylori 26695	39%	37°C	NC_000915
Helicobacter pylori J99	39%	37°C	NC_000921
Cyanobacteria			
Synechocystis PCC6803	47.7%	25°C	NC_000911

3.0 Results

The results are divided into two sections. Section 3.1 contains the results of the intra-genomic correspondence analysis, which permits us to investigate synonymous codon usage variation within species, allowing us to detect subsets of genes that vary in synonymous codon usage within a genome. However, this type of analysis will fail to detect factors that influence synonymous codon usage in all genes in a genome. Take the *B.burgdorferi* genome for example. We have shown, using the intra-genomic analysis, that the major factor influencing codon usage in this genome is strand bias. However, as we will show in the trans-genomic analysis, the *B.burgdorferi* genome as a whole is a very A+T rich species and as such is biased towards using A or T ending codons. Since this A+T bias affects all genes within the genome, the effect cannot be detected by correspondence analysis of synonymous codon usage variation within the genome. To overcome this problem correspondence analysis was also implemented on RSCU values for all identified ORFs from each of the forty genomes as one dataset, consisting of 84,162 coding sequences. This allows investigation of synonymous codon usage variation among species. These results are shown in section 3.2. Tables and figures are located at the end of each section.

3.1 Intra-genomic Analysis

Correspondence analysis was implemented on RSCU values for all identified genes from each of the forty genomes. The two most important axes, axis 1 and axis 2, generated from the correspondence analysis were plotted for each of the forty genomes (Appendix Figures 7a – 7jj). Axis 1 is the axis which accounts for the greatest proportion of variation in the dataset, with each subsequent axis accounting for a diminishing proportion of variation in the dataset. To investigate the relationship between nucleotide composition and synonymous codon usage variation in each genome, A, T, C, G, G+T, G+C and C+T contents at the third codon positions (A3, T3, C3, G3, GT3, GC3, and CT3) were calculated using a Perl program. Regression and correlation analyses were carried out between axis 1 position, and axis 2 position, and base composition at the third codon positions. This resulted in 560 analyses, the results of which are summarized in Table 2a and Table 2b (at the end of this section). To investigate whether translational selection has an effect on each of the genomes, a subset of highly expressed genes, defined as ribosomal proteins, elongation factors and ribosomal subunits were highlighted on each of the correspondence analysis plots. If translational selection is operating on the highly expressed genes in a genome to select for the optimal codons, then the highly expressed genes should have a codon usage pattern that varies from the rest of the lowly or moderately expressed genes. Correspondence analysis will thus result in the separation of these two groups. To test the statistical

significance of the separation of the highly expressed genes from the rest of the genes in the genome, on axis 1 or axis 2, T-tests were performed on the data (see below).

A sample of four of the genomes investigated, *B.burgdorferi*, *P.aeruginosa*, *H.influenzae* and *R.prowazekii* is shown figures 1a-1d (see end of section), while plots of axis 1 versus axis 2 for the other genomes can be found in the appendix (Figures 7a – 7jj). The results of correspondence analysis of RSCU values from the *B.burgdorferi* genome are shown in figure 1a. This plot reveals that the genes separate into two distinct categories on axis 1, those located on the G+T rich leading strand and those found on the A+C rich lagging strand. Regression and correlation analyses of axis 1 position and base composition at the third codon positions revealed a strong correlation ($r^2 = 0.86$) between axis 1 and GT3 content (Table 2a) while the regression analysis determined that axis 1 was significantly dependent on GT3 content ($P < 0.0001$). Variation on axis 2 in *B.burgdorferi* was not significantly related to any of the nucleotide biases investigated (Table 2b) or to the separation of the highly expressed genes (Table 3). Variation on axis 2 in the *B.burgdorferi* genome has previously been determined to be mainly due to two outliers. These have been identified as hypothetical proteins in a previous study and probably do not represent real genes. (McInerney, 1998). In *B.burgdorferi* most of the highly expressed genes separate on axis 1 along with the leading strand genes.

To investigate the effects of strand bias on synonymous codon usage in other genomes regression and correlation analyses between axis 1 and axis 2

positions and GT3 content were carried out (Tables 2a and 2b). If strand bias is a major influence on codon usage in a particular genome, a strong correlation between axis 1 position and GT3 content is expected. Out of the other thirty-nine genomes analyzed, ten were found to have a correlation between axis 1 and GT3 content with an $r^2 > 0.35$ (Table 2a). The regression analysis revealed that axis 1 was significantly dependent on GT3 content in these species ($P < 0.0001$). A further twelve genomes had a weak correlation ($r^2 \leq 0.2$, Table 2a).

Correspondence analysis plots for the *C.muridarum* and *C.trachomatis* genomes clearly show the separation of the genes into two distinct clusters on axis 1 (Figures 7k and 7l). As in *B.burgdorferi* this separation is related to GT3 content (Table 2a). This leads us to propose that these clusters represent the distinctive codon usage of genes located on leading strand compared to those on the lagging strand. The three strains of *C.pneumoniae* (Figures 7h-j), *C.jejuni* (Figure 7g), *T.pallidum* (Figure 7gg) and *X.fastidiosa* (Figure 7jj) all have codon usage patterns that are related to G+T content (Table 2a) although the separation of genes located on the leading strand from those located on the lagging strand is not immediately obvious. The highly expressed genes in these species tend to favor being located on the leading strand. Organisms whose codon usage is influenced by strand bias are widely distributed across eubacteria, although this phenomenon is not particularly evident in archaea.

Correspondence analysis of the *P.aeruginosa* genome is presented in figure 1b. Genes from this genome cluster together along axis 1 according to their GC3 content. This results in a graph in the shape of a comet consisting of

the majority of genes and a tail of G+C poor genes behind it. Correlation and regression analyses between axis 1 position and GC3 content in *P.aeruginosa* results in a strong correlation (r^2 = 0.81) and revealed that axis 1 was significantly dependent on GC3 content (P <0.0001). The highly expressed genes in the *P.aeruginosa* genome separate from the majority of genes in the genome on axis 1 and axis 2. This separation of these two groups was determined to be statistically significant at the P <0.0001 level on both axes. Axis 2 variation is not obviously related to any of the nucleotide biases (Table 2b) and appears to be primarily due to the separation of the highly expressed genes. Axis 1 in another 30 genomes was found to be significantly dependent on GC3 (P < 0.0001). Of these fifteen had a strong correlation (r^2 >0.6) between axis 1 and GC3 (Table 2a).

A plot of axis 1 versus axis 2 for the *H.influenzae* genome is shown in figure 1c. In the *H.influenzae* genome variation in synonymous codon usage on axis 1 is not well correlated with any of the nucleotide biases investigated. The major source of variation in this genome is the separation of the highly expressed genes both on axis 1 and axis 2 from the majority of other genes in the genome. Again the separation was highly significant on both axes (P< 0.0001). Axis 2 is correlated with GC3 content (r^2 = 0.47). In a number of other genomes axis 1 was not correlated or was very weakly correlated with GC3 content. These include *B.burgdorferi*, *C.muridarum*, *L.lactis*, *M.jannaschii*, *P.multocida* and *U.urealyticum* (Table 2a). Unlike what was found in the

H.influenzae genome, axis 2 was also poorly correlated with GC3 content (Table 2b).

A plot of the two most important axes for *R.prowazekii* (Figure 1d) reveals that the separation of the highly expressed genes is not obvious in this genome. The level of significance was only at the $P < 0.05$ level. Furthermore, only very weak relationships between variation in synonymous codon usage and any of the nucleotide biases investigated could be found in this genome (Tables 2a and 2b).

The results of correspondence analysis of RSCU values from each of the other thirty-six genomes are shown in appendix figures 7a – 7jj. These figures show the highly expressed genes highlighted with respect to the rest of the genes to determine if an effect of translational selection is evident. The results of regression analyses between axis 1 and axis 2 positions and the various nucleotide biases investigated, for each genome, are summarized in table 2a and table 2b respectively.

A statistically significant separation of the highly expressed genes on either axis 1 or axis 2 was evident in all but the *T.acidophilum* genome. On axis 1 a separation of the highly expressed genes that was significant at the $P < 0.0001$ level was evident in twenty-eight species; *A.aeolicus, A.pernix, B.burgdorferi, B.halodurans, B.subtilis, C.jejuni, C.muridarum, C.pneumoniae AR39, C.pneumoniae CWL029, C.pneumoniae J138, C.trachomatis, D.radiodurans, E.coli K12, E.coli 0157, H.influenzae, L.lactis, M.genitalium, M.jannaschii, M.tuberculosis, P.abyssi, P.aeruginosa, P.horikoshii, P.multocida,*

Synechocystis sp., T.pallidum, U.urealyticum, V.cholerae and *X.fastidiosa*. A separation of the highly expressed genes on axis 1 that was significant at the P<0.05 level was found in eight species; *A.fulgidus, H.pylori 26695, H.pylori J99, M.pneumoniae, N.meningitidis MC58, N.meningitidis Z2491, R.prowazekii* and *T.maritima*. A non-significant separation of the highly expressed genes on axis 1 was found in four species; *Buchnera sp., Halobacterium sp., M.thermoautotrophicum* and *T.acidophilum*.

On axis 2 a separation of the highly expressed genes that was significant at the P< 0.0001 level was evident in twenty-one species; *B.subtilis, Buchnera sp., C.jejuni, C.muridarum, C.pneumoniae AR39, C.trachomatis, D.radiodurans, E.coli K12, E.coli 0157, H.influenzae, L.lactis, M.jannaschii, M.pneumoniae, M.thermoautotrophicum, M.tuberculosis, N.meningitidis MC58, N.meningitidis Z2491, P.aeruginosa, P.multocida, Synechocystis sp.*, and *V.cholerae*. A separation of the highly expressed genes on axis 2 that was significant at the P<0.05 level was observed in ten of the species investigated; *A.pernix, C.pneumoniae CWL029, C.pneumoniae J138, Halobacterium* sp., *H.pylori J99, M.genitalium, P.abyssi, P.horikoshii, R.prowazekii* and *T.pallidum*. A non-significant separation of the highly expressed genes on axis 2 was found in ten species; *A.aeolicus, A.fulgidus, B.burgdorferi, B.halodurans, H.pylori 26695, M.tuberculosis, T.acidophilum, T.maritima, U.urealyticum, and X.fastidiosa*.

Correspondence analysis of RSCU values from each genome also resulted in some observations that could not be explained and require further analysis. The results of correspondence analysis of RSCU values from the

A.pernix genome are shown in figure 5a. This plot reveals that variation in axis 1 is primarily due to the separation of genes into two distinct clusters, one of which contains the majority of genes including the highly expressed ones. The reason for the separation of these genes has yet to be determined. In the *P.abyssi*, *T.acidophilum* and *T.maritima* genomes, plots of axis 1 and axis 2 generated from the correspondence analysis reveals that the genes separate into three distinct clusters on axis 2 (Figures 7bb, 7ee and 7ff). This pattern is not due to phylogenetic relatedness, G+C content, G+T content, hydrophobicity, expression level, or any obvious functional relationship between the genes located together.

Table 2a. A summary of the correlation and regression analyses between axis 1 position and nucleotide content at the third codon positions in forty completely sequenced bacterial genomes.

Organism	GT3	GC3	CT3	A3	T3	G3	C3
Archaea:							
Crenarchaeota							
Aeropyrum pernix	0.01**	0.44***	0.49***	0.01***	0.62***	0.61***	NS
Euryarchaeota							
Archaeoglobus fulgidus	NS	0.72***	NS	0.44***	0.4***	0.31***	0.23***
Halobacterium sp.	0.06***	0.90***	0.05***	0.71***	0.83***	0.07***	0.41***
Methanobacterium thermoautotrophicum	0.06***	0.65***	0.01*	0.23***	0.46***	0.22***	0.5***
Methanococcus jannaschii	0.16***	NS	0.1***	0.14***	0.15***	0.005*	0.01*
Pyrococcus abyssi	NS	0.63***	0.02**	0.22***	0.45***	0.34***	0.3***
Pyrococcus horikoshii	NS	0.12***	0.45***	0.37***	0.07***	0.07***	0.39***
Thermoplasma acidophilum	0.02**	0.81***	0.02**	0.41***	0.62***	0.46***	0.44***
Eubacteria:							
Aquificales							
Aquifex aeolicus	0.15***	0.65***	0.005*	0.12***	0.5***	0.19***	0.47***
Firmicutes							
Bacillus halodurans	NS	0.35***	0.02***	0.19***	0.08***	0.1***	0.23***
Bacillus subtilis	0.02***	0.78***	NS	0.44***	0.38***	0.5***	0.34***
Lactococcus lactis	0.05***	NS	0.16***	0.005**	0.01**	0.21***	0.14***
Mycoplasma genitalium	NS	0.80***	NS	0.35***	0.35***	0.55***	0.68***
Mycoplasma pneumoniae	NS	0.71***	NS	0.44***	0.35***	0.55***	0.53***
Urealplasma urealyticum	NS	0.02**	NS	0.01**	NS	0.02**	NS
Mycobacterium tuberculosis	0.1***	0.78***	0.05***	0.38***	0.62***	0.02***	0.39***
Spirochaetales							
Borrelia burgdorferi	0.86***	NS	0.42***	0.72***	0.71***	0.61***	0.70***
Treponema pallidum	0.86***	0.12***	0.21***	0.1***	0.49***	0.39***	0.82***
Thermotogales							
Thermotoga maritima	0.15***	0.69***	NS	0.12***	0.53***	0.19***	0.49***
Deinococcus group							
Deinococcus radiodurans	0.18***	0.83***	0.08***	0.57***	0.72***	0.02***	0.52***
Chlamydia							
Chlamydia trachomatis	0.64***	0.05***	0.33***	0.01**	0.11***	0.54***	0.76***
Chlamydia muridarum	0.7***	0.01*	0.28***	0.07***	0.15***	0.59***	0.73***
Chlamydophila pneumoniae CWL029	0.75***	0.12***	0.07***	0.08***	0.4***	0.38***	0.72***
Chlamydophila pneumoniae AR39	0.65***	0.07***	0.09***	0.06***	0.27***	0.37***	0.63***
Chlamydophila pneumoniae J138	0.74***	0.13***	0.07***	0.08***	0.41***	0.38***	0.73***

Table 2a. *(Continued)*

Organism	GT3	GC3	CT3	A3	T3	G3	C3
Proteobacteria							
Rickettsia prowazekii	NS	0.01**	NS	NS	NS	NS	0.02**
Neisseria meningitidis Z2491	0.38***	0.82***	0.22***	0.34***	0.75***	NS	0.79***
Neisseria meningitidis MC58	0.37***	0.85***	0.22***	0.39***	0.75***	0.02**	0.79***
Buchnera sp.	0.01*	NS	NS	0.02*	NS	NS	NS
Escherichia coli K12	NS	0.52***	0.06***	0.47***	0.23***	0.14***	0.43***
Escherichia coli 0157	NS	0.65***	0.05***	0.58***	0.30***	0.22***	0.5***
Haemophilus influenzae	0.11***	NS	0.07***	NS	NS	0.1***	0.16***
Pasteurella multocida	0.07***	0.04***	0.04***	NS	0.03***	0.02***	0.14***
Psuedomonas aeruginosa	0.23***	0.81***	0.08***	0.53***	0.72***	0.02***	0.53***
Xylella fastidiosa	0.85***	0.31***	0.12***	0.09***	0.77***	0.28***	0.87***
Vibrio cholerae	NS	0.44***	0.05***	0.32***	0.13***	0.11***	0.33***
Campylobacter jejuni	0.42***	0.04***	NS	0.13***	0.23***	0.24***	0.46***
Helicobacter pylori 26695	0.02*	0.43***	0.05***	0.25***	0.03***	0.13***	0.21***
Helicobacter pylori J99	0.005*	0.40***	0.02*	0.20***	0.07***	0.13***	0.19***
Cyanobacteria							
Synechocystis PCC6803	NS	0.75***	0.13***	0.60***	0.25***	0.22***	0.58***

Values shown are r^2 values from the correlation analysis.

P values are generated from the regression analysis.

* = P< 0.01; ** = P< 0.001; *** = P< 0.0001; NS = Non-significant.

Table 2b. A summary of the correlation and regression analyses between axis 2 position and nucleotide content at the third codon positions in forty completely sequenced bacterial genomes.

Organism	GT3	GC3	CT3	A3	T3	G3	C3
Archaea:							
Crenarchaeota							
Aeropyrum pernix	0.01***	0.40***	0.15***	0.48***	0.07***	0.02***	0.48***
Euryarchaeota							
Archaeoglobus fulgidus	0.08***	0.02***	0.22***	0.04***	NS	0.08***	0.20***
Halobacterium sp.	0.73***	0.01***	0.62***	0.06***	NS	0.72***	0.40**
Methanobacterium thermoautotrophicum	0.28***	0.07***	0.12***	0.37***	0.06***	NS	0.13***
Methanococcus jannaschii	0.14***	NS	0.02***	0.07***	0.08***	0.03***	0.05***
Pyrococcus abyssi	NS	0.01***	NS	0.01***	NS	NS	0.02***
Pyrococcus horikoshii	NS	0.57***	NS	0.15***	0.27**	0.27***	0.14***
Thermoplasma acidophilum	NS	0.02***	NS	0.03***	0.006*	0.02***	0.008**
Eubacteria:							
Aquificales							
Aquifex aeolicus	0.07***	0.02***	0.02***	0.02***	NS	0.10***	0.01***
Firmicutes							
Bacillus halodurans	0.03***	0.17***	0.04***	0.04***	0.09***	0.23***	NS
Bacillus subtilis	0.10***	NS	0.09***	NS	0.002*	0.07***	0.12***
Lactococcus lactis	0.17***	0.07***	NS	0.008***	0.11***	0.02***	0.19***
Mycoplasma genitalium	NS	NS	NS	NS	NS	NS	NS
Mycoplasma pneumoniae	NS	0.10***	0.02***	0.03***	0.09***	0.12***	0.04***
Ureaplasma urealyticum	0.07***	NS	0.04***	0.06***	0.04***	NS	NS
Mycobacterium tuberculosis	0.49***	0.03***	0.17***	0.23***	0.03***	0.02***	0.19***
Spirochaetales							
Borrelia burgdorferi	NS	NS	NS	NS	NS	NS	NS
Treponema pallidum	0.06***	0.62***	NS	0.61***	0.06***	0.27***	0.06***
Thermotogales							
Thermotoga maritima	0.005**	NS	NS	NS	0.007**	NS	NS
Deinococcus group							
Deinococcus radiodurans	0.49***	0.03***	0.50***	0.04***	0.01***	0.63***	0.26***
Chlamydia							
Chlamydia trachomatis	NS	0.12***	NS	0.05***	0.01**	0.03***	0.02**
Chlamydia muridarum	NS	0.05***	0.02**	0.02***	NS	NS	0.04***
Chlamydophila pneumoniae CWL029	0.02***	0.04**	0.05***	0.10***	0.01**	0.007*	0.01**
Chlamydophila pneumoniae AR39	0.05***	0.09***	0.04***	0.17***	0.01**	0.04***	0.01**
Chlamydophila pneumoniae J138	0.02***	0.05***	0.06***	0.11***	0.01**	0.008*	0.02***

Table 2b. *(Continued)*

Organism	GT3	GC3	CT3	A3	T3	G3	C3
Proteobacteria							
Rickettsia prowazekii	0.02***	0.03***	NS	0.03***	NS	0.04***	NS
Neisseria meningitidis Z2491	0.47***	0.06***	0.29***	0.28***	NS	0.77***	0.13***
Neisseria meningitidis MC58	0.48***	0.03***	0.26***	0.25***	0.01***	0.73***	0.13***
Buchnera sp.	0.13***	NS	0.04***	0.07***	0.09***	0.03**	0.09***
Escherichia coli K12	0.07***	0.20***	0.11***	0.1***	0.17***	0.31***	0.003**
Escherichia coli 0157	0.07***	0.11***	0.14***	0.04***	0.11***	0.25***	NS
Haemophilus influenzae	NS	0.47***	NS	0.14***	0.17***	0.18***	0.21***
Pasteurella multocida	0.06***	0.40***	NS	0.23***	0.07***	0.31***	0.04***
Psuedomonas aeruginosa	0.003***	0.05***	0.1***	NS	0.1***	0.1***	NS
Xylella fastidiosa	0.08***	0.45***	0.01***	0.62***	0.02***	0.34***	0.03***
Vibrio cholerae	0.04***	0.17***	0.04***	0.09***	0.08***	0.23***	0.01***
Campylobacter jejuni	0.02***	0.09***	NS	NS	0.04***	NS	0.09***
Helicobacter pylori 26695	0.07***	0.02***	0.04***	0.1***	NS	0.14***	0.04***
Helicobacter pylori J99	0.03***	0.03***	0.06***	NS	0.02***	0.12***	0.02***
Cyanobacteria							
Synechocystis PCC6803	0.12***	0.02***	0.35***	0.004**	0.09***	0.36***	0.10***

Values shown are r^2 values from the correlation analysis.

P values are generated from the regression analysis.

* = P< 0.01; ** = P< 0.001; *** = P< 0.0001; NS = Non-significant.

Table 3. P values generated from T-tests comparing the separation of Highly expressed genes, on axis 1 and axis 2, from the rest of the genes in each of the forty genomes.

Organism	Axis 1 P-value	Axis 2 P-value
Archaea:		
Crenarchaeota		
Aeropyrum pernix	3.71E-20	6.00E-03
Euryarchaeota		
Archaeoglobus fulgidus	2.00E-02	6.00E-02
Halobacterium sp.	7.00E-02	6.00E-03
Methanobacterium thermoautotrophicum	7.50E-02	7.51E-21
Methanococcus jannaschii	2.44E-26	1.92E-14
Pyrococcus abyssi	2.16E-11	3.53E-03
Pyrococcus horikoshii	4.09E-09	2.50E-03
Thermoplasma acidophilum	5.61E-01	2.47E-01
Eubacteria:		
Aquificales		
Aquifex aeolicus	7.06E-05	1.30E-01
Firmicutes		
Bacillus halodurans	2.14E-15	3.68E-01
Bacillus subtilis	1.35E-17	8.18E-15
Lactococcus lactis	9.73E-31	1.80E-18
Mycoplasma genitalium	9.46E-05	4.00E-02
Mycoplasma pneumoniae	7.50E-03	6.26E-07
Ureaplasma urealyticum	7.89E-27	5.78E-02
Mycobacterium tuberculosis	5.41E-11	4.17E-01
Spirochaetales		
Borrelia burgdorferi	2.26E-12	9.25E-01
Treponema pallidum	5.90E-09	3.83E-03
Thermotogales		
Thermotoga maritima	5.92E-03	7.02E-02
Thermus/Deinococcus group		
Deinococcus radiodurans	2.18E-17	2.48E-23
Chlamydia		
Chlamydia trachomatis	5.64E-06	1.79E-18
Chlamydia muridarum	1.53E-07	7.02E-22
Chlamydophila pneumoniae CWL029	3.72E-06	1.37E-02
Chlamydophila pneumoniae AR39	1.36E-07	1.76E-17
Chlamydophila pneumoniae J138	1.03E-06	2.24E-02

Table 3. *(Continued)*

Organism	Axis 1 P-value	Axis 2 P-value
Proteobacteria		
Rickettsia prowazekii	4.11E-02	2.03E-02
Neisseria meningitidis Z2491	1.03E-03	4.44E-24
Neisseria meningitidis MC58	1.31E-02	2.13E-16
Buchnera sp.	6.90E-01	3.28E-10
Escherichia coli K12	2.82E-31	1.22E-31
Escherichia coli 0157	6.24E-18	1.42E-24
Haemophilus influenzae	4.45E-29	7.76E-12
Pasteurella multocida	1.51E-32	2.29E-19
Pseudomonas aeruginosa	1.60E-08	1.32E-31
Xylella fastidiosa	1.33E-12	6.01E-01
Vibrio cholerae	5.89E-43	8.31E-30
Campylobacter jejuni	7.09E-08	3.69E-12
Helicobacter pylori 26695	2.00E-02	2.90E-01
Helicobacter pylori J99	1.68E-04	2.00E-02
Cyanobacteria		
Synechocystis PCC6803	1.29E-25	4.57E-14

E = X 10

E.G. 2.03E-02 = 0.0203

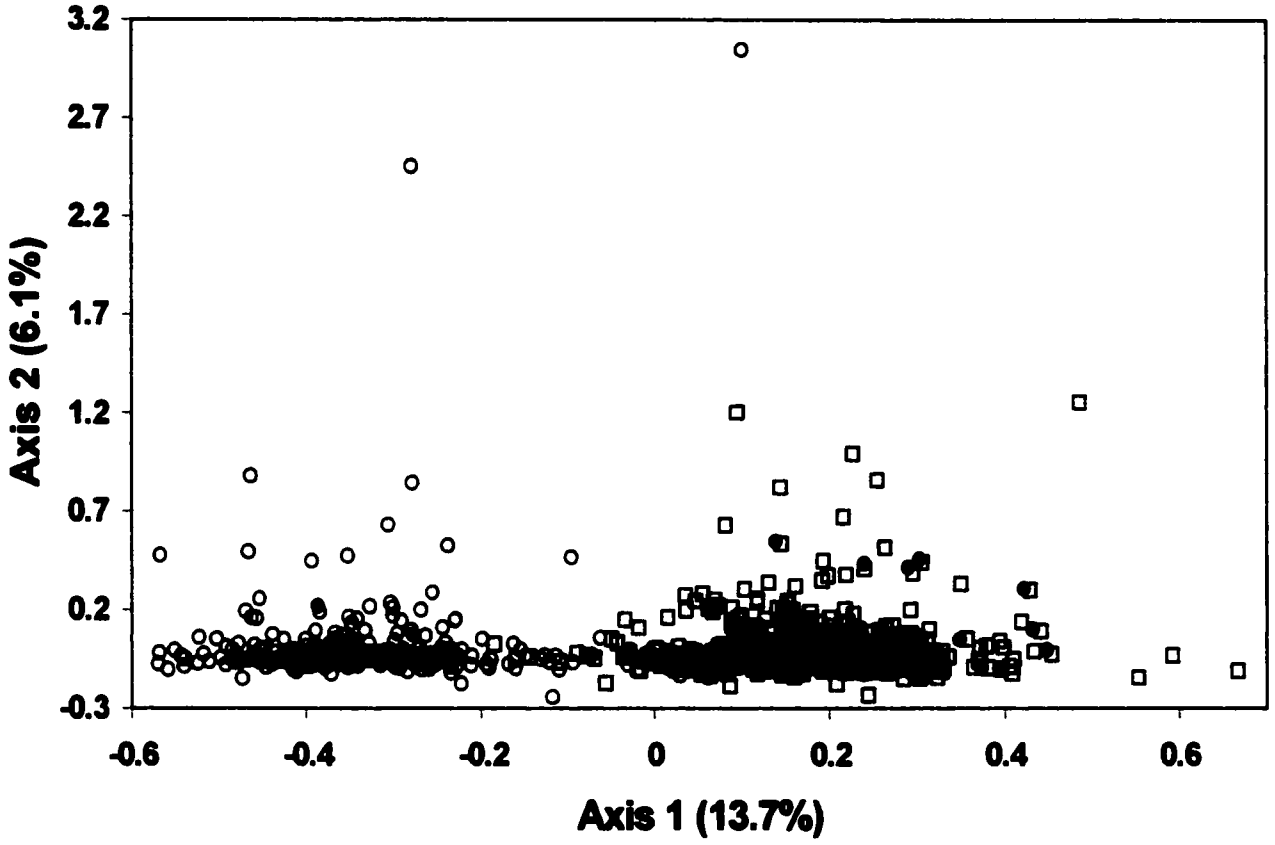

Figure 1a. A plot of axis 1 against axis 2 generated from correspondence analysis of RSCU values from the *B.burgdorferi* genome. Note the distinct clustering of genes located on the GT-rich leading strand (open squares) from those located on the AC-rich lagging strand (open circles). Highly expressed genes are shown in red. The percentage variation explained by each axis is shown in brackets.

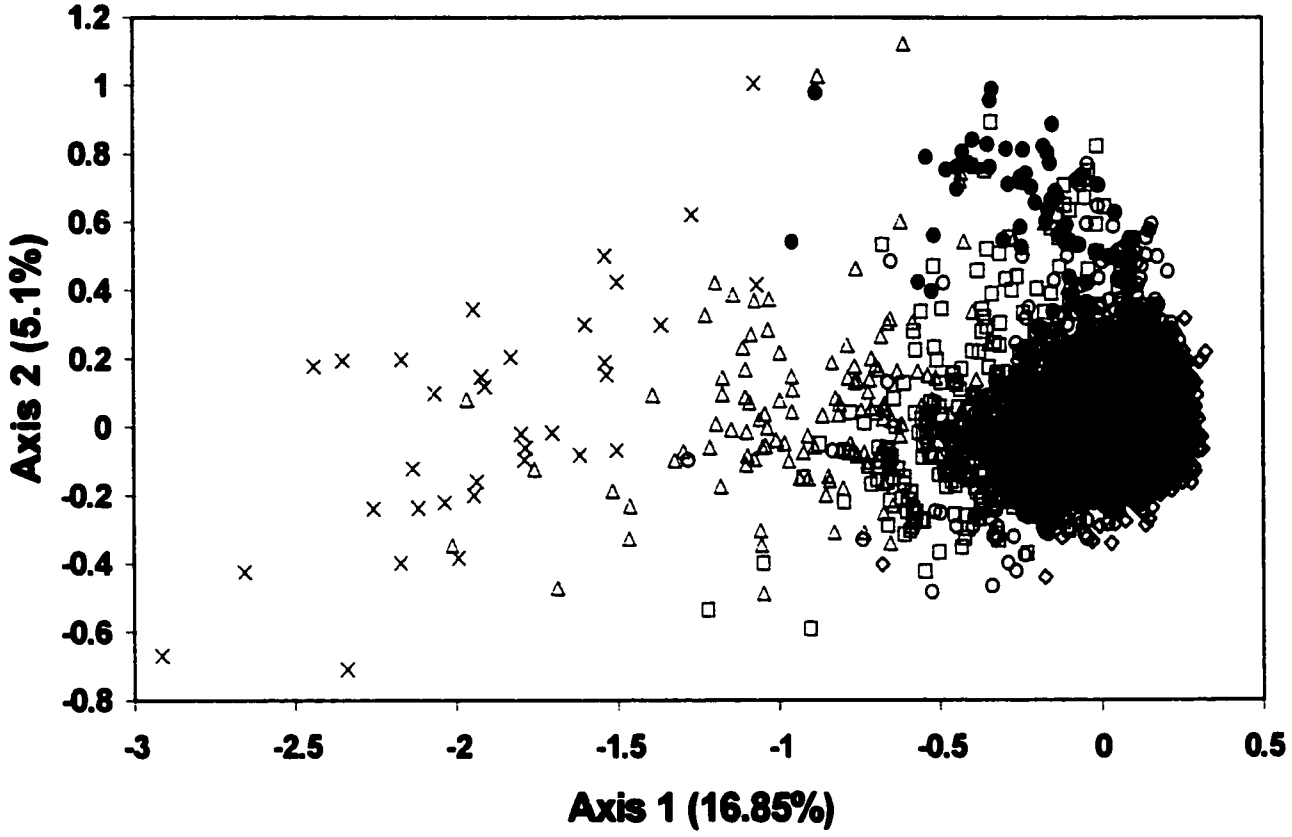

Figure 1b. A plot of axis 1 against axis 2 generated from correspondence analysis of RSCU values from the *P.aeruginosa* genome. Genes with a GC3 content > 90% are represented by dark blue open diamonds; 80-90% light blue open circles; 70-80% light blue open squares; 50-70% turquoise open triangles; < 50% green crosses. Highly expressed genes are shown in red. The percentage variation explained by each axis is shown in brackets.

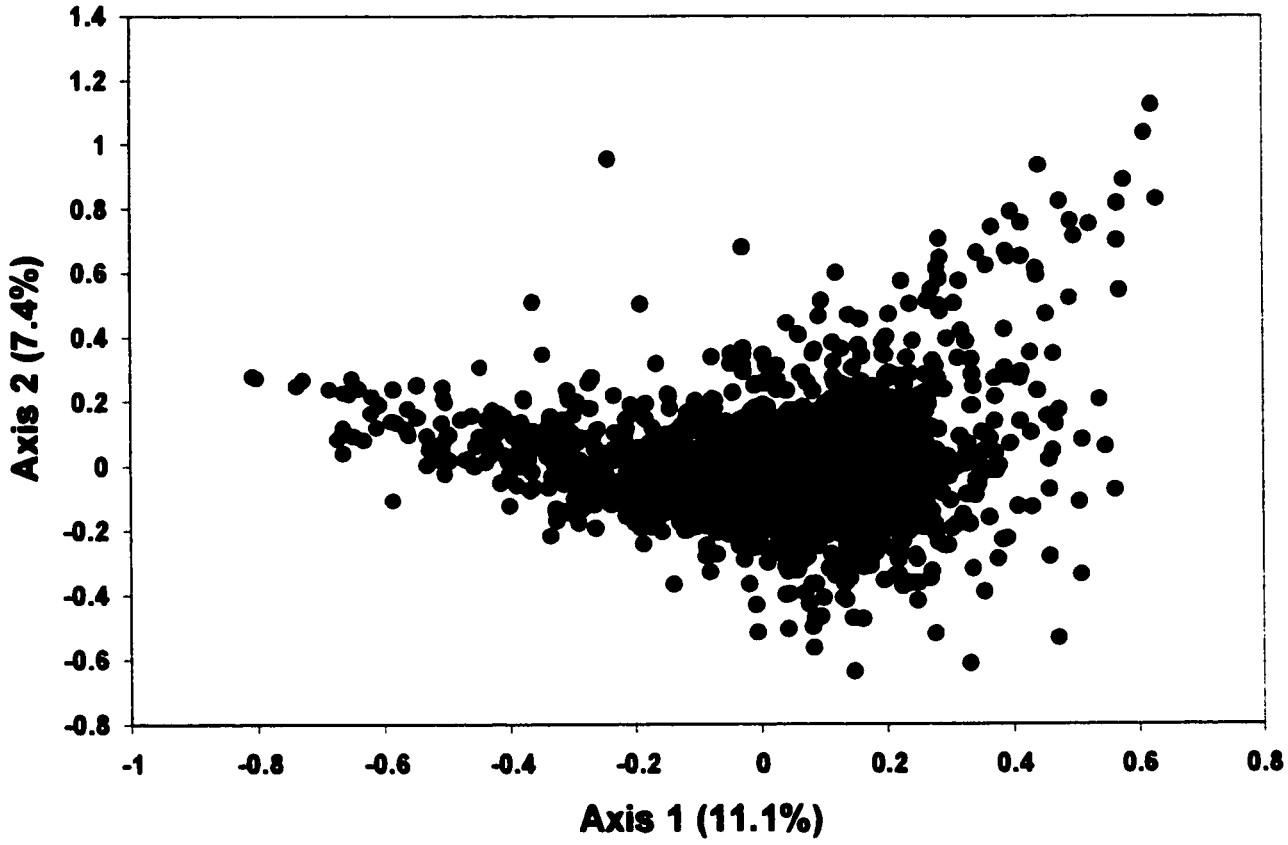

Figure 1c. A plot of axis 1 against axis 2 generated from correspondence analysis of RSCU values from the *H.influenzae* genome. The highly expressed genes (red dots) separate from the majority of genes in the genome (black dots). The percentage variation explained by each axis is shown in brackets.

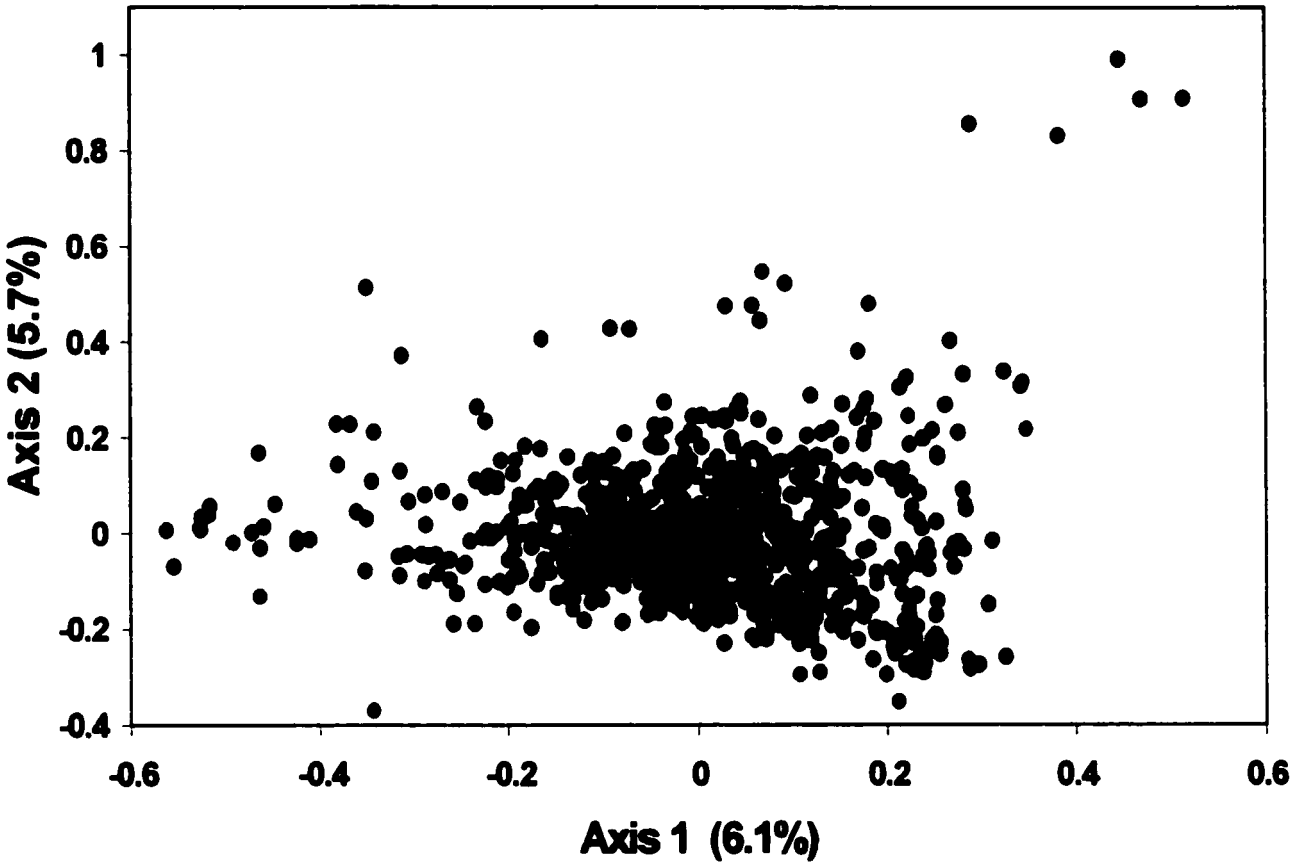

Figure 1d. A plot of axis 1 against axis 2 generated from correspondence analysis of RSCU values from the *R.prowazekii* genome. Highly expressed genes are shown in red. The percentage variation explained by each axis is shown in brackets.

3.2 Trans-genomic Analysis

As previously discussed, the intra-genomic correspondence analysis permits us to investigate synonymous codon usage variation within species, allowing us to detect subsets of genes that vary in synonymous codon usage within a genome. However, this type of analysis will fail to detect factors that influence synonymous codon usage in all genes in a genome. To overcome this problem correspondence analysis was also implemented on RSCU values for all identified ORFs from each of the forty genomes as one dataset, consisting of 84,162 coding sequences. This allows for simultaneous investigation of synonymous codon usage variation both within and among species.

A plot of axis 1 versus axis 2 generated from the correspondence analysis is shown in figure 2a (see end of section). Although this graph is quite complex with a large amount of overlap among genes from different genomes, one can still get the impression that genes from a particular genome tend to cluster together. On axis 1 The G+C rich species (Table 1, Materials and Methods) are located toward one end, the A+T rich species are grouped together at the other end, and the more G+C neutral species are located along the center of the axis. Correlation and regression analysis of axis 1 and axis 2 positions and nucleotide content at the third codon positions (Table 4, see end of section) reveals a strong correlation with GC3 content ($r^2 = 0.95$) and the regression analysis revealed that axis 1 was significantly dependent on GC3 content ($P < 0.0001$).

The relative positions of each of the genomes are more clearly observed by plotting the mean axis 1 and axis 2 coordinates for each genome (Figure 3). The error bars on figure 3 indicate the 99.9% confidence intervals. This reveals that the mean positions of all but the most closely related species are statistically significantly different from each other. Figure 4 also shows a plot of the mean axis 1 and axis 2 positions for each genome. This figure color-codes species from the same taxonomic group. From this graph it is observed that codon usage can be very different in species from the same taxonomic groups. For example codon usage in the species of proteobacteria is quite divergent, with the genomes separating right along axis 1. It is also evident that the eubacteria and archaea do not have distinct patterns of codon usage. In species that are closely related such as the two *E.coli* strains, the two *N.meningitidis* strains or the *Chlamydia*, the codon usage patterns are similar indicating that the codon usage has not yet had enough time to evolve.

Figure 4 also shows a plot of the mean axis 1 and 2 coordinates for a subset of highly expressed genes (as previously defined) from each genome. T-tests were carried out to determine if there is a statistically significant separation of the highly expressed genes from the majority of genes in each of the genomes (Table 5). A statistically significant separation of the highly expressed genes on either axis 1 or axis 2 was evident in all but the two *H.pylori* species. On axis 1 a separation of the highly expressed genes that was significant at the $P< 0.0001$ level was evident in nineteen species; *B.burgdorferi, B.subtilis, C.muridarum, C.pneumoniae AR39, C.pneumoniae CWL029, C.pneumoniae*

J138, D.radiodurans, E.coli K12, E.coli 0157, L.lactis, M.genitalium, M.tuberculosis, N.meningitidis Z2491, P.aeruginosa, P.multocida, R.prowazekii, Synechocystis sp., V.cholerae and *X.fastidiosa*. A separation of the highly expressed genes on axis 1 that was significant at the P<0.05 level was found in fifteen of the species investigated; *A.aeolicus, A.fulgidus, A.pernix, B.halodurans, Buchnera sp., C.trachomatis, Halobacterium* sp., *H.influenzae, M.pneumoniae, N.meningitidis MC58, P.abyssi, T.acidophilum, T.maritima, T.pallidum and U.urealyticum.* A non-significant separation of the highly expressed genes on axis 1 was found in six species; *C.jejuni, H.pylori 26695, H.pylori J99, M.jannaschii, M.thermoautotrophicum* and *P.horikoshii.*

On axis 2 a separation of the highly expressed genes that was significant at the P< 0.0001 level was evident in twenty-one of the species investigated; *A.aeolicus, A.fulgidus, A.pernix, B.halodurans, B.subtilis, Buchnera sp., C.muridarum, C.pneumoniae AR39, E.coli K12, E.coli 0157, H.influenzae, L.lactis, N.meningitidis MC58, N.meningitidis 26695, P.abyssi, P.aeruginosa, P.horikoshii, P.multocida, T.acidophilum, T.maritima and X.fastidiosa*. A separation of the highly expressed genes on axis 2 that was significant at the P<0.05 level was observed in ten species; *C.jejuni, C.pneumoniae CWL029, D.radiodurans, M.genitalium, M.jannaschii, M.pneumoniae, M.thermoautotrophicum, Synechocystis sp., U.urealyticum* and *V.cholerae*. A non-significant separation of the highly expressed genes on axis 2 was found in nine species; *B.burgdorferi, C.pneumoniae J138, C.trachomatis, Halobacterium sp., H.pylori 26695, H.pylori J99, M.tuberculosis, R.prowazekii* and *T.pallidum.*

Note also the direction of the arrows in figure 3. It appears that much of the variation between the highly expressed genes and their respective genomes is along axis 2 and that this variation is in opposite directions in the thermophiles compared to the non-thermophiles.

The considerable variation on axis 2 provides an interesting insight into synonymous codon usage variation between species. Figure 2b and figure 4 show the separation of genes from the thermophilic species of bacteria, *A.aeolicus*, *A.fulgidus*, *A.pernix*, *M.jannaschii*, *M.thermoautotrophicum*, *P.abyssi*, *P.horikoshii*, *T.acidophilum*, and *T.maritima* from the other non-thermophilic bacteria. We define bacteria as thermophilic based on their optimal growth temperature (See Table 1). The separation of the thermophiles from the non-thermophiles was determined to be significant using T-tests on axis 2 ($P < 0.001$) but not on axis 1 ($P = 0.39$).

Regression analysis between axis 2 position and base composition at the third codon positions did not reveal a strong correlation between nucleotide bias and the variation in synonymous codon usage between the thermophiles and non-thermophiles (Table 4). Since nucleotide bias did not seem to be responsible for the variation between the two groups we decided to investigate which codons accounted for the variation between the two groups. It is possible using codonW to specify the particular subset of codons ones wishes to analyze. Analysis of the two-fold synonymous codons or the four-fold synonymous codons revealed that the division between the thermophiles and other species was greatly reduced but not eliminated, indicating that these groups of codons

accounted for a minor component of the variation in codon usage between these species (Figures 5a and 5b). Analysis of the six-fold synonymous codons revealed an even greater separation of the two groups than when all codons were analyzed (Figure 5c). This suggested that it was variability in six-fold synonymous codon usage that was mostly responsible for the separation of the thermophiles from the other species on the correspondence analysis plots. *M.jannashii* appears to be the exception to this. When arginine codons were removed from the dataset of six-fold synonymous codons the division between the thermophiles and non-thermophiles was greatly reduced (Figure 5d). This implied that a variation in arginine usage was largely accounted for the difference in codon usage between the two groups. To determine what the variation in arginine usage was we compared synonymous codon usage tables for thermophilic species to those of non-thermophilic species of a similar G+C content. We found that thermophiles show a strong preference for the AGA and AGG arginine codons, while non-thermophiles prefer CGC and CGT. This difference was determined to be statistically significant.

In correspondence analysis the same set of axes used to investigate variation among genes can also be used to examine variation among codons (Figure 6). On Axis 1 the codons separate into two groups, those ending in G and C and those ending in A and T. By comparing this graph to figure 3 which shows the mean position of each genome on axis 1 and axis 2, it is clear that on axis 1 the G+C rich species preferentially use codons ending in G and C while the A+T rich species prefer A and T ending codons. On axis 2 it is observed that

the arginine codons AGG and AGA are preferentially used in the thermophiles, as was previously described. The non-thermophiles prefer the CGT, CGA, CGC and CGG arginine codons. We also identify that the ATA isoleucine codon is preferred by the thermophiles. Together these three codons account for much of the variation between the thermophiles and non-thermophiles on axis 2.

Table 4 A summary of the correlation and regression analyses between axis 1 and axis 2 positions and nucleotide content at the third codon positions in the combined dataset of forty completely sequenced bacterial genomes.

Base Composition	Axis 1	Axis 2
A3	0.78***	0.001***
G3	0.62***	0.04***
C3	0.85***	0.03***
T3	0.78***	0.11***
GC3	0.95***	0.04***
GT3	0.09***	0.05***
CT3	0.06***	0.07***

Values shown are r^2 values from the correlation analysis.

P values are generated from the regression analysis.

*** = P< 0.0001

Table 5. P values generated from T-tests comparing the separation of Highly expressed genes, on axis 1 and axis 2, from the rest of the genes in each of the forty genomes in the combined dataset.

Organism	Axis 1 P-value	Axis 2 P-value
Archaea:		
Crenarchaeota		
Aeropyrum pernix	2.05E-02	2.00E-20
Euryarchaeota		
Archaeoglobus fulgidus	3.32E-03	8.85E-06
Halobacterium sp.	4.14E-02	3.61E-01
Methanobacterium thermoautotrophicum	1.58E-01	4.73E-04
Methanococcus jannaschii	4.38E-01	1.72E-03
Pyrococcus abyssi	6.80E-03	2.02E-08
Pyrococcus horikoshii	5.73E-01	2.37E-06
Thermoplasma acidophilum	6.18E-03	3.76E-05
Eubacteria:		
Aquificales		
Aquifex aeolicus	4.20E-02	1.05E-04
Firmicutes		
Bacillus halodurans	4.92E-03	5.89E-12
Bacillus subtilis	8.78E-15	1.16E-13
Lactococcus lactis	1.05E-14	5.91E-06
Mycoplasma genitalium	1.04E-05	1.72E-03
Mycoplasma pneumoniae	1.39E-02	3.73E-04
Urealplasma urealyticum	8.33E-04	2.94E-02
Mycobacterium tuberculosis	1.02E-11	5.17E-01
Spirochaetales		
Borrelia burgdorferi	6.12E-05	5.78E-02
Treponema pallidum	3.47E-02	3.52E-01
Thermotogales		
Thermotoga maritima	2.08E-02	2.68E-08
Thermus/Deinococcus group		
Deinococcus radiodurans	2.67E-14	1.88E-02
Chlamydia		
Chlamydia trachomatis	1.33E-02	1.29E-01
Chlamydia muridarum	3.20E-09	2.06E-06
Chlamydophila pneumoniae CWL029	2.07E-12	4.40E-04
Chlamydophila pneumoniae AR39	7.02E-22	4.56E-08
Chlamydophila pneumoniae J138	2.95E-12	3.74E-01

Table 5. (continued)

Organism	Axis 1 P-value	Axis 2 P-value
Proteobacteria		
Rickettsia prowazekii	2.56E-07	7.02E-01
Neisseria meningitidis Z2491	4.64E-06	4.44E-21
Neisseria meningitidis MC58	1.22E-04	3.66E-22
Buchnera sp.	3.82E-02	8.40E-08
Escherichia coli K12	1.55E-06	2.82E-07
Escherichia coli 0157	8.20E-09	9.44E-16
Haemophilus influenzae	2.18E-02	9.00E-10
Pasteurella multocida	2.09E-08	5.08E-10
Pseudomonas aeruginosa	9.40E-11	2.31E-23
Xylella fastidiosa	1.57E-09	1.49E-13
Vibrio cholerae	1.81E-06	1.66E-04
Campylobacter jejuni	7.01E-01	6.27E-03
Helicobacter pylori 26695	4.83E-01	4.90E-01
Helicobacter pylori J99	1.55E-01	8.35E-01
Cyanobacteria		
Synechocystis PCC6803	3.00E-19	2.76E-02

E = X 10

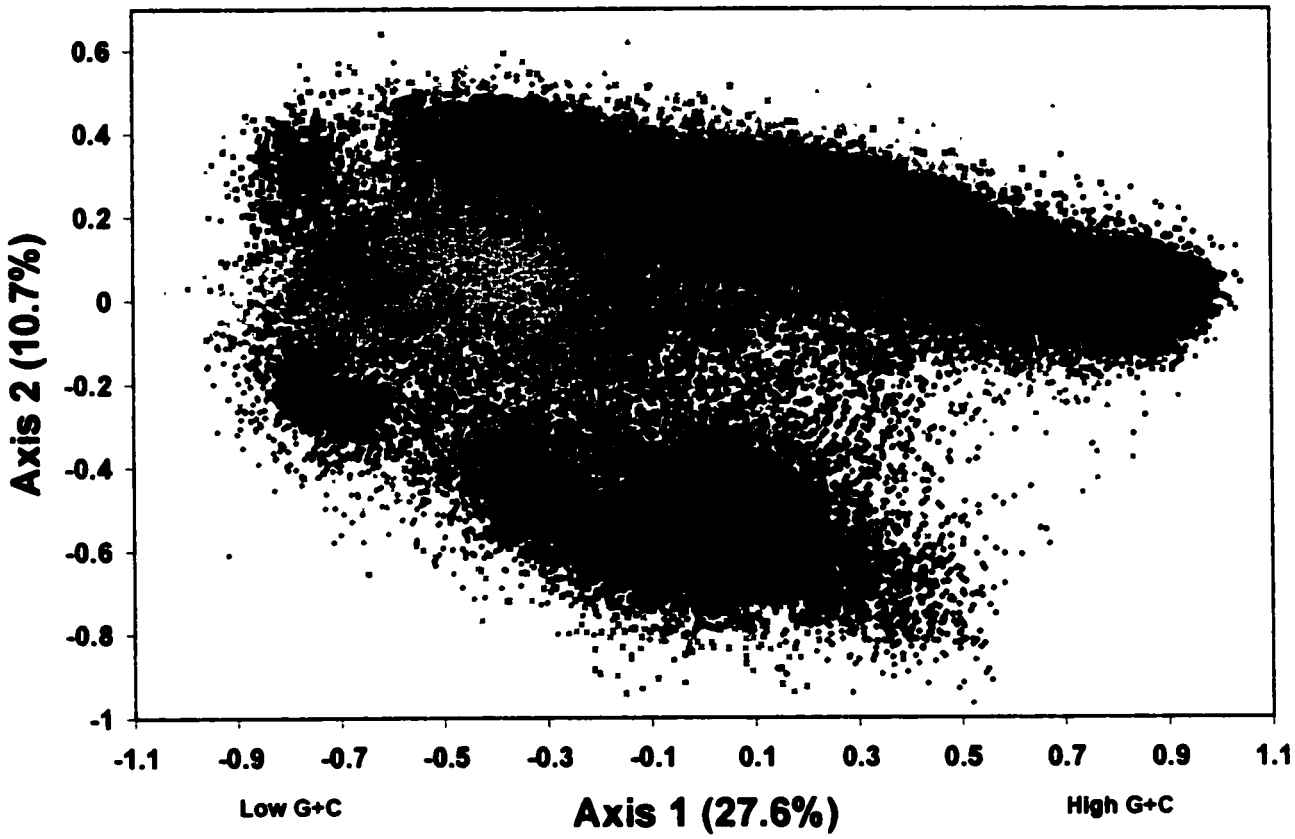

Figure 2a. Plot of axis 1 Vs axis 2 generated from correspondence analysis of RSCU values from forty completely sequenced bacterial genomes. Each genome is represented by a different colour. See Fig. 3 for the mean position of each genome.

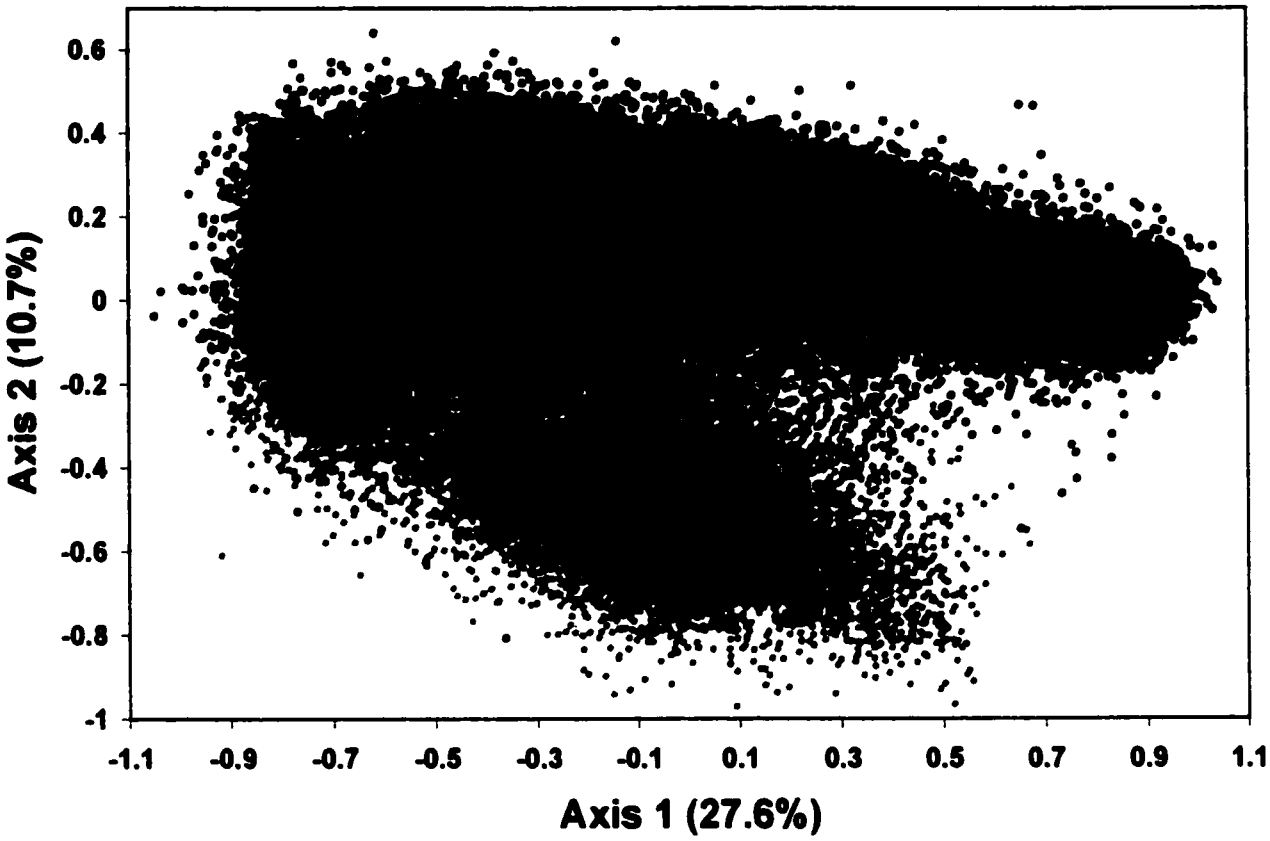

Figure 2b. A plot of axis 1 Vs axis 2 generated from correspondence analysis of RSCU values from 40 completely sequenced bacterial genomes. Genes from thermophilic bacteria are shown as orange dots and those from the non-thermophilic species are shown as black dots.

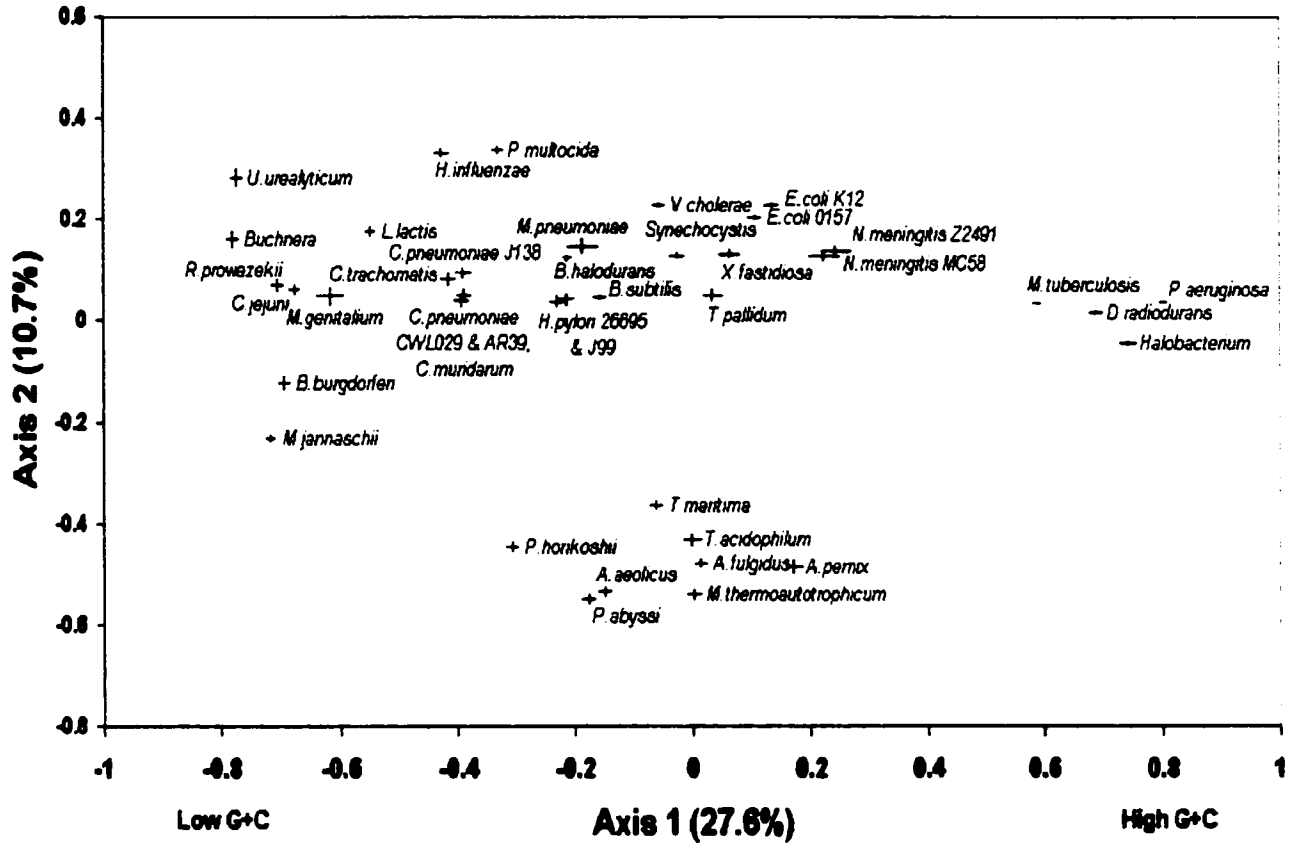

Figure 3. A plot of the mean axis 1 and axis 2 coordinates generated by correspondence analysis of RSCU values for each of the forty genomes. The error bars indicate the 99.99% confidence intervals.

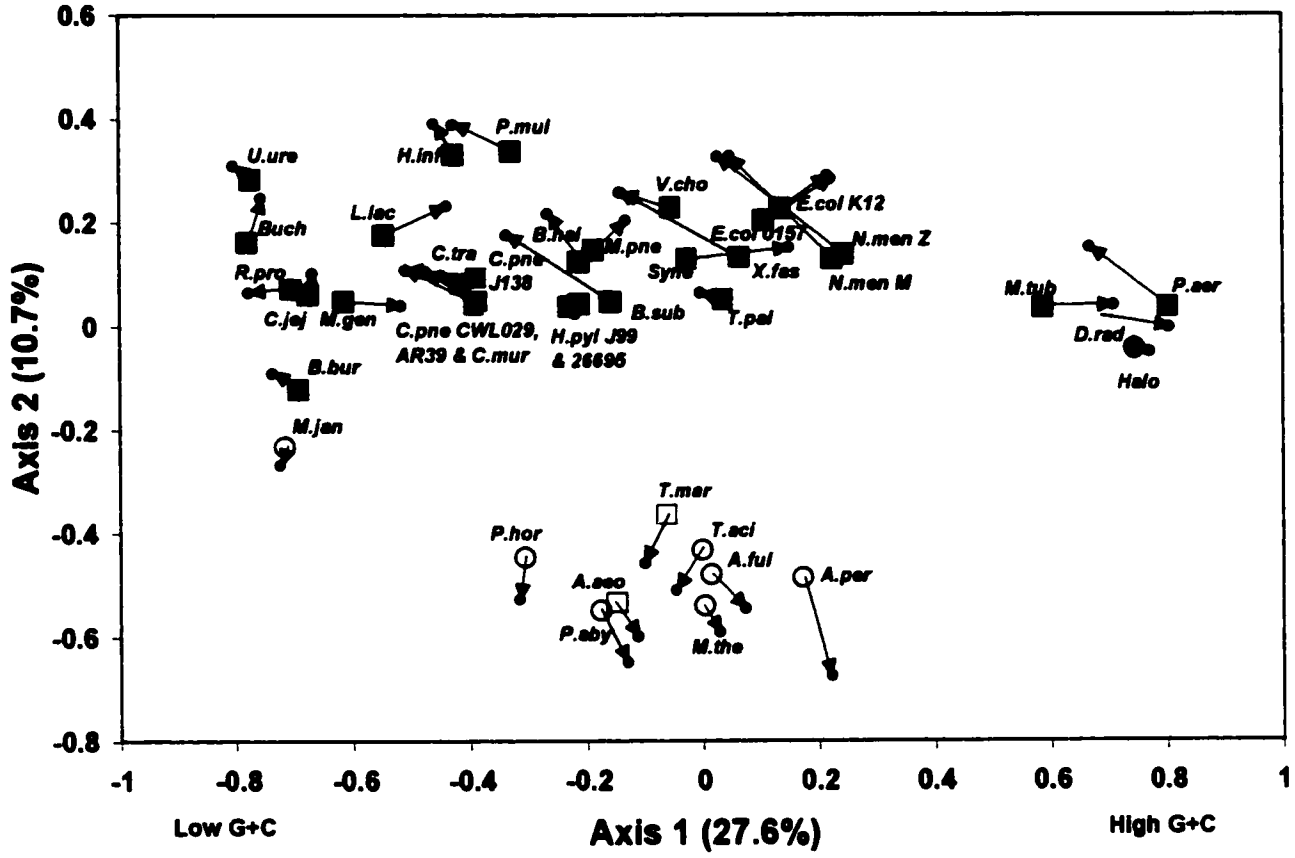

Figure 4. A plot of the mean axis 1 and axis 2 coordinates generated by correspondence analysis of RSCU values for each of the forty genomes and the mean axis 1 and axis 2 coordinates of the highly expressed genes from each genome (small red circles). The arrows link each genome to its subset of highly expressed genes. The percentage variation in the data explained by axis 1 and 2 is shown in brackets. Eubacteria are shown as square symbols; Archaea are shown as black circles. Thermophiles are shown as open symbols; Non-thermophiles are filled. The eubacteria are colour-coded according to their classification at NCBI as follows: pink, *proteobacteria*; blue, *firmicutes*; orange, *Chlamydia*; dark green, *spirochaetales*; yellow, *deinococcus*; red, *aquificales*; turquoise, *cyanobacteria*; bright green, *thermotogales*. See appendix for abbreviations.

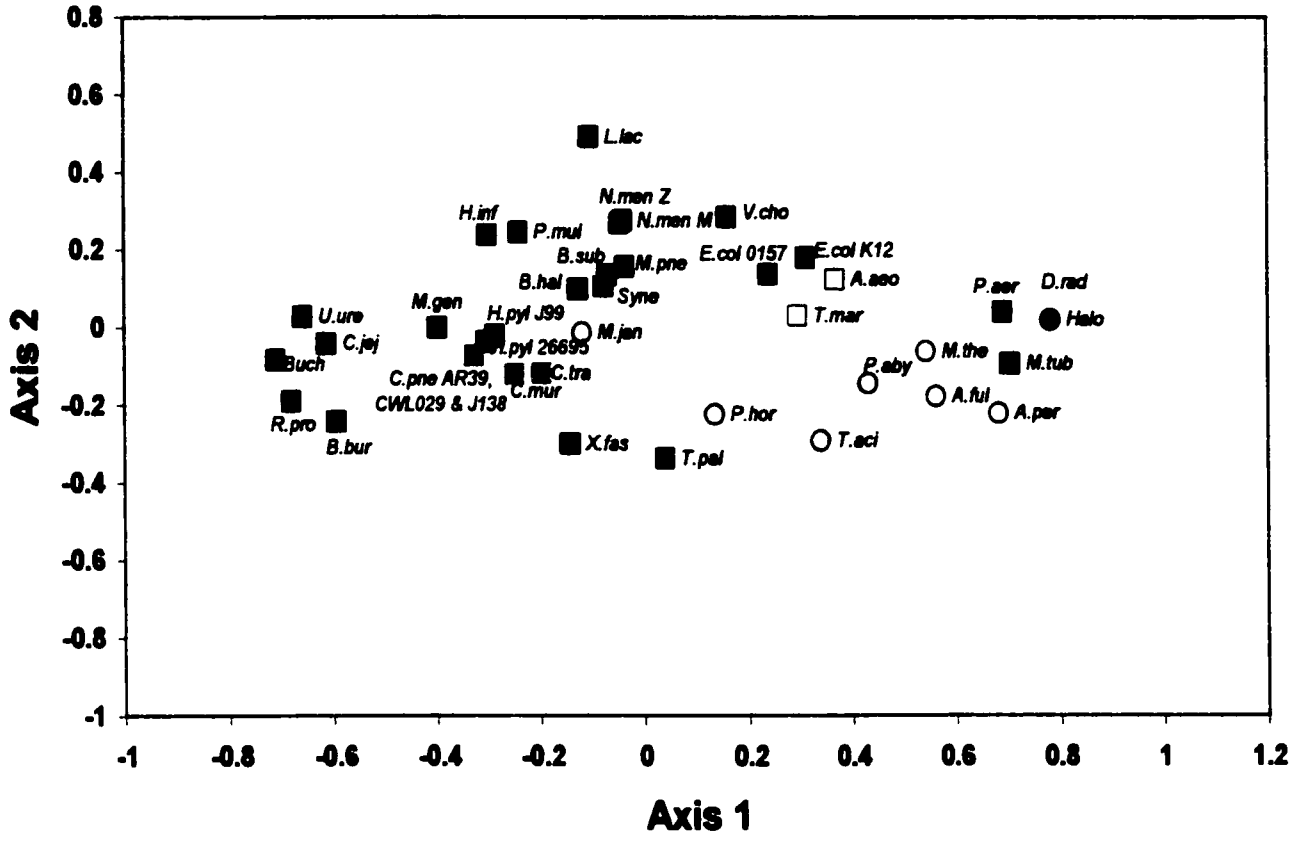

Figure 5a. A plot of the mean axis 1 and axis 2 coordinates generated by correspondence analysis of two-fold synonymous codons from each of the forty genomes. Eubacteria are shown as square symbols; Archaea are shown as black circles. Thermophiles are shown as open symbols; Non-thermophiles are filled. The eubacteria are colour-coded according to their classification at NCBI as follows: pink, *proteobacteria*; blue, *firmicutes*; orange, *Chlamydia*; dark green, *spirochaetales*; yellow, *deinococcus*; red, *aquificales*; turquoise, *cyanobacteria*; bright green, *thermotogales*. See appendix for abbreviations.

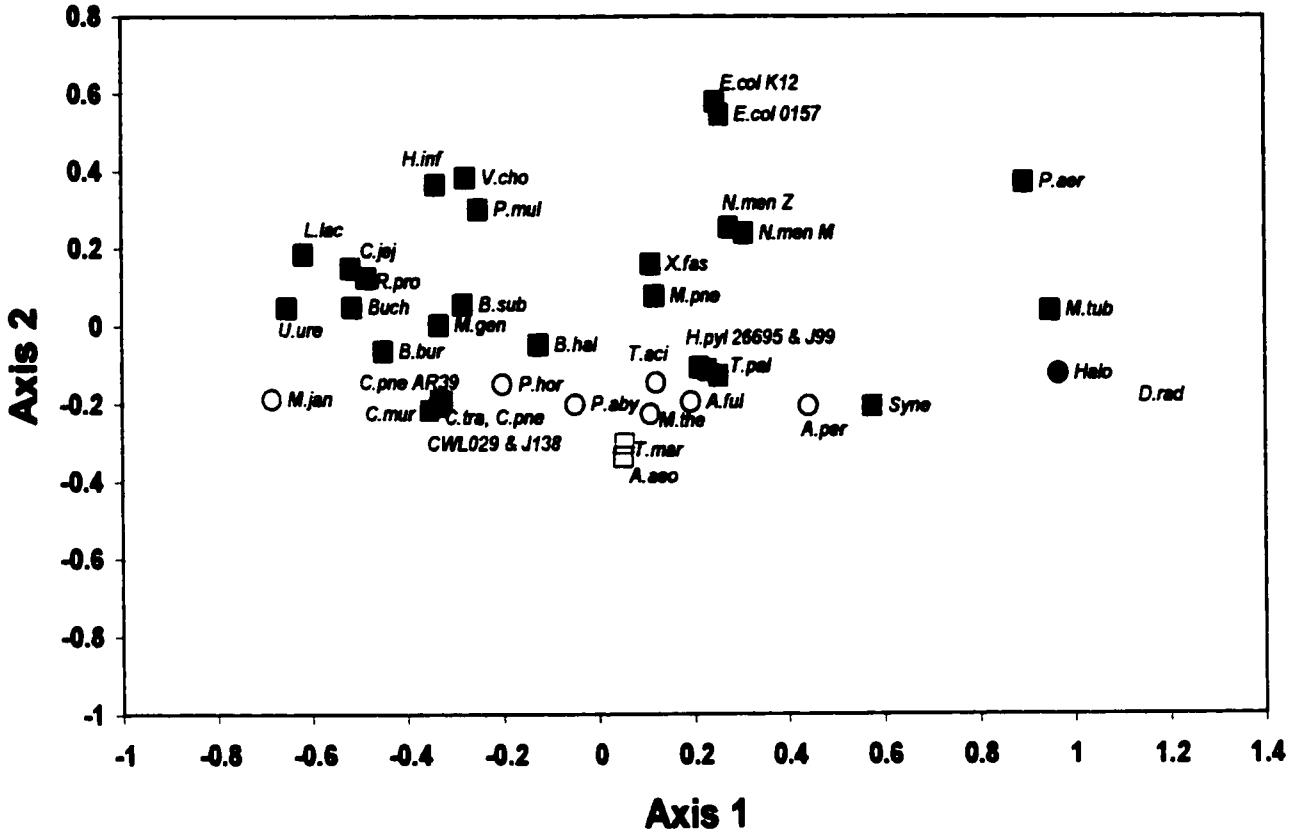

Figure 5b. A plot of the mean axis 1 and axis 2 coordinates generated by correspondence analysis of four-fold synonymous codons from each of the forty genomes. Eubacteria are shown as square symbols; Archaea are shown as black circles. Thermophiles are shown as open symbols; Non-thermophiles are filled. The eubacteria are colour-coded according to their classification at NCBI as follows: pink, *proteobacteria*; blue, *firmicutes*; orange, *Chlamydia*; dark green, *spirochaetales*; yellow, *deinococcus*; red, *aquificales*; turquoise, *cyanobacteria*; bright green, *thermotogales*. See appendix for abbreviations.

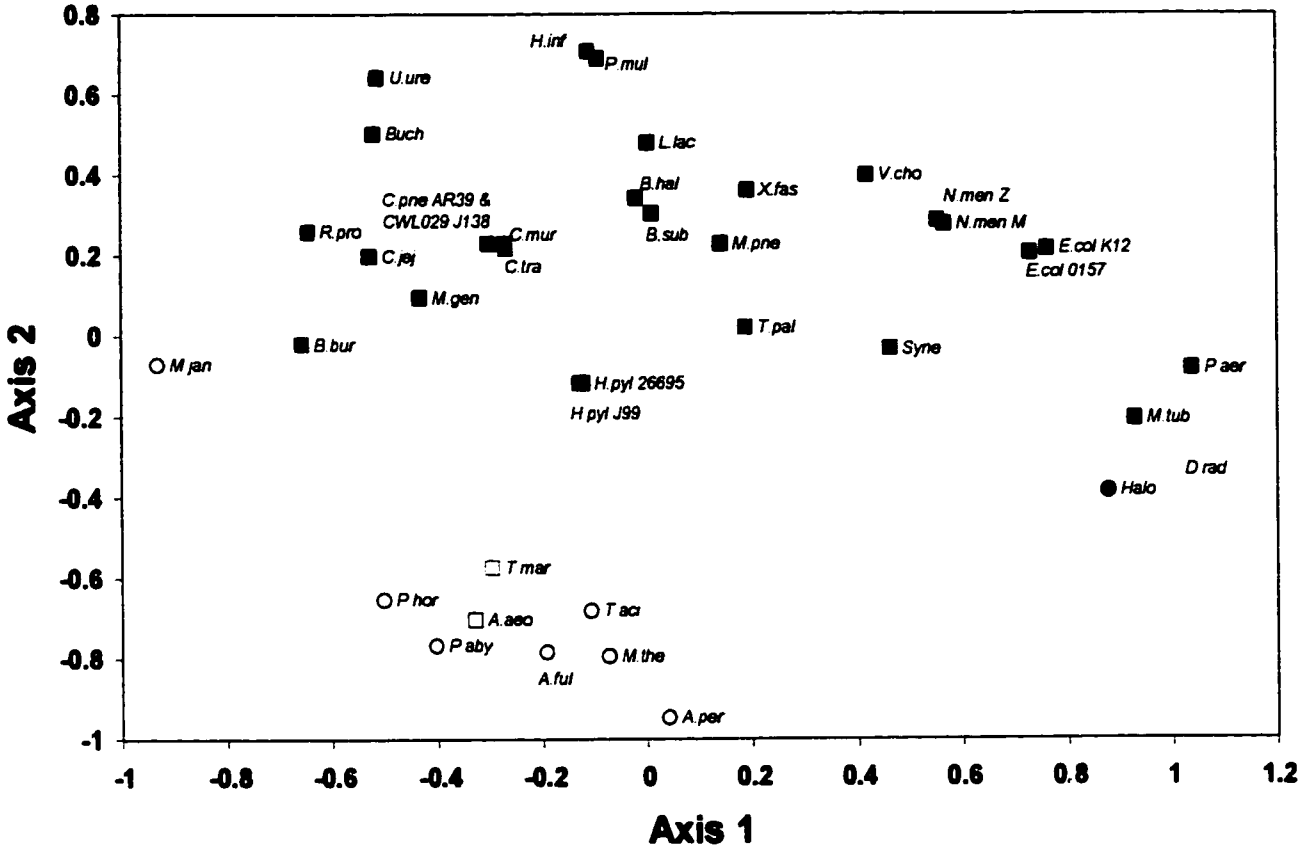

Figure 5c. A plot of the mean axis 1 and axis 2 coordinates generated by correspondence analysis of six-fold synonymous codons from each of the forty genomes. Eubacteria are shown as square symbols; Archaea are shown as black circles. Thermophiles are shown as open symbols; Non-thermophiles are filled. The eubacteria are colour-coded according to their classification at NCBI as follows: pink, *proteobacteria*; blue, *firmicutes*; orange, *Chlamydia*; dark green, *spirochaetales*; yellow, *deinococcus*; red, *aquificales*; turquoise, *cyanobacteria*; bright green, *thermotogales*. See appendix for abbreviations.

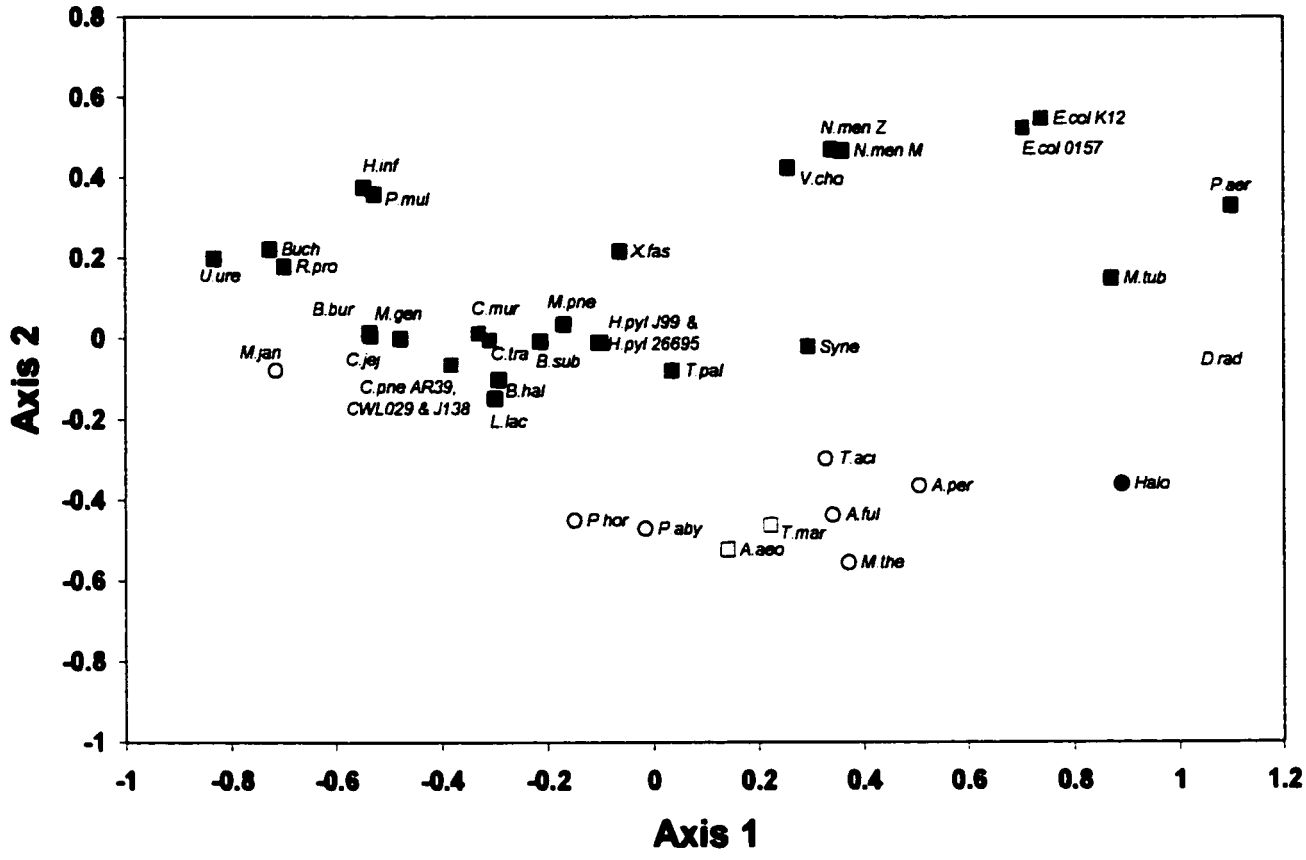

Figure 5d. A plot of the mean axis 1 and axis 2 coordinates generated by correspondence analysis of six-fold synonymous codons with arginine codons removed from each of the forty genomes. Eubacteria are shown as square symbols; Archaea are shown as black circles. Thermophiles are shown as open symbols; Non-thermophiles are filled. The eubacteria are colour-coded according to their classification at NCBI as follows: pink, *proteobacteria*; blue, *firmicutes*; orange, *Chlamydia*; dark green, *spirochaetales*; yellow, *deinococcus*; red, *aquificales*; turquoise, *cyanobacteria*; bright green, *thermotogales*. See appendix for abbreviations.

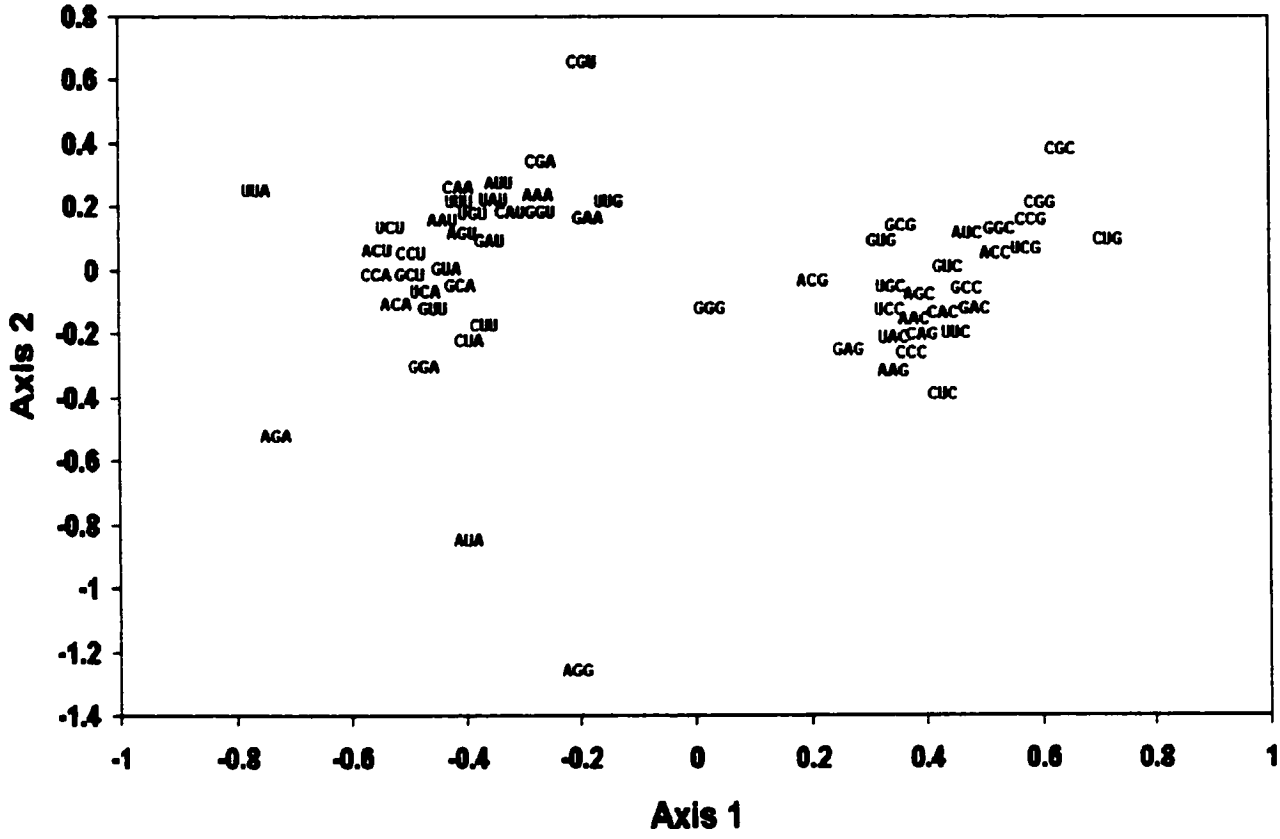

Figure 6. Correspondence analysis of codons from the combined dataset of forty completely sequenced bacterial genomes.

4.0 Discussion

4.1 Single Genome Analyses

To investigate whether translational selection has an effect on each of the genomes, a subset of highly expressed genes were highlighted on each of the correspondence analysis plots. If translational selection is operating on the highly expressed genes in a genome to select for the optimal codons, then the highly expressed genes should have a codon usage pattern that varies from the majority of genes in the genome. Correspondence analysis will thus result in the separation of these two groups. This method of detecting translational selection has been successfully applied in a number of codon usage studies (Shields and Sharp, 1987; Sharp and Devine, 1989; Lloyd and Sharp, 1991; Médigue *et al.*, 1991; Lloyd and Sharp, 1992; Lloyd and Sharp, 1994; Lafay and Sharp, 1999; Musto *et al.*, 1999; Romero *et al.*, 2000). We use T-tests to test the statistical significance of the separation of the highly expressed genes from the majority of genes in the genome. In this study regression analysis between axis 1 position and GC3 content allowed us to investigate the role of mutational bias in shaping codon usage variation in these genomes. Since changes in third codon positions are synonymous there should be no selection operating on them, thus base composition at these positions should be entirely due to mutational pressures. Again this method has been widely implemented (Chiapello, 1998; Lafay and Sharp, 1999; Musto *et al.*, 1999; Romero *et al.*, 2000).

Correspondence analysis of synonymous codon usage revealed the separation of the highly expressed genes from the rest of the genes in the genome in almost all of the species investigated. This separation was determined to be statistically significant on either axis 1 or axis 2 or both in all genomes except for *T.acidophilum*. In this species synonymous codon usage was primarily determined by mutational bias. This does not mean that there is no selection for optimal codons, only that the effect may be insufficient to be observed with regard to the first two axes. Regression analysis of axis 1 and axis 2 position and base composition at the third codon positions revealed significant and often strong correlations with GC3 content in most of the genomes investigated. We thus can conclude that synonymous codon usage patterns within species are due to a balance of selection favoring optimal codons and mutational bias in almost all of the genomes investigated.

Investigation of synonymous codon usage in *B.burgdorferi*, *C.muridarum*, *L.lactis*, *M.jannaschii*, *P.multocida* and *U.urealyticum* uncovered evidence for the effect of translational selection, however the effects of mutational bias were less obvious. Only very weak correlations between axis 1 or axis 2 position and GC3 content were found. This is not to say that there is no influence of mutational bias, as GC3 could be correlated with one of the other 57 less significant axes. It does mean that mutational bias is not the major factor in determining synonymous codon usage variation in these species.

Even some of the highly biased genomes appear to be under the influence of selection. For example *P.aeruginosa*, which has a GC content of

66.6% (Table 1), clearly shows a separation of the highly expressed genes from the rest in the genome. It has previously been proposed, based on a small sample of genes, that translational selection is not evident in highly biased genomes (Wright and Bibb, 1992; Ohkubo, 1987). We are now able to show that with a large enough dataset and effective analysis tools it is possible to detect selection in biased genomes.

In most G+C rich genomes, such as *P.aeruginosa*, *Halobacterium sp.*, *M.tuberculosis* and *D.radiodurans* GC3 content is strongly related to variation in codon usage. However, in A+T rich species, *M.jannaschii, M.genitalium, U.urealyticum, B.burgdorferi, R.prowazekii, Buchnera sp.* and *C.jejuni* there is little or no relationship found. In these A+T rich genomes it appears that almost all the genes have evolved the same base content at the third codon positions. Why the variation persists in the G+C rich species is unknown.

It has been widely known for the past number of years that there is asymmetrical bias between the leading and lagging strand of replication in most eubacterial genomes (Lobry, 1996a; Blattner *et al.*, 1997; McLean *et al.*, 1998). The bias is such that the leading strand tends to be G+T rich while the lagging strand is conversely A+C rich. Only recently however, has the effect of this bias on codon usage been investigated (McInerney, 1998; Lafay *et al.*, 1999; Romero *et al.*, 2000). Although the majority of genomes analysed by us exhibited some correlation between axis 1 position and GT3 content, synonymous codon usage in a number of genomes, such as *B.burgdorferi*, *T.pallidum*, the Chlamydia species, the two *N.meningitis* strains, *X.fastidiosa* and *C.jejuni* stand out as

being the most influenced by strand bias. Statistical analysis between axis 1 position and GT3 content in genomes that resulted in weak correlations cannot be used as evidence of strand bias, as the correlation could be due to the influence of the single nucleotides, G or T. Strand bias in the *B.burgdorferi*, *T.pallidum*, and *C.trachomatis* genomes has previously been shown to be a primary factor in shaping codon usage in these species (McInerney, 1998; Lafay *et al.*, 1999; Romero *et al.*, 2000). We confirm the results of these previous studies and have found that in a further six species strand bias is also the most significant factor affecting codon usage variation within these genomes. This challenges the view that the selection-mutational model is sufficient to explain codon usage variation in all prokaryotes.

4.2 Multi-Genome Analysis

As previously discussed, correspondence analysis of synonymous codon usage within species will only detect variation within a genome and will fail to detect factors that influence synonymous codon usage in all genes in a genome. To overcome this problem and to provide an insight in synonymous codon usage variation among the different species of bacteria we analyzed RSCU values from all forty genomes as one dataset. The separation of the genomes along axis 1 according to their G+C content and the highly significant regression between axis 1 and GC3 leads us to conclude that mutational bias is the most important factor influencing trans-genomic codon usage variation.

Correspondence analysis of codons revealed that G+C rich species favor G and C ending codons while A+T rich species favor A and T ending codons. Some other interesting observations were made from this trans-genomic analysis. Although it is clear that codon usage varies considerably among species that belong to the same taxonomic groups, organisms that are closely related, such as the two different strains of *E.coli*, do share almost identical patterns of codon usage, in agreement with previous observations (Sharp *et al.*, 1988). This indicates that synonymous codon usage evolves over time.

To investigate the separation of the highly expressed genes from the majority of genes in each genome in this combined dataset, we plotted the mean axis 1 and axis 2 coordinates of the highly expressed subset relative to the mean position of each genome. T-tests were again used to determine the statistical significance of the separation. Using this method we were able to demonstrate the influence of translational selection in almost all of the genomes investigated. Only in the two *H.pylori* species was a non-significant result obtained.

One of the most interesting findings of this study, demonstrated by the multi-genome analysis, is the difference in codon usage between the thermophilic bacteria and the other non-thermophilic species. It has been well characterized that amino acid substitutions in proteins of thermophiles compared to species living at moderate temperatures are not symmetrical. Certain amino acids, such as glutamate, arginine and lysine, are more abundant in the thermophiles (McDonald *et al.*, 1999; Haney, *et al.*, 1999; Cambillau *et al.*,

2000). In this study we have found that selection also influences codon choice for the six-fold synonymous amino acids. The difference in synonymous codon usage between thermophiles and the non-thermophiles is mostly due to a disparity in arginine and isoleucine usage. Thermophiles show a strong preference for the usage of AGA and AGG arginine codons and the ATA isoleucine codon, while non-thermophiles prefer CGC, CGT, CGA and CGG arginine codons and the ATC and ATT isoleucine codons. Since the group of thermophilic species includes both eubacteria and archaea the unique pattern of codon usage is not due to phylogenetic relatedness nor is it due to G+C content, therefore there must be a common selective pressure on these codons in thermophiles for thermal adaptation. An explanation of how a subset of codons could confer a selective advantage for life in extreme temperatures remains elusive, but perhaps more efficient aminoacyl tRNA synthetases (Cambillau *et al.*, 2000), or more heat stable tRNAs themselves result in this selective advantage (Hurst *et al.*, 2001).

5.0 Conclusions

This study of synonymous codon usage within and among forty completely sequenced bacterial genomes reveals that although certain factors, such as strand bias, mutational bias or translational selection, may influence codon choice to a greater or lesser extent depending on the genome in question, these and other factors working in parallel or in opposition are responsible for shaping the overall codon usage pattern in a particular species. Why certain genomes appear to be more influenced by a particular bias than another remains unknown. These data have implications for algorithms that use codon usage indices to identify ORFs. Those that do not account for the variation within and among different genomes will be less effective in identifying novel genes.

This study indicates that the variation in synonymous codon usage among the forty completely sequenced bacterial genomes is primarily due mutational bias. We also find a difference in synonymous codon usage between thermophilic and non-thermophilic species of bacteria. This variation appears to be due to alternative usage of synonymous codons, primarily arginine and isoleucine usage.

6.0 References

Akashi H (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**: 927-935

Akashi H, Kliman RM, Eyre-Walker A (1998). Mutation pressure, natural selection, and the evolution of base composition in *Drosophila. Genetica* **102-103**: 49-60

Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, Carmel G, Tummino PJ, Caruso A, Uria-Nickelsen M, Mills DM, Ives C, Gibson R, Merberg D, Mills SD, Jiang Q, Taylor DE, Vovis GF, Trust TJ (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori. Nature* **397**:176-180

Andersson GE, Sharp PM (1996a). Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* **142**: 915-925

Andersson SG, Sharp PM (1996b) Codon usage and base composition in *Rickettsia prowazekii. J Mol Evol* **42**: 525-536

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**: 133-140

Berg OG, Kurland CG (1997) Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* **270**: 544-550

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. *J Mol Evol* **24**: 1-11

Bernardi G (2000) The compositional evolution of vertebrate genomes. *Gene* **259**: 31-43

Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The

complete genome sequence of *Escherichia coli K-12. Science* **277**: 1453-1474

Bolotin,A , Wincker,P., Mauger,S., Jaillon,O., Malarme,K., Weissenbach,J., Ehrlich,S.D. and Sorokin,A. (2001) *Genome Res.* In press

Brewer BJ (1988) When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**: 679-686

Bulmer M (1990) The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res* **18**: 2869-2873

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897-907

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM,Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273**: 1058-1073

Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. *J Biol Chem* **275**: 32383-32386

Chiapello H, Lisacek F, Caboche M, Henaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana. Gene* **209**: GC1-GC38

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S,Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537-544

Curran JF, Yarus M (1989) Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol* **209**: 65-77

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE,

Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**: 353-358

de Miranda AB, Alvarez-Valin F, Jabbari K, Degrave WM, Bernardi G (2000) Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in Mycobacterium tuberculosis and *Mycobacterium leprae*. *J Mol Evol* **50**: 45-55

Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol* **260**: 649-663

Duret L (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**: 287-289

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

Emilsson V, Kurland CG (1990) Growth rate dependence of transfer RNA abundance in *Escherichia coli*. *EMBO J* **9**:4359-4366

Eyre-Walker A (1996) Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol* **13**: 864-872

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.*, (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science* **269**: 496-512

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. *Trends Genet* **13**:240-245

Frank AC, Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**: 65-77

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C, Cotton MD, Venter JC, et al. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**: 375-388

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Venter JC, *et al.*, (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi. Nature* **390**: 580-586

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, *et al.*, (1995) The minimal gene complement of *Mycoplasma genitalium. Science* **270**: 397-403

French S (1992) Consequences of replication fork movement through transcription units in vivo. *Science* **258**: 1362-1365

Gautier C (2000) Compositional bias in DNA. *Curr Opin Genet Dev* **10**:656-661

Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum. Nature* **407**: 757-762

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10**: 7055-7074

Grantham R, Gautier C, Gouy M (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res* **8**:1893-1912

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalogue usage is a genome strategy for genome expressivity. *Nucleic Acids Res* **9**: r43-r75

Greenacre, M. J. (1984) Theory and applications of correspondence analysis. London, Academic Press.

Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* **26**: 2286-2290

Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**: 199-209

Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc Natl Acad Sci U S A* **96**: 3578-3583

Hartl DL, Moriyama EN, Sawyer SA (1994) Selection intensity for codon bias. *Genetics* **138**: 227-234

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleishmann RD, Nierman WC, White O (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**: 477-483

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**: 4420-4449

Hurst LD, Merchant AR (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc R Soc Lond B Biol Sci* **268**: 493-497

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* **151**: 389-409

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and

Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**: 573-597

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13-34

Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS (1999) Comparative genomes of *Chlamydia pneumoniae* and *C.trachomatis*. *Nat Genet* **21**: 385-389

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**: 143-155

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain *PCC6803*. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**: 109-136

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, Nagai Y, Sakai M, Ogura K, Otsuka R, Nakazawa H, Takamiya M, Ohfuku Y, Funahashi T, Tanaka T, Kudoh Y, Yamazaki J, Kushida N, Oguchi A, Aoki K, Kikuchi H (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii OT3*. *DNA Res* **5**: 55-76

Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, Kosugi H, Hosoyama A, Fukui S, Nagai Y, Nishijima K, Nakazawa H, Takamiya M, Masuda S, Funahashi T, Tanaka T, Kudoh Y, Yamazaki J, Kushida N, Oguchi A, Kikuchi H, et al. (1999) Complete genome sequence of an aerobic hyper-

thermophilic crenarchaeon, *Aeropyrum pernix K1. DNA Res* **6**: 83-101, 145-52

Kerr AR, Peden JF, Sharp PM (1997) Systematic base composition variation around the genome *of Mycoplasma genitalium*, but *not Mycoplasma pneumoniae* [letter]. *Mol Microbiol* **25**: 1177-1179

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**: 364-70

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A, *et al.*, (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**: 249-56

Lafay B, Atherton JC, Sharp PM (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* **146**: 851-60

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res* **27**: 1642-1649

Lafay B, Sharp PM (1999) Synonymous codon usage variation among *Giardia lamblia* genes and isolates. *Mol Biol Evol* **16**: 1484-1495

Lloyd AT, Sharp PM (1991) Codon usage in *Aspergillus nidulans*. *Mol Gen Genet* **230**: 288-294

Lloyd AT, Sharp PM (1992) Evolution of codon usage patterns: the extent and

nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. *Nucleic Acids Res* **20**: 5289-5295

Lloyd AT, Sharp PM (1993) Synonymous codon usage *in Kluyveromyces lactis*. Yeast **9**: 1219-1228

Lobry JR (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**: 660-665

Lobry JR (1996b) Origin of replication of *Mycoplasma genitalium*. *Science* **272**: 745-746

Lopez P, Philippe H, Myllykallio H, Forterre P (1999) Identification of putative chromosomal origins of replication in Archaea. *Mol Microbiol* **32**: 883-886

Malumbres M, Gil JA, Martin JF (1993) Codon preference in *corynebacteria*. *Gene* **134**: 15-24

Marais G, Duret L (2001) Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J Mol Evol* **52**: 275-280

Martin AP (1995) Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. *Mol Biol Evol* **12**: 1124-1131

May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V (2001) Complete genomic sequence of *Pasteurella multocida,Pm70*. *Proc Natl Acad Sci U S A* **98**: 3460-3465

McDonald JH, Grasso AM, Rejto LK (1999) Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Mol Biol Evol* **16**: 1785-1790

McInerney JO (1997) Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. *Microb Comp Genomics* **2**: 1-10

McInerney JO (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc Natl Acad Sci U S A* **95**: 10698-10703

McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**: 691-696

Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A (1991) Evidence for

horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222**: 851-856

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in Drosophila. *J Mol Evol* **45**: 514-23

Moriyama EN, Powell JR (1998) Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res* **26**: 3188-93

Morton BR (1999) Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis. Proc Natl Acad Sci U S A* **96**: 5123-5128

Moszer I (1998) The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. *FEBS Lett* **430**: 28-36

Mrazek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A* **95**: 3720-3725

Musto H, Romero H, Zavala A, Jabbari K, Bernardi G (1999) Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. *J Mol Evol* **49**: 27-35

Myllykallio H, Lopez P, Lopez-Garcia P, Heilig R, Saurin W, Zivanovic Y, Philippe H, Forterre P (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* **288**: 2212-2215

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM, *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima. Nature* **399**: 323-9

Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW,

Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL, Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ebhardt H, Lowe TM, Liang P, Riley M, Hood L, DasSarma S (2000) Genome sequence of *Halobacterium species NRC-1. Proc Natl Acad Sci U S A* **97**: 12176-12181

Ohama T, Muto A, Osawa S (1990) Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. *Nucleic Acids Res* **18**: 1565-1569

Ohkubo S, Muto A, Kawauchi Y, Yamao F, Osawa S (1987) The ribosomal protein gene cluster of *Mycoplasma capricolum. Mol Gen Genet* **210**: 314-322

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. *Gene* **215**: 405-13

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV, Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream MA, Rutherford KM, van Vliet AH, Whitehead S, Barrell BG (2000a) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**:665-668

Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels K, Leather S, Moule S, Mungall K, Quail MA, Rajandream MA, Rutherford KM, Simmonds M, Skelton J, Whitehead S, Spratt BG, Barrell BG (2000b) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis Z2491. Nature* **404**: 502-506

Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET,

Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli O157:H7. Nature* **409**: 529-533

Perriere G, Lobry JR, Thioulouse J (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. *Comput Appl Biosci* **12**: 519-524

Picardeau M, Lobry JR, Hinnebusch BJ (1999) Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. *Mol Microbiol* **32**: 437-445

Pouwels PH, Leunissen JA (1994) Divergence in codon usage of *Lactobacillus* species. *Nucleic Acids Res* **22**: 929-936

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci U S A* **94**: 7784-7790

Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, Bass S, Linher K, Weidman J, Khouri H, Craven B, Bowman C, Dodson R, Gwinn M, Nelson W, DeBoy R, Kolonay J, McClarty G, Salzberg SL, Eisen J, Fraser CM (2000) Genome sequences of *Chlamydia trachomatis MoPn* and *Chlamydia pneumoniae AR39. Nucleic Acids Res* **28**: 1397-1406

Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G Codon usage can affect efficiency of translation of genes in *Escherichia coli. Nucleic Acids Res* **12**: 6663-6671

Rocha EP, Danchin A, Viari A (1999) Universal replication biases in bacteria. *Mol Microbiol* **32**: 11-16

Romero H, Zavala A, Musto H (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res* **28**: 2084-2090

Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum. Nature* **407**: 508-513

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon Usage Patterns in *Esherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster and Homo sapiens*: a review of the considerable within-species diversity. *Nucleic Acids Res* **16**: 8207-8211

Sharp PM, Devine KM (1989) Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. *Nucleic Acids Res* **17**: 5029-5039

Sharp PM, Li WH (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14**: 7737-7749

Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281-1295

Sharp PM, Rogers MS, McConnell DJ (1984) Selection pressures on codon usage in the complete genome of *bacteriophage T7*. *J Mol Evol* **21**: 150-160

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* **21**: 835-841

Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res* **15**:8023-8040

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* **5**: 704-716

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp. APS*. *Nature* **407**: 81-86

Shirai M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shiba T, Ishii K, Hattori M, Kuhara S, Nakazawa T (2000) Comparison of whole genome

sequences of *Chlamydia pneumoniae J138* from Japan and *CWL029* from USA. *Nucleic Acids Res* **28**: 2311-2314

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, Araya JE, Baia GS, Baptista CS, Barros MH, Bonaccorsi ED, Bordin S, Bove JM, Briones MR, Bueno MR, Camargo AA, Camargo LE, Carraro DM, Carrer H, Colauto NB, Colombo C, Costa FF, Costa MC, Costa-Neto CM, Coutinho LL, Cristofani M, Dias-Neto E, Docena C, El-Dorry H, Facincani AP, Ferreira AJ, Ferreira VC, Ferro JA, Fraga JS, Franca SC, Franco MC, Frohme M, Furlan LR, Garnier M, Goldman GH, Goldman MH, Gomes SL, Gruber A, Ho PL, Hoheisel JD, Junqueira ML, Kemper EL, Kitajima JP, Marino CL (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature* **406**: 151-157

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Reeve JN, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum deltaH*: functional analysis and comparative genomics. *J Bacteriol* **179**: 7135-7155

Sorensen MA, Kurland CG, Pedersen S (1989) Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol* **207**: 365-377

Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res* **22**: 2437-2446

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**: 754-759

Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT (2000) Complete genome sequence of *Pseudomonas aeruginosa PA01*, an opportunistic pathogen. *Nature* **406**: 959-964

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* **85**: 2653-2657

Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res* **28**: 4317-4331

Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH, Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS, Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, Vamathevan J, Gill J, Scarlato V, Masignani V, Pizza M, Grandi G, Sun L, Smith HO, Fraser CM, Moxon ER, Rappuoli R, Venter JC (2000) Complete genome sequence of *Neisseria meningitidis serogroup B* strain *MC58. Science* **287**: 1809-1815

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Venter JC, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**: 539-547

White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, Moffat KS, Qin H, Jiang L, Pamphile W, Crosby M, Shen M, Vamathevan JJ, Lam P, McDonald L,

Utterback T, Zalewski C, Makarova KS, Aravind L, Daly MJ, Fraser CM, et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans R1*. *Science* **286**:1571-1577

Wolfe KH (1991) Mammalian DNA replication: mutation biases and the mutation rate. *J Theor Biol* **149**: 441-451

Wright F, Bibb MJ (1992) Codon usage in the G+C-rich *Streptomyces* genome. *Gene 113*: 55-65

Xia X (1998) How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics* **149**: 37-44

Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S (1991) Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Res* **19**:6119-6122

7.0 Appendix

Figure 4 and figure 5 abbreviations (see page 68 and 69): A.per, *Aeropyrum pernix*; A.aeo, *Aquifex aeolicus*; A.ful, *Archaeoglobus fulgidus*; B.hal, *Bacillus halodurans*; B.sub, *Bacillus subtilis*; B.bur, *Borrelia burgdorferi*; Buch, *Buchnera sp.*; C.jej, *Campylobacter jejuni*; C.pne CWL029, *Chlamydophila pneumoniae CWL029*; C.pne AR39, *Chlamydophila pneumoniae AR39*; C.pne J138, *Chlamydophila pneumoniae J138*; C.tra, *Chlamydia trachomatis*; C.mur, *Chlamydia muridarum*; D.rad, *Deinococcus radiodurans*; E.col K12, *Escherichia coli K-12*, E.col 0157, *Escherichia coli 0157:H7*; H.inf, *Haemophilus influenzae*; Halo, *Halobacterium sp.*; H.pyl 26695, *Helicobacter pylori 26695*; H.pyl J99, *Helicobacter pylori J99*; L.lac, *Lactococcus lactis*; M.the, *Methanobacterium thermoautotrophicum*; M.jan, *Methanococcus jannaschii*; M.tub, *Mycobacterium tuberculosis*; M.gen, *Mycoplasma genitalium*, M.pne, *Mycoplasma pneumoniae*; N.men M, *Neisseria meningitis MC58*; N.men Z, *Neisseria meningitis Z2491*; P.mul, *Pasteurella multocida*; P.aer, *Pseudomonas aeruginosa*; P.aby, *Pyrococcus abyssi*; P.hor, *Pyrococcus horikoshii*; R.pro, *Rickettsia prowazekii*; Syne, *Synechocytis sp.*; T.aci, *Thermoplasma acidophilum*; T.mar, *Thermotoga maritime*; T.pal *Treponema pallidum*; U.ure, *Ureaplasma urealyticum*; V.cho, *Vibrio cholerae*; X.fas, *Xylella fastidiosa*.

Figure 7a - 7jj (below). A plot of axis 1 against axis 2 generated from correspondence analysis of RSCU values for each of the thirty-six genomes not previously shown. Highly expressed genes are shown in red.

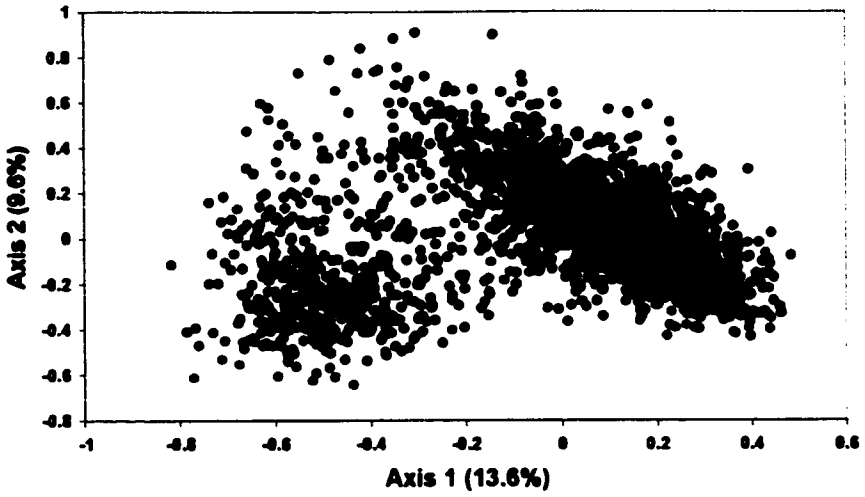

Figure 7a. *Aeropyrum pernix*

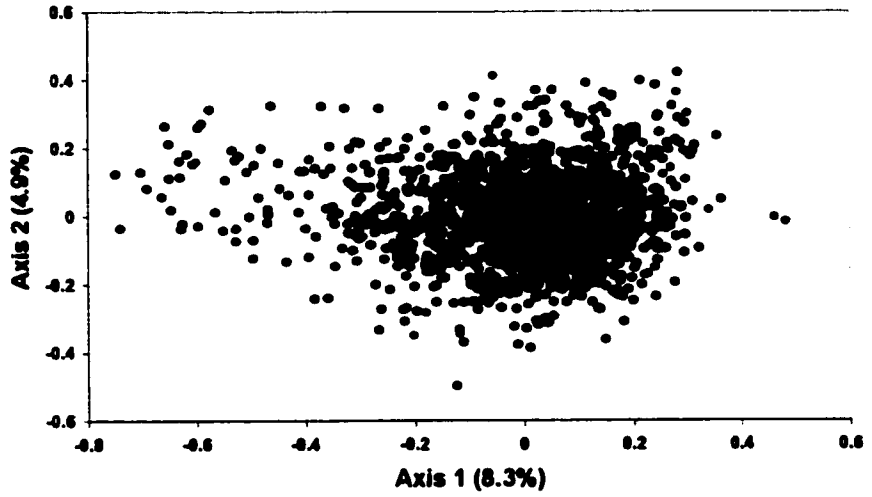

Figure 7b. *Aquifex aeolicus*

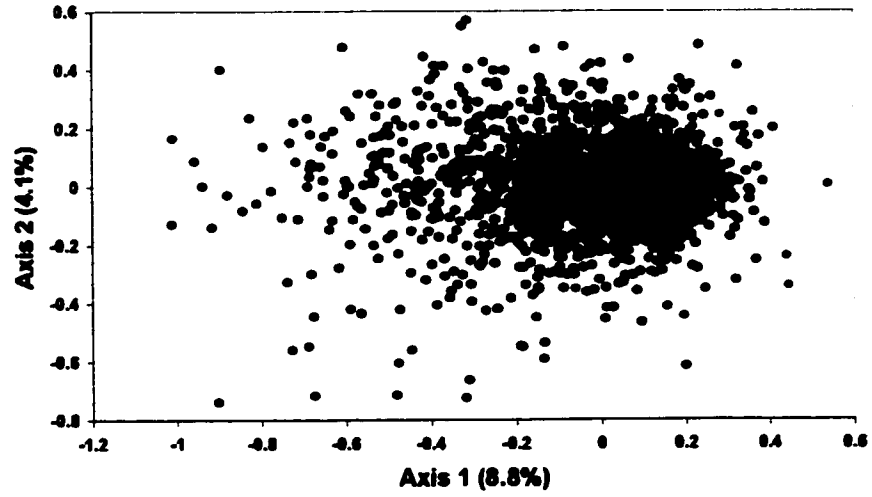

Figure 7c. *Archaeoglobus fulgidus*

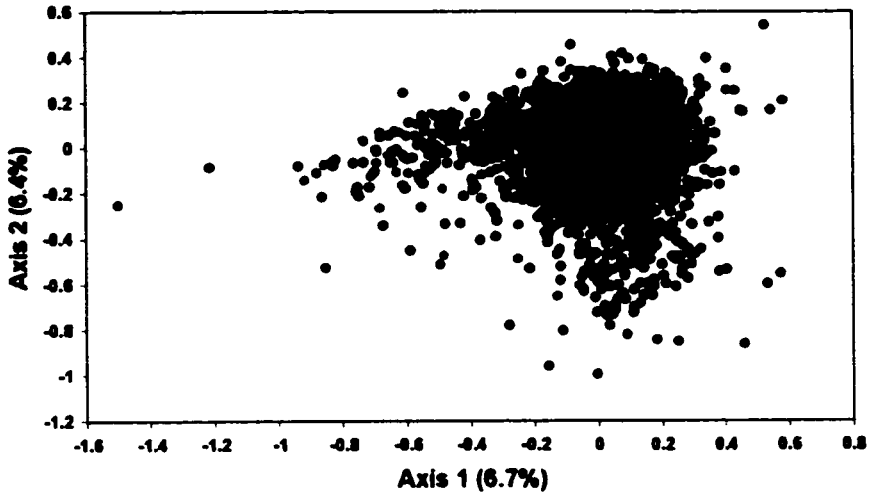

Figure 7d. *Bacillus halodurans*

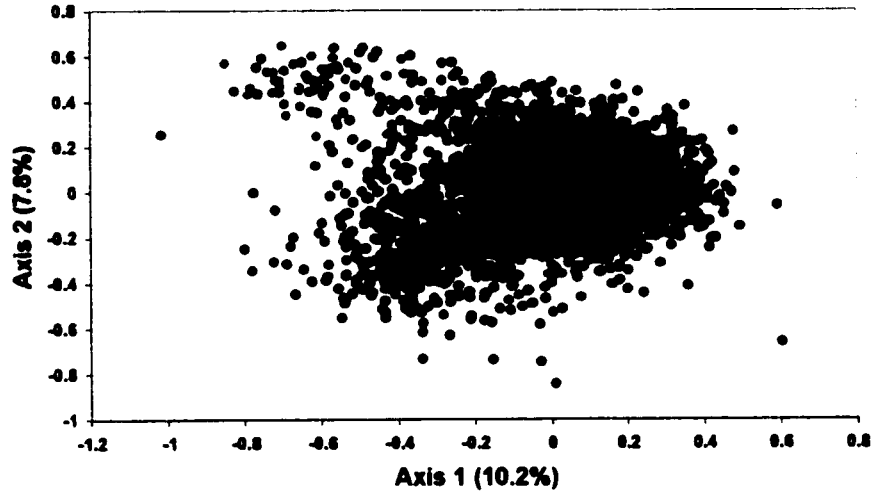

Figure 7e. *Bacillus subtilis*

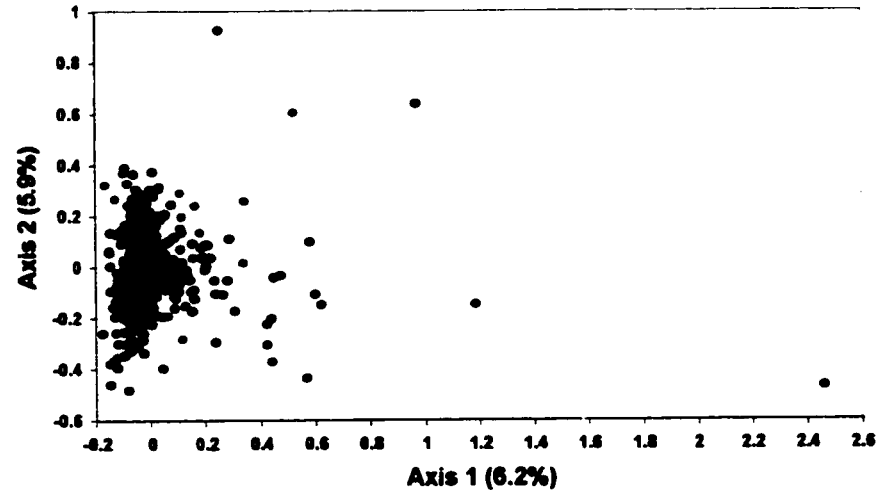

Figure 7f. *Buchnera sp. APS*

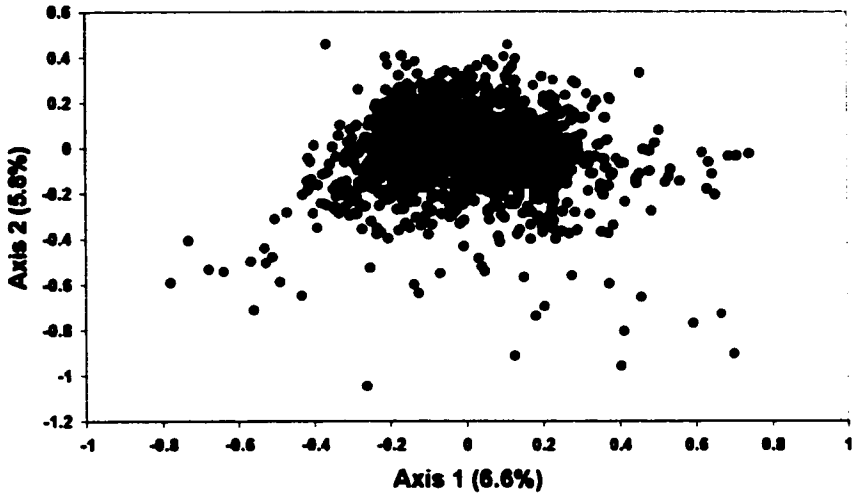

Figure 7g. *Campylobacter jejuni*

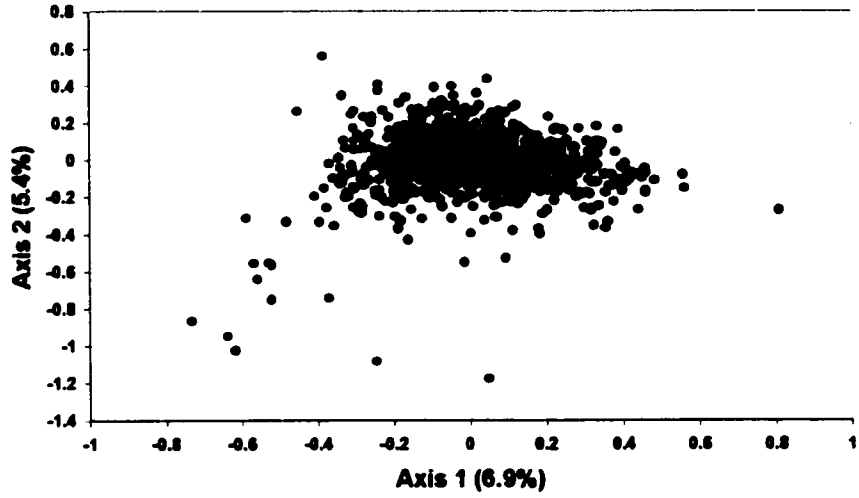

Figure 7h. *Chlamydophila pneumoniae AR39*

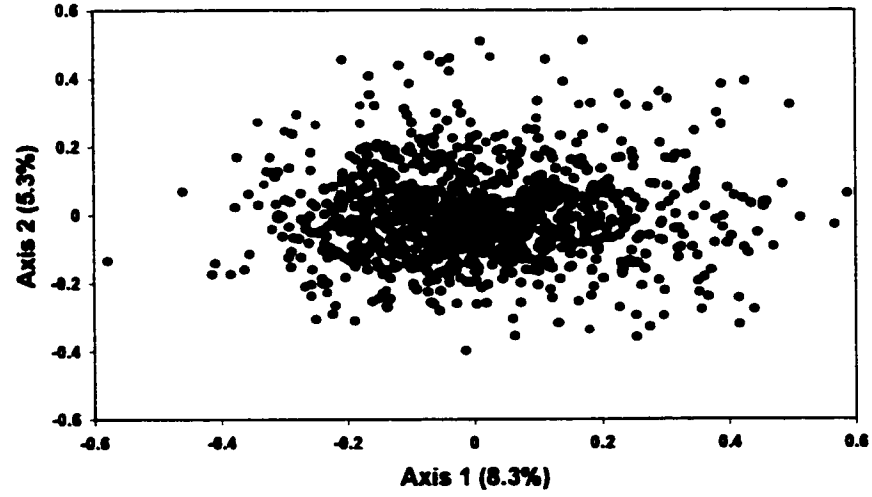

Figure 7i. *Chlamydophila pneumoniae CWL029*

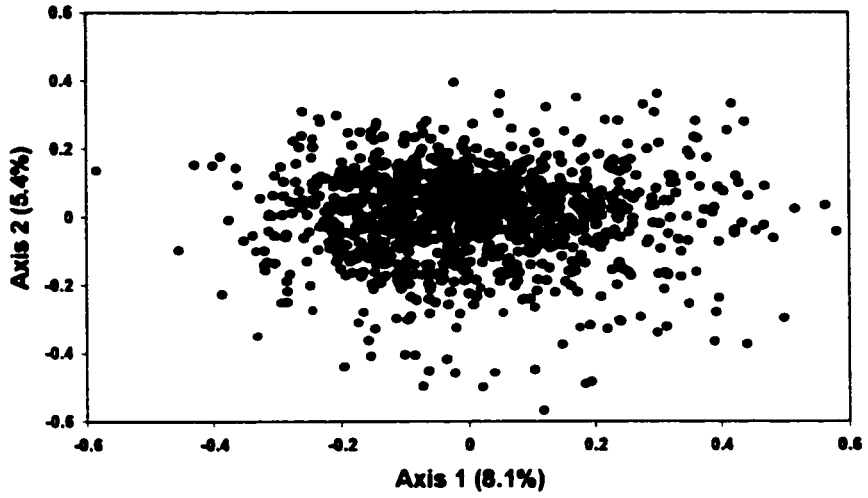

Figure 7j. *Chlamydophila pneumoniae J138*

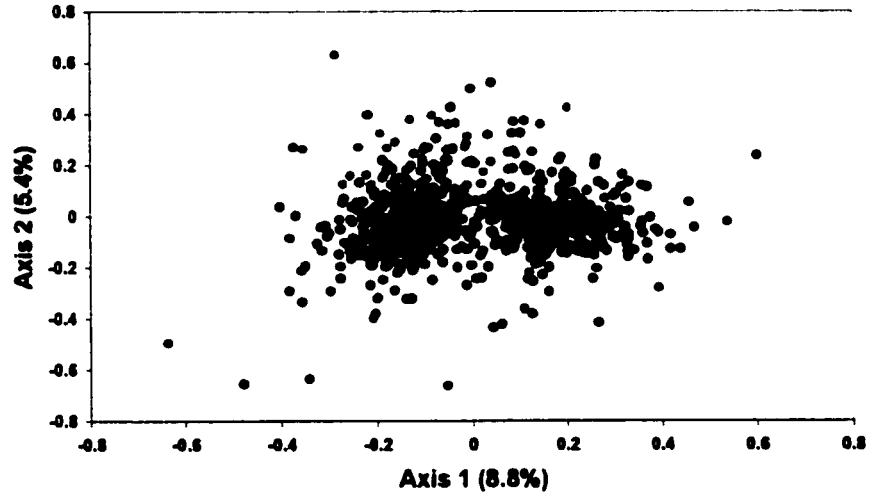

Figure 7k. *Chlamydia muridarum*

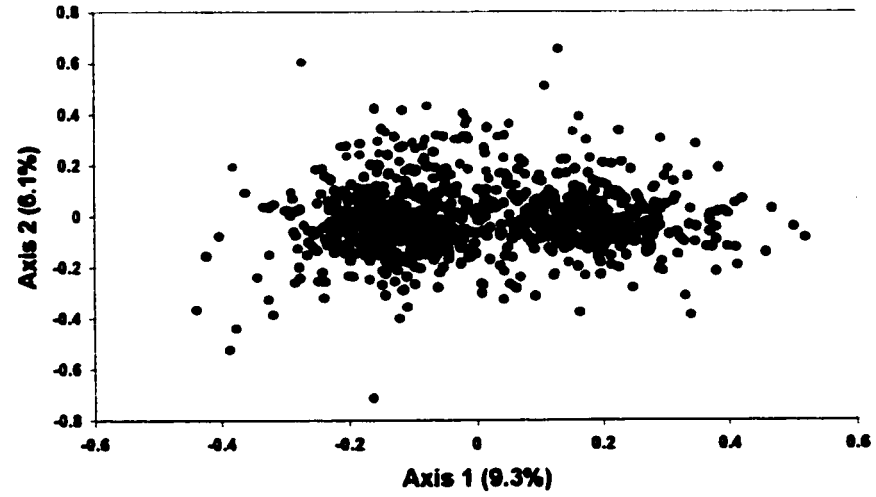

Figure 7l. *Chlamydia trachomatis*

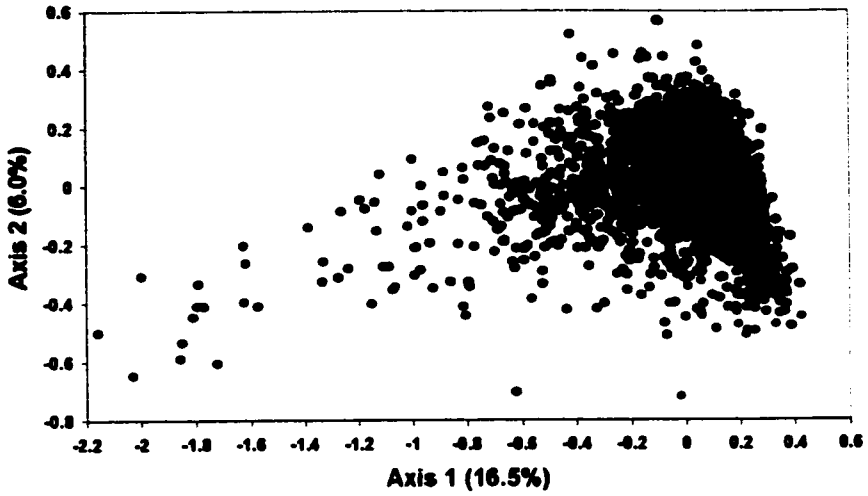

Figure 7m. *Deinococcus radiodurans*

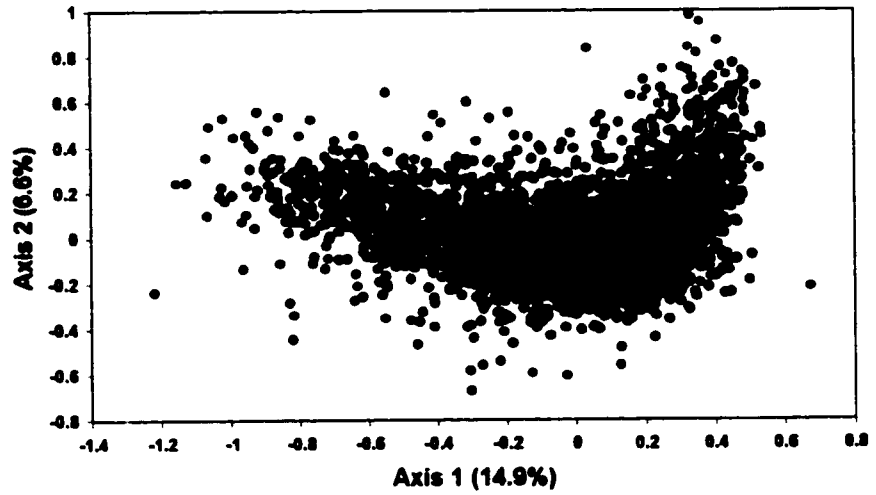

Figure 7n. *Escherichia coli 0157*

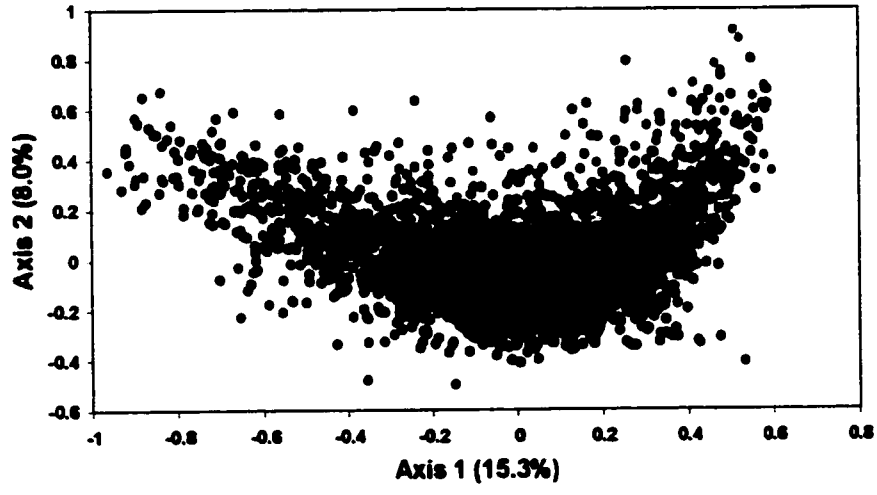

Figure 7o. *E.coli K12*

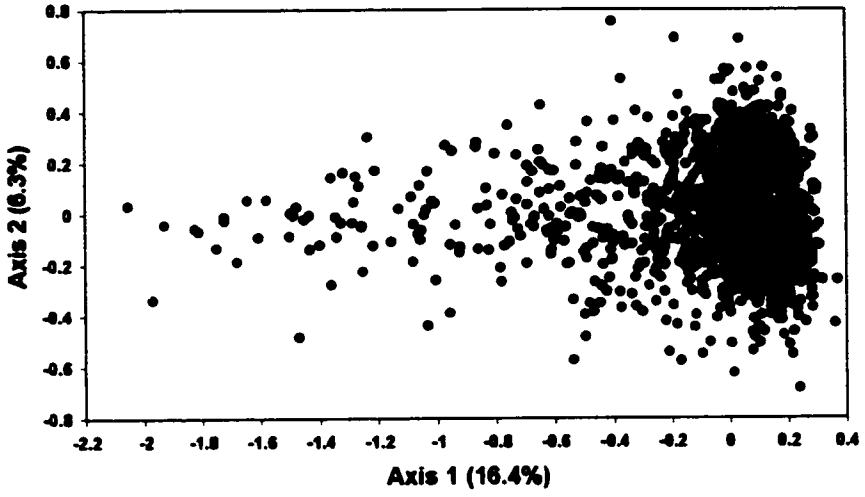

Figure 7p. *Halobacterium sp.*

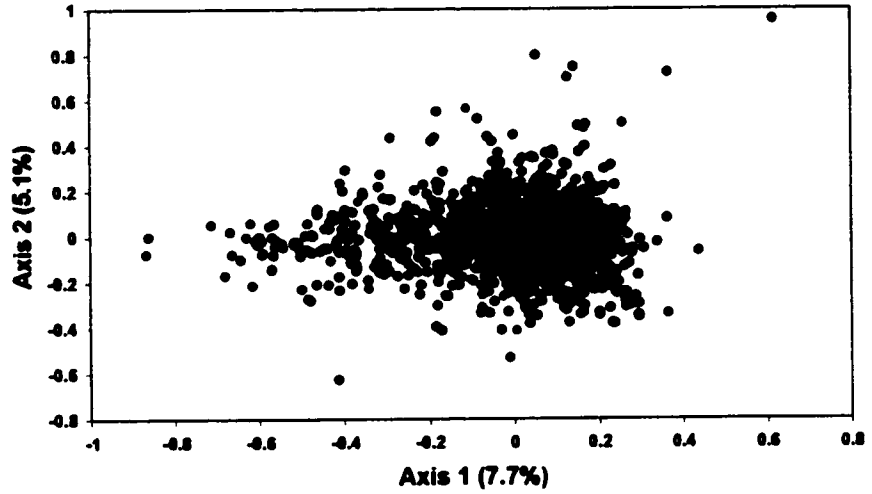

Figure 7q. *Helicobacter pylori 26695*

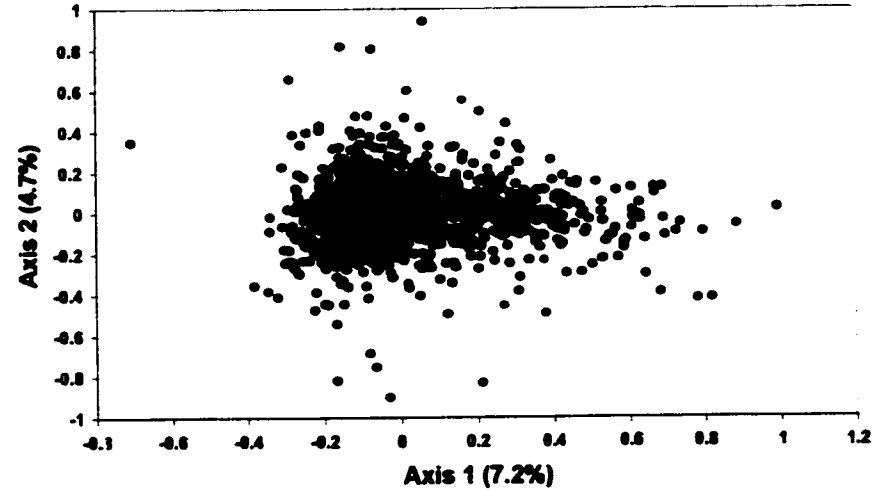

Figure 7r. *Helicobacter pylori J99*

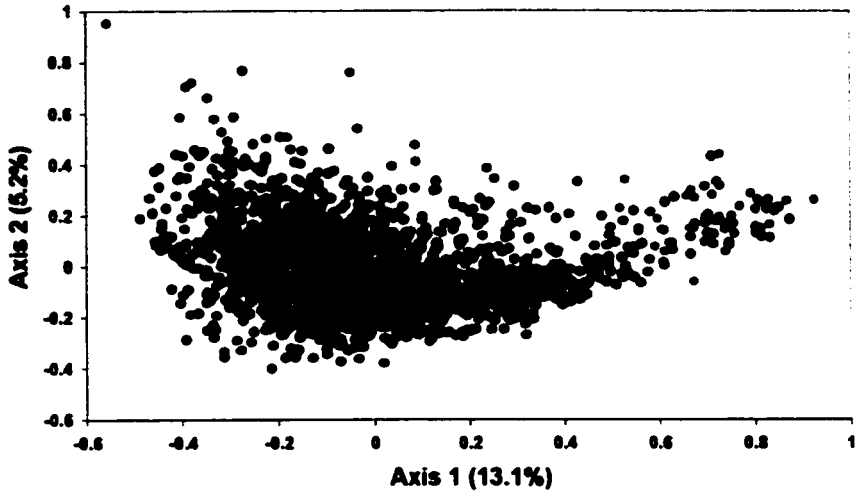

Figure 7s. *Lactococcus lactis*

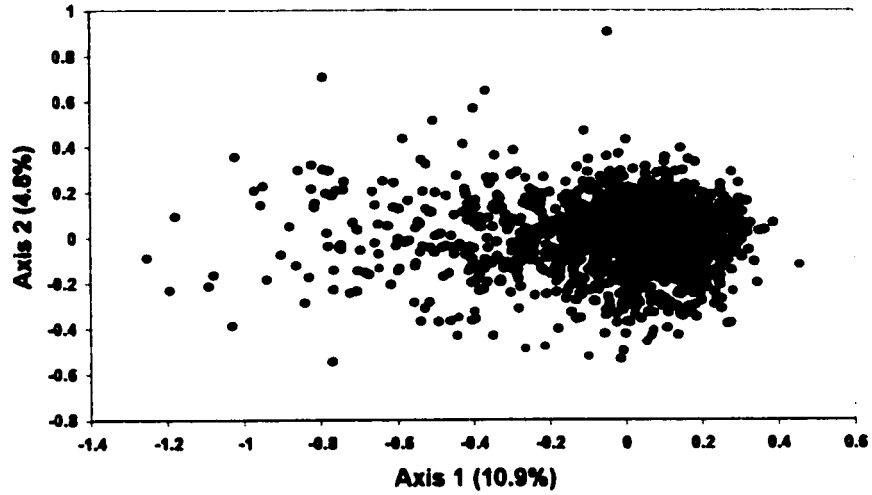

Figure 7t. *Methanobacterium thermoautotrophicum*

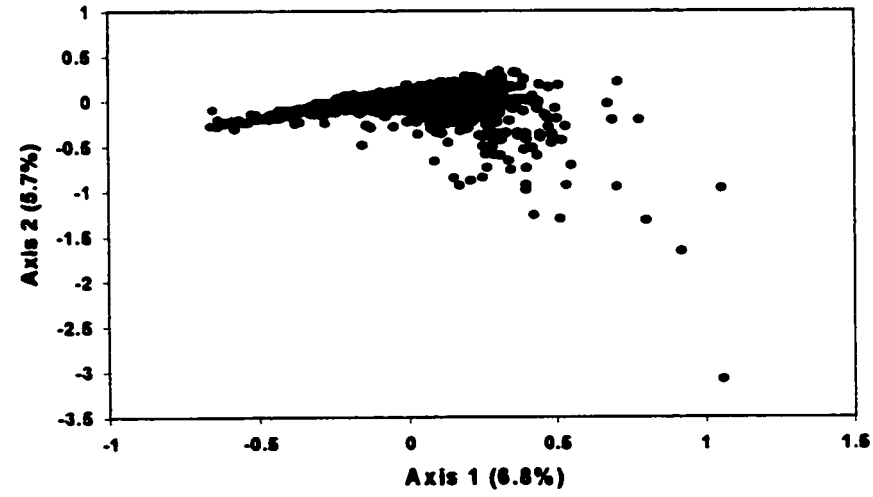

Figure 7u. *Methanococcus jannaschii*

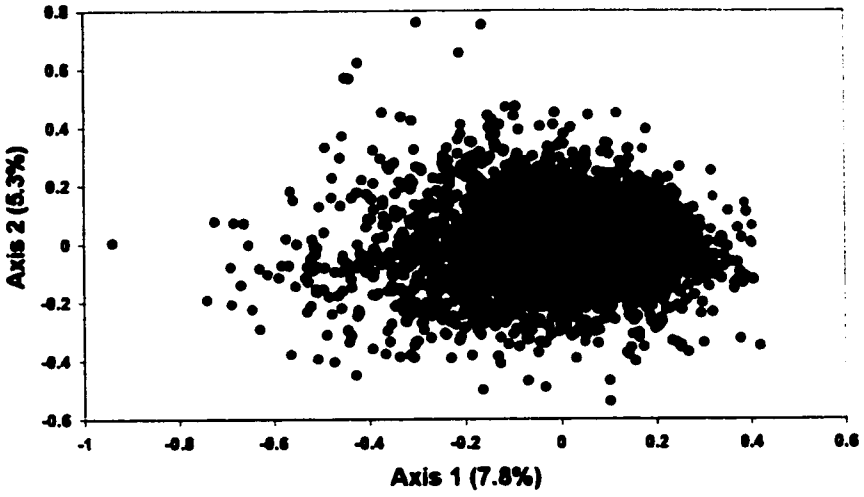

Figure 7v. *Mycobacterium tuberculosis*

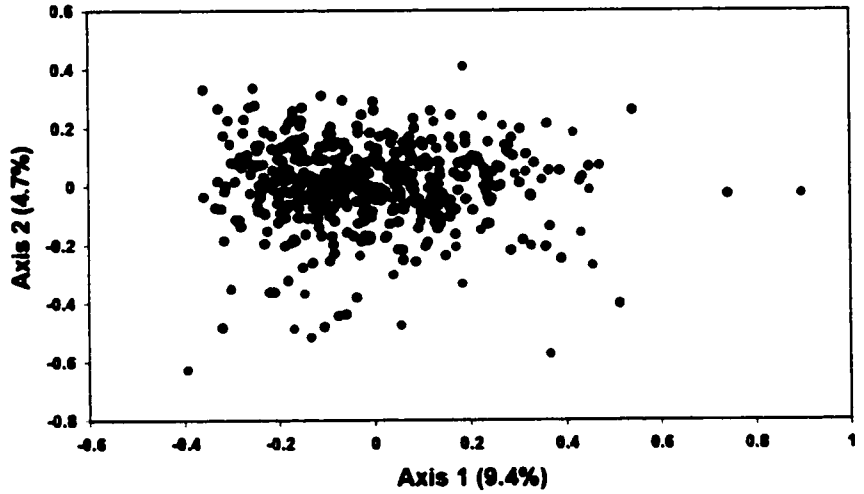

Figure 7w. *Mycoplasma genitalium*

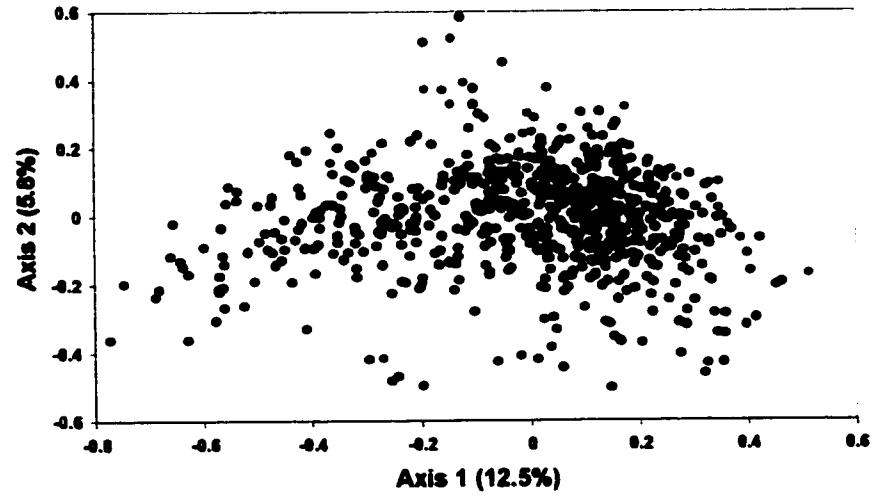

Figure 7x. *Mycoplasma pneumoniae*

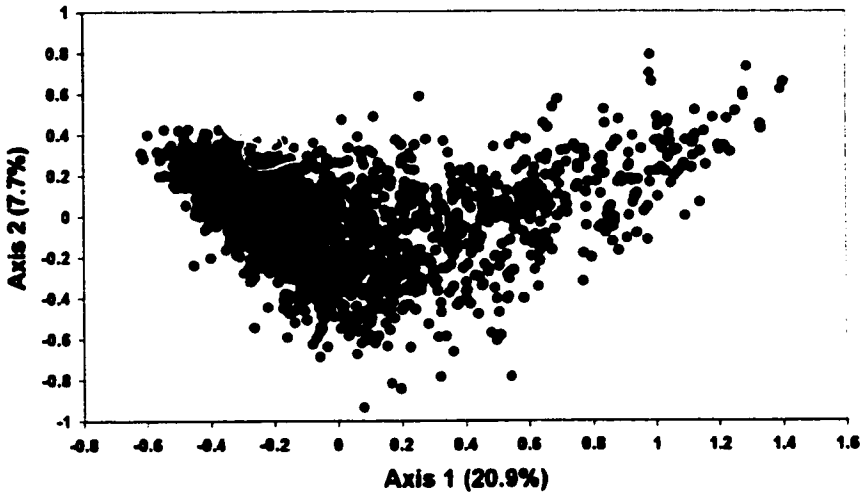

Figure 7y. *Neisseria meningitidis MC58*

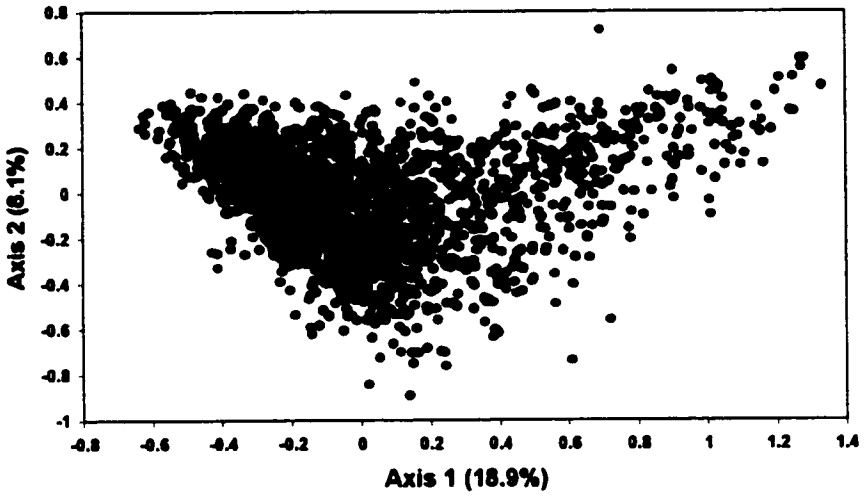

Figure 7z. *Neisseria meningitidis Z2491*

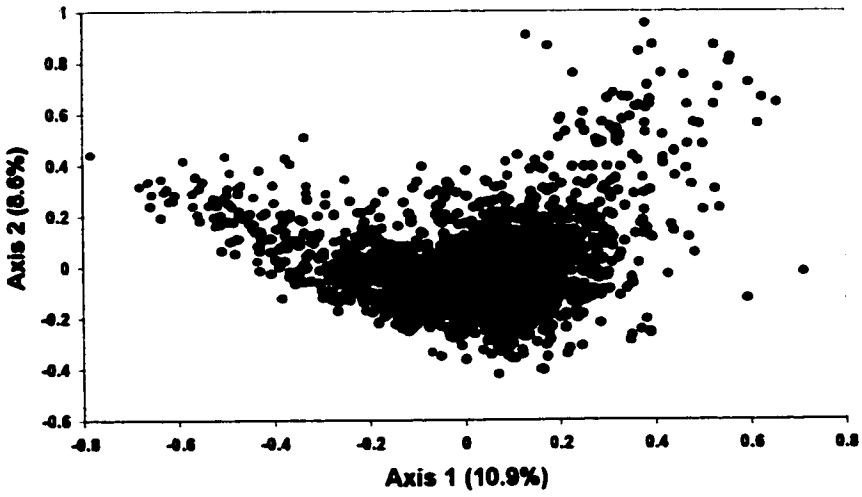

Figure 7aa. *Pasteurella multocida*

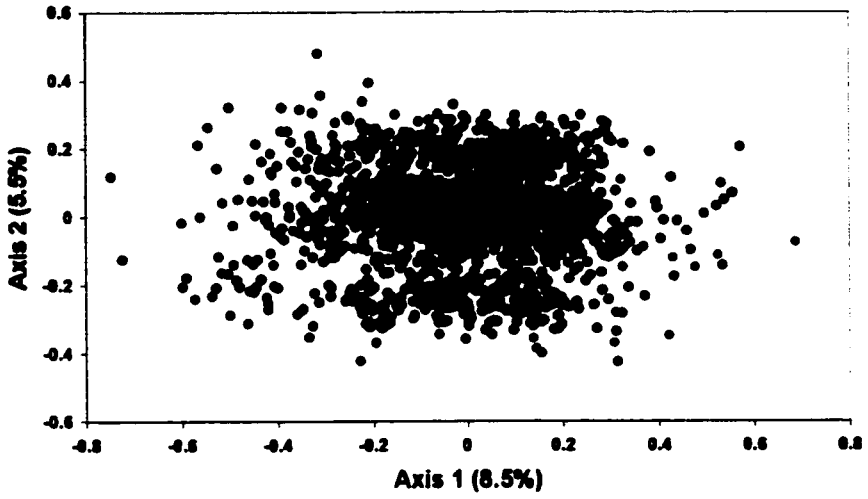

Figure 7bb. *Pyrococcus abyssi*

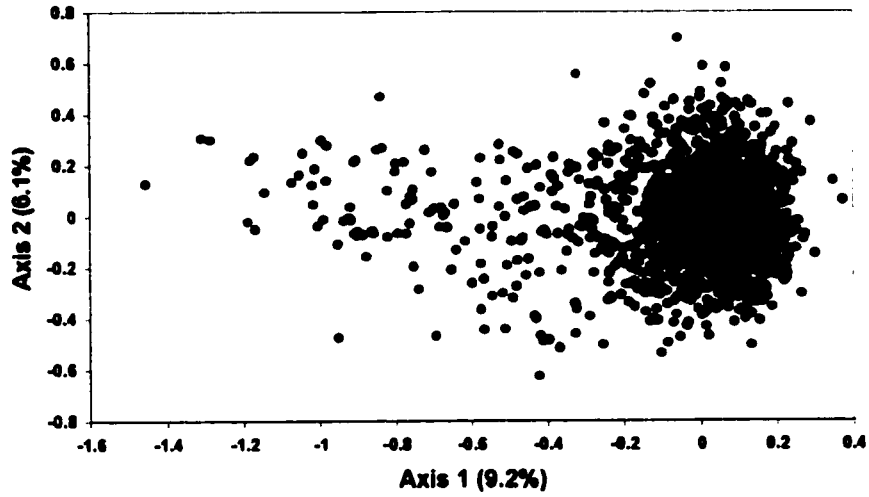

Figure 7cc. *Pyrococcus horikoshii*

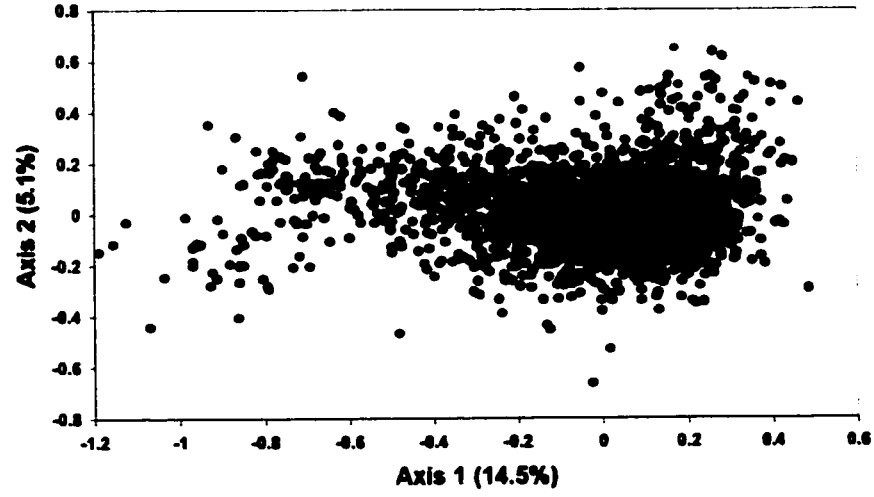

Figure 7dd. *Synechocytis sp. PCC6803*

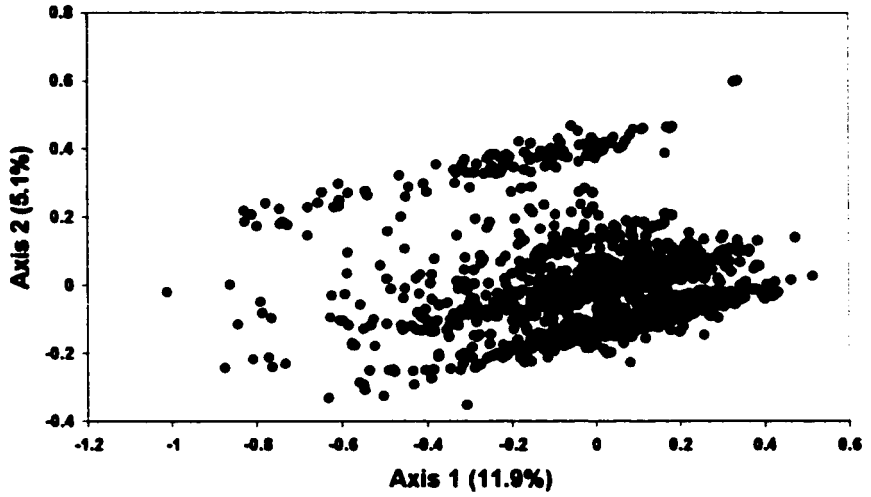

Figure 7ee. *Thermoplasma acidophilum*

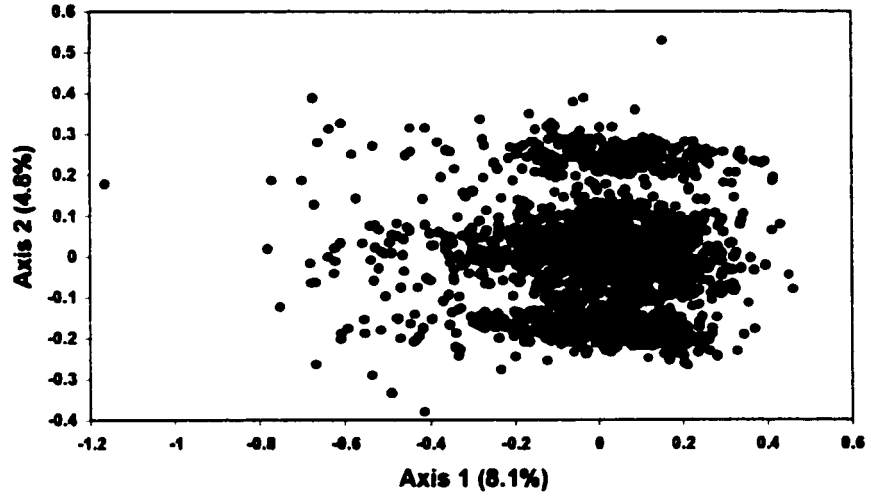

Figure 7ff. *Thermotoga maritima*

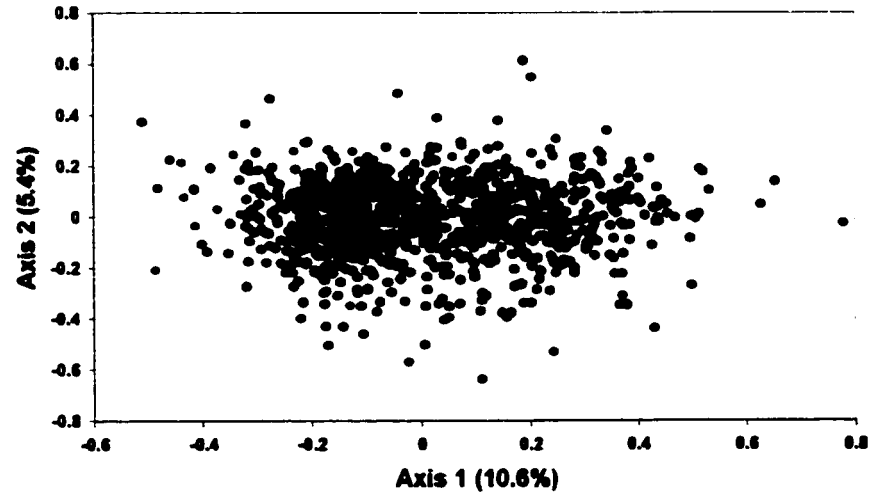

Figure 7gg. *Treponema pallidum*

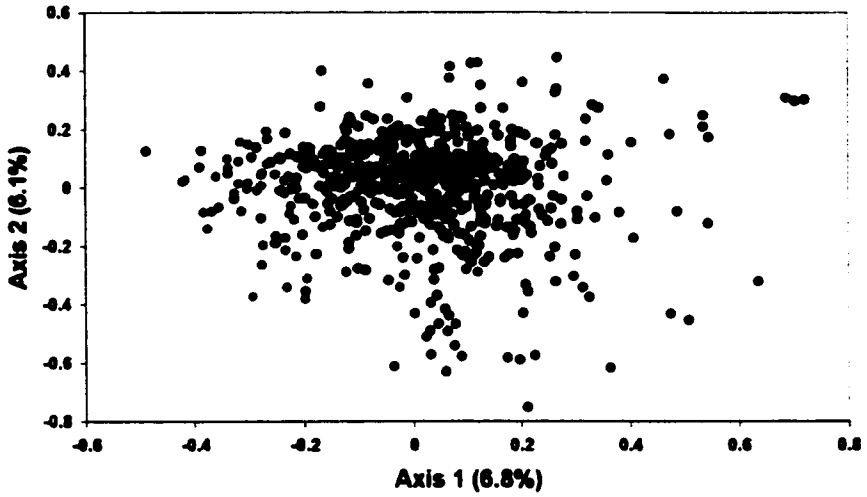

Figure 7hh. *Ureaplasma urealyticum*

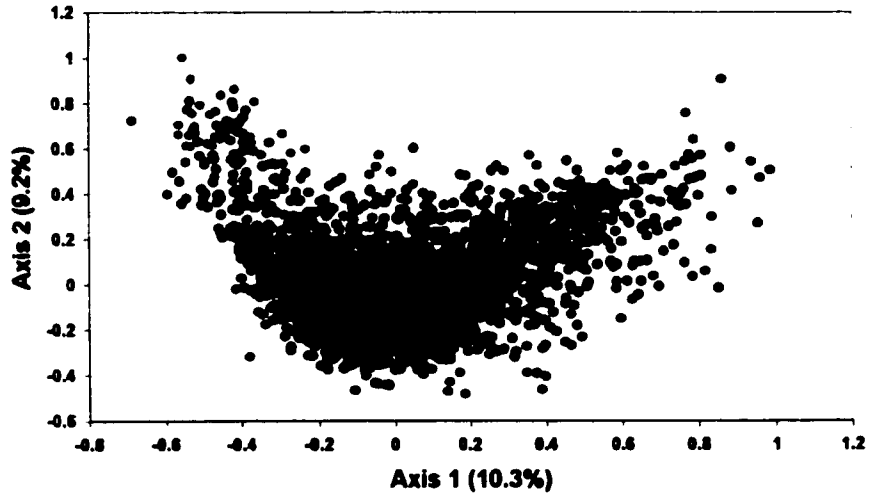

Figure 7ii. *Vibrio cholerae*

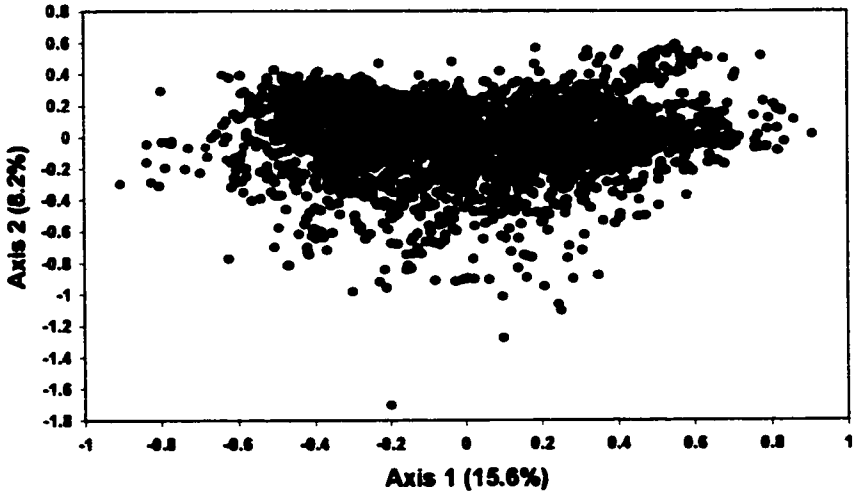

Figure 7jj. *Xylella fastidiosa*