MICROBIAL & COMPARATIVE GENOMICS
Volume 2, Number 1, 1997
Mary Ann Liebert, Inc.

Prokaryotic Genome Evolution as Assessed by Multivariate Analysis of Codon Usage Patterns

JAMES O. McINERNEY

ABSTRACT

This study analyzed the codon usage patterns in three completely sequenced prokaryotic genomes. The major influences on codon usage biases were identified. It was predicted that most of the as yet unidentified open reading frames will be shown to encode functional proteins. It was also shown that translational selection in *Mycoplasma genitalium* has been unable to overcome the effects of random genetic drift. The dual influences of mutational pressure and translational selection are at play in the genomes of *Haemophilia influenzae* and *Methanococcus jannaschii*. In addition, there is considerable variation in base composition around the genome of *M. genitalium*, and this variation is gradual. These data support the theory that regional variation in base composition can be attributed to alterations in the dNTP pool during DNA replication.

INTRODUCTION

Historically, the first microbe whose complete genome sequence was reported was *Haemophilus influenzae* (Fleischmann et al., 1995). This is a free-living member of the Pasteurellaceae, which is in the gamma subdivision of the Proteobacteria (formerly, purple bacteria). Its complete genome was described as 1,830,137 bp with an overall genomic G + C base composition of 38% (Fleischmann et al., 1995). The genome of *Mycoplasma genitalium* is the smallest known in any free-living organism and was found to be 580,070 bp in length, with an overall G + C base composition of 32% (Fraser et al., 1995). The mycoplasmas are a diverse group of wall-less parasites that are specifically related to the low G + C group of the gram-positive bacteria. The wall-less phenotype and small genome are thought to be manifestations of this organism's parasitic lifestyle. The third genome sequence that is available is from *Methanococcus jannaschii*. Unlike the first two prokaryotes, this organism is not a member of the Bacteria (Woese et al., 1990) but of the Archaea, an evolutionarily distinct domain, thought to be more closely related to Eukarya (Iwabe et al., 1989). A circular chromosome and two extrachromosomal elements comprise the genome of this organism. The chromosome is 1,664,976 bp in length and is similar in base composition to the other two, with a G + C content of 31.4% (Bult et al., 1996).

An unusual situation was apparent following the sequencing of the third genome. Whereas a large number of predicted open reading frames (ORFs) from the *H. influenzae* and *M. genitalium* genomes revealed matches in the public repositories (78% in the case of *H. influenzae* and 83% in the case of *M. genitalium*), only 38% of the archaeon's predicted ORFs revealed matches. This situation existed despite the fact that there were approximately 50 recognizable gene families in this genome that had no database matches.

Department of Zoology, The Natural History Museum, London, England.

Codon usage variation is a well-studied phenomenon whose origins can be traced back to when the first nucleotide sequence repositories appeared (Grantham et al., 1981; Ikemura, 1981). Almost all of the 59 codons for which there is a synonymous alternative have been observed to be rarely used in some organisms and frequently used in others. To date, no organisms have been identified where the pattern of codon usage is completely random in all of the genes. The two principal forces that affect codon usage are translational selection and mutational bias. It has been shown that the pattern of codon usage in the highly expressed genes of *Escherichia coli* and *Saccharomyces cerevisiae* correlates very strongly with the known abundances of the isoaccepting transfer-RNAs (tRNAs) (Ikemura, 1981, 1982; Bennetzen and Hall, 1982; Sharp and Cowe, 1991). The advantage of this system (translational selection) is self-evident. Using a codon for which there is an abundant cognate tRNA can speed up the process of mRNA translation. Mutational pressure (the force that influences the base composition of the genome) is also a potent force in fashioning codon usage patterns.

There is often a mutation-selection balance in operation that shapes the overall frequency with which each codon is used (for review, see Sharp et al., 1993). The selective difference between using a codon that will expedite translation and one that will not is thought to be very small. Selection for a preferred codon, therefore, will only happen in a situation where random genetic drift is small. This means that translational selection will be observed only in species where the effective population size is large enough for the effects of random genetic drift to be subdued (Muto and Osawa, 1987). Even for organisms with large effective population sizes, not all genes acquire a rich supply of preferred codons. A gene whose product is required in large amounts or quickly at particular times can exert a greater amount of selective pressure than a gene whose product is not required in large amounts. The situation is seen in many prokaryotes and yeast, where highly expressed genes use a small subset of the 59 available synonymously degenerate codons in their highly expressed genes, whereas the lowly expressed genes use codons in a less restricted fashion (Sharp et al., 1986; Sharp and Devine, 1989; Lloyd and Sharp, 1991, 1993).

MATERIALS AND METHODS

The predicted ORFs for each of the three genomes were retrieved from The Institute for Genomic Research (TIGR) worldwide web server (http://www.tigr.org/tdb/mdb/mdb.html). There were no attempts made to further analyze the ORFs to see if any recent entries in the public repositories would reveal similarities to the unknown ORFs. Correspondence analysis of the relative synonymous codon usage (RSCU) values for each of the codons was carried out. RSCU values are calculated by dividing the observed number of times that a particular codon is used by the number of times it should be observed if codon usage for that amino acid was random. Analyzing RSCU values reduces any biases associated with amino acid composition of the proteins.

To identify translational selection as a force that influences codon usage bias, it is necessary to identify gene products that are expected to be expressed at a high level. In each organism, there are approximately 50 ribosomal protein genes. These gene products are usually expressed at a high level in most organisms because of the requirement for large numbers of functional ribosomes (Srivastava and Schlessinger, 1990). In addition, the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene, which is a key enzyme in glycolysis, is also expected to be expressed at a high level. The translation elongation factor genes are also known to be required in large quantities.

The preferred method of analyzing codon usage variation is correspondence analysis, which involves plotting genes according to their usage of codons in a 59-dimensional space. The axis that displays the greatest amount of variation is then identified, as are subsequent axes. The axis of greatest dispersion identifies the major source of codon usage variation in the dataset (Greenacre, 1984). Other important statistics that can help identify trends in a dataset include the effective number of codons (ENC) and the G + C base composition at synonymously degenerate third positions (GC3s). ENC is a measure that ranges from 61 for a gene that is using all codons with equal frequency to 20 for a gene that is selectively choosing a single codon for each amino acid. For details of the calculation, see Wright, 1990. Correspondence analysis was carried out on all of the predicted ORFs whether or not they had been associated with a cellular role. The four axes of greatest dispersion were identified in each case. A separate analysis was carried out on the known and unknown ORFs. This was done to test whether the unknown ORFs followed the same general patterns as the known proteins.

The software that was used for these analyses was kindly provided by Andrew T. Lloyd (Irish National Centre for BioInformatics) and included CODONS (Lloyd and Sharp, 1992), CU, RSCU, DECOFO, DECOR, HILO, CUTAB, CUSORT, TIDY, ADD2CU, ENC, BIAS, and CORRCHI.

RESULTS

Correspondence analysis of the known and the unknown ORFs revealed that in all cases the trends were quite similar (data not shown). This is an indication that the majority of the remaining ORFs will, in time, be proven to encode functional proteins. In all three genomes, the amount of variation attributable to the first four axes was comparable. In *H. influenzae*, the first four axes account for 12.6%, 9%, 6.8%, and 5.4% of the variation, respectively. In *M. genitalium*, the values were 13%, 8.3%, 6.6%, and 6.1%. In *M. jannaschii*, the four principal axes explained 12.3%, 8.7%, 6.9%, and 6.2% of the variation. These values are much lower than those reported for other studies [e.g., in a study on *Kluyveromyces lactis* (Lloyd and Sharp, 1993), the principal axis was seen to account for 44% of the variation in codon usage]. It is likely that such high figures were due in some way to gene sampling.

Figure 1(A, B, and C) shows a plot of axis 1 (the axis showing the greatest amount of dispersion) against axis 2 (the next most important axis). Some proteins that are expected to be expressed at a high level are shown as open squares. These include ribosomal proteins (a total of 54 for *H. influenzae*, 52 for *M. genitalium*, and 62 for *M. jannaschii*), elongation factors (a total of 5 for *H. influenzae*, 4 for *M. genitalium*, and 2 for *M. jannaschii*), and a single GAPDH gene in each organism. The rest of the data is shown as closed diamonds.

The overall shapes of Figure 1A and C are strikingly similar. In both cases, the highly expressed genes tend to accumulate at the left side of the graphs. This is the part of axis 1 that has the most strongly biased genes. In addition, the graphs can be fitted quite well with a binomial curve. In contrast, Figure 1B, which is the plot of the genes for *M. genitalium*, is remarkably different. In this instance, the binomial shape is not evident, and the highly expressed genes are spread throughout the dataset in a seemingly random fashion. The dataset as a whole resembles a circular plot, loosely centred around the origin of the graph. This appears to indicate that a different force governing codon usage is at play in the genome of *M. genitalium.* It was noticed that axis 1 for the *M. jannaschii* genome was unusual in that a small cluster of genes was found on the right side of the axis (Fig. 1C). An investigation of the probable cause of this clustering was carried out, and the only factor that appeared to be common to all of these genes is lack of (or very low levels of) proline residues in their amino acid sequences. Some of these genes have been identified, but the rest are unassigned with respect to function. Because of this clustering, which leads to axis 1 being identified as the major source of variation, axis 2 (the next most important trend) was chosen for subsequent analyses.

An investigation of the correlation of position on the axis of greatest dispersion and the G + C base composition at the third position of codons for which there is a synonymous alternative (GC3s) was carried out (Fig. 1D,E,F). Again, the *H. influenzae* and the *M. jannaschii* genomes showed a remarkable degree of similarity. The overall shape of the plots is amorphous, indicating a random process. The highly expressed gene sets are found virtually throughout the range of base compositions (in Fig. 1D the axis of greatest dispersion is used, and in Fig. 1F the next most important axis is used; this explains why the highly expressed genes are on opposite sides of the graphs). In the case of *M. genitalium*, however, a very different situation is evident. The GC3s of the genes in the dataset are very tightly associated with the variation in codon usage. The highly expressed genes are evenly distributed along the curve, and the regression coefficient for the fit of the data is 0.901. This is perhaps the strongest evidence that codon usage in this organism is influenced very strongly by mutational pressure and that translational selection has been unable to overcome this influence. The range of GC3s values indicates that the *M. genitalium* genome as a whole is strongly biased, with even the most equivalent values rarely venturing higher than 40% G + C. The genome of *M.*

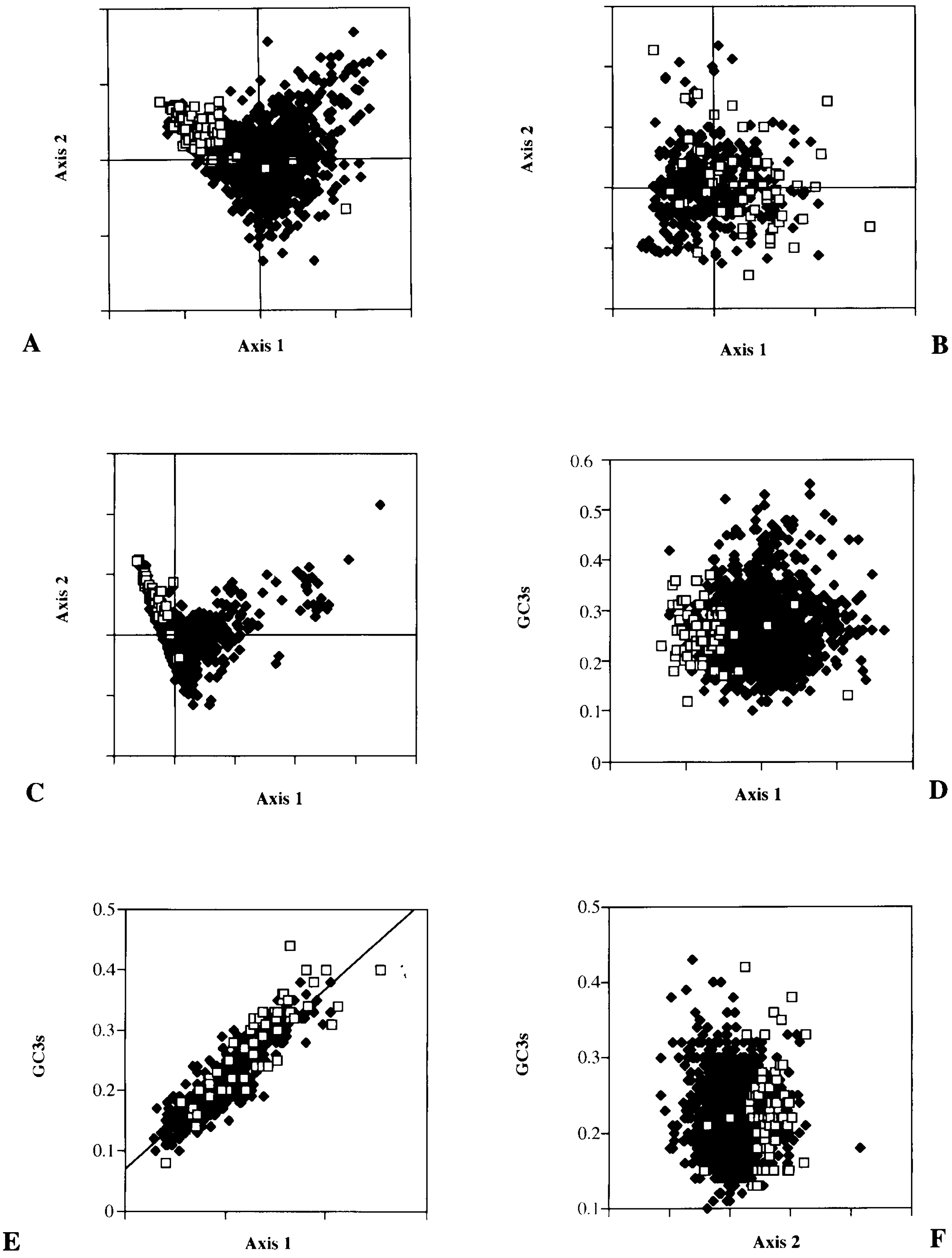

FIG. 1. The genomes are ordered: *H. influenzae* (A,D,G,J), *M. genitalium* (B,E,H,K), and *M. jannaschii* (C,F,I,L). **A,B,C.** Position on axis 1 plotted against position on axis 2. The absolute values for a correspondence analysis are meaningless. However, the lines define the center of the axes. **D,E,F.** Position on the axis of greatest dispersion plotted against G + C base composition at the third position of codons for which there is a synonymous alternative. **E.** Regression curve fitted to the points (y = 0.003x + 0.219; R = 0.901).

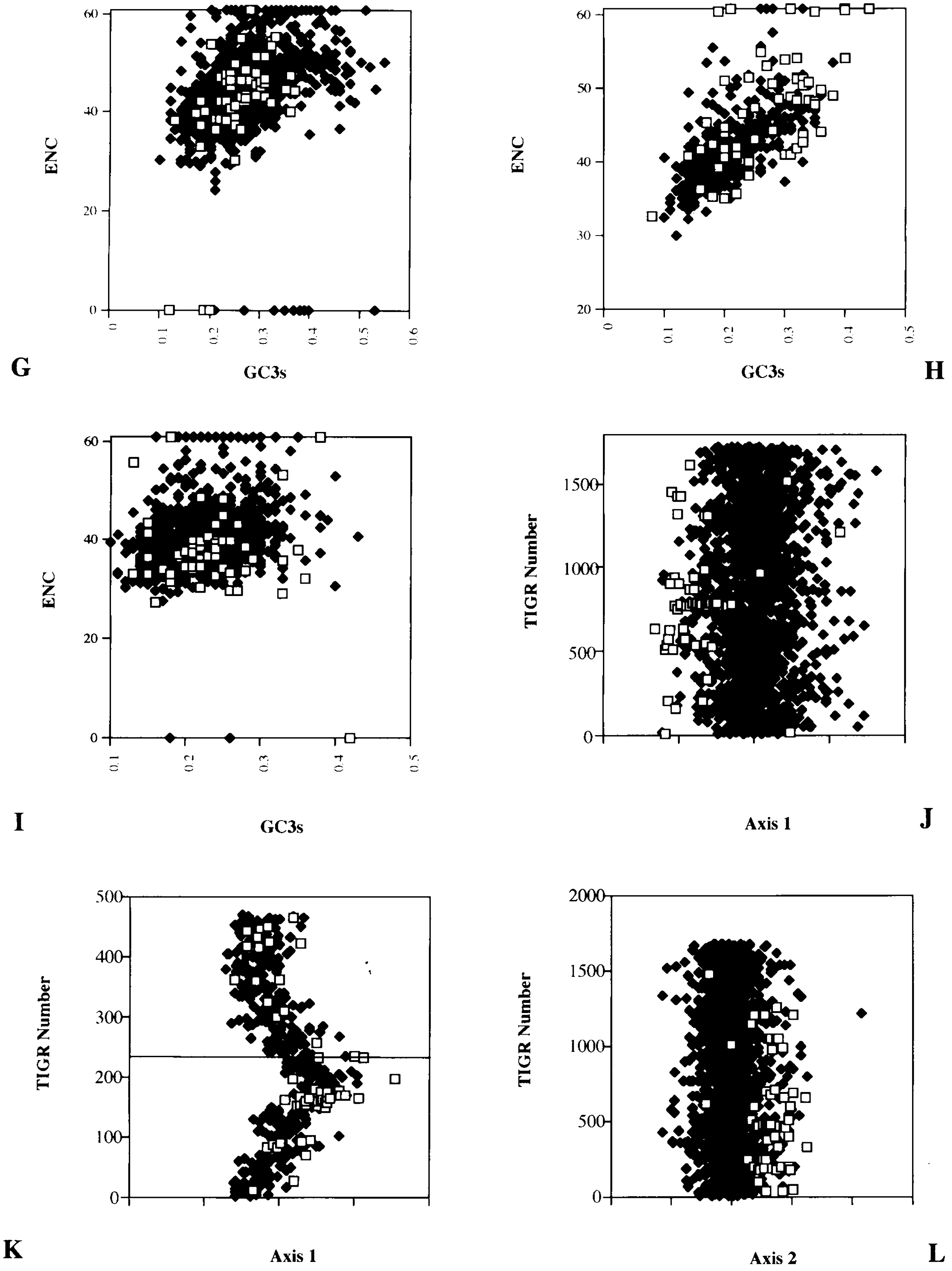

FIG. 1 (cont.) **G,H,I.** GC3s vs the effective number of codons (ENC) measure for the gene. **J,K,L.** Position on the axis of greatest dispersion plotted against TIGR number. The horizontal line denotes the location on the genome diametrically opposite the proposed origin.

jannaschii, however, appears to have a similar bias, with *H. influenzae* being only slightly less biased. It is not, therefore, likely that the sole reason for such an atypical mode of codon selection in *M. genitalium* is a strong mutational pressure, one that simply cannot be overcome by translational selection.

The GC3s values were plotted against the ENC value for that gene (Fig. 1G, H, I). In all three diagrams, the genes that are expected to be expressed at high levels are scattered throughout the graphs. The diagrams for *H. influenzae* and *M. genitalium* show the most similarities in this instance. The trend is most obvious in the *M. genitalium* genome, where there is a very obvious correlation between GC3s and ENC. *H. influenzae* does show some signs of correlation between the two factors. The graph of the *M. jannaschii* genome does not show an obvious trend. Low ENC values are as likely to be found in genes with relatively high GC3s values as they are to be found in genes with relatively low GC3s values.

To see whether some other factors are having a major effect on codon usage, a number of other statistics were examined to see if there was a correlation. First, the major axes of dispersion were compared with the abundances of each of the amino acids. There were no apparent trends in any of these analyses (unpublished observations). However, an analysis of the correlation between position on the axis of greatest dispersion and position on the genome revealed an unexpected situation. The results of this analysis are shown in Figure 1J, K, L. A precursory analysis of the variation in gene length around the genome revealed that there were no significant pockets of long or short genes in any chromosome and that genes of all lengths were quite evenly distributed around the genome (unpublished observations). For this reason, it is legitimate to use the TIGR number (the number assigned to each ORF by The Institute for Genomic Research), as this is approximately equivalent to position on the genome, in minutes, from the postulated origin of replication.

As can be seen in Figure 1J, K, L, the highly expressed genes are distributed quite well, although not entirely randomly, along the genome of each organism. The behavior of all of the genes in the *H. influenzae* and *M. jannaschii* genomes is, once again, strikingly similar. There is no correlation between position along the genome and position on the axis of greatest dispersion. The situation is quite different in *M. genitalium*, where significant variation is seen with respect to position on the genome. Although the diagram is linear, it is a representation of a circular genome. The genes located at the top and bottom of the graph are closest to the presumed origin of replication. With a slight offset from the origin, there are two points, almost diametrically opposite one another, where codon usage patterns differ most on the genome.

DISCUSSION

The plots of the first and second most important axes following correspondence analysis showed two different types of pattern. For *H. influenzae* and *M. jannaschii*, the plot is similar to what is termed a "rabbit head" plot. This has been seen previously in *E. coli* (Médigue et al., 1991). The rabbit head plot of *E. coli* genes was interpreted to mean that there were three distinct groups of genes present in the genome. These groups of genes could be classified both on the basis of codon usage bias and in terms of the biologic properties of the genes. One group was composed almost entirely of genes that were expressed at constitutively high levels. It is of note, therefore, that the plots for *H. influenzae* and *M. jannaschii* show the highly expressed genes clustering together. It is likely that a picture similar to that seen in *E. coli* will emerge from these two genomes. The *M. genitalium* genome plot, however, does not show any of the characteristics of the other two. The rabbit head plot is not evident, and the genes expected to be highly expressed are spread throughout the graph.

The small cluster of points that are slightly separated from the rest on the *M. jannaschii* plot (Fig. 1C) remains enigmatic. The only factor that was seen to correlate with the position of these genes on axis 1 was low levels or complete lack of proline residues. In general, *M. jannaschii* proteins have some of the lowest levels of proline for any available large genomic dataset (data not shown). The use of RSCU values is quite effective in reducing the effects of amino acid composition on a correspondence analysis. Médigue et al. (1991) found that some genes on the end of rabbit head graphs, opposite the highly expressed genes, were laterally transferred. This is another theory, but in the absence of information about the identities of most of the genes in this cluster, it is difficult to ascribe a reason for their position.

When GC3s values were plotted against the most important axes following correspondence analysis, a very unusual feature was evident. Whereas the *H. influenzae* and *M. jannaschii* genomes did not appear to show any correlation between variation in base composition and position on the axis of greatest dispersion, the *M. genitalium* genome showed a very significant trend (Fig. 1D, E, F). Although the plot of the two most important axes following correspodence analysis was largely uninformative about the forces governing codon usage variation in the genome (Fig. 1B), this plot clearly shows that variation in mutational bias is a very strong force. It is also significant that the lowest and highest values for GC3s are comparable for all three genomes. There is no reason to suggest from this analysis that there is any unique mutational pressure in the genome of *M. genitalium*, nor is there an unusual amount of variation in base composition.

Codon usage in a genome is thought to be subject to the influences of mutation, selection, and random genetic drift. The net codon usage in the genome reflects the balance that has been reached by these pressures. The mutational bias of the genes in *M. genitalium* is not excessive (at least no more than in the other two genomes). Therefore, the selective pressure exerted by translation in this organism is not particularly strong. Advantageous changes can only become fixed in a population if either the advantage is considerably large or the population is sufficiently large that random genetic drift is negligible. In a situation where the long-term effective population size of an organism is small, the selective differences between alternative synonymous codons is not sufficient to overcome the effects of random genetic drift. The manifestation of this situation is usually seen in larger animals (whose long-term effective population sizes are known to be small), where codon usage of a gene is governed largely by the base composition of that gene and not by its level of expression. Prokaryotes are expected to have large long-term effective population sizes. In such organisms as *E. coli* (with a large long-term effective population size), translational selection is known to have a major influence on codon usage bias, and the correlation between codon usage bias and gene expression is very strong.

It has been reported recently that codon selection in the intracellular parasite *Rickettsia prowazekii* (Bacteria: alpha-proteobacteria) does not appear to be subject to translational selection pressures (Andersson and Sharp, 1996). This is particularly interesting in light of the similarities in lifestyle between *R. prowazekii* and *M. genitalium*. Both are parasitic, with reduced genome sizes. Also of interest is that these organisms are only distant relatives, both genomes contain only a single copy of each of the rRNA genes, and both organisms have relatively slow generation times. It is conceivable, therefore, that under certain situations, where a prokaryote has evolved by a degenerative process, with a reducing genome size and a lifestyle with very little change in environmental conditions (in this instance, both organisms are parasites and so the external environment is not constantly changing), perhaps the pressure exerted by translational selection may not be sufficiently strong to become the dominant force in shaping codon selection.

In Figure 1G, H, I, the ENC values are plotted against GC3s values. The purpose of this plot is to investigate if variation in base composition (mutational bias) is having an effect on the number of codons that are used by a gene. It is expected that in instances where mutational bias is having a strong effect on codon selection, there will be a correlation between the two factors [in fact, the relationship between ENC and GC3s under the null hypothesis of no selection and G + C bias being due to mutation can be approximated mathematically (Wright, 1990)]. It can be seen from Figure 1 G and H that there is a correlation between the two factors. As the mutational bias in a gene becomes less pronounced, the effective number of codons in the genes becomes greater. This is an indication that mutational bias is playing a strong role in shaping the codon selection in both of these genomes. The graph of the *M. jannaschii* genome does not appear to display the same strength of correlation. If variation in mutational biases are indeed shaping codon choice in this genome, the effects are not so pronounced. In *E. coli* and *S. cerevisiae* there is also a lack of correlation between ENC and GC3s (Wright, 1990). However, these organisms have very low levels of variation in GC3s values and also the genomic G + C base composition is close to 0.5. Another aspect of these plots that is noteworthy is the positions of the highly expressed gene sets. In each case, the highly expressed genes are spread throughout the rest of the data. Although it appears that the majority of the highly expressed genes in the *M. jannaschii* genome are seen to have low ENC values (highly biased), some of the genes have ENC values of 61.

The results of the plots in Figure 1G and I are unusual in light of the graphs in Figure 1A and C. In the correspondence analysis graphs, the highly expressed gene dataset clustered together very tightly. This is an indication that their codon usages are similar and that one of the principal forces governing codon usage in both of these genomes is level of expression. In the graphs of GC3s vs ENC, however, there is an obvious spread of base compositions and biases. Highly expressed genes in these genomes have a wide range of GC3s values, and the effective number of codons that they use is also wide ranging. These factors taken together suggest that the highly expressed datasets are preferentially using a set of codons that differ in some way from the rest of the genes in the organisms and that this codon usage is not based specifically on G + C base composition, nor are they distinguished by extremes of ENC values. The same is not true for *M. genitalium*, where the highly expressed genes are not distinguished by the correspondence analysis (Fig. 1B).

It was shown in Figure 1E that there was a strong correlation between GC3s and position on axis 1 in *M. genitalium*. Figure 1K shows that this correlation is also related to genome location. Position on the axis of greatest dispersion was not correlated with genome location in either of the other two genomes. GC3s and position on the genome were also plotted against one another (data not shown), with similar results to those seen in Figure 1J, K, L. This is one of the most startling results of these analyses and suggests some kind of physical reason for codon usage variation in *M. genitalium*. Traditionally, codon usage variation in prokaryotes has been viewed as the result of mutational bias or translational selection, with no suggestion that position on the genome has an influence. The opposite is the case in mammals. In warm-blooded vertebrates there are discrete regions of base compositional bias called isochores (Bernardi, 1993). There are two suggestions concerning the reasons for variation in base composition in isochores. Whereas Bernardi (1993) argues that there is a selective advantage conferred by this mode of evolution, Wolfe et al. (1989) argue that it is merely a reflection of mutational bias under varying dNTP pools during DNA replication. Although *M. genitalium* does not display the discreteness that is characteristic of isochores, it does seem to support the theory of Wolfe et al. It is possible that *M. genitalium* experiences a depletion of adenosine and thymidine nucleotides as DNA replication proceeds. This leads to pressure toward misincorporation of guanine and cytosine residues as DNA replication approaches completion.

Whatever the explanation for the variation in base composition around the genome of *M. genitalium*, the phenomenon has a number of repercussions. Most maximum likelihood models of sequence evolution assume that erroneous misincorporation during DNA replication proceeds in a way that is dependent on the concentration of the available precursor nucleotides (Felsenstein, 1981). These data also seem to justify this assumption.

These data underline the importance of genome sequencing projects in furtherning our knowledge of biological processes. None of the genomes in this study have evolved identically. The diversity of genomic evolutionary patterns is in contrast to our expectations. In addition, although a total of more than 1000 ORFs from these genomes still remain enigmatic with respect to function, their codon usage is largely indistinguishable from the known proteins, and this is a strong indication that in the future most of these will be shown to be functional. These data also show that not all prokaryotic codon usage is governed largely by the need to optimize translational efficiency in highly expressed genes. It is probable that a great many prokaryotic genomes have evolved in completely different ways, and completion of the current sequencing projects will prove to be the best source of this information.

ACKNOWLEDGMENTS

I am indebted to Andrew Lloyd and the Irish National Centre for Bioinformatics for provision of software and to two anonymous reviewers whose comments helped to greatly improve the quality of this manuscript. I would also like to thank The National History Museum for providing facilities.

REFERENCES

ANDERSSON, S.G.E., and SHARP, P.M. (1996). Codon usage and base composition in *Rickettsia prowazekii*. J Mol Evol **42,** 525–536.

BENNETZEN, J.L., and HALL, B.D. (1982). Codon selection in yeast. J Biol Chem **257,** 3026–3031.

BERNARDI, G. (1993). The isochore organisation of the human genome and its evolutionary history—a review. Gene **135,** 57–66.

BULT, C.J., WHITE, O., OLSEN, G.J., ZHOU, L., FLEISCHMANN, R.D., SUTTON, G.C., et al. (1996). Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii.* Science **273,** 1058–1073.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17,** 368–376.

FLEISCHMANN, R.D., ADAMS, M.D., WHITE, O., CLAYTON, R.A., KIRKNESS, E.F., KERLAVAGE, A.R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269,** 496–512.

FRASER, C.M., GOCAYNE, J.D., WHITE, O., ADAMS, M.D., CLAYTON, R.A., FLEISCHMANN, R.D., et al. (1995). The minimal gene complement of *Mycoplasma genitalium.* Science **270,** 397–403.

GRANTHAM, R., GAUTIER, C., GOUY, M., JACOBZONE, M., and MERCIER, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res, **9,** r43–r74.

GREENACRE, M.J. (1984). *Theory and Applications of Correspondence Analysis.* (Academic Press, London).

IKEMURA, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein sequence: 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. J Mol Biol **158,** 573–597.

IWABE, N., KUMA, K.-I., HASEGAWA, M., OSAWA, S., and MIYATA, T. (1989). Evolutionary relationships of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA **86,** 9355–9359.

LLOYD, A.T., and SHARP, P.M. (1991). Codon usage in *Aspergillus nidulans.* Mol Gen Genet **230,** 288–294.

LLOYD, A.T., and SHARP, P.M. (1992). CODONS: a microcomputer program for codon usage analysis. J Hered **83,** 239–240.

LLOYD, A.T., and SHARP, P.M. (1993). Synonymous codon usage in *Kluyveromyces lactis.* Yeast **9,** 1219–1228.

MÉDIGUE, C., ROUXEL, T., VIGIER, P., HÉNAUT, A., and DANCHIN, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol **222,** 851–856.

MUTO, A., and OSAWA, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA **84,** 166–169.

SHARP, P.M., and COWE, E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae.* Yeast **7,** 657–678.

SHARP, P.M., and DEVINE, K.M. (1989). Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do "prefer" optimal codons. Nucleic Acids Res **17,** 5029–5039.

SHARP, P.M., STENICO, M., PEDEN, J.F., and LLOYD, A.T. (1993). Codon usage: mutational bias, translational selection or both? Biochem Soc Trans **21,** 835–841.

SHARP, P.M., TUOHY, T.M.F., and MOSURSKI, K.R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res **14,** 5125–5143.

SRIVASTAVA, A.K., and SCHLESSINGER, D. (1990). Mechanism and regulation of bacterial ribosomal RNA processing. Annu Rev Microbiol **44,** 105–129.

WOESE, C.R., KANDLER, O., and WHEELIS, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc Natl Acad Sci USA **87,** 4587.

WOLFE, K.H., SHARP, P.M., and LI, W.-H. (1989). Mutation rates differ among regions of the mammalian genome. Nature **337,** 283–285.

WRIGHT, F. (1990). The "effective number of codons" used in a gene. Gene **87,** 23–29.

Address reprint requests to:
James O. McInerney
Department of Zoology
The Natural History Museum
Cromwell Road
London SW7 5BD
UK

This article has been cited by:

1. Richard Herrmann, Berta Reiner. 1998. Mycoplasma pneumoniae and Mycoplasma genitalium: a comparison of two closely related bacterial species. *Current Opinion in Microbiology* **1**:5, 572-579. [CrossRef]