Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases.

Bénédicte Lafay, Andrew T. Lloyd, Michael J. Mclean, Kevin M. Devine, Paul M. Sharp, Kenneth H. Wolfe

1642–1649 *Nucleic Acids Research, 1999, Vol. 27, No. 7*

Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases

Bénédicte Lafay, Andrew T. Lloyd[1], Michael J. McLean[1], Kevin M. Devine[1], Paul M. Sharp and Kenneth H. Wolfe[1,*]

Division of Genetics, University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK and
[1]Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

Received December 23, 1998; Revised and Accepted February 15, 1999

ABSTRACT

The genomes of the spirochaetes *Borrelia burgdorferi* and *Treponema pallidum* show strong strand-specific skews in nucleotide composition, with the leading strand in replication being richer in G and T than the lagging strand in both species. This mutation bias results in codon usage and amino acid composition patterns that are significantly different between genes encoded on the two strands, in both species. There are also substantial differences between the species, with *T.pallidum* having a much higher G+C content than *B.burgdorferi*. These changes in amino acid and codon compositions represent neutral sequence change that has been caused by strong strand- and species-specific mutation pressures. Genes that have been relocated between the leading and lagging strands since *B.burgdorferi* and *T.pallidum* diverged from a common ancestor now show codon and amino acid compositions typical of their current locations. There is no evidence that translational selection operates on codon usage in highly expressed genes in these species, and the primary influence on codon usage is whether a gene is transcribed in the same direction as replication, or opposite to it. The *dnaA* gene in both species has codon usage patterns distinctive of a lagging strand gene, indicating that the origin of replication lies downstream of this gene, possibly within *dnaN*. Our findings strongly suggest that gene-finding algorithms that ignore variability within the genome may be flawed.

INTRODUCTION

The complete genome sequences of two pathogenic spirochaete bacteria, *Borrelia burgdorferi* and *Treponema pallidum*, have been reported by Fraser and colleagues (1,2). Although classified in the same family (Spirochaetaceae), these species have very different genome G+C contents, respectively 28.6 and 52.8%, and are not particularly closely related in terms of those bacteria whose genomes have been completely sequenced. Their small subunit ribosomal RNA sequences share only 79% identity, which is similar to that between the proteobacterium *Escherichia coli* and the Gram-positive *Bacillus subtilis* (77%) or the cyanobacterium *Synechocystis* sp. PCC6803 (77%). Only 46% of *T.pallidum* ORFs have identifiable homologues in *B.burgdorferi* (2), and the average level of protein sequence identity in the 229 sequence pairs analysed here is only 44%.

The spirochaete genomes are notable for their unusual base compositions. Fraser *et al*. (1) and Grigoriev (3) showed that the left and right halves of the linear *B.burgdorferi* chromosome have very different values of GC skew (the quantity G–C/G+C), and used this as an argument that the origin of replication is located in the centre of the genome. We showed that *B.burgdorferi* has strong AT skews as well as GC skews when measured at third positions of codons, and that these skews are also present in the circular *T.pallidum* genome (4). The skews in the spirochaetes are much more severe than in other prokaryotes, but as in most other bacteria they switch sign at the probable origin and terminus of replication and the leading strand in replication is comparatively G+T-rich (1–7). Indeed, a recent analysis of codon usage in *B.burgdorferi* genes (8) showed that the major cause of variation was the location of a gene (on the leading or lagging strand in replication) rather than its expression level. This is unusual among bacteria, both in terms of the apparent lack of selection for efficient translation, and in terms of the effect of a gene's chromosomal position on its codon usage. The only previous reports of chromosomal position affecting codon usage in bacteria have been in some species (particularly *Mycoplasma genitalium*) where G+C content varies in a cyclic fashion around the genome (4,9–11). In *B.burgdorferi* it is a gene's orientation relative to the direction of DNA replication, not its location on the chromosome, that determines its codon usage pattern (8).

Here we have compared codon usage in *T.pallidum* to that in *B.burgdorferi*, and investigated the effect on amino acid usage in both species, focusing in particular on changes that have occurred in orthologous genes that are replicated on different DNA strands in the two species. We show that, despite having different G+C content and chromosome structure and little conservation of gene order, codon usage in *T.pallidum* varies similarly to that in

*To whom correspondence should be addressed. Tel: +353 1 608 1253; Fax: +353 1 679 8558; Email: khwolfe@tcd.ie

B.burgdorferi. Furthermore, this variation is also reflected in amino acid usage. Consequently, orthologous genes from the two spirochaetes show species- and strand-specific trends in codon and amino acid usage.

MATERIALS AND METHODS

DNA and protein sequences were obtained from The Institute for Genomic Research (TIGR; ftp://ftp.tigr.org). Annotation tables were obtained from the NCBI Entrez Genomes Division WWW site (http://www.ncbi.nlm.nih.gov) in August 1998, and listed 850 *B.burgdorferi* chromosomal genes and 1031 *T.pallidum* genes. The origin of replication was initially assumed to be upstream of *dnaA*, and then changed to the *dnaA–dnaN* intergenic spacer in both species (see Results). The terminus in *T.pallidum* was assumed to lie between genes TP0515 and TP0516.

Codon bias was measured using relative synonymous codon usage (RSCU) values (12). The RSCU value for a codon is a measure of its usage, relative to other codons for that amino acid. RSCU values are scaled such that, if codon usage was uniform within each amino acid group, the RSCU values would all be 1. Thus, RSCU values higher than 1 indicate codons used frequently within their amino acid group (12).

To look for trends in the data, correspondence analysis (13) was carried out on RSCU values. Correspondence analysis is a multivariate statistical analysis method particularly appropriate for contingency data, in which the values in the dataset are not independent. The raw data for this analysis were a matrix containing 59 RSCU values (i.e., for all codons except Met, Trp and Stop) for each gene in the genome. Analysis of RSCU values rather than actual codon counts minimises the effects of different amino acid usage among genes. Analyses were carried out separately on each species, and on a dataset comprising all genes from both genomes (1881 genes).

Orthologous *B.burgdorferi* and *T.pallidum* genes were identified using BLASTP (14) searches. Only 1:1 orthology relationships were considered: a pair of genes were regarded as orthologues if they were each other's only hit above a significance threshold, which was set at a BLASTP score of 200 [using the BLOSUM62 substitution matrix (15) and SEG filter (16)]. Protein pairs identified in this way were then aligned using the Gap program in the GCG package (with default parameters) to examine amino acid substitutions.

RESULTS

Correspondence analysis of codon usage

Codon usage patterns in *B.burgdorferi* and *T.pallidum* were analysed by correspondence analysis of RSCU values in a dataset of 1881 genes pooled from the two species. Correspondence analysis reveals trends in the data that may be impossible to tease out on a gene-by-gene comparison. It defines a series of orthogonal (uncorrelated) axes through the data, ordered so that Axis 1 is the axis describing the largest fraction of the variation in the data, Axis 2 describes the second-largest trend, and so on with each subsequent axis describing a progressively smaller amount of variation. Genes that have similar codon usage will appear close together in the multi-dimensional hyperspace that correspondence analysis describes. For the set of 1881 genes from the two species Axes 1 and 2 account for 34.8 and 8.6%, respectively, of the variability among the data (Fig. 1), and the subsequent axes account for <3.7% each.

The major trend in codon usage in the pooled dataset, identified as Axis 1 in the correspondence analysis, is the species of origin of each gene. Codon usage is manifestly different between the two species with *B.burgdorferi* genes being distributed on the right (positive values on Axis 1) and *T.pallidum* genes on the left in Figure 1. Codon usage tabulated across all genes in *B.burgdorferi* is biased towards U- and A-ending codons as is expected for a species with an A+T-rich genome (Table 1). In *T.pallidum* codon usage is more random, and the trend is to use more U than C, and more G than A, at codon third positions. Chi-squared tests show that, for almost every amino acid, *B.burgdorferi* uses significantly more A- and/or U-ending codons, and fewer G- and/or C-ending codons, than does *T.pallidum* (Table 1). The only exceptions to this are the glycine codon GGU, which is used approximately equally in the two species, and three arginine codons. For arginine, the composition difference between the two species governs base choice at codon position 1 among the six synonyms, with all four CGN codons being used considerably more in *T.pallidum*, and the two AGR codons preferred in *B.burgdorferi* (Table 1).

Axis 2 represents the second-most significant source of variation in the data, and differentiates genes according to the strand on which they are located (Fig. 1). As reported recently by McInerney (8), genes from the leading and lagging strands in *B.burgdorferi* (referred to as Bb-lead and Bb-lag genes) form two distinct clusters (Fig. 1). In *T.pallidum* there is less discrimination between the leading and lagging strand clusters (referred to as Tp-lead and Tp-lag genes) but the trend is clearly the same. Significantly, leading strand genes from both species are located at the same (negative) end of Axis 2, indicating that the differences between leading and lagging strand genes involve similar types of change in codon usage in the two species.

Strand-specific biases in codon usage

In most bacterial genomes there is a tendency for the leading-strand to be richer in G and T than the lagging strand (6,7). This becomes exaggerated if only silent codon positions are considered (4,17). For the spirochaetes, plotting the G+T content of silent codon positions of genes versus their chromosomal location (results not shown) clearly distinguishes the genes on the two strands in each species but does not suggest any other variation in G+T content with chromosomal location, in contrast to the systematic variation of G+C content observed in the *M.genitalium* genome (9,10).

In both spirochaetes, the synonymous codon usage is different between the two strands (Table 1). There are significant ($P < 0.001$) differences in leading versus lagging strand genes, for 49 of the 59 synonymously variable sense codons in *B.burgdorferi*, and for 54 codons in *T.pallidum*. Almost all the changes involve increased use of G- and U-ending codons, and decreased use of C- and A-ending codons, on the leading strand relative to the lagging (Table 1). The only significant exceptions to this trend are the leucine codons CUG (decreased on the Bb-lead) and UUA (increased on Tp-lead). Because the *B.burgdorferi* genome is so A+T-rich, the leading versus lagging strand effect is manifested most clearly in the relative frequencies of A- versus U-ending codons (more U-ending codons in Bb-lead genes, and A-ending codons in Bb-lag genes; Table 1) and some differences between

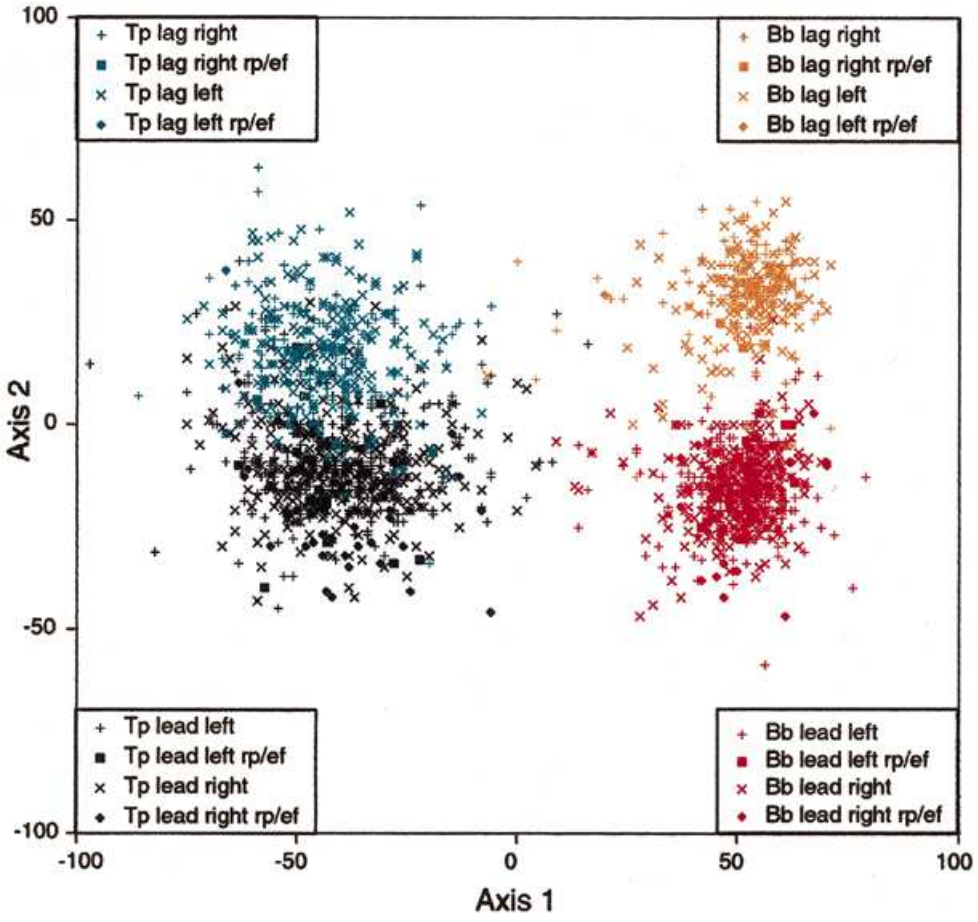

Figure 1. Correspondence analysis of RSCU values from 1881 genes in a pooled dataset (850 from *B.burgdorferi* and 1031 from *T.pallidum*). Genes are colour-coded as follows: red, *B.burgdorferi* leading strand; orange, *B.burgdorferi* lagging strand; dark blue, *T.pallidum* leading strand; light blue; *T.pallidum* lagging strand. Filled symbols indicate ribosomal proteins and translation elongation factors. 'Left' and 'right' refer to genes located in the left and right halves of the two genomes.

the strands for G- or C-ending codons do not achieve statistical significance.

If correspondence analysis is carried out on each species separately, the inter-strand differences reappear as Axis 1 but the other axes reveal further differences between the two species. In correspondence analysis, the same sets of axes can be used to examine variation among codons or among genes (13,18). When the locations of codons are superimposed upon those of genes (Fig. 2), G- and U-ending codons are found on one side of Axis 1, with C- and A-ending codons on the other, due to the strand effect. For *B.burgdorferi*, while the major trend is clearly caused by G+T content, the second axis is less simple to explain. It is dominated by usage of a single arginine codon, CGC (Fig. 2). It has been proposed that the inclusion of a small number of artefactual ORFs in the dataset was responsible (8); however, the same trend (CGC usage) also appears on Axis 2 in an analysis run on a reduced dataset of known genes. Furthermore, if CGC is excluded, Axis 2 appears to be related to another Arg codon (CGA), and so on with CGG. Thus the second source of variation in *B.burgdorferi* seems to be the CGN family of Arg codons, which are only marginally used in that species. The unusual position on Axis 1 of the CUG codon in *B.burgdorferi* correspondence analysis (Fig. 2) is also notable; although little used among genes on either strand, this codon is less frequent in Bb-lead genes than Bb-lag genes, which is unexpected for a G-ending codon (Table 1) and may be related to the existence of six synonyms for this amino acid.

For *T.pallidum*, the codons separate neatly into four classes based on the synonymous site (Fig. 2), indicating that the second major trend is related to silent-site G+C content of genes; G- and C-ending codons are found on one side of Axis 2, U- and A-ending codons on the other. The purine-ending codons are more separated on Axis 2 than the pyrimidines. However, the genes at the extremities of Axis 2 do not correspond to any particular functional group and do not have any preferential location along the chromosome.

Intraspecific codon usage patterns and the location of the origin of replication

In Figure 1 there is some overlap between the leading and lagging strand clusters on Axis 2, differing in magnitude in the two species. For example, if a cutoff line is drawn horizontally through Figure 1 corresponding to an Axis 2 value of +5, the Axis 2 values of 70 Tp-lead and 47 Tp-lag genes place them on the 'wrong' side of this cutoff (i.e., on the opposite side to most of the other genes from the same strand), whereas only nine Bb-lead and seven Bb-lag genes are misplaced. Most of these 16 *B.burgdorferi*

Table 1. Codon usage (RSCU values) in *B.burgdorferi* (Bb) and *T.pallidum* (Tp)

		RSCU								
		Both strands			Bb			Tp		
Amino Acid	Codon	Bb	χ^{2a}	Tp	lead	χ^{2b}	lag	lead	χ^{2b}	lag
Phe	UUU	1.81	>>	1.39	1.88	>>	1.64	1.49	>>	1.20
	UUC	0.19	<<	0.61	0.12	<<	0.36	0.51	<<	0.80
Leu	UUA	2.42	>>	0.51	2.39	<	2.49	0.54	>>	0.47
	UUG	1.07	<<	1.18	1.30	>>	0.59	1.44	>>	0.73
Leu	CUU	1.76	>>	1.39	1.91	>>	1.46	1.40	ns	1.38
	CUC	0.12	<<	1.25	0.07	<<	0.24	0.99	<<	1.69
	CUA	0.49	>>	0.34	0.23	<<	1.03	0.29	<<	0.43
	CUG	0.13	<<	1.33	0.10	<<	0.20	1.35	ns	1.30
Ile	AUU	1.67	>>	1.38	2.00	>>	1.20	1.51	>>	1.17
	AUC	0.22	<<	1.06	0.14	<<	0.33	0.94	<<	1.27
	AUA	1.11	>>	0.56	0.86	<<	1.48	0.55	ns	0.56
Met	AUG	-	-	-	-	-	-	-	-	-
Val	GUU	2.35	>>	0.90	2.59	>>	1.45	0.93	>>	0.84
	GUC	0.18	<<	0.59	0.13	<<	0.35	0.49	<<	0.84
	GUA	1.08	>>	0.76	0.87	<<	1.83	0.70	<<	0.88
	GUG	0.40	<<	1.75	0.41	ns	0.37	1.87	>>	1.44
Ser	UCU	2.12	>>	1.45	2.38	>>	1.50	1.58	>>	1.21
	UCC	0.27	<<	0.96	0.23	<<	0.35	0.80	<<	1.25
	UCA	1.40	>>	0.73	1.15	<<	1.98	0.69	<<	0.80
	UCG	0.21	<<	0.97	0.21	ns	0.19	1.08	>>	0.77
Pro	CCU	1.79	>>	1.25	2.03	>>	1.39	1.33	>>	1.12
	CCC	0.61	<<	0.91	0.59	ns	0.64	0.76	<<	1.14
	CCA	1.41	>>	0.62	1.18	<<	1.81	0.56	<<	0.72
	CCG	0.19	<<	1.22	0.21	ns	0.16	1.35	>>	1.03
Thr	ACU	1.52	>>	0.73	1.88	>>	1.06	0.78	>>	0.67
	ACC	0.52	<<	1.10	0.52	ns	0.53	0.93	<<	1.34
	ACA	1.76	>>	0.79	1.35	<<	2.28	0.73	<<	0.88
	ACG	0.20	<<	1.37	0.25	>>	0.14	1.56	>>	1.11
Ala	GCU	1.80	>>	0.64	2.08	>>	1.25	0.66	>	0.61
	GCC	0.43	<<	0.62	0.39	<<	0.51	0.52	<<	0.79
	GCA	1.57	>>	1.21	1.29	<<	2.12	1.14	<<	1.34
	GCG	0.20	<<	1.53	0.24	>>	0.12	1.68	>>	1.26
Tyr	UAU	1.60	>>	0.97	1.77	>>	1.27	1.07	>>	0.77
	UAC	0.40	<<	1.03	0.23	<<	0.73	0.93	<<	1.23
ter	UAA	1.89	-	0.74	1.71	-	2.24	0.65	-	0.90
ter	UAG	0.56	-	1.16	0.69	-	0.30	1.21	-	1.06
His	CAU	1.49	>>	0.85	1.67	>>	1.22	0.99	>>	0.66
	CAC	0.51	<<	1.15	0.33	<<	0.78	1.01	<<	1.34
Gln	CAA	1.64	>>	0.62	1.51	<<	1.83	0.56	<<	0.73
	CAG	0.36	<<	1.38	0.49	>>	0.17	1.44	>>	1.27
Asn	AAU	1.63	>>	1.05	1.80	>>	1.38	1.17	>>	0.85
	AAC	0.37	<<	0.95	0.20	<<	0.62	0.83	<<	1.15
Lys	AAA	1.58	>>	0.91	1.42	<<	1.82	0.82	<<	1.08
	AAG	0.42	<<	1.09	0.58	>>	0.18	1.18	>>	0.92
Asp	GAU	1.64	>>	1.16	1.75	>>	1.33	1.29	>>	0.91
	GAC	0.36	<<	0.84	0.25	<<	0.67	0.71	<<	1.09
Glu	GAA	1.48	>>	0.95	1.31	<<	1.76	0.88	<<	1.07
	GAG	0.52	<<	1.05	0.69	>>	0.24	1.12	>>	0.93
Cys	UGU	1.34	>>	1.07	1.54	>>	0.87	1.21	>>	0.80
	UGC	0.66	<<	0.93	0.46	<<	1.13	0.79	<<	1.20
ter	UGA	0.55	-	1.10	0.60	-	0.46	1.14	-	1.04
Trp	UGG	-	-	-	-	-	-	-	-	-
Arg	CGU	0.33	<<	1.72	0.40	>>	0.13	1.92	>>	1.31
	CGC	0.16	<<	1.92	0.17	ns	0.16	1.64	<<	2.51
	CGA	0.31	<<	0.46	0.30	ns	0.32	0.42	<<	0.54
	CGG	0.08	<<	1.01	0.09	ns	0.05	1.09	>>	0.83
Ser	AGU	1.17	>>	0.97	1.34	>>	0.77	1.06	>>	0.80
	AGC	0.83	<<	0.92	0.67	<<	1.21	0.79	<<	1.17
Arg	AGA	3.92	>>	0.41	3.67	<<	4.69	0.40	ns	0.42
	AGG	1.20	>>	0.49	1.38	>>	0.64	0.53	>>	0.39
Gly	GGU	1.07	ns	1.10	1.30	>>	0.51	1.18	>>	0.91
	GGC	0.62	<<	0.85	0.62	ns	0.63	0.71	<<	1.14
	GGA	1.71	>>	0.93	1.40	<<	2.44	0.88	<<	1.03
	GGG	0.60	<<	1.13	0.68	>>	0.42	1.23	>>	0.91

[a]Chi-squared tests: << and >> denote $P < 0.001$ and indicate the direction of the difference; ns, not significant. White-on-black symbols indicate differences that are not in the expected direction, given the high A+T content and low G+C content of the *B.burgdorferi* genome relative to the *T.pallidum* genome.

[b]Chi-squared tests: << and >> denote $P < 0.001$ and indicate the direction of the difference; < and > denote $P < 0.01$; ns, not significant. White-on-black symbols indicate differences that are not in the expected direction, given the high G+T content and low A+C content of the leading strand ('lead') relative to the lagging strand ('lag'), in both species.

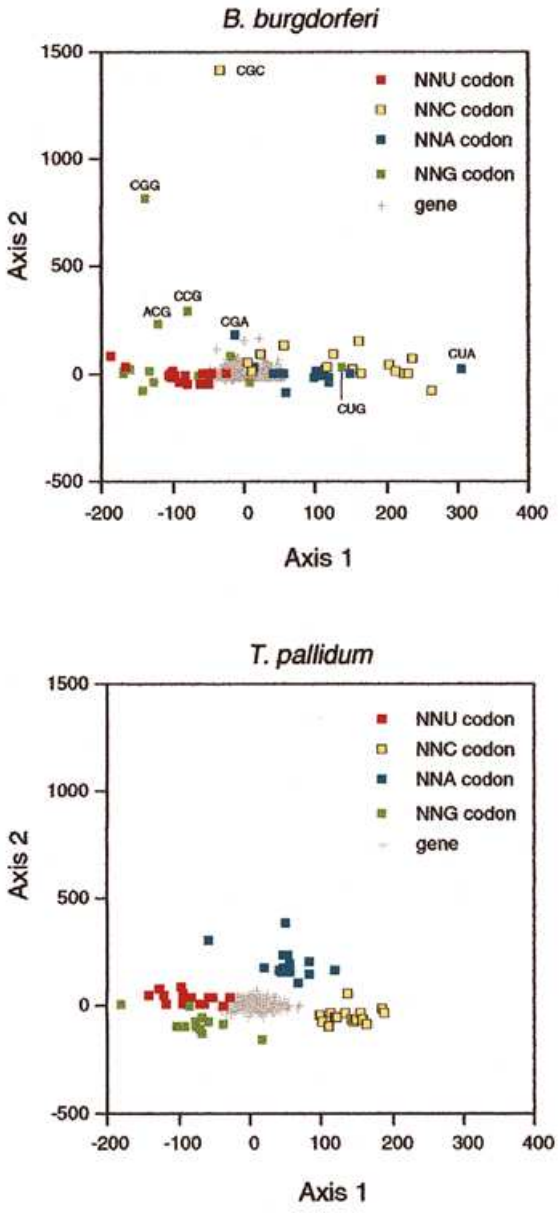

Figure 2. Correspondence analysis of codons (coloured points) superimposed on genes (grey points).

'outliers' are short genes. Only four of them (BB0001, BB0844, BB0437 and BB0438) are more than 150 codons long, and for each of these a possible explanation is apparent. BB0001 and BB0844 are located very close to the telomeres of the linear chromosome and could have become inverted during the proposed telomeric exchanges between the chromosome and linear plasmids (2).

The other two *B.burgdorferi* genes with unusual Axis 2 values, BB0437 and BB0438, are *dnaA* and *dnaN*. In both *B.burgdorferi* and *T.pallidum* these genes are located very close to the presumed origin of replication (1,2). The location of the origin has not been determined experimentally in either species, but has been inferred both from analysis of base composition skew (19), and from the realisation that *dnaA* is close to the origin of replication in many bacteria (20–22). Fraser *et al.* (2) used a numbering scheme for the *T.pallidum* genome which placed nucleotide number 1 upstream of

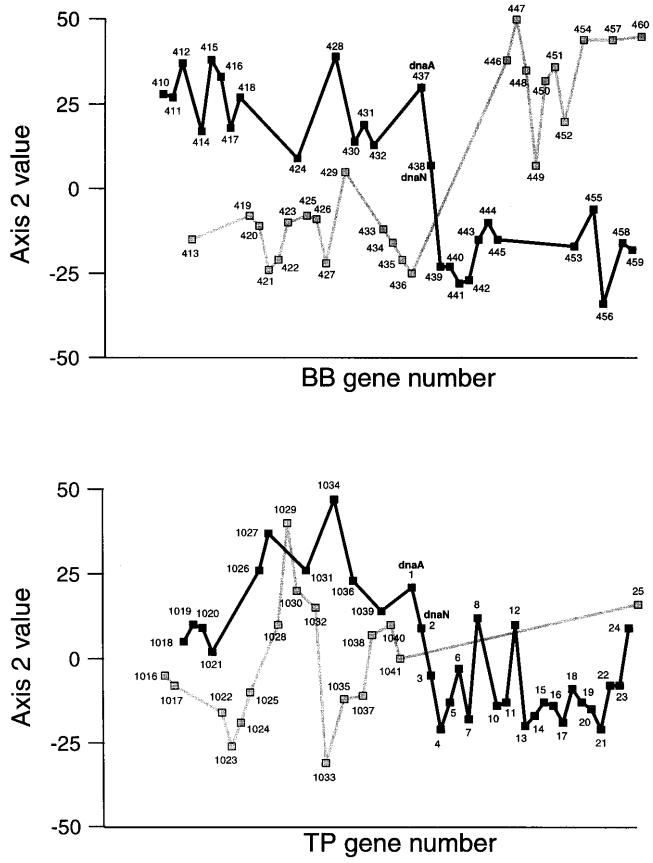

Figure 3. Plot of Axis 2 values (from the correspondence analysis shown in Fig. 1) of genes flanking the origins of replication in *B.burgdorferi* and *T.pallidum*. Black lines connect genes transcribed rightwards, and grey lines connect genes transcribed leftwards. To the left of the origin, leading strand genes are grey and lagging strand genes are black. The situation is reversed on the right of the origin (leading strand genes are black and lagging strand genes are grey).

dnaA, suggesting that it is a leading-strand gene, and in our initial analysis we assumed that *dnaA* and *dnaN* were leading-strand genes in both *B.burgdorferi* and *T.pallidum*. However, the Axis 2 position of *dnaA* in *B.burgdorferi* suggests very strongly that it is a lagging strand gene (Fig. 3), so that the origin of replication must be downstream of *dnaA*.

In both *B.burgdorferi* and *T.pallidum*, *dnaN* is immediately downstream of *dnaA*, and in the same orientation. The Axis 2 value of *dnaN* in *B.burgdorferi* is +7, which is intermediate between typical values of leading and lagging strand genes (Fig. 3). If *dnaN* is regarded as a leading-strand gene, its Axis 2 value ranks it eighth highest of all Bb-lead genes; alternatively, if it is regarded as a Bb-lag gene, it has the tenth lowest Axis 2 value among Bb-lag genes. One likely explanation is that the origin of replication is located within *dnaN*. This is supported by the observation that if *B.burgdorferi dnaN* is divided into two halves, the Axis 2 value of the 5′ end is +34 (typical of a Bb-lag gene) and of the 3′ end is –8 (typical of a Bb-lead gene). The Axis 2 values of *T.pallidum* genes located near the origin are also consistent with an origin within *dnaN*, though the distinction between the two strands is less clear-cut (Fig. 3). Alternative possible explanations are that there are multiple closely-spaced origins in this region (20), or that the origin has moved recently. For the other analyses in this paper, we assumed *dnaN* to be a leading-strand gene in both species. These findings are borne out when correspondence analyses are carried out on each species separately.

Codon usage in highly expressed genes

Studies on many bacterial species have shown that highly expressed genes use a subset of 'optimal' codons due to selection for efficient translation of their mRNAs, whereas genes with lower expression levels have more random codon usage. This occurs, for example, in *E.coli* (23,24), *B.subtilis* (25), *Mycobacterium tuberculosis* (26) and *Haemophilus influenzae* (9). Remarkably, in both *B.burgdorferi* and *T.pallidum*, genes expected to be highly expressed (such as ribosomal proteins and elongation factors) do not have a codon usage pattern distinct from the majority of genes (Fig. 1). Nor did the putative high-expression genes appear exceptional when additional axes from single-species correspondence analyses were examined, or in terms of their fit to the tRNA anticodon sets encoded by these genomes (data not shown). This suggests that selection for efficient translation is either absent or ineffective in these spirochaetes, in agreement with McInerney's analysis of *B.burgdorferi* (8).

Fraser *et al*. (2) calculated codon adaptation index (CAI) (12) values for *T.pallidum* and *B.burgdorferi* genes, using genes that are universally highly expressed as a reference set, and noted that genes with high CAI values were disproportionately frequent on the leading strand. The CAI measures the extent to which a gene uses a particular subset of codons, normally those that are translationally optimal. In species where translational selection is effective, highly expressed genes are characterised by strong codon bias towards those optimal codons and have high CAI values (12). In the spirochaetes, however, the combination of having most of the highly-expressed genes (the reference set) on the leading strand together with the apparent lack of effective translational selection, means that the CAI as used by Fraser *et al*. merely detects leading strand genes. It does not have further implications for the levels of expression of genes.

Codon usage changes in orthologous genes

Using a stringent criterion for BLAST searches (see Materials and Methods) we identified 229 pairs of orthologous genes in the two species. Dot-matrix plots of the chromosomal locations of these genes showed that gene order is poorly conserved between these species, with the exception of a number of large operons (results not shown). The 229 orthologue pairs comprise 133 Bb-lead/Tp-lead genes, 28 Bb-lag/Tp-lag genes, 35 Bb-lead/Tp-lag genes and 33 Bb-lag/Tp-lead genes. The latter two categories are genes that have been relocated from the leading strand to the lagging, or vice versa, in one of the species since their divergence from a common ancestor. In the correspondence analysis of RSCU values the positions of these four classes of orthologous gene pairs are distinct (Fig. 4). It is apparent that genes that have switched strand during spirochaete evolution now have codon usage patterns typical of their current strands.

Amino acid composition

The unusual nucleotide composition of the spirochaete genomes also affects the amino acid composition of their proteins. There are significant differences between the two species (Table 2).

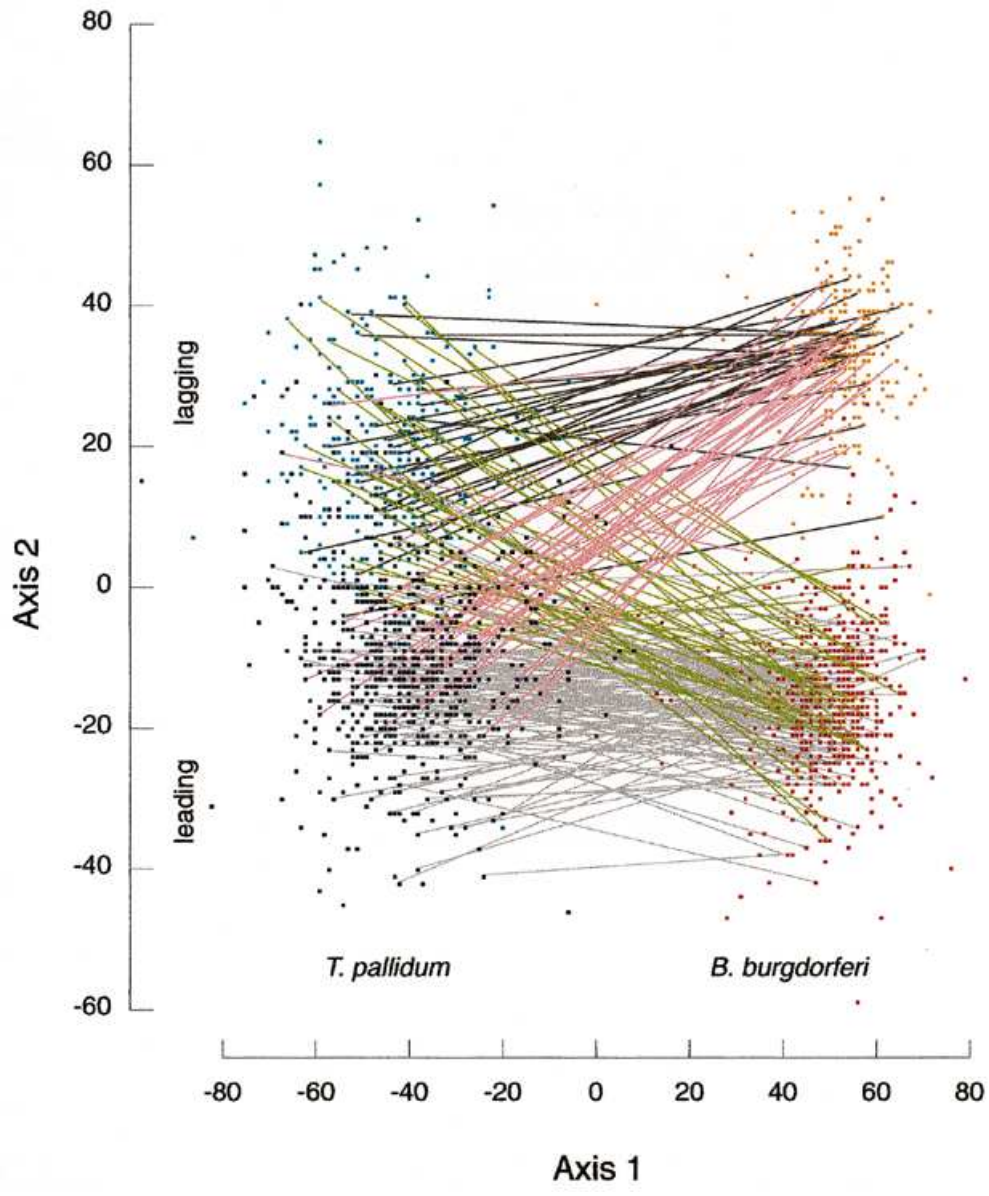

Figure 4. Correspondence analysis of RSCU values from 1881 genes (as in Fig. 1), with orthologous gene pairs linked by lines. Lines connecting pairs of genes that are on different strands in the two species are green (Bb-lead/Tp-lag) or pink (Bb-lag/Tp-lead). Lines connecting pairs whose strand is conserved are light grey (Bb-lead/Tp-lead) or dark grey (Bb-lag/Tp-lag). Points without lines did not have homologues above the BLASTP cutoff used.

Borrelia burgdorferi is comparatively rich in the six amino acids having A or U in codon positions 1 and 2 (column 'W' in Table 2), and poor in the three amino acids with G+C-containing codons (column 'S'). Chi-squared tests show significant differences between the two species for eight of these nine amino acids (all except Met), when genes located on both DNA strands are considered together. There are also significant interspecies differences for six other amino acids, making a total of 14 (Table 2). There appear to be some trade-offs between the species in amino acid choice, with *B.burgdorferi* having abundant Asn (AAY codons), Lys (AAR) and Ile (AUH), whereas *T.pallidum* uses other amino acids that are chemically similar but have more G+C-rich codons: Gln (CAR), Arg (CGN and AGR) and Val (GUN). *Borrelia burgdorferi* proteins have more than twice as much lysine, but less than half as much arginine, as *T.pallidum* proteins.

The amino acid composition of proteins encoded by leading-strand genes is also different from that of lagging-strand genes. These differences are statistically significant ($P < 0.001$) for 16 amino acids in *B.burgdorferi* and 11 amino acids in *T.pallidum* (Table 2). In both species these amino acid usage trends are almost universally in the directions consistent with a G+T-rich leading strand and an A+C-rich lagging strand. From base composition at codon positions 1 and 2, five amino acids (Phe, Trp, Cys, Val and Gly) would be expected to be more common in leading-strand genes, and six amino acids (Asn, Lys, Pro, Gln, His and Thr) would be expected to be less common (Table 2, columns 'K' and 'M'). Statistically significant differences are seen for 10

Table 2. Amino acid composition in *B.burgdorferi* (Bb) and *T.pallidum* (Tp)

Amino acid (codons)	Codon positions 1 and 2[a] S	W	K	M	Percent amino acid composition: Both strands Bb	χ²[b]	Tp	Bb lead	χ²[c]	Bb lag	Tp lead	χ²[c]	Tp lag
Asn (AAY)		+		+	7.3	>>	2.5	6.7	<<	8.3	2.3	<<	2.8
Lys (AAR)		+		+	10.3	>>	4.0	9.4	<<	12.0	4.0	ns	4.0
Ile (AUH)		+			10.8	>>	4.9	9.6	<<	13.0	4.7	<<	5.2
Met (AUG)		+			1.8	ns	2.0	1.8	ns	1.7	2.0	>>	1.9
Tyr (UAY)		+			4.2	>>	3.0	4.3	ns	4.2	3.0	ns	3.1
Phe (UUY)		+	+		6.3	>>	4.5	6.9	>>	5.2	4.5	ns	4.4
Trp (UGG)			+		0.5	ns	1.0	0.5	<	0.6	1.0	ns	1.0
Cys (UGY)			+		0.7	<<	1.9	0.7	>>	0.6	2.0	ns	1.8
Val (GUN)			+		5.4	<<	8.4	6.4	>>	3.4	9.1	>>	7.1
Gly (GGN)	+		+		5.2	<<	7.0	5.6	>>	4.5	7.3	>>	6.4
Ala (GCN)	+				4.5	<<	10.1	4.5	ns	4.4	10.0	ns	10.3
Pro (CCN)	+			+	2.5	<<	4.2	2.4	<<	2.8	4.0	<<	4.6
Gln (CAR)				+	2.3	<<	3.8	2.1	<<	2.6	3.9	ns	3.7
His (CAY)				+	1.2	<<	2.8	1.1	<<	1.4	2.5	<<	3.2
Thr (ACN)				+	3.9	<<	5.3	3.3	<<	5.1	4.8	<<	6.1
Asp (GAY)					5.2	ns	4.5	5.8	>>	4.0	4.6	>>	4.3
Glu (GAR)					6.8	ns	6.0	6.5	<<	7.3	6.0	ns	5.8
Arg (CGN+AGR)					3.2	<<	7.4	3.7	>>	2.3	7.7	>>	7.0
Leu (UUR+CUN)					10.4	ns	10.2	10.6	>>	10.0	10.0	<<	10.6
Ser (UCN+AGY)					7.5	ns	6.6	8.0	>>	6.5	6.6	ns	6.7

[a]Plus signs indicate codons in which the bases at positions 1 and 2 can both be described by the IUPAC ambiguity codes S (G or C), W (A or U), K (G or U) or M (A or C). Amino acids with six codons were not classified.
[b]Chi-squared tests: << and >> denote $P < 0.001$ and indicate the direction of the difference; ns, not significant. White-on-black symbols indicate differences that are in the expected direction, given the high W content and low S content of the *B.burgdorferi* genome relative to the *T.pallidum* genome.
[c]Chi-squared tests: << and >> denote $P < 0.001$ and indicate the direction of the difference; < denotes $P < 0.01$; ns, not significant. White-on-black symbols indicate differences that are in the expected direction, given the high K content and low M content of the leading strand ('lead') relative to the lagging strand ('lag'), in both species.

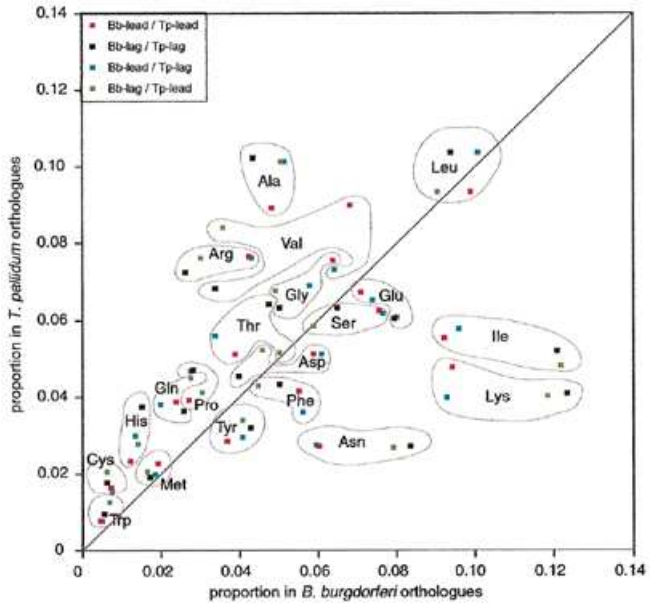

Figure 5. Mean amino acid compositions of orthologous *B.burgdorferi* and *T.pallidum* proteins, in each of the four possible species/DNA strand combinations. The four points for each amino acid are ringed. The diagonal line represents equal proportions in the two species.

of these 11 comparisons in *B.burgdorferi*, and for six comparisons in *T.pallidum* (Table 2). The only amino acid that significantly bucks this trend is tryptophan, which is unexpectedly common in Bb-lag genes ($P < 0.01$).

Comparison of aligned protein sequences from the 229 orthologous pairs confirms that amino acid substitutions of the expected types have occurred. As an extreme example, the lagging strand is Arg-rich but Lys-poor in *T.pallidum*, whereas it is Arg-poor and Lys-rich in *B.burgdorferi*. In the sequences of the 28 orthologous genes located on the two lagging strands, amino acid sites that are Arg in *T.pallidum* are more often substituted to Lys in *B.burgdorferi* (28.8%) than conserved as Arg (22.4%).

Amino acid composition in the 229 orthologous genes also indicates that genes that have switched DNA strand during spirochaete evolution now have amino acid compositions typical of their current locations (Fig. 5), in a manner similar to the changes in codon usage shown in Figure 4. This is clearest for those amino acids where there are large differences both between species and between strands, such as Val, Ile, Leu and Thr. For these amino acids, the points for the four classes of gene (Fig. 5) form an approximate square, with the Bb-lead/Tp-lead genes (red symbols) and Bb-lag/Tp-lag genes (black symbols) in two opposite corners, and the switched-strand genes (blue and green symbols) in the other two corners. Thus, for example, leucine content is approximately the same in all Bb-lead proteins in this dataset, regardless of whether they have Tp-lead or Tp-lag orthologues (9.8 and 10.0%, respectively). Approximately half of the *B.burgdorferi* genes in the Bb-lead/Tp-lag set have, presumably, been inverted at some stage during *B.burgdorferi* evolution but their amino acid composition is now typical of Bb-lead genes.

DISCUSSION

Dramatic differences are seen in amino acid and codon usage in these spirochaetes, both between species and between DNA strands. Interstrand differences occur in both the linear genome of *B.burgdorferi* (1,8) and the circular genome of *T.pallidum*. Because of the apparent lack of translational selection on codon choice, and the existence of two different patterns within each genome, it seems unlikely that the different compositions of genes and proteins are the result of natural selection. It is more probable that they represent neutral sequence change that has been driven by the strong mutation biases in these species. Species-specific mutation has previously been shown to affect amino acid usage in very G+C-rich or G+C-poor bacterial species (27–29) but this is the first report of mutational heterogeneity within a genome causing compositional heterogeneity within the corresponding proteome. The mutational pressures have undoubtedly contributed to the extensive sequence divergence between *B.burgdorferi* and *T.pallidum* proteins.

Having genome sequences from two spirochaetes makes it possible to identify genes that have become inverted during evolution but does not allow us to deduce when the inversions occurred, or in which lineage. Our approach grouped the inverted genes into two classes and found that these genes now seem assimilated into their current chromosomal environments. It is not possible to measure the speed of this assimilation without sequence data from other related species, which would allow the

approximate dates of the inversions to be inferred. The existence of two distinct classes of genes in bacterial genomes such as these may create problems for computer programs that attempt to distinguish between real genes and artefactual ORFs in genome sequences. Current programs that use a single model of nucleotide composition (30,31) are probably inappropriate. One approach to a new genome sequence could be to first examine codon and amino acid usage patterns in the known genes (those with homologues), and then to search for the remaining genes using multiple different models of the properties of the known genes as necessary.

ACKNOWLEDGEMENTS

This study was supported by the Fourth Framework programme of the European Commission (BIO4-CT95-0130) and the BBSRC (G04905).

REFERENCES

1 Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R., Lathigra,R., White,O., Ketchum,K.A., Dodson,R., Hickey,E.K. *et al.* (1997) *Nature*, **390**, 580–586.
2 Fraser,C.M., Norris,S.J., Weinstock,G.M., White,O., Sutton,G.G., Dodson,R., Gwinn,M., Hickey,E.K., Clayton,R., Ketchum,K.A. *et al.* (1998) *Science*, **281**, 375–388.
3 Grigoriev,A. (1998) *Nucleic Acids Res.*, **26**, 2286–2290.
4 McLean,M.J., Wolfe,K.H. and Devine,K.M. (1998) *J. Mol. Evol.*, **47**, 691–696.
5 Perrière,G., Lobry,J.R. and Thioulouse,J. (1996) *Comput. Appl. Biosci.*, **12**, 519–524.
6 Francino,M.P. and Ochman,H. (1997) *Trends Genet.*, **13**, 240–245.
7 Lobry,J.R. (1996) *Mol. Biol. Evol.*, **13**, 660–665.
8 McInerney,J.O. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 10698–10703.
9 McInerney,J.O. (1997) *Micro. Comp. Genom.*, **2**, 1–10.
10 Kerr,A.R.W., Peden,J.F. and Sharp,P.M. (1997) *Mol. Microbiol.*, **25**, 1177–1179.
11 Deschavanne,P. and Filipski,J. (1995) *Nucleic Acids Res.*, **23**, 1350–1353.
12 Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
13 Greenacre,M.J. (1984) *Theory and Applications of Correspondence Analysis.* Academic Press, London.
14 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
15 Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
16 Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
17 Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) *Science*, **277**, 1453–1474.
18 Holm,L. (1986) *Nucleic Acids Res.*, **14**, 3075–3087.
19 Lobry,J.R. (1996) *Science*, **272**, 745–746.
20 Old,I.G., Margarita,D. and Saint Girons,I. (1993) *FEMS Microbiol. Lett.*, **111**, 109–114.
21 Calcutt,M.J. and Schmidt,F.J. (1992) *J. Bacteriol.*, **174**, 3220–3226.
22 Ogasawara,N. and Yoshikawa,H. (1992) *Mol. Microbiol.*, **6**, 629–634.
23 Gouy,M. and Gautier,C. (1982) *Nucleic Acids Res.*, **10**, 7055–7074.
24 Medigue,C., Rouxel,T., Vigier,P., Henaut,A. and Danchin,A. (1991) *J. Mol. Biol.*, **222**, 851–856.
25 Shields,D.C. and Sharp,P.M. (1987) *Nucleic Acids Res.*, **15**, 8023–8040.
26 Andersson,S.G.E. and Sharp,P.M. (1996) *Microbiology*, **142**, 915–925.
27 Osawa,S. (1995) *Evolution of the Genetic Code.* Oxford University Press, Oxford.
28 Li,W.-H. (1997) *Molecular Evolution.* Sinauer, Sunderland, MA.
29 Gu,X., Hewett-Emmett,D. and Li,W.-H. (1998) *Genetica*, **102–103**, 383–391.
30 Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) *Nucleic Acids Res.*, **26**, 544–548.
31 Lukashin,A.V. and Borodovsky,M. (1998) *Nucleic Acids Res.*, **26**, 1107–1115.