Codon usage and bias among individual genes of the coccidia and piroplasms

J. T. ELLIS[1], D. A. MORRISON[2], D. AVERY *and* A. M. JOHNSON[1]

[1] *Department of Microbiology and* [2] *Department of Applied Biology, University of Technology Sydney, St Leonards Campus, P.O. Box 123 Broadway, New South Wales 2007, Australia*

(*Received 8 December 1993; revised 31 January 1994; accepted 17 February 1994*)

SUMMARY

Codon usage has been analysed in individual gene sequences, derived from a variety of parasitic protozoa in the class Sporozoa of the phylum Apicomplexa using metric multidimensional scaling. The two groups of codon usage patterns detected reflect the two main subgroups of organisms studied (the coccidia and the piroplasms), and it is the pattern of usage of synonymous codons that has the largest influence on overall codon usage in the individual genes, rather than being the pattern of amino acid composition of the gene product. The magnitude of the codon usage bias in the sequences was determined using three commonly used indices – N_C, GC_{3S} and B. In general, although relatively low levels of codon usage bias were detected in these gene sequences, codon usage bias does explain at least some of the codon usage patterns observed. Codon usage bias was observed to be dependent on the overall base composition of the genes analysed, which in turn was reflected in the types of codons that were either over- or under-represented in the nucleotide sequences. In keeping with observations on prokaryotic organisms, it is speculated that the codon usage patterns detected in these parasitic protozoa are the result of directional mutation pressure on the base composition of the genomic DNA.

Key words: *Babesia*, *Eimeria*, *Theileria*, *Toxoplasma*, piroplasms, coccidia, codon usage.

INTRODUCTION

Dinucleotide frequency and codon usage have been analysed in gene sequences from five parasitic protozoa (*Babesia bovis*, *Babesia rodhaini*, *Theileria parva*, *Toxoplasma gondii* and *Eimeria tenella*) of the class Sporozoa (Ellis *et al.* 1993). The analysis showed that in keeping with the 'genome hypothesis', codon usage in these organisms was non-random and species specific. Metric multidimensional scaling (Belbin, 1989) of the codon usage data was used to investigate overall similarities in the pattern of codon usage between taxa (i.e. average codon usage across all gene sequences for each taxon), and this analysis separated the taxa into groups which were in general agreement with the current classification of these organisms into the subclasses Coccidiasina (*Eimeria*, *Toxoplasma*) and Piroplasmasina (*Babesia*, *Theileria*) (Levine, 1985; Ellis *et al.* 1993). Several low-usage codons were identified, and their usage also appeared to be related to the taxonomic position of the organisms under study.

It is, however, possible that the differences observed between taxa in this study may simply reflect codon usage bias that is unrelated to the phylogeny of the organisms. Bias in codon usage may result from the influence of a number of factors, such as the amino acid composition of the gene product, bias in codon usage of synonymous codons within amino acids, or directional mutation pressure on the base composition of the genome (reviewed by Andersson & Kurland, 1990; Osawa *et al.* 1992). In addition, genes of unicellular organisms that are highly expressed normally show a differential use of synonymous codons compared with other genes expressed at much lower levels. Since no data were available concerning the levels of expression of the genes studied by Ellis *et al.* (1993), it was not possible to analyse the relationship between gene expression and codon usage. However, it is possible to investigate differences in the codon usage of the overall gene sequences that are attributable to bias in the amino acid composition or to bias in synonymous codon usage (Long & Gillespie, 1991).

Here, we present a more extensive analysis of codon usage and codon usage bias in the individual gene sequences derived from organisms of the genera *Babesia*, *Eimeria*, *Theileria* and *Toxoplasma*.

MATERIALS AND METHODS

The data set (shown in Table 1) contains 36 genes of the coccidia and the piroplasms. The gene sequences used in this study were primarily the data set previously described (Ellis *et al.* 1993), along with the following sequences. (1) *T. gondii*, TOXGRA4A (Mevelec *et al.* 1992); (2) the complete gene sequence of 5401 of *E. tenella* (Anderson *et al.* 1990) replaces the partial sequence of 5401 (Danforth *et al.* 1989); (3) BBOBTB (GenBank accession number L00978),

Table 1. Sequences used in this study

Source/locus*	Accession number†	Length‡	GC	N_c	GC_{3S}	B	Code§
Babesia							
BB0225AA	M80466	1722	0·44	49·65	0·36	0·164	1
BBOBABR12	K02833	601	0·43	60·59	0·46	0·131	2
BBOMER60	M38218	1698	0·43	51·66	0·46	0·098	3
BBOMSA	M80467	956	0·42	46·86	0·47	0·211	4
BBOPBV42A	M77192	858	0·44	60·81	0·37	0·130	5
BBOBTB	L00978	1326	0·49	41·06	0·51	0·144	6
BBWC11MR	X63723	4311	0·58	46·97	0·59	0·126	7
BBO80KDC1B	M93126	1824	0·46	50·00	0·41	0·154	8
BBOSAGS	M19145	1014	0·40	33·42	0·34	0·317	9
BBOSAGS	M19145	981	0·39	37·88	0·32	0·391	10
Eimeria							
EIMSPORAN	X15898	651	0·69	37·41	0·83	0·441	11
EIMET100	M73495	2145	0·55	57·20	0·51	0·063	12
pTCD26	—	706	0·52	50·66	0·49	0·196	13
5401	—	6567	0·60	50·94	0·65	0·323	14
SP59	—	681	0·61	54·57	0·67	0·191	15
S0311-29	—	785	0·54	55·05	0·55	0·147	16
37 kDa antigen	—	792	0·62	60·44	0·58	0·186	17
200 kDa antigen	—	3003	0·55	50·64	0·55	0·124	18
GX3276	—	633	0·73	46·56	0·68	0·407	19
Theileria							
THECK11AS	M92084	1263	0·35	48·96	0·38	0·274	20
THET70HSP	J04653	1941	0·47	40·27	0·49	0·262	21
THETACP	M86659	1326	0·36	43·94	0·30	0·196	22
THE104MRA	M29954	2775	0·41	47·60	0·28	0·138	23
THECYSPTS	M37791	1320	0·42	47·01	0·41	0·193	24
THEHSP90	M57386	2166	0·49	49·13	0·66	0·170	25
THEP67PRT	M67476	789	0·43	42·34	0·40	0·223	26
THEP67PRT	M67476	2130	0·43	46·50	0·29	0·191	27
Toxoplasma							
TOXANT28K	J04018	759	0·53	52·76	0·54	0·154	28
TOXANTMS	M23658	960	0·53	59·64	0·53	0·112	29
TOXANTP	M26007	573	0·53	61·00	0·57	0·170	30
TOXNTP	M33472	572	0·52	59·85	0·59	0·205	31
TOXP22	M33572	561	0·55	61·00	0·54	0·150	32
TOXROP1A	M71274	1260	0·58	57·95	0·50	0·089	33
TOXTUBAA	M20024	1362	0·58	38·72	0·75	0·370	34
TOXTUBBA	M20025	1350	0·59	39·96	0·77	0·314	35
TOXGRA4A	M76432	1038	0·56	60·27	0·51	0·075	36

* Description of the gene sequence.
† GenBank or EMBL accession number used for the retrieval of the gene sequences.
‡ Length of gene sequence in base pairs.
§ Sequence code used in the figures.

BBWC11MR (GenBank accession number X63723) and BBO80KDC1B (GenBank accession number M93126) of *B. bovis*; and (4) THECK11AS (ole-Moi Yoi *et al.* 1992) and the two coding sequences of THEP67 (Nene *et al.* 1992) of *T. parva*. DNA sequences in GenBank were extracted from Release 76·0 using the accession numbers given in Table 1.

Codon usage tables were constructed using the routine CODON FREQUENCY in the GCG software package (Devereux, Haeberli & Smithies, 1984) run on a SunSparc computer through the Australian National Genomic Information Service. Heterogeneity in codon usage between genes was investigated by metric multidimensional scaling (MMDS, Ellis *et al.* 1993; Belbin, 1989). This involves displaying codon usage bias among genes as a multivariate analysis ordination which visualizes the relationships among entities as a two-dimensional graph, where the distance between the entities on the graph represents their differences in codon usage. A two-dimensional ordination was used, with the Euclidean distance measure (Faith, Minchin & Belbin, 1987).

The computer program CODONS (Lloyd & Sharp, 1992*a*) was used to calculate two of the indicators of codon usage bias for the nucleotide sequences shown in Table 1, namely N_c (Wright, 1990) and GC_{3S} (Sharp & Devine, 1989). N_c

(commonly referred to as the 'effective number' of codons used by a gene) is a general measure of non-uniformity of codon usage and can take values between 20 (for genes which are highly biased and use only 1 codon for each amino acid) and 61 (for unbiased genes), whereas GC_{3S} is defined as the frequency of G plus C at silent (i.e. synonymously variable) third positions of sense codons (excluding codons for Trp, Met and stop codons). A standardized measure of bias in synonymous codon preference (B) was also calculated (Long & Gillespie, 1991). B can take values between 0 and 1 (for biased genes).

A χ^2-goodness-of-fit test was used to identify codons that were over- or under-represented in the data set. The expected codon usage was calculated for synonymous codons on the basis that all synonymous codons might be expected to be used with equal frequency in an unbiased gene.

RESULTS

Patterns of codon usage

Metric multidimensional scaling was used to analyse the variability between genes in overall codon usage. This analysis separated the gene data set into two groups (separated by the dashed line in Fig. 1). Most of the gene sequences from *Babesia* and *Theileria* group together (to the left of the line), and similarly most of the gene sequences from *Eimeria* and *Toxoplasma* group together (to the right of the line). This means that the overall patterns of codon usage shown by each of the genes of the coccidia are more similar to each other than they are to those of the piroplasms. This observation is in accordance with the results for the overall codon usage for each taxon that has been reported previously (Ellis *et al.* 1993). However, some of the nucleotide sequences appear to show unusual codon usage relative to their related genes (e.g. BBO225AA and BBOMSA of *B. bovis*, and EIMSPORAN, GX3276 and the 37 kDa antigen fragment of *E. tenella*), and three genes do not follow the general trend (BBOBTB and BBO80KDC1B of *B. bovis*, and the 5401 sequence of *E. tenella*).

Codon usage patterns shown by a gene can result from a bias in the amino acid composition of the gene product. Such a situation could potentially lead to an over-representation of such codons (and therefore resulting in codon usage bias) simply because of the over-representation of these amino acids in the gene product. Fig. 2 shows the result obtained when the method used includes a procedure which removes the effect of the amino acid composition of the gene product from the measure of codon usage. Any variability detected by this analysis must result from bias in the usage of synonymous codons.

As in Fig. 1, the ordination of Fig. 2 shows two main groups of nucleotide sequences, originating

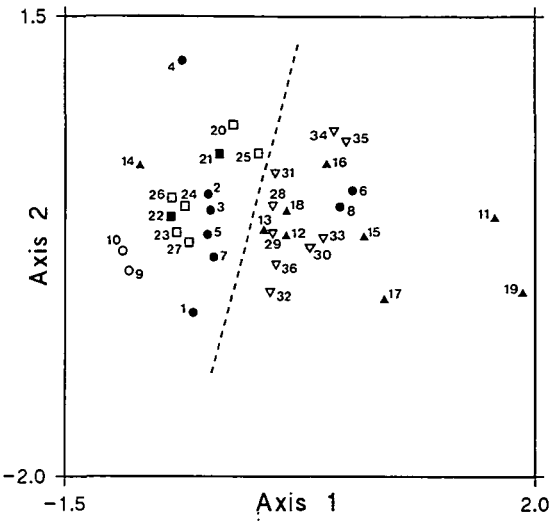

Fig. 1. Metric multidimensional scaling (Belbin, 1989) was used to investigate the overall pattern of codon usage shown by gene sequences of coccidia and piroplasms (●, *B. bovis*; ○, *B. rodhaini*; ▲, *E. tenella*; ■, *T. annulata*; □, *T. parva*; ▽, *T. gondii*). For each gene sequence, the usage of each codon was expressed as a frequency/1000 codons. A two-dimensional ordination was used, with the manhattan distance measure (Faith *et al.* 1987). Most of the piroplasm sequences are to the left of the line, and most of the coccidia sequences are to the right of the line. Sequence codes are as in Table 1.

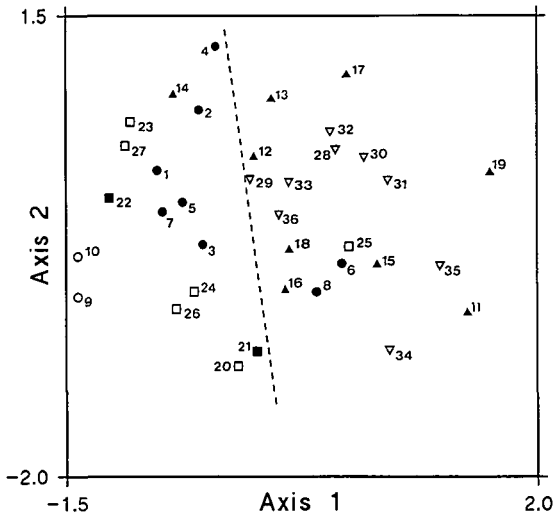

Fig. 2. Metric multidimensional scaling (Belbin, 1989) was used to investigate the pattern of codon usage among synonymous codons shown by gene sequences of coccidia and piroplasms. For each gene sequence, the usage of each codon was expressed as a frequency relative to the other synonymous codons. A two-dimensional ordination was used, with the manhattan distance measure (Faith *et al.* 1987). Most of the piroplasm sequences are to the left of the line, and most of the coccidia sequences are to the right of the line. Sequence codes are as in Table 1. Symbols are as in Fig. 1.

from either the coccidia or the piroplasms (separated by the dashed line in Fig. 2). This indicates that, while the amino acid composition of the product

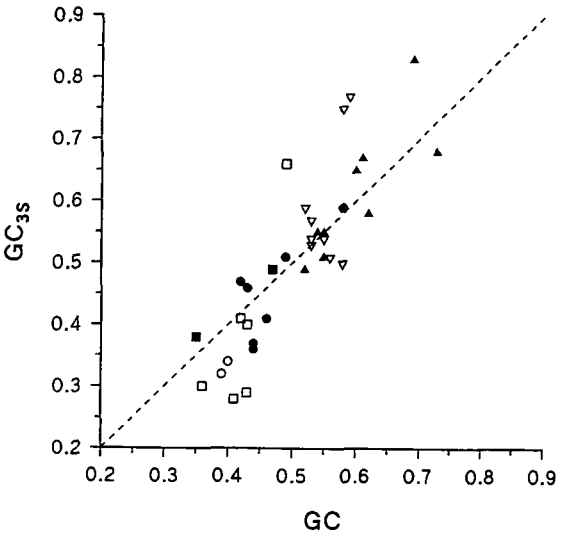

Fig. 3. Frequency of G+C at silent third positions of synonymous codons (excluding codons for Trp, Met and stop codons) (GC_{3S}, Wright (1990)) relative to the G+C frequency of the whole gene for each of the nucleotide sequences. The line is GC_{3S} = GC. Symbols are as in Fig. 1.

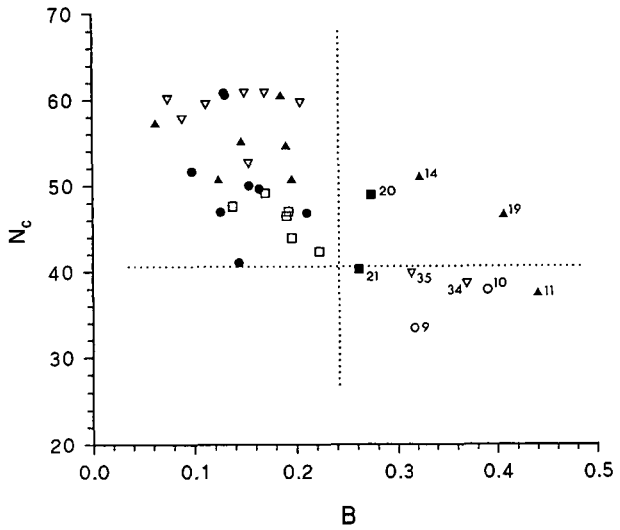

Fig. 4. 'Effective' number of codons used by the gene sequence (N_c, Wright (1990)) relative to the standardized synonymous codon bias (B, Long & Gillespie (1991)). Gene sequences below the horizontal line are indicated by N_c as having higher codon usage bias than those above the line, while gene sequences to the right of the vertical line are indicated by B as having higher codon usage bias than those of the left of the line. Sequence codes are as in Table 1. Symbols are as in Fig. 1.

of these genes clearly affects the pattern of codon usage shown by the nucleotide sequences, it is the pattern of usage of synonymous codons that has the largest influence on overall codon usage. Furthermore, the BBO225AA, BBOMSA, EIMSPORAN, GX3276, 37 kDa antigen fragment, BBOBTB, BBO80KDC1B, and 5401 genes are now more similar to the related sequences, indicating that it is partly an unusual amino acid composition that makes them different from their relatives. In addition, in this analysis the THEHSP90 sequence clusters with the majority of the coccidial sequences, showing that it has an unusual usage of synonymous codons relative to the other genes of the piroplasms.

Codon usage bias

The computer program CODONS was used to calculate two indicators of codon usage bias for the nucleotide sequences shown in Table 1, namely N_C and GC_{3S}. B was also calculated, and the results obtained are presented in Table 1.

This analysis shows that the main codon usage difference between the coccidia and the piroplasms is related to a significant difference in their GC content (analysis of variance, $F = 59{\cdot}13$, $P < 0{\cdot}001$) – the GC content of the piroplasm genes ranges from 0·35 to 0·58, while the GC content of the coccidial genes ranges from 0·52 to 0·69. Furthermore, the GC_{3S} values for the nucleotide sequences under study are strongly related to the overall GC values (see Fig. 3), indicating that the variation in silent third codon positions is not random.

A plot of N_C versus B for the coccidia and the piroplasms is shown in Fig. 4. The plot reveals that six nucleotide sequences are significantly biased according to both bias measures (both coding sequences of BBOSAGS, EIMSPORAN, THET70HSP, TOXTUBAA and TOXTUBBA), as they have N_C values of less than 40 and a B that is greater than 0·25. Bias in a further three sequences is detected by B alone (5401, GX3276, and THECK11AS). Five of these nine biased sequences have peripheral positions on the MMDS ordinations (both coding sequences of BBOSAGS, EIMSPORAN, 5401, and GX3276), indicating that their unusual codon usage pattern is a result of codon usage bias.

These analyses show that the unusual codon usage patterns shown by BBO225AA, BBOMSA, BBOTB, BBO80KDC1B, the 37 kDa antigen fragment, and THEHSP90 are probably not the result of codon usage bias. They must therefore be the result of unusual choices of codons, rather than very unequal usage of the codons.

Bias in nucleotide sequences of coccidia

The ordination analyses separate the nucleotide sequences of *E. tenella* into two subsets – the first group contains those sequences that are more widely distributed on the ordination, and includes four nucleotide sequences whose average base composition (as %GC) exceeds 60 (EIMSPORAN, SP59, GX3276, and the nucleotide sequence encoding the

Table 2. Codons over- or under-represented in gene sequences of *Eimeria tenella*

Sequence	Over-represented codons*	Under-represented codons*
Group 1		
SP59	GGC, GTG, TCC, CAG, CCC	CAA
37 kDa antigen	CAG, CGG	CAA
EIMSPORAN	GGC, CAG, CCC, GAG, AGC, TGC, CTG	CAA, CTA, GAA, GGA, TGT, CTT
GX3276	GTG, CCC, CGG, CTG, GCA	CTA
Group 2		
pTCD26	TCT	
EIMET100	TGC	TGT
SO311-29		CTA
200 kDa antigen	TCT, TTG, TTC, AAG, AAC, ATT, GTC, GCT, CGC, CTT	CTA, GTA, ATA, TTA, CGG, CGA, GCG, AAA, TTT, AAT

* χ^2 values (not shown) are significant at $P < 0{\cdot}05$.

37 kDa antigen; designated here as group 1). The other group contains all the remaining nucleotide sequences of *E. tenella* (except 5401), as well as all the genes of *T. gondii*. This observation raises the possibility that the codon usage shown by the two subsets of *E. tenella* genes may be related to the base composition of the gene sequences, which in turn should be evident as an over-representation of codons containing G and/or C in one subset compared with the other. Consequently, codon usage tables compiled from the individual nucleotide sequence data of *E. tenella* were subject to analysis using a χ^2-goodness-of-fit test in order to identify those codons that were either over- or under-represented in the sequences under study. The results obtained are shown in Table 2.

It is evident that the group 1 sequences contain a significant number of codons that are over-represented and nearly all of them contain G and/or C at two or more of the codon base positions and either G or C at the third base position. Like these nucleotide sequences, 5401 (which stands alone on the ordination) also contains 11 over-represented codons which contain G or C at the third base position and 17 under-represented codons, 14 of which contain A or T at the third base position.

Three of the four sequences in group 2 (pTCD26, EIMET100 and SO311-29) show few codons that are over- or under-represented (Table 2), and this correlates with the unbiased nature of these sequences predicted from the N_C and B values (Table 1). However, the nucleotide sequence encoding the 200 kDa antigen has N_C and B values that are indicative of a relatively un-biased gene, but it contains 10 over-represented codons and 10 under-represented codons. Three gene sequences of *T. gondii* (TOXANT28K, TOXANTMS and TOXANTP) that clustered around the group 2 sequences of *E. tenella* were also analysed by a χ^2-goodness-of-fit test. In general, they contained few over- or under-represented codons, with only low values of χ^2 (not shown). These results therefore seem to show that although exceptions may well exist in this group, the nucleotide sequences that cluster around (and include) the group 2 sequences represent a group that contain a relatively unbiased pattern of codon usage.

Bias in nucleotide sequences of piroplasms

The ordinations also reveal that the group of nucleotide sequences derived from the piroplasms form a cluster that runs almost parallel to the dashed line that has been drawn to separate the sequences of the piroplasms from those of the coccidia. Therefore a χ^2-goodness-of-fit analysis of the codon usage shown by groups of genes that appear at the extreme edges of this cluster should reveal the extremes of bias shown by the group. The first group analysed contained the sequences BBOMSA, THECK11AS and THEHSP90. All three of these genes had N_C and B values that indicated they should show a significant degree of codon usage bias. The χ^2-goodness-of-fit test revealed several codons that were significantly over- or under-represented (see Table 3). At the other extreme edge of the piroplasm group of sequences, BBO225AA and the two coding regions of BBOSAGS were also analysed. They contained a much larger number of over- and under-represented codons. The over-represented codons identified in all three nucleotide sequences showed a significant bias towards T at the third base position of over-represented codons, and G or A at the third base position of under-represented codons.

Table 3. Codons over- or under-represented in gene sequences of the piroplasms

Sequence	Over-represented codons*	Under-represented codons*
BBOMSA	GGA, ACA	
THECK11AS	ACA, GTA, CTA	
THEHSP90	GAG, GAC, GCA, AAC, TCA, GTC, AAG	GGG, GAA, GAT, AAT, AAA, ACG
BBO225AA	CTT, TCT, ATT, GAT, GGT, GTC	ATA, CTG, GAC, GGG, GTA, ATC, TCA, CTA
BBOSAGS (26)	CTT, CAA, TCT, AGT, GCT, GTT, GAA, GGT, AAG, CCT	CAG, ATA, GCA, GTG, GAG, GCG, AAA, ACG, ACA
BBOSAGS (17)	CTT, CAA, TCT, ACT, ATT, GCT, GTT, GAT, GAA	CAG, ATA, GCA, GCG, GTG, GAC, GAG
BBWC11MR	GTG, GCC, GGT, ATC, GAG, CTG, CTC, CCC	GTA, GTT, GCG, ATA, ATT, ACG, TCG, CAA, CTA, CCG
THE104MRA	GTT, GCA, AGT, ATT, ACA, AGT, TTT, CGT, CTT, CCA, CCT, GGA, GAA	GGG, GAG, GCG, GTC, AGC, ATC, ACG, TTC, CCG, CCC
BBOPBV42A	GTT, GCT, TCT	GCG, TCG
THECYSPTS	GGT, GTT, TCT, ACT, CGT	GAA, GGG, GAG, GTG, ACG, TCG, ACA
BBOMER60	GCT, GTT, ACT, CGT, CCT	GCG, ACG, CCG
BBOBABR12	CCA	
BBOBTB	GGC, GTT, GCT, ATT, ACT, CCC	CCG
BBO80KDC1B	GGA, GGT, GAA, GCC, ATC, TTC, CTC, CCA	GGG, CAG, GGC, GTG, ACG, TTT, TCA, CGC, CTG

* χ^2 values (not shown) are significant at $P < 0{\cdot}05$.

Similarly, analysis of a randomly chosen set of nucleotide sequences (BBWC11MR, THE104MRA, BBOPBV42A, THECYSPTS, BBOMER60 and BBOBABR12) also revealed a preference for T in the third base position of over-represented codons (22 of the 34 identified had T in the third base position). Nineteen out of 30 under-represented codons in this data set contained G at the third base position. Therefore, we conclude that although the multivariate analysis technique used here reveals substantial variability in the pattern of codon usage used by the genes of the piroplasms, generally these sequences show an over-representation of T at the third base position of over-represented codons and G at the third base position of under-represented codons.

In both ordinations, BBOBTB and BBO80KDC1B cluster close to the majority of the coccidial genes. Consequently, the pattern of codon usage shown by BBOBTB and BBO80KDC1B were investigated further. BBOBTB contained 10 over-represented codons (ending in either C or A, except GGT) and 8 under-represented codons. BBO80KDC1B contained 6 over-represented codons (ending in T or C), but only CCG was under-represented. There was no obvious trend in the identity of the over- and under-represented codons, although T (and to a lesser extent C) was obviously the base of choice at the third base position of over-represented codons.

DISCUSSION

In contrast to the study of Ellis *et al.* (1993), which investigated the overall similarity in codon usage between taxa (i.e. average codon usage across all nucleotide sequences for each taxon) the study presented here compares differences in codon usage between individual genes of parasitic protozoa representing the subclasses Coccidiasina and Piroplasmasina of the class Sporozoa in the phylum Apicomplexa. This study has shown that although the patterns of codon usage of the gene sequences from members of the same subclass are generally more similar to each other than they are to members of the other subclass, the nucleotide sequences derived from any one taxon do not have identical patterns of codon usage. These observations are in keeping with those made for most other eukaryotic and prokaryotic organisms (e.g. Sharp & Devine, 1989; Lloyd & Sharp, 1992*b*). In addition, it has

been shown that the largest influence on the overall pattern of codon usage shown by the individual genes is the pattern of usage of synonymous codons, rather than the pattern of amino acid composition of the gene product.

In genes of both the coccidia and the piroplasms, considerable variability was detected in codon usage bias, and this does explain some of the unusual patterns of codon usage detected in the gene sequences. In the coccidia codon usage bias is generally associated with an over-representation of G and/or C-rich codons. Furthermore, the observation that the data set of nucleotide sequences from *E. tenella* contains a group of sequences that are biased in GC composition explains the observed 'preference' shown by codons of *E. tenella* for a G or C at the first and third base positions reported previously (Ellis *et al.* 1993). The sequences of coccidia and piroplasms differ in their GC composition, and the GC content of the silent third codon positions is non-random with respect to this. In the piroplasms a 'preference' for codons containing T at the third base position was detected. In addition, the overall codon usage pattern detected in two gene sequences of *Babesia rodhaini* shows little similarity to the overall codon usage patterns of other piroplasms (Ellis *et al.* 1993).

Codon usage in the malarial parasite *Plasmodium falciparum* has been described previously (Hyde & Sims, 1987; Weber, 1987; Saul & Battistutta, 1988). Of relevance to the study reported here are the observations that the 'preferred' codons (described in this study as over-represented codons) used by *P. falciparum* contain T and/or A. The genome of *P. falciparum* has a GC content (as %GC) of 18, and therefore like the coccidia and the piroplasms, the codons over-represented in genes of *P. falciparum* are probably the result of directional mutation pressure on the base composition of the genomic DNA. This theory is based on the assumption that the effect of mutation is not random but has a directionality that drives the GC content of DNA towards either a higher or lower %GC.

Analysis of DNA sequence data has revealed that probably all organisms are subject to directional mutation pressure (Sueoka, 1988). In bacteria, for example, the GC content of various parts of the genome reveals a positive linear relationship with the GC content of their genomic DNA (Muto & Osawa, 1987). It therefore appears reasonable to conclude that directional mutation pressure has also had a major influence on the genome of sporozoan parasites, since the %GC of the nucleotide sequences under study is also similar in magnitude to the base composition of the genome from which they are derived. For example, the %GC of genomic DNA from *T. parva* and *T. gondii* have been reported as approximately 38 (Irwin, 1987) and 50 (Johnson, Dubey & Dame, 1986), respectively, while our sampled sequences show 42 and 55, respectively. In the class Sporozoa, therefore, codon usage appears to be partly determined by the genome base composition. Such a conclusion was drawn for data derived from vertebrates, other eukaryotes and prokaryotes (Bernardi & Bernardi, 1985).

REFERENCES

ANDERSON, D., McCANDLISS, R. J., STRAUSBERG, S. L. & STRAUSBERG, R. L. (1990). Genetically engineered coccidiosis vaccine. *Patent Corporation Treaty* WO90/00403.

ANDERSSON, S. G. E. & KURLAND, C. G. (1990). Codon preferences in free-living micro-organisms. *Microbiology Reviews* **54**, 198–210.

BELBIN, L. (1989). PATN – Pattern Analysis Package Technical Reference. Canberra: CSIRO.

BERNARDI, G. & BERNARDI, G. (1985). Codon usage and genome composition. *Journal of Molecular Evolution* **22**, 363–5.

DANFORTH, H. D., AUGUSTINE, P. C., RUFF, M. D., McCANDLISS, R., STRAUSBERG, R. L. & LIKEL, M. (1989). Genetically engineered antigen confers partial protection against avian coccidial parasites. *Poultry Science* **68**, 1643–52.

DEVEREUX, J., HAEBERLI, P. & SMITHIES, O. (1985). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 216–23.

ELLIS, J., GRIFFIN, H., MORRISON, D. & JOHNSON, A. M. (1993). Analysis of dinucleotide frequency and codon usage in the phylum Apicomplexa. *Gene* **126**, 163–70.

FAITH, D. P., MINCHIN, P. R. & BELBIN, L. (1987). Compositional dissimilarity as a robust measure of ecological distances: a theoretical model and computer simulations. *Vegetatio* **69**, 57–68.

HYDE, J. E. & SIMS, P. F. G. (1987). Anomalous dinucleotide frequencies in both coding and non-coding regions from the genome of the human malaria parasite *Plasmodium falciparum*. *Gene* **61**, 177–87.

IRWIN, A. D. (1987). Characterisation of species and strains of *Theileria*. *Advances in Parasitology* **26**, 145–97.

JOHNSON, A. M., DUBEY, J. P. & DAME, J. B. (1986). Purification and characterisation of *Toxoplasma gondii* tachyzoite DNA. *Australian Journal of Experimental Biology and Medical Science* **64**, 351–5.

LEVINE, N. D. (1985). Phylum II. Apicomplexa Levine, 1970. In *Illustrated Guide to the Protozoa* (ed. Lee, J. J., Hutner, S. H. & Bovee, E. C.), pp. 322–374. Kansas: Society of Protozoologists.

LLOYD, A. T. & SHARP, P. M. (1992*a*). CODONS: a microcomputer program for codon usage analysis. *Journal of Heredity* **83**, 239–40.

LLOYD, A. T. & SHARP, P. M. (1992*b*). Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. *Nucleic Acids Research* **20**, 5289–95.

LONG, M. & GILLESPIE, J. H. (1991). Codon usage divergence of homologous vertebrate genes and codon usage clock. *Journal of Molecular Evolution* **32**, 6–15.

MEVELEC, M., CHARDES, T., MERCEREAU-PIUJALON, O., BOURGUIN, I., ARCHBAROU, A., DUBREMETZ, J. & BOUT, D. (1992). Molecular cloning of GRA4, a *Toxoplasma*

gondii dense granule protein, recognised by mucosal IgA antibodies. *Molecular and Biochemical Parasitology* **56**, 227–38.

MUTO, A. & OSAWA, S. (1987). The G+C content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Scienes, USA* **84**, 166–9.

NENE, V., IAMS, K. P., GOBRIGHT, E. I. & MUSOKE, A. J. (1992). Characterisation of the gene encoding a candidate vaccine antigen of *Theileria parva* sporozoites. *Molecular and Biochemical Parasitology* **51**, 17–28.

OLE-MOI YOI, O. K., SUGIMOTO, C., CONRAD, P. A. & MACKLIN, M. D. (1992). Cloning and characterisation of the casein kinase II alpha subunit gene from the lymphocyte-transforming intracellular protozoan parasite *Theileria parva*. *Biochemistry* **31**, 6193–202.

OSAWA, S., JUKES, T. H., WATANABE, K. & MUTO, A. (1992). Recent evidence for evolution of the genetic code. *Microbiology Reviews* **56**, 229–64.

SAUL, A. & BATTISTUTTA, D. (1988). Codon usage in *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* **27**, 35–42.

SHARP, P. M. & DEVINE, K. M. (1989). Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. *Nucleic Acids Research* **17**, 5029–39.

SUEOKA, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences, USA* **85**, 2653–7.

WEBER, J. L. (1987). Analysis of sequences from the extremely A+T-rich genome of *Plasmodium falciparum*. *Gene* **52**, 103–9.

WRIGHT, F. (1990). The effective number of codons used in a gene. *Gene* **87**, 23–9.