Gene, 126 (1993) 163–170

GENE 07001

Analysis of dinucleotide frequency and codon usage in the phylum Apicomplexa

(Low-usage codons; phylogeny; parasitic protozoa; *Babesia*; *Eimeria*; *Theileria*; *Toxoplasma*; piroplasms; coccidia)

John Ellis[a], Hugh Griffin[b], David Morrison[c] and Alan M. Johnson[a]

[a]*Department of Microbiology, University of Technology, Sydney, St. Leonards Campus, Broadway, New South Wales, 2007, Australia;* [b] *Genetics and Microbiology Department, AFRC Institute of Food Research, Norwich Laboratory, Norwich, Norfolk, NR4 7UA, UK. Tel. (44-603) 56122;* and [c] *Department of Applied Biology, University of Technology, Sydney, St. Leonards Campus, Broadway, New South Wales, 2007, Australia. Tel. (61-2) 330-4159*

Received by: P.A. Manning: 15 July 1992; Revised/Accepted: 7 October/12 October 1992; Received at publishers: 17 December 1992

SUMMARY

Dinucleotide frequency (DiF) and codon usage (cu) were analysed in gene sequences from four parasitic protozoa, *Babesia bovis*, *Theileria parva*, *Toxoplasma gondii* and *Eimeria tenella*, of the phylum Apicomplexa. In keeping with the 'genome hypothesis', cu was found to be non-random and species specific in these organisms, although cu among members of the same subclass was found to be very similar. Several low-usage (lu) codons were identified, and the usage of lu codons appears to be related to the taxonomic position of the organisms under study. A comparison of the observed/expected DiF ratios obtained from gene coding regions revealed a low frequency of the TA and CG dinucleotides in all organisms studied. A comparison of these DiF ratios with those found in rRNA-encoding genes and in introns, showed that in the parasites, *B. bovis* and *Th. parva* (representing the piroplasms), the low frequency of dinucleotides appeared to be the result of coding pressure alone. In *T. gondii* and *E. tenella* (representing the coccidia), however, coding pressure could not completely explain differences in DiF.

INTRODUCTION

Recent studies comparing 18S rRNA sequences from parasites within the phylum Apicomplexa have suggested that, although at least some of the current classifications are incorrect, (Ellis et al., 1992; Johnson et al., 1991; Tenter et al., 1992) all of these organisms are probably monophyletic (Barta et al., 1991). Since many of them cause substantial economic and reproductive loss in animals and/or humans (Ellis and Johnson, 1992), it is essential that we ascertain as much about the comparative molecular biology of these organisms as possible.

The genome hypothesis of Grantham et al. (1981) states that cu is nonrandom and species specific. However, it was recognised that amongst taxonomically related species, the pattern of codon usage may be similar (Ikemura, 1985; Anderson and Kurland, 1990; Osawa et al., 1992). Until recently, there has been insufficient nt sequence data available from gene sequences derived from apicomplexan parasites to allow comparisons to be made in order to assess the genome hypothesis and its application to these organisms. A preliminary study which involved a comparison of DiF and cu between *T. gondii* and *Plasmodium falciparum* (Johnson, 1990)

Correspondence to: Dr. A.M. Johnson, Department of Microbiology, University of Technology, Sydney, St. Leonards Campus, P.O. Box 123, Broadway, New South Wales, 2007, Australia. Tel. (61-2) 330-4075; Fax (61-2) 330-4003.

Abbreviations: aa, amino acid(s); *B.*, *Babesia*; bp, base pair(s); cu, codon usage; DiF, dinucleotide frequency; *E.*, *Eimeria*; E.P.O., European Patent Office; GCG, Genetic Computer Group, Madison, WI, USA; lu, low usage; N, any nucleoside; nt, nucleotides(s); o/e, ratio of observed to expected DiF values, ORF, open reading frame; *P*, probability; *P.*, *Plasmodium*; P.C.T., Patent Corporation Treaty; R, purine (A or G); r, ribosomal; *T.*, *Toxoplasma*; *Th.*, *Theileria*; W, A or T; Y, pyrimidine (C or T).

revealed striking differences, indicating no especially close relationships between them. Furthermore, since the amount of genetic divergence among some members of the Apicomplexa is as great as that between vertebrates and invertebrates (Ellis et al., 1992, Johnson et al., 1991) it is not unreasonable to expect that there may be large differences in cu among these organisms. This study extends the comparison of DiF and cu between *Plasmodium* and *Toxoplasma* to include representatives of three additional genera that are present in the phylum Apicomplexa, namely *Babesia*, *Eimeria* and *Theileria*. In addition we provide a re-evaluation and update of DiF and cu found in gene sequences of *T. gondii*.

RESULTS AND DISCUSSION

The nt sequences analysed in this study are shown in Table I. They represent almost a complete, albeit relatively small, set of gene sequence data that is currently available for these organisms. Many of them code for antigens, which introduces a potential source of bias into the data. However, they were chosen for study because they are all probably transcribed by RNA polymerase II; greater than 500 bp long; characterised by a single, major ORF; and they do not contain a large amount of repetitive nt sequence. The nt sequences present in either GenBank or the EMBL databases were extracted using the accession Nos. given in Table I.

(a) Dinucleotide frequency

The DiF in coding regions of the sequences listed in Table I was determined and the results are presented in Table II. The significance of the deviation of the observed from the expected DiF calculated from the overall nt composition of the sequences under study was assessed using the χ^2 test. In all four species TA was found to be significantly under represented in this data set ($P<0.05$). All species also showed an under representation of the CG dinucleotide ($P<0.05$), although in the data derived from *T. gondii* the magnitude of this under representation was small. This result is consistent with earlier observations (Johnson, 1990). The finding that the DiF of CG and TA dinucleotides is low is in keeping with the fact that both are under represented in genes of a wide range of different organisms such as bacteria, yeast, drosophila and primates (Zhang et al., 1991). Both *B. bovis* and *E. tenella* also showed a significant overrepresentation of the TG dinucleotide. These results were of interest since it has been reported that the CG dinucleotide occurs at low frequency in the genome of *P. falciparum* (Weber, 1987; Hyde and Sims, 1987). Until recently however there has been no evidence to indicate the presence of methylated analogues of cytosine in the genome of *P. falciparum* (Pollack et al., 1991). Such analogues are known to mutate in the genomes of vertebrates into TG, resulting in an over representation of this dinucleotide. Although it is not known whether methylated analogues of cytosine are present in DNA of *Babesia*, *Eimeria*, *Theileria* or *Toxoplasma*, the under representation of CG would be consistent with this possibility.

It has been known for some time that many parasite genes contain major biases in cu, which are frequently manifest at the DNA level as simple repetitive sequences within the gene sequence (Ellis and Tomley, 1991; Schofield, 1990; Enea et al., 1984). Since it is possible that specific individual genes may contribute more than others to the under or over representation of dinucleotides in the data set, the deviations of the observed from the expected frequency were determined for each individual gene sequence (not shown). In the case of the CG dinucleotide, nearly all of the gene sequences analysed had a significant under representation ($P\leqslant 0.05$) of this dinucleotide. Only in the data from *B. bovis* was there any evidence of a major contribution from a subset of genes to the under representation of CG. In this instance all of the genes had an under representation of CG which was significant at $P\leqslant 0.1$, with three (BBO225AA, BBOMER60 and BBOBABR12) being significant at $P\leqslant 0.01$. In addition all genes analysed in all of these organisms studied had a low frequency of the TA dinucleotide which was significant at $P\leqslant 0.01$. Similar observations were made in two additional gene sequences of *Th. annulata* (THETACP, THE7OHSP) where CG and TA were again both under represented ($P\leqslant 0.01$). Accordingly it is possible to conclude that the under representation of CG and TA does not result from major biases within individual genes, but is a feature of all of them.

(b) Low usage codons

Codon usage in both eukaryotes and prokaryotes is known to be predominantly influenced by directional mutation pressure on the nt composition of the genome (Osawa et al., 1992). The evolution of a genome may therefore result in the under or over representation of some dinucleotides and consequently it has been argued that DiF may influence codon usage (Phillips et al., 1987; Shields and Sharp, 1987; Zhang et al., 1991). Therefore dinucleotides under represented in coding regions may be apparent as codons which are present at low frequency. Table III shows a comparison of codon usage for *B. bovis*, *Th. parva*, *T. gondii* and *E. tenella* derived from the sequences listed in Table I. Several of the codons are used infrequently (lu codons) and we arbitrarily assigned codons to this class if their frequency of usage was less than 10 per 1000 codons. Three codons (CGA,

TABLE I

Sequences used in this study

Source/locus[a]	Description of gene product[b]	Accession numbers[c]	Location of ORF (bp)[d]	Reference[e]
B. bovis				
BBO225AA	225-kDa antigen	M80466	323-2044	Jasmer et al. (1992)
BBOBABR12	Rearranging locus	K02833	122-723	Cowman et al. (1984)
BBOMER60	60-kDa antigen	M38218	122-1819	Suarez et al. (1991)
BBOPBV42A	Surface protein	M77192	3-860	GenBank
BBOMSA	44-kDa antigen	M80467	1-956	GenBank
B. rodhaini				
BBOSAGS	26-kDa antigen	M19145	73-1086	Snary et al. (1988)
BBOSAGS	17-kDa antigen	M19145	1395-2375	Snary et al. (1988)
Th. parva				
THE CYSPTS	Cysteine protease	M37791	96-573 606-1447	Nene et al. (1990)
THE104MRAA	104-kDa antigen	M29954	452-3226	Iams et al. (1990)
THEHSP90	Heat-shock protein	M57386	130-2295	GenBank
Th. annulata				
THETACP	Cysteine protease	M86659	133-1458	Baylis et al. (1992)
THET70HSP	Heat-shock protein	J04653	445-2385	Mason et al. (1989)
T. gondii				
TOXANT28K	28-kDa antigen	J04018	171-520 706-1174	Prince et al. (1989)
TOXANTMS	30-kDa antigen	M23658	362-1321	Burg et al. (1988)
TOXANTP	24-kDa antigen	M26007	613-1185	Cesbron-Delauw et al. (1989)
TOXNTP	Nucleoside triphosphate hydrolase	M33472	7-578	Johnson et al. (1989)
TOXP22	22-kDa antigen	M33572	182-742	Prince et al. (1990)
TOXTUBAA	α Tubulin	M20024	387-448 960-1080 1231-2409	Nagel and Boothroyd (1988)
TOXTUBBA	β Tubulin	M20025	202-297 763-1856 1998-2085 2186-2257	Nagel and Boothroyd (1988)
TOXROP1A	Antigen	M71274	201-1460	Ossorio et al. (1992)
Eimeria tenella				
EIMSPORAN	Antigen	X15898	66-716	Liberator et al. (1989a)
EIMET100	Antigen	M73495	78-180 698-2278 2506-2828 2983-3120	Tomley et al. (1991)
pTCD26	Antigen		6-698	Newman et al. (1987)
5401	Antigen		1-864	Danforth et al. (1989)
SP59	Antigen		1-681	Liberator et al. (1989b)
SO311-29	Antigen		93-878	Liberator et al. (1989b)
	37-kDa antigen		1-1000	Altenburger et al. (1989)
	200-kDa antigen		1-3094	Altenburger et al. (1989)
GX 3276	Antigen		1-633	Anderson et al. (1990)

[a] Database designation of nt sequence (see footnotes c and d).

[b] Brief description for gene product.

[c] GenBank or EMBL accession Nos. used for the retrieval of nt sequences.

[d] Position of open reading frame (ORF) analysed within the nt sequence according to the numbering recorded in databases (see footnotes a, c and e).

[e] Publication reference.

TABLE II

Comparison of observed/expected dinucleotide frequency ratios found in gene coding regions of four apicomplexan parasites

Dinucleotide	*B. bovis*	*Th. parva*	*T. gondii*	*E. tenella*
	o/e values[a]			
TT	1.07	1.15	1.16	1.11
TC	1.09	1.17	1.20	0.99
TA	0.61*	0.69*	0.46*	0.43*
TG	1.43*	1.18	1.15	1.31*
CT	1.18	1.18	0.99	1.31*
CC	1.01	1.07	0.92	0.90
CA	1.11	1.06	1.15	1.30*
CG	0.55*	0.66*	0.95*	0.74*
AT	0.90	0.81	0.82*	0.75*
AC	0.95	1.06	0.97	0.76*
AA	1.12	1.04	1.09	1.08
AG	1.01	1.07	1.09	1.25*
GT	0.85	0.94	1.08	0.83*
GC	0.95	0.77	0.95	1.21*
GA	1.23	1.19	1.11	1.05
GG	0.90	1.03	0.86	0.87

[a]This analysis of DiF was carried out using a routine in the sequence analysis package of Staden (1986) run on a Sun Sparc computer. The observed/expected (o/e) values obtained were weighted according to the lengths of the sequences by multiplying the o/e values by the length of that sequence, summing these figures, and then dividing by the total sum length of the gene sequences (Hyde and Sims, 1987; Johnson, 1990). The significance of the deviations of observed from expected values obtained was measured using the χ^2 test. In order to analyse the sequences from genes containing introns, a single ORF was constructed for each gene using the ASSEMBLE routine in the GCG package (Devereux et al., 1984). Asterisks mark values which are significant at $P<0.05$.

Arg; CGG, Arg; TGT, Cys) are infrequently used by all the organisms. Strikingly twelve codons (ACG, CGA, CCG, CGC, CGG, CGT, GGG, GCG, TCG, TGC, TGG and TGT) are used infrequently by both *B. bovis* and *Th. parva* four of which code for arginine. Three of these, TGT, TGC (coding for Cys) and TGG (Trp), encode aa that are not abundant in proteins, and so the infrequent use of these codons merely reflects the low abundance of these aa in these proteins. It is interesting to note that like the piroplasms, *Saccharomyces cerevisiae* and primates show a low usage of codons containing the CG dinucleotide (Zhang et al., 1991). Ten codons (CGG, CGA, GTA, ATA, TGT, TTA, TCA, CTA, CAT and TAT) are used infrequently by *T. gondii* and *E. tenella*. Six of these contain either the TA or AT dinucleotide, and two others contain the CG dinucleotide. On this basis, it would seem that the genes of piroplasms are characterised by the low frequency of codons containing the CG dinucleotide, whereas the coccidial genes are characterised by the low frequency of a different class of codons, those containing the AT or TA dinucleotide. It is interesting to note that if one considers the o/e DiF ratios in the piroplasms AT or TA > CG, whereas in the coccidia CG > TA or AT. Such results therefore imply that in these organisms codons that are used infrequently generally contain dinucleotides that are under represented.

As stated by Zhang et al. (1991) it is of course necessary to carefully distinguish between cause and effect, i.e., it is necessary to determine whether low DiF is the cause of low codon frequency or the other way around. One approach used by others (e.g., Zhang et al. 1991) is to consider dinucleotide frequencies in transcribed, but untranslated coding regions (such as r genes or introns) in order to answer the question: what is the effect upon DiF when coding pressure is removed? Table IV shows the o/e DiF values derived from the small subunit rRNA gene sequences from four piroplasms (*B. bovis*, *B. bigemina*, *B. rodhaini* and *Th. annulata*). In all four of these species, the o/e values for both CG and TA revert to values nearing 1.0, which suggests that the under representation of these two dinucleotides in coding regions which are translated in piroplasms probably results from the coding pressure. Unfortunately complete gene sequences for the small subunit rRNA of *T. gondii* and *E. tenella* are not yet available. Consequently, dinucleotide frequencies in introns of these species were analysed in order to assess the effect of coding pressure on DiF. In both *T. gondii* and *E. tenella* the o/e ratios for CG and AT in the introns studied were similar to those values reported earlier (section **a**, Table II). Consequently, it would seem that coding pressure is unlikely to be responsible for the under representation of AT in the genes studied. Intriguingly, the ratios for TA rose from 0.46 to 0.73 and from 0.43 to 0.74 in *T. gondii* and *E. tenella*, respectively, indicating that coding pressure is partially responsible for the very low levels of this dinucleotide in the gene sequences we have studied. This is not surprising since this dinucleotide is found in stop codons.

(c) Comparison of the pattern of codon usage among taxa

Metric multidimensional scaling (Belbin, 1989) was used to investigate overall similarities in the pattern of cu among the piroplasms (*B. bovis*, *B. rodhaini*, *Th. parva* and *Th. annulata*), the coccidia (*T. gondii* and *E. tenella*) and *P. falciparum*. A two-dimensional ordination was used, with the standardised euclidean distance measure (Faith et al., 1987), and the results are displayed in Fig. 1. This analysis separated the taxa into three groups. One group contained the coccidia, and their pattern of cu was distinct from the other two groups. The second group contained three of the piroplasms (*B. bovis*, *Th. parva* and *Th. annulata*). These results are in agreement with the current classification of these organisms into the subclasses, Coccidiasina and Piroplasmasina respectively (Levine, 1985). The third much looser group contained

TABLE III

Codon usage[a]

aa	Codon	*B. bovis*	*Th. parva*	*T. gondii*	*E. tenella*	aa	Codon	*B. bovis*	*Th. parva*	*T. gondii*	*E. tenella*
		cu values[a]						cu values[a]			
Gly	GGG	2.57	2.87	10.96	17.46	Trp	TGG	7.72	8.62	5.68	13.02
Gly	GGA	7.72	15.33	15.42	33.15	End	TGA	7.72	0.00	4.06	0.59
Gly	GGT	15.43	18.21	30.03	28.42	Cys	TGT	5.14	6.23	6.49	6.22
Gly	GGC	7.20	10.06	24.76	42.03	Cys	TGC	4.63	5.27	12.58	25.46
Glu	GAG	32.92	41.69	39.77	48.54	End	TAG	3.09	0.00	0.41	1.18
Glu	GAA	47.84	45.04	30.84	35.52	End	TAA	4.63	1.44	3.25	0.89
Asp	GAT	36.01	31.62	17.05	13.02	Tyr	TAT	17.49	17.25	5.28	5.03
Asp	GAC	22.12	38.81	30.84	17.17	Tyr	TAC	13.37	21.08	14.20	13.91
Val	GTG	11.83	13.90	21.92	21.31	Leu	TTG	25.21	17.73	18.67	15.39
Val	GTA	10.29	14.85	8.12	8.58	Leu	TTA	10.80	11.98	3.25	2.96
Val	GTT	18.52	25.40	18.26	21.31	Phe	TTT	17.49	23.96	11.36	7.70
Val	GTC	16.46	14.37	28.81	24.57	Phe	TTC	23.15	16.29	32.06	18.06
Ala	GCG	5.14	4.31	19.48	17.17	Ser	TCG	6.69	4.79	15.02	8.88
Ala	GCA	6.69	19.65	17.45	34.93	Ser	TCA	14.92	24.44	9.74	8.58
Ala	GCT	16.98	11.02	17.86	37.59	Ser	TCT	22.63	17.25	10.55	19.24
Ala	GCC	11.83	10.54	23.13	29.01	Ser	TCC	10.29	11.50	16.23	15.10
Arg	AGG	17.49	17.25	10.96	12.73	Arg	CGG	3.09	0.00	8.52	7.99
Arg	AGA	13.37	19.17	12.99	8.58	Arg	CGA	2.06	0.00	7.31	2.66
Ser	AGT	14.40	13.42	12.58	9.18	Arg	CGT	8.74	7.67	15.83	5.62
Ser	AGC	15.43	8.15	16.23	16.87	Arg	CGC	4.63	4.31	16.23	14.80
Lys	AAG	30.86	54.62	28.81	19.54	Gln	CAG	14.40	11.02	33.69	31.08
Lys	AAA	31.38	42.17	12.18	11.25	Gln	CAA	21.60	10.54	14.61	14.50
Asn	AAT	23.15	14.85	8.93	10.66	His	CAT	16.98	10.54	4.87	5.03
Asn	AAC	24.18	27.31	25.97	15.98	His	CAC	14.40	10.06	13.39	6.51
Met	ATG	26.23	20.60	23.13	21.31	Leu	CTG	22.12	10.54	22.73	21.90
Ile	ATA	11.32	8.15	3.65	3.85	Leu	CTA	13.37	9.10	6.90	5.33
Ile	ATT	34.47	17.25	10.15	12.73	Leu	CTT	20.06	18.21	8.93	13.91
Ile	ATC	13.89	15.33	19.48	11.84	Leu	CTC	16.98	20.12	22.32	18.94
Thr	ACG	6.69	3.83	14.20	9.18	Pro	CCG	8.23	7.19	17.45	11.54
Thr	ACA	18.52	19.17	9.74	10.95	Pro	CCA	19.03	26.35	18.67	11.84
Thr	ACT	21.60	23.96	10.96	15.10	Pro	CCT	19.03	23.00	9.33	16.58
Thr	ACC	13.37	12.46	19.48	10.36	Pro	CCC	14.40	8.15	16.23	18.65

[a]The cu values given are the frequency per 1000 codons.

P. falciparum and *B. rodhaini*. The level of dissimilarity that exists between *B. rodhaini* and the other three piroplasms is as great as that detected between these piroplasms and *P. falciparum* or between *B. rodhaini* and *P. falciparum*. Of interest here are the results obtained from a phylogenetic analysis of the genus *Babesia* using rRNA gene sequence comparisons (Ellis et al 1992) which revealed that *B. rodhaini* lay outside of a clade consisting of *B. bovis*, *B. bigemina* and *Th. annulata*. The results obtained from this analysis of cu support the view that *B. rodhaini* may not be as closely related to the other piroplasms as initially thought.

In unicellular organisms, genes that are highly expressed normally show a differential use of synonymous codons compared to other genes expressed at much lower levels. It is therefore possible that the differences observed between taxa may simply reflect biases in cu of genes which have different levels of expression. Unfortunately there are no data available concerning the levels of expression of the genes studied here and so it is not currently possible to analyse the relationship between genes expression and cu. However the level of bias in synonymous cu was determined using the measure (B) of Long and Gillespie (1991). All of the genes included in this study had a B value of less than 0.44, showing that they represented a set which showed little bias in synonymous cu. This suggests therefore that the differences observed in cu between taxa by metric multidimensional scaling,

TABLE IV

Observed/expected dinucleotide frequency ratios in transcribed/untranslated gene sequence

Genes/introns	Dinucleotide		
	CG o/e values[c]	AT	TA
ssRNA genes[a]			
B. bovis	0.90 (0.55)	0.85 (0.90)	0.92 (0.61)
B. bigemina	0.87	0.8	0.77
B. rodhaini	0.94	0.93	0.89
Th. annulata	0.93 (0.66)	0.9 (0.81)	0.88 (0.69)
Introns[b]			
T. gondii	0.98 (0.95)	0.84 (0.87)	0.73 (0.46)
E. tenella	0.74 (0.74)	0.73 (0.75)	0.74 (0.43)

[a]The sequences analysed have accession Nos. M87566, *B. bovis*; M87565, *B. rodhaini*: X59604, *B. bigemina* and M34845, *Th. annulata* (Ellis et al., 1992). ssRNA, small subunit ribosomal RNA.
[b]The introns analysed were in the sequences TOXANT28K, TOXTUBAA, TOXTUBBA, EIMET100 and EIMANTP (accession No. M21088; Files et al., 1987) (see Table I).
[c]Values in parentheses are for coding regions (see Table II).

albeit small, probably represent true differences between taxa.

(d) Other factors affecting codon usage

It has been suggested that reading frames within coding regions have a bias for the consensus sequence RNY which may be the residue of a primitive code (Shepherd et al., 1981; Zhang et al., 1991). However, another study provided evidence against such a primitive residuum in coding sequences (Wong and Cedergren, 1986). In order to assess the application of the RNY concept to coding regions of the coccidia and piroplasms the frequency of mononucleosides at each nt position within codons was analysed (Table V). All of these species had a preference for a purine in the first position, especially *E. tenella* where there was a major preference for G. Like *P. falciparum* (Saul and Battistutta, 1988) both *B. bovis* and *Th. parva* showed a slight preference for W at the third position, where as *T. gondii* and *E. tenella* favoured G or C. The general rule for a Y at this position was however also satisfied for all four organisms. At the second base position, both the two piroplasms and *T. gondii* showed a preference for A or a Y, whereas the data from *E. tenella* implied that any base was suitable in this position. Consequently these data loosely support the RNY concept, and its application to ORFs of coccidia and piroplasms from the phylum Apicomplexa. In comparison it should be noted that the consensus for *P. falciparum* is RNW (Saul and Battistutta, 1988) and that in keeping with the observations by Zhang et al. (1991) on a variety of different organisms, all of the codons used infrequently by these piroplasms or coccidia are not of the RNY type.

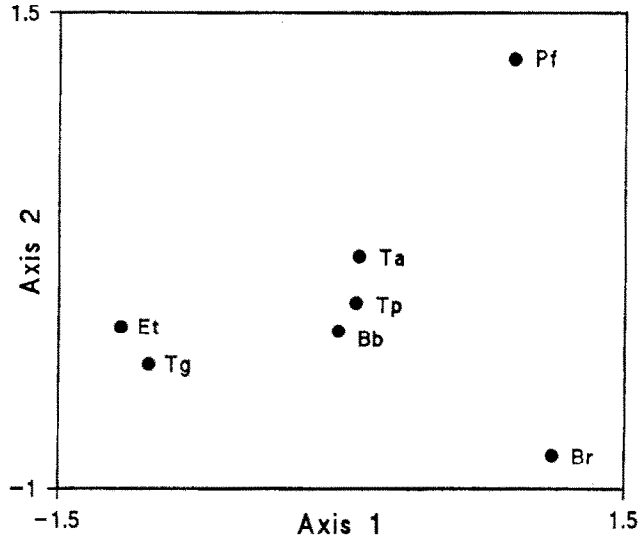

Fig. 1. Metric multidimensional scaling (Belbin, 1989) was used to investigate the pattern of codon usage among the taxa (Bb, *B. bovis*; Tp, *Th. parva*; Ta, *Th. annulata*; Tg, *T. gondii*; Et, *E. tenella*; Pf, *P. falciparum*; Br, *B. rodhaini*). A two dimensional ordination was used, with the euclidean distance measure (Faith et al., 1987). Data for *P. falciparum* were taken from Wada et al. (1991).

(e) Conclusions

(*1*) DiF and cu have been determined and compared in four species representing two subclasses of the phylum Apicomplexa. In general, cu of members of the same sub-

TABLE V

Frequency of nt at each nt position within codons

Position in codon[a]	*B. bovis*	*Th. parva*	*T. gondii*	*E. tenella*
	Frequency of nt[b]			
First				
T	19.5	19.9	16.0	15.0
C	19.0	16.9	20.3	23.5
A	29.0	31.2	25.6	17.1
G	32.5	32.0	38.1	44.4
Second				
T	25.4	26.2	26.2	22.3
C	25.0	21.9	26.0	29.5
A	38.2	37.9	28.2	23.7
G	11.4	14.1	19.6	24.5
Third				
T	33.4	27.9	21.5	20.3
C	22.5	24.7	32.7	29.7
A	25.7	25.6	17.8	19.2
G	18.4	21.8	28.0	30.8

[a]The nt position within codon.
[b]Frequencies (in %) were calculated for each position in all codons. Each column of four adds to 100%.

class were more similar to each other than to members of the other subclass.

(2) Several lu codons were identified. The identity of these lu codons varies between the two different subclasses and appears to be related to the taxonomic position of these organisms.

(3) The TA and CG dinucleotides were under represented in genes of the four organisms studied. This finding is consistent with results obtained from a wide variety of other organisms such as bacteria, yeast and primates (Zhang et al., 1991). It was reasoned that in the piroplasms coding pressure appears to determine the observed low frequency of the CG and TA dinucleotides, whereas in the coccidia other factors may be involved.

REFERENCES

Altenburger, W., Binger, M.H., Chizzonite, R.A., Kramer, R.A., Lomedico, P.T. and McAndrew, S.J.: Recombinant coccidiosis vaccines. E.P.O. 0344808 A1 (1989).

Anderson, D., McCandliss, R.J., Strausberg, S.L. and Strausberg, R.L.: Genetically engineered coccidiosis vaccine. P.C.T. WO90/00403 (1990)

Andersson, S.G.E. and Kurland, C.G.: Codon preferences in free-living micro-organisms. Microbiol. Rev. 54 (1990) 198–210.

Barta, J.R., Jenkins, M.C. and Danforth, H.D.: Evolutionary relationships of avian *Eimeria* species among other apicomplexan protozoa based on partial small subunit ribosomal RNA sequences: monophyly of the Apicomplexa is supported. Mol. Biol. Evol. 8 (1991) 345–355.

Baylis, H.A., Megson, A., Mottram, J.C. and Hall R.: Characterisation of the gene for a cysteine protease from *Theileria annulata*. Mol. Biochem. Parasitol. 54 (1992) 105–108.

Belbin, L.: PATN—Pattern Analysis Package Technical Reference. CSIRO, Canberra, 1989.

Burg, J.L., Perlman, D., Kasper, L.H., Ware, P.L. and Boothroyd J.C.: Molecular analysis of the gene encoding the major surface antigen of *Toxoplasma gondii*. J. Immunol. 141 (1988) 3584–3591.

Cesbron-Delauw, M.F., Guy, B., Torpier, G., Pierce, R.J., Lenzen, G., Cesbron, J.Y., Charif, H., Lepage, P., Darcy, F., Lecocq, J.-P. and Capron A.: Molecular characterisation of a 23-kilodalton major antigen secreted by *Toxoplasma gondii*. Proc. Natl. Acad. Sci. USA 86 (1989) 7537–7541.

Cowman, A.F., Bernard, O., Stewart, N. and Kemp, D.J.: Genes of the protozoan parasite *Babesia bovis* that rearrange to produce RNA species with different sequences. Cell 37 (1984) 653–660.

Danforth, H.D., Augustine, P.C., Ruff, M.D., McCandliss, R., Strausberg, R.L. and Likel, M.: Genetically engineered antigen confers partial protection against avian coccidial parasites. Poultry Sci. 68 (1989) 1643–1652.

Devereux, J., Haeberli, P. and Smithies, O.: A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12 (1984) 216–223.

Ellis, J., Hefford, C., Baverstock, P.R., Dalrymple, B.P. and Johnson, A.M.: Ribosomal DNA sequence comparison of *Babesia* and *Theileria*. Mol. Biochem. Parasitol. 54 (1992) 87–96.

Ellis, J. and Johnson, A.M.: Control of intracellular parasites: The Coccidia. In: Yong, W.K. (Ed.), Control of Animal Parasites Using Biotechnology, CRC Press, Boca Raton, FL, 1992, pp 241–272.

Ellis, J. and Tomley, F.: Development of a genetically engineered vaccine against poultry coccidiosis. Parasitol. Today 7 (1991) 344–346.

Enea, V., Ellis, J., Zavala, F., Arnot, D., Asavanich, A., Masuda, A., Quaki, I. and Nussenzweig, R.: DNA cloning of *Plasmodium falciparum* circumsporozoite gene: amino acid sequence of repetitive epitope. Science 225 (1984) 628–629.

Faith, D.P., Minchin, P.R. and Belbin, L.: Compositional dissimilarity as a robust measure of ecological distances: a theoretical model and computer simulations. Vegetatio 69 (1987) 57–68.

Files, J.G., Paul, L.S. and Gabe, J.D.: Identification and characterisation of the gene for a major surface antigen of *Eimeria tenella*. In: Agabian N., Goodman, H. and Noquiera, N. (Eds.), Molecular Strategies of Parasitic Invasion, U.C.L.A. Symposia on Molecular and Cellular Biology. Liss, New York, 1987, pp 713–723

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R.: Codon catalogue usage is a genome strategy modulated for gene expressivity. Nucleic Acids. Res. 9 (1981) r43—r74.

Jasmer, D.P., Reduker, D.W., Perryman, L.E. and McGuire, T.C.: A *Babesia bovis* 225 kDa protein located on the cytoplasmic side of the erythrocyte membrane has sequence similarity with a region of glycogen phosphorylase. Mol. Biochem. Parasitol. 52 (1992) 263–270.

Johnson, A.M.: Comparison of dinucleotide frequency and codon usage in *Toxoplasma* and *Plasmodium*: evolutionary implications. J. Mol. Evol. 30 (1990) 383–387.

Johnson, A.M., Fielke, R., Ellis, J., O'Donoghue, P.J. and Baverstock, P.R.: The phylogenetic relationships of the genus *Eimeria* based on comparison of partial sequences of 18S rRNA. Syst. Parasitol. 18 (1991) 1–8.

Johnson, A.M., Illana, S., McDonald, P.J. and Asai, T.: Cloning, expression, and nucleotide sequence of the gene fragment encoding an antigenic portion of the nucleoside triphosphate hydrolase of *Toxoplasma gondii*. Gene 85 (1989) 215–220.

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

Hyde, J.E. and Sims, P.F.G.: Anomalous dinucleotide frequencies in both coding and non-coding regions from the genome of the human malaria parasite *Plasmodium falciparum*. Gene 61 (1987) 177–187.

Iams, K.P., Young, J.R., Nene, V., Desai, J., Webster, P., Ole-Moiyoi, O. K. and Musoke, A.J.: Characterisation of the gene encoding a 104-kilodalton microneme-rhoptry protein of *Theileria parva*. Mol. Biochem. Parasitol. 39 (1990) 47–60.

Levine, N.D.: Phylum II. Apicomplexa Levine, 1970. In: Lee, J.J., Hunter, S.H. and Bovee, E.C. (Eds.), Illustrated Guide to the Protozoa, Society of Protozoologists, Kansas, USA, 1985, pp. 322–374.

Liberator, P.A., Hsu, J. and Turner, M.J.: Tandem trinucleotide repeats throughout the nucleotide sequence of a cDNA encoding an *Eimeria tenella* sporozoite antigen. Nucleic Acids. Res. 17 (1989a) 7104.

Liberator, P.A., Nollstadt, K.H., Turner, M.J., Crane, M.S.J., Karkhanis, Y.D., Chakraborty, P.R. and Profous-Juchelka, H.: Recombinant and native group B *Eimeria tenella* immunogens useful as coccidiosis vaccines. E.P.O. 0337589 A3 (1989b)

Long, M. and Gillespie J.H.: Codon usage divergence of homologous vertebrate genes and codon usage clock. J. Mol. Evol. 32 (1991) 6–15.

Mason, P.J., Shields, B.R., Tait, A., Beck, P. and Hall, R.: Sequence and expression of a gene from *Theileria annulata* coding for a 70-kilodalton heat-shock protein. Mol. Biochem. Parasitol. 37 (1989) 27–35.

Nagel, S.D. and Boothroyd, J.C.: The alpha- and beta- tubulins of *Toxoplasma gondii* are encoded by single copy genes containing multiple introns. Mol. Biochem. Parasitol. 29 (1988) 261–273.

Nene, V., Gobright, E. and Musoke, A.J. and Lonsdale-Eccles, J.D.: A single exon codes for the enzyme domain of a protozoan cysteine protease. J. Biol. Chem. 265 (1990) 18047–18050.

Newman, K.Z., Gore, T.C., Tedesco, J.L., Petersen, G.R., Brothers,

V.M., Files, J.G., Paul, L.S., Chang, R.J., Andrews, W.H., Kuhn, I, McCaman, M., Sias, S.R., Nordgren, R.M. and Dragon, E.A.: Antigenic proteins and vaccines containing them for prevention of coccidiosis. E.P.O. 0231537 A2. (1987)

Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A.: Recent evidence for evolution of the genetic code. Microbiol. Rev. 56 (1992) 229–264.

Ossorio, P.N., Schwartzman, J.D. and Boothroyd, J.C.: Rhoptry protein associated with host cell penetration has unusual charge symmetry. Mol. Biochem. Parasitol. 50 (1992) 1–15.

Phillips, G.J., Arnold, J. and Ivarie, R.: The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over and under presented sequences by Markov chain analysis. Nucleic Acids. Res. 15 (1987) 2627–2638.

Pollack, Y., Kogan, N. and Golenser, J.: *Plasmodium falciparum* evidence for a DNA methylation pattern. Exp. Parasitol. 72 (1991) 339–344.

Prince, J.B., Arujo, F.G., Remington, J.S., Burg, J.L., Boothroyd, J.C. and Sharma, S.D.: Cloning of cDNAs encoding a 28 kilodalton antigen of *Toxoplasma gondii*. Mol. Biochem. Parasitol. 34 (1989) 3–14.

Prince, J.B., Auer, K.L., Huskinson, J., Parmley, S.F., Arujo, F.G. and Remington, J.S.: Cloning, expression, and cDNA sequence of the surface antigen P22 from *Toxoplasma gondii*. Mol. Biochem. Parasitol. 43 (1990) 97–106.

Saul, A. and Battistutta, D.: Codon usage in *Plasmodium falciparum*. Mol. Biochem. Parasitol. 27 (1988) 35–42.

Schofield, L.: The circumsporozoite protein of *Plasmodium*: a mechanism of immune evasion by the malaria parasite? Bull. World Health Organ. 68 Suppl. (1990) 66–73.

Shepherd, J.C.W. : Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA 78 (1981) 1596–1600.

Shields, D.C. and Sharp, P.M.: Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. 15 (1987) 8023–8040.

Snary, D. and Smith, M.A.: Sequence homology of surface membrane proteins of *Babesia rodhaini*. Mol. Biochem. Parasitol. 27 (1988) 303–312.

Staden, R.: The current status and portability of our sequence handling software. Nucleic Acids Res. 14 (1986) 217–231.

Suarez, C.E., Palmer, G.H., Jasmer, D.P., Hines, S.A., Perryman, L.E. and McElwain, T.F.: Characterisation of the gene encoding a 60-kilodalton *Babesia bovis* merozoite protein with conserved and surface exposed epitopes. Mol. Biochem. Parasitol. 46 (1991) 45–52.

Tenter, A.M., Baverstock, P.R. and Johnson, A.M.: Phylogenetic relationships of *Sarcocystis* species based on ribosomal RNA sequence comparison. Int. J. Parasitol. 22 (1992) 503–514.

Tomley, F.M., Clarke, L., Kawazoe, U., Dijkema, R. and Kok, J.J.: Sequence of the gene encoding an immunodominant microneme protein of *Eimeria tenella*. Mol. Biochem. Parasitol. 49 (1991) 277–288.

Wada, K., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T. and Ikemura, T.: Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. 19 (1991) 1981–1986.

Weber, J.L. : Analysis of sequences from the extremely A+T-rich genome of *Plasmodium falciparum*. Gene 52 (1987) 103–109.

Wong, J. and Cedergren R.: Natural selection versus primitive gene structure as determinant of codon usage. Eur. J. Biochem. 159 (1986) 175–180.

Zhang, S., Zubay, G. and Goldman, E.: Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. Gene 105 (1991) 61–72.