Europ. J. Protistol. 33, 266–273 (1997)
August 29, 1997

European Journal of
PROTISTOLOGY

Codon Usage Patterns in *Trichomonas vaginalis*

James O. McInerney

Department of Zoology, The Natural History Museum, London, UK

Abstract

Codon usage variation was studied in the parabasilid protozoan *Trichomonas vaginalis*. Correspondence analysis was employed to identify the single largest source of variation in the dataset. The G+C content at the third position of synonymously variable codons was also evaluated. A strong positive correlation was observed between the effective number of codons measure and the position of the genes on the axis of greatest dispersion following correspondence analysis. The correlation between base composition at the third position and effective number of codons was weaker. The results indicate that the major influence on codon usage bias is translational selection, with a moderate effect attributable to mutational pressure. In addition, twenty codons were identified whose use in biased genes is statistically significantly greater than in unbiased genes.

Key words: Codon usage bias; Correspondence analysis; Effective number of codons; Mutational bias; Translational selection

Introduction

The genetic code is composed of triplets of nucleotides that encode the primary information required for production of functional proteins. Of a total of 64 possible combinations of the four bases, 5 are non-degenerate (three encode termination codons, one encodes methionine and one encodes tryptophan). This leaves 59 codons, each of which has an alternative synonym (another codon that encodes the same amino acid). The frequency with which the synonymously variable codons are used, varies widely between genes. In almost every organism in which codon usage has been studied, at least some genes have been shown to possess codon usage patterns that deviate from random expectation [23]. It has been noted that virtually every single codon has been used as a major codon by some organisms and as a minor codon by others, thus underlying the importance of determining codon usage bias in every organism [1]. Mutational pressure (governing the G+C content of the genome or portion of the genome) and translational selection are the most important factors known to influence the selection of synonymous codons [23].

During studies on the abundance of tRNA species in *Saccharomyces cerevisiae* and *Escherichia coli* Ikemura [10, 11] deduced that the preferred codons in highly expressed genes correlated very strongly with the abundance of iso-accepting tRNA. This suggests a mechanism whereby selection of a small subset of codons can greatly expedite the production of abundantly-required proteins. The selective difference between using an 'optimal' codon in a gene sequence and using a non-optimal codon is thought to be very small. In highly expressed genes, the advantage of using an optimal codon is greater than in lowly expressed genes. In organisms with a large long-term effective population size, small selective differences can overcome the effects of random genetic drift. In this situation, highly expressed genes tend to use a small subset of codons. It is thought that the cognate tRNA species are also to be found in abundance in those cells (although evidence for this is still only based on a few model organisms) [10, 11]. In the genes that are poorly expressed, the selective advantage of using these 'optimal' codons is not great enough, and so, these genes tend to have a less biased codon usage pattern and are influenced to a greater extent by mutational pressures. Some organisms, such as *Mycoplasma capricolum* and *Micrococcus luteus* are so strongly influenced by mutational pressure, that identification of optimal codons for highly expressed genes is very difficult [23]. In fact it is not clear if there is any translational selection effect on the highly expressed genes.

In contrast to the prokaryotic situation, in larger multicellular organisms, with smaller effective population sizes, the selective advantage of using optimal subsets of codons does not have the same effect. In these organisms, the small selective advantage offered by this

mechanism is not capable of overcoming genetic drift [2]. Warmblooded vertebrates possess "isochores" of DNA with similar base compositions (called neogenome and paleo-genome by Bernardi [3]). Consequently, genes found to reside in G+C-rich isochores possess codons that end predominantly in either guanine or cytosine, and conversely in A+T-rich isochores the codons tend to end in adenine or thymidine [3]. In the Atlantic salmon also, codon usage bias is also largely governed by base composition of the genes [16]. This is not to say that these situations are characteristic of all multicelluar organisms. Codon usage in the fruit fly *Drosophila melanogaster* an the nematode *Caenorhabditis elegans* are governed by the same rules as are seen for *E. coli* and *S. cerevisiae* [25].

For practical purposes there are a number of reasons why determining the codon usage pattern in an organism is important. The polymerase chain reaction (PCR) [18] has become an invaluable tool with which to manipulate DNA. It is frequently desirable to be able to back-translate a protein sequence to its corresponding DNA sequence without prior knowledge of the DNA sequence. This necessitates the introduction of nucleotide positions of degenerate composition. If the codon usage pattern of the organism is known, then some *a priori* assumptions may be made concerning the identity of the synonymously variable positions. Consequently, the degeneracy of the oligonucleotide primers may be reduced and also the chance of mis-priming. Codon usage tables have been used to predict the likelihood that an open reading frame encodes a protein [4]. They have also been used to identify those that do not encode proteins [14]. Furthermore, codon usage patterns have been used to examine evolutionary relationships among organisms [17]. Once the pattern of codon usage has been determined in an organism, certain predictions may be made for newly isolated genes. For example, in prokaryotes and yeast, the level of codon usage bias may indicate the level of expression of the gene.

The preferred method of identifying the major trends governing the use of synonymous codons is correspondence analysis [7, 12, 14, 20, 21]. This method identifies the most important factors of dispersion, i.e. the major trends, in the data. A series of orthogonal axes in a 59-dimensional space are created (each axis corresponds to one of the codons for which there is a synonymously degenerate alternative). The axis that represents the greatest amount of variation is identified and then subsequent axes in diminishing order. Correlations may then be made with the trends along the first and consequent axes. The distance between any two genes on these axes is a measure of their dissimilarity with respect to these trends.

Other indices that are valuable when analysing the pattern of codon usage include the effective number of codons (Nc) [26] and the G+C composition at the third position of synonymous codons (GC3s), i.e. excluding the Trp, Met and termination codons. The Nc value of a gene reflects the deviation from random usage of all codons. The value ranges from 61 (for a gene without a preference for any subset of codons) to 20 (for a gene with an extreme codon usage bias, where only one codon is used for each amino acid).

T. vaginalis is a member of the order Trichomonadaceae within the Class Parabasalia of the Phylum Zoomastigina. The Trichomonads are flagellates, generally found in parasitic or endosymbiotic associations in oxygen-poor environments. *T. vaginalis* is widespread and is well studied. It is a common commensal inhabitant of human genital tracts and is transferred by sexual intercourse. It is estimated that this organism is carried, without any pathogenicity, by 3.5% of people worldwide. The parasite can, on occasion however, become invasive and is a confirmed pathogen. It is one of the few Parabasalians that can be cultured. On phylogenetic trees based on small and large subunit ribosomal RNAs, this organism is normally placed close to the base of the eukaryotic tree [19].

The cellular organisation of *T. vaginalis* is substantially different from the student textbook definition of the eukaryotic cell. Trichomonads do not possess mitochondria and their metabolism is anaerobically fermentative. *T. vaginalis* does, however, accommodate a hydrogenosome. Within this organelle, pyruvate or malate are converted to acetate, CO_2 and H_2. ATP is produced in this pathway via substrate level phosphorylation. A number of genes contributing to processes within the hydrogenosome have been sequenced in *T. vaginalis*. One such gene is pyruvate: ferredoxin oxidoreductase [9]. Although there are no published phylogenetic trees inferred from this protein, Hrdý and Müller concluded that its "[...] sequence is quite similar to its eubacterial homologues". Also, the glyceraldehyde-3-phosphate dehydrogenase gene from this species has been sequenced and both least squares [5] and parsimony phylogenetic trees derived from amino acid identities place the sequences within the Bacteria [15]. The position of this gene indicates that it does not share recentness of common ancestry with other eukaryotic homologues. In addition, the hydrogenosome can be said to be carrying out a function in *T. vaginalis* that is analogous to mitochondria in crown-group eukaryotes.

The purposes of this study were twofold. Firstly, codon usage bias was investigated for the purposes of identifying the preferred codons and the major influences on codon usage bias. To this end, correspondence analysis was employed and the single most important trend of codon usage bias was identified. In addition, given the hypothesis that the hydrogenosomal origin is independent of the rest of the cell, it was investigated

whether genes that are implicated in hydrogenosome function have codon usage patterns that differ from other *T. vaginalis* genes.

Methods

All sequences used in this study were retrieved from GenBank (release 92) using the ACNUC sequence retrieval system [6]. The features table was used to identify coding regions. Potential duplicate entries were aligned using CLUSTALW 1.6 and analysed using Li's 1993 method for divergence at synonymously degenerate sites. Sequences were used in the study if the divergence at these sites exceeded 0.2. It was decided that at this level of difference, there would be room for changes in codon usage patterns. If divergence was less than 0.2, only one copy was retained. Codon usage was calculated for each of the genes in this study by the CODONS program [3] and other software written by the Irish National Bioinformatics Centre (INCBI). The frequency with which each codon is used was calculated. However, because these values may vary greatly between codons (depending on the amino acid composition of the protein sequence), it was decided that Sharp's Relative-Synonymous Codon Usage (RSCU) measure would be used in all analyses [22]. RSCU values were computed by dividing the observed frequency of a codon by the expected frequency if all synonyms for the corresponding amino acid were used equally. An RSCU value that is close to 1.0 indicates that this codon is being used at the expected frequency. Values that are higher indicate that this codon is used more frequently than the expectation. The G+C base composition of 'silent' third positions were calculated and are expressed as the GC3s value. The effective number of codons index was also calculated using CODONS (for detailed claculations of this measure see [26]).

Multivariate statistical analysis was carried out using a modified version of the DECORANA program written in FORTRAN by M. O. Hill (see [14]). The chosen method was correspondence analysis [8], which has been widely used in studies of codon usage patterns [7, 24, 25]. The axis that accounted for the most variation was identified and the four subsequent axes. The eigenvectors of these four axes were analysed to identify the amount of the total variation that is being accounted for by each axis.

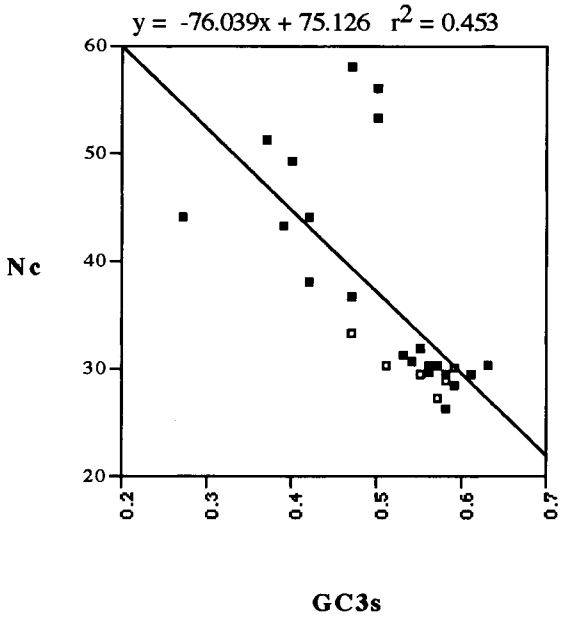
Table 1. The genes used in this study listed in order of their appearance on axis 1 following correspondence analysis. The first column lists the genes according to their CDS name in the ACNUC databases. The second column contains the gene name. The third column details the position of the gene along axis 1 following correspondence analysis. The gene length in amino acids (Laa) is in the next column. The G+C content of the genes calculated from all positions in the gene is in the next column. The G+C content at third positions of codons that are synonymously degenerate is in the next column, these positions may change without affecting the encoded amino acid. The last column contains the value calculated for the effective number of codons used by the gene.

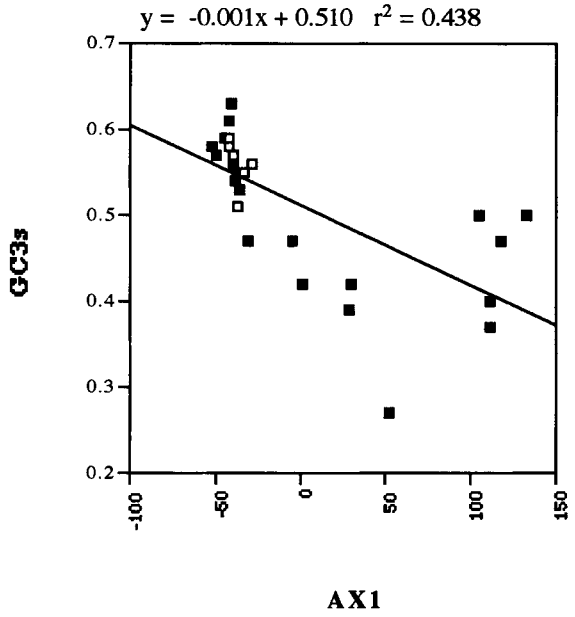
MNEMONIC	Gene name	AX1	Laa	GC	GC3S	N_C
TRITUBA. BTUB1	Beta-tubulin	–52	446	0.53	0.58	26.3
TRIGPDH	Glyceraldehyde-3-P-Dehyd mRNA	–50	350	0.52	0.57	27.3
TVU18347.AP65-2	AP65-2 adhesin	–45	567	0.53	0.59	28.4
TVCP1.CP1	Cysteine proteinase CP1 mRNA	–43	309	0.52	0.61	29.5
TVU16822.PFOA	Pyruvate-ferroxid-red	–43	1157	0.52	0.58	28.8
TRIALPHB.PE1	Succinyl CoA subunit B gene	–42	309	0.54	0.58	29.4
TVU16838.MAEC	Hydrogenosomal malic C	–42	228	0.54	0.59	30.0
TVCP2.CP2	Cysteine proteinase CP2 mRNA	–41	314	0.53	0.63	30.2
TVU07203.PE1	Hydrogenosomal adenyl. kinase	–40	229	0.51	0.56	30.2
TVU38692.PE1	Cytosolic malate dehyd.	–40	332	0.54	0.57	30.2
TRISUCBETA.PE1	Succinyl CoA Synth. beta gene	–39	407	0.52	0.54	30.7
TRIFERRA.PE1	Ferredoxin gene	–37	100	0.52	0.55	29.4
TVU16837.MAEB	Hydrogenosomal malic B	–37	567	0.51	0.51	30.3
TVU18346.AP65-1	Adhesin mRNA	–36	567	0.51	0.53	31.3
TVCP3.CP3	Cysteine proteinase CP3 mRNA	–34	278	0.49	0.55	31.8
TRIALPHA.PE1	Succinyl CoA subunit A gene	–31	309	0.51	0.47	33.2
TVU16839.MAED	Hydrogenosomal Malic D	–29	73	0.53	0.56	29.7
TRIVALS	valyl tRNA synthase gene	–5	489	0.47	0.47	36.7
TVU35878.NANA	N-acetylneuraminate lyase	1	318	0.42	0.42	38.1
TVPGP1	p-glycoprotein-like gene	29	589	0.41	0.39	43.3
TVCYSP.PE1	putative cyst. protease mRNA	30	223	0.42	0.42	44.0
TRISIMM	Surface immunogen mRNA	53	132	0.42	0.27	44.0
TVSAP.PE1	Secreted adhesive protein mRNA	105	426	0.52	0.50	56.0
TRITUBB	Beta tubulin gene	111	141	0.52	0.40	49.2
TRITUBC	Beta tubulin gene	111	141	0.43	0.37	51.3
TRISIMMA	Surface immunogen gene	117	110	0.42	0.47	58.1
TVCP4.CP4	Cysteine protease CP4 mRNA	132	290	0.53	0.50	53.2

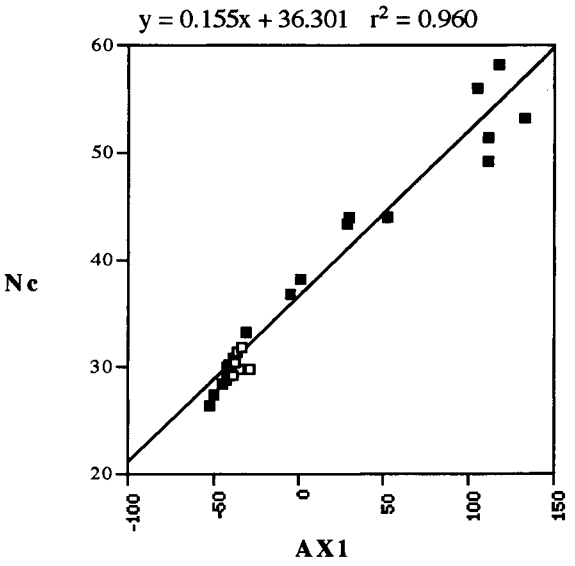
Results

The database search and screening procedure identified 27 genes from *T. vaginalis*. These genes are listed in Table 1 according to their relative positions on the axis of greatest dispersion. Table 1 also gives the G+C content of each of the genes as a whole and also the G+C content at synonymously variable third positions of codons. The column N_C on this Table refers to the 'effective number of codons' [26].

The base composition of the genes, considering all positions ranged from 41% G+C for the pgp1 gene to 54% for succinyl CoA subunit B, hydrogenosomal malic subunit C and cytosolic malate dehydrogenase (ΔGC: 13%). The range of base compositions at the synonymously degenerate third positions is much wider. The sequence with the highest GC3s value (61%) is the cysteine proteinase 1 gene and the lowest value (27%) is seen in the surface immunogen (ΔGC: 34%). The GC3s and N_C values were plotted against one another to evaluate the effect of base composition on codon usage. The results of this analysis are shown in Figure 1 (a). The beta tubulin gene uses the least number of codons (N_C: 26.3) and the surface immunogen gene (TRISIMMA) uses all codons with almost equal frequency (N_C: 58.1). A linear regression curve







1a | 1b
1c

Fig. 1a. Plot of Nc versus GC3s. A linear regression line was computed for the dataset. The formula of the line is given and its regression co-efficient.

b. Plot of GC3s values versus those found on axis 1 following correspondence analysis. The formula for the linear regression line is given along with the regression co-efficient.

c. Plot of Nc values versus those found along axis 1 following correspondence analysis. The formula for the linear regression line is given along with its regression co-efficient.

was calculated from the data and plotted onto the graph. The regression coefficient (r^2) of this plot is 0.453. Although this regression fit is not particularly high, most of the deviation from the line is due to four genes. The TRISIMM gene has a GC3s base composition that is substantially lower than the others, whilst its N_C value is moderate. Three genes with moderate GC3s values have N_C values close to random expectation. With the exception of these genes, visual inspection of the plot indicates that there is a definite trend of increasing bias (decrease in the effective number of codons) with base composition.

Six genes are indicated in Figure 1a by open boxes. These genes are known to be involved in hydrogenosome function and include the ferredoxin gene, the hydrogenosomal adenylate kinase gene, the pyruvate: ferredoxin oxidoreductase gene and hydrogenosomal malic enzyme subunits C, B and D. These genes are seen to be interspersed among the more biased genes. All of these enzymes appear to be highly biased with respect to the number of codons being used by each of the genes (as evidenced by low N_C values). The positions of these genes on Axis 1 following correspondence analysis also identifies the pattern of codon usage for these genes to be typical of the rest of the dataset.

Correspondence analysis was utilised to identify the variation in codon usage among the genes. Four axes from the total of 59 were examined, the variation in

Table 2. Codon usage table calculated from the five most biased and the five least biased genes. The first column indicates the codon as designated by its three letter abbreviation. The second column indicates the decoding triplet for 32 codons. The third column (N) refers to the number of codons used in the calculation for the highly biased dataset. The fourth column is the corresponding relative synonymous codon usage value. The fifth and sixth columns contain the same information for the lowly biased dataset. Columns seven to twelve contain the same information for the next 32 codons.

Codon		N	RSCU	N	RSCU	Codon		N	RSCU	N	RSCU
Phe	UUU	1	0.02	14	0.82	Ser	UCU*	34	1.40	23	0.85
	UUC*	105	1.98	20	1.18		UCC*	89	3.66	21	0.77
Leu	UUA	4	0.10	12	0.54		UCA	21	0.86	52	1.91
	UUG	0	0.00	26	1.17		UCG	0	0.00	26	0.96
Leu	CUU*	82	2.11	23	1.04	Pro	CCU	1	0.03	22	1.09
	CUC*	147	3.79	19	0.86		CCC	0	0.00	13	0.64
	CUA	0	0.00	17	0.77		CCA*	140	3.97	20	1.48
	CUG	0	0.00	36	1.62		CCG	0	0.00	16	0.79
Ile	AUU	13	0.24	10	0.68	Thr	ACU	6	0.14	26	1.11
	AUC*	150	2.74	18	1.23		ACC	6	0.14	25	1.06
	AUA	1	0.02	16	1.09		ACA*	160	3.72	27	1.15
Met	AUG	69	1.00	22	1.00		ACG	0	0.00	16	0.68
Val	GUU*	78	1.67	4	0.43	Ala	GCU*	190	3.10	33	1.74
	GUC*	108	2.31	7	0.76		GCC	45	0.73	8	0.42
	GUA	0	0.00	11	1.19		GCA	10	0.16	21	1.11
	GUG	1	0.02	15	1.62		GCG	0	0.00	14	0.74
Tyr	UAU	11	0.20	13	0.96	Cys	UGU	0	0.00	18	1.38
	UAC*	99	1.80	14	1.04		UGC*	66	2.00	8	0.62
ter	UAA	5	0.00	12	0.00	ter	UGA	0	0.00	20	0.00
ter	UAG	0	0.00	4	0.00	Trp	UGG	37	1.00	21	1.00
His	CAU	5	0.16	32	1.14	Arg	CGU*	50	2.61	7	0.38
	CAC*	59	1.84	24	0.86		CGC*	55	2.87	11	0.59
Gln	CAA	20	0.38	32	0.98		CGA	0	0.00	21	1.12
	CAG*	86	1.62	33	1.02		CGG	0	0.00	11	0.59
Asn	AAU	6	0.10	11	0.85	Ser	AGU	1	0.04	21	0.77
	AAC*	120	1.90	15	1.15		AGC	1	0.04	20	0.74
Lys	AAA	3	0.03	12	0.65	Arg	AGA	8	0.42	30	1.61
	AAG*	196	1.97	25	1.35		AGG	2	0.10	32	1.71
Asp	GAU	89	1.24	6	0.92	Gly	GGU	77	1.41	9	1.09
	GAC	54	0.76	7	1.08		GGC*	129	2.36	10	1.21
Glu	GAA	107	1.18	5	1.25		GGA	13	0.24	8	0.97
	GAG	72	0.82	3	0.75		GGG	0	0.00	6	0.73

subsequent axes was expected to be too small. The first axis, i.e. the major factor of dispersion of the dataset, accounted for 39.3% of the variation in the dataset. This is a high percentage of the variation and compares with 44% seen for *Kluyveromyces lactis* [14] and 36% for *Aspergillus nidulans* [12]. The three subsequent axes only account for 12.75%, 8.6% and 7.9%, respectively. The implication is that a single factor is responsible for shaping the codon usage bias of the organism.

G+C base composition at synonymously degenerate third positions was plotted against the values obtained by correspondence analysis along axis 1. The results of this analysis are seen in Figure 1b. Again the r^2 value of the linear regression line is quite low (0.438). On this occasion, however, the low regression coefficient does not appear to be due to a small number of points that do not conform with the expectation. Rather, the fit generally is not particularly good. When N_C values were plotted against the values from axis 1 (Fig. 1c), the correlation was extremely strong. A linear regression line was fitted to the data and the r^2 value was 0.960. This indicates that correspondence analysis has identified the factor leading to bias in codon usage and this factor is not particularly well correlated with base composition.

Codon usage Table

Codon usage can be measured in terms of the number of times that a particular codon is used. This measurement can appear to be biased, however, if certain amino acids are used more frequently than others. It is preferable to represent the data in terms of relative synonymous codon usage (RSCU) [22]. This measures the relative number of times a particular codon is used with respect to how often it would be used in the absence of codon usage bias. The expectation for each codon is that its RSCU value is close to 1.00. Codon usage scores in excess of this value indicate that it is being used more frequently than expected and lower values have the opposite meaning. This has the effect of normalising the codon usage values. The RSCU values were used to calculate the correspondence analysis factors of dispersion. Five genes were taken from each end of the axis 1 and their RSCU values were compared. For each codon, the RSCU values were compared and a chi squared test with 41 degrees of freedom. The values were scaled by division by the number of codons in the gene, excluding Trp and Met. This follows the procedure of Shields and Sharp (1987) [24]. By carrying out such an analysis, codons can be identified whose usage is statistically different in different genes. The procedure employed here firstly necessitated the summation of the codon usage values for each of the two sets of five genes. This was followed by conversion of these scores to RSCU values. The chi-squared test was performed and Table 2 shows the results of this analysis.

Visual inspection of the table clearly reveals that the highly biased genes are extremely selective in their choice of codons. An example of the difference in codon usage between the two datasets can be seen with phenylalanine. This amino acid is used 106 times in the highly biased dataset, the UUC codon is used 105 times whilst the UUU codon is only used once. This contrasts with the lowly biased dataset. The phenylalanine amino acid appears 34 times in these five genes. Here, however, the ratio of usage of the UUC codon compared to UUU is approximately 3:2. This trend continues throughout the dataset for nearly all amino acids. A total of 20 codons were identified from the dataset that were used at a statistically greater frequency in the highly biased genes than in the genes of more moderate bias and these are denoted by asterisk.

Discussion

The extreme differences observed in base composition of synonymously degenerate third positions of the genes in this dataset might suggest that mutational pressures are having a strong effect on the choice of codon and this is true to a certain extent. The highly biased genes generally tend to be more GC-rich at the synonymously degenerate third positions. For five amino acids, however, there is evidence that a U-ending codon is used significantly more often in highly biased than in lowly biased genes. The preference for these 'U'-ending codons may be accounted for by the 'wobble' hypothesis, were a single tRNA species is recognising both of these codons. It is significant also in this regard that in three of these instances (Ser, Leu and Arg) the amino acid is six-fold redundant and in the other two cases (Val and Ala), the amino acid is four-fold redundant. For four of these amino acids, the C-ending codon is a preferred codon also. For Ala however, only the U-ending codon is used significantly more often in the more biased genes. It is interesting to note that the U-ending codon for this amino acid is used more often than the expectation even in the lowly biased dataset, an indication that its cognate tRNA is the most abundant isoaccepting tRNA species for this particular amino acid. The U-ending codons for amino acids that are encoded by only two or three codons do not seem to be "preferred" in any instance in the highly biased dataset.

There have been speculations concerning the influence of the middle nucleotide of a triplet in codon selection. The hypothesis concerns the roles played by bonding during codon-anticodon interactions. In this model, an ideal codon contains a mixture of weak-bonding (Adenine or Thymidine) and strong-bonding

(Guanine or Cytosine) nucleotides with a particular order (e.g. Strong-Weak-Strong) within the triplet. In this dataset, there does not appear to be a very strong trend towards a particular nucleotide combination.

Whilst there is little data concerning the level of expression of the genes from *T. vaginalis*, there are some genes at the biased end of axis 1 whose expression levels are expected to be high. These include the glyceraldehyde-3-phosphate dehydrogenase gene, pyruvate:ferredoxin oxidoreductase and succinyl CoA synthase subunit B. These genes are essential for metabolic processes and are generally produced in large quantities. In this dataset there is an observable trend seen in the relationships between GC3s, the axis of greatest dispersion (axis 1) and the effective number of codons (N_c). However, the strongest relationship (as judged by the regression coefficient from the plot in Figure 1c is seen between N_c and axis 1. The decay in the relationship of the other two plots suggest that there is another force acting on codon usage that is independent of base composition. The role of translational selection in shaping codon usage patterns is well documented (see [23] for review). In the absence of RNA expression data and protein abundance information, it may only be speculated that translational selection is the 'other' force.

There are a number of genes in this dataset that are known to be involved in hydrogenosome function. These include adenylate kinase, ferredoxin, hydrogenosomal malic acid enzymes B, C and D and pyruvate:ferredoxin oxidreductase. In Figure 1a, b and c these genes are represented by open boxes. A cursory examination of codon usage in these genes indicates that they are highly biased (as evidenced by the low N_c value). In addition, the GC3s values for each of these genes is quite high (>50%), another statistic they have in common with the other *T. vaginalis* genes. These statistics alone are not conclusive evidence of their conformity to 'typical' *T. vaginalis* codon usage. Although this might seem to indicate that the hydrogenosome-associated proteins may have similar codon usage with the non-hydrogenosome-associated genes this apparent similarity could be due to any number of convergences. The situation is clarified however by correspondence analysis. On the axis of greatest dispersion (and subsequent axes: personal observation) the hydrogenosome-associated genes are interspersed with the other genes. This indicates that the pattern of codon usage is similar and indicates that the same forces are influencing codon selection in both types of genes.

If the hydrogenosome-associated genes do not share a common history with the more typically eukaryotic genes and were acquired from some mitochondrion-like symbiont, then two scenarios may be envisioned. The first suggests that the codon usage of this symbiont was identical to that of its host and the genes have co-evolved and maintained the same codon usage. The second, and more likely situation is that a sufficient length of time has expired since the symbiotic event for these genes to have adopted the same patterns as *T. vaginalis* and also these genes are under the same influences as the native genes. In light of the fact that these genes are using the host translation machinery and also the 'genome hypothesis' of Grantham and co-workers [7] of an organism-specific codon usage choice, the co-evolution theory is not very realistic.

In conclusion, the codon usage patterns of *T. vaginalis* have been identified. The relatively small dataset used in this analysis is probably quite an accurate representation of the overall pattern, as indicated by the high regression in Figure 1c. Codons whose use is obviously favourable were identified and a hypothesis concerning the evolution of codon usage in hydrogenosomally linked proteins was constructed. Also, the major influences on codon usage bias were discussed.

Acknowledgements: I am indebted to Andrew T. Lloyd, Irish National Centre for Bioinformatics for provision of and advice on codon usage statistics and programs. I would also like to thank T. Martin Embley, D. Horner and P. Dyal for their advice. The author would also like to thank two anonymous referees for their comments.

References

1 Andersson S. G. E. and Kurland C. G. (1990): Codon preferences in free-living microorganisms. Microbiol. Rev. 54, 198–210.
2 Aota S.-I. and Ikemura T. (1986): Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nuc. Acids Res. 14, 6345–6355.
3 Bernardi G. (1993): The isochore organisation of the human genome and its evolutionary history – a review. Gene 135, 57–66.
4 Fickett J. W. (1982): Recognition of protein coding regions in DNA sequences. Nuc. Acids Res., 10, 5303–5318.
5 Fitch W. M. and Margoliash E. (1967): Construction of phylogenetic trees. Science 155, 279–284.
6 Gouy M., Gautier C., Attimonelli M., Lanave C., and Di Paola G. (1985): ACNUC-a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. CABIOS 1, 167–172.
7 Grantham R., Gautier C., Gouy M., Jacobzone M., and Mercier R. (1981): Codon catalog usage is a genome strategy modulated for gene expressivity. Nuc. Acids Res. 9, r43–r74.
8 Greenacre M. J. (1984): Theory and applications of correspondence analysis. Academic Press, London.
9 Hrdý I. and Müller M. (1995): Primary structure and eubacterial relationships of the pyruvate:ferredoxin oxidoreductase of the amitochondriate eukaryote *Trichomonas vaginalis*. J. Mol. Evol. 40, 388–396.
10 Ikemura T. (1981): Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurance of the

respective codons in its protein sequence: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. 151, 389–409.
11 Ikemura T. (1982): Correlation between the abundance of yeast transfer RNAs and the occurence of the respective codons in protein genes. J. Mol. Biol. 158, 573–597.
12 Lloyd A. T. and Sharp P. M. (1991): Codon usage in *Aspergillus nidulans*. Mol. Gen. Genet. 230, 288–294.
13 Lloyd A. T. and Sharp P. M. (1992): CODONS: a microcomputer program for codon usage analysis. J. Hered. 83, 239–240.
14 Lloyd A. T. and Sharp P. M. (1993): Synonymous codon usage in *Kluyveromyces lactis*. Yeast 9, 1219–1228.
15 Markoš A., Miretsky A., and Müller M. (1993): A glyceraldehyde-3-phosphate dehydrogenase with eubacterial features in the amitochondriate eukaryote, *Trichomonas vaginalis*. J. Mol. Evol. 37, 631–643.
16 McInerney J. O. (1996): Codon usage patterns in the Atlantic Salmon (*Salmo salar L.*). Mol. Mar. Biol. Biotechnol. 5, 344–351.
17 Nesti C., Poli G., Chica M., Ambrosino P., Scapoli C., and Barrai I. (1995): Phylogeny inferred from codon usage pattern in 31 organisms. CABIOS 11, 167–171.
18 Saiki R. K., Gelfand D. H., Stoffel S., Scharf S. F., Higuchi R., Horn G. T., Mullis K. B., and Erlich H. A. (1988): Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487–491.
19 Schlegel M. (1994): Molecular phylogeny of eukaryotes. Trends in Ecology and Evolution 9, 330–335.
20 Sharp P. M. (1989) Evolution at 'silent' sites in DNA. Evolution and animal breeding. In: Hill W. G. and Mackay T. F. C. (eds.), pp. 23–32. Wallingford, UK, CAB International.
21 Sharp P. M. and Devine K. M. (1989): Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. Nuc. Acids Res. 17, 5029–5039.
22 Sharp P. M. and Li W.-H. (1986): An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.
23 Sharp P. M., Stenico M., Peden J. F., and Lloyd A. T. (1993): Codon usage: mutational bias, translational selection or both? Biochem. Soc. Trans. 21, 835–841.
24 Shields D. C. and Sharp P. M. (1987): Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nuc. Acids Res. 15, 8023–8040.
25 Stenico M., Lloyd A. T., and Sharp P. M. (1994): Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nuc. Acids Res. 22, 2437–2446.
26 Wright F. (1990): The 'effective number of codons' used in a gene. Gene 87, 23–29.

Address for correspondence: J. O. McInerney, Department of Zoology, The Natural History Museum, London, UK.
E-mail: j.mcinerney@nhm.ac.uk

Note added in proof:

It has been pointed out to me that the sequences of some of the outliers on the correspondence analysis graphs might be incorrect (M. Müller, personal communication). This does not affect the overall conclusions from this paper. Also, while this manuscript was in press, a phylogenetic analysis of the pyruvate: ferredoxin oxidoreductase was published (Kletzin and Adams 1996).

Kletzin A. and Adams M. W. W. (1996): Molecular and phylogenetic characterization of pyruvate and 2-ketoisovalerate ferredoxin oxidoreductases from *pyrococcus furiosus* and pyruvate ferredoxin oxidoreductase from *Thermotoga maritima*. J. Bacteriol. 178, 248–257.