Schistosoma mansoni: patterns of codon usage and bias

J. T. ELLIS *and* D. A. MORRISON

School of Biological and Biomedical Sciences, University of Technology Sydney, Westborne Street, Gore Hill, NSW 2065, Australia

(*Received 16 March 1994; accepted 12 May 1994*)

SUMMARY

Codon usage and bias has been examined in 20 genes of *Schistosoma mansoni*. Significant heterogeneity was detected in the patterns of codon usage and bias among genes by metric multidimensional scaling and three general indictors of bias (GC_{3S}, N_c and B). In keeping with observations on sporozoan parasites, 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 over or under-represented in the sequences.

Key words: *Schistosoma mansoni*, codon usage, bias.

INTRODUCTION

The phylum Platyhelminthes contains the various dorso-ventrally flattened animals known as flatworms (Mehlhorn & Walldorf, 1988). Most of the flatworms are acoelomate; have bodies that are covered by a tegument, and they are hermaphroditic, in that both male and female reproductive systems occur in the same individual. The class Digenea of the phylum Platyhelminthes contains some 6000 individual species of trematodes (also known as flukes) and probably the most well known of these are the schistosomes because they cause the severe debilitating disease known as schistosomiasis or bilharzia in many of the tropical areas of the world. The genus *Schistosoma* contains at least 19 recognized species and most of these are pathogens although only seven species are recognized as parasites of man (Johnston *et al.* 1993). Recently, it has been estimated that in excess of 200 million people may be affected by schistosomiasis (WHO, 1985, 1990).

The search for a vaccine against schistosomiasis has resulted in the accumulation of a reasonable body of gene sequence data that is amenable for analysis. Indeed, analyses of codon usage in 19 schistosome sequences (Meadows & Simpson, 1989) and 21 genes of *Schistosoma mansoni* have been reported (Wada *et al.* 1990). However, analyses such as these are flawed in several important aspects, because these studies considered only the total codon usage summed over all the genes that were studied, and did not take into account any heterogeneity within the data set (Sharp & Devine, 1989; Lloyd & Sharp, 1992*a*). Since it is now well documented that the extent to which alternative synonymous codons are used by an organism is non-random, and that the pattern of codon usage varies not only between species but also between genes of the same species, it is obvious that such approaches can result in 'serious misconceptions' about the patterns of codon usage and bias in *S. mansoni* (Lloyd & Sharp, 1992*a*).

Bias in codon usage may result from the influence of a number of factors such as the amino acid composition of the gene product or directional mutation pressure on the base composition of the genome (Andersson & Kurland, 1990; Osawa *et al.* 1992). In addition, genes of unicellular organisms (such as *Saccharomyces cerevisiae*) that are highly expressed normally show a differential use of synonymous codons compared with other genes expressed at much lower levels (Lloyd & Sharp, 1992*a*).

In the study presented here, a complete investigation of codon usage and bias in gene sequences of *S. mansoni* is presented. The magnitude of the codon usage bias detected in these gene sequences was determined using three commonly used indices: the G+C frequency at silent third codon positions (GC_{3S}) (Sharp & Devine, 1989), the 'effective' number of codons (N_c) (Wright, 1990) and the standardized measure of bias in synonymous codon preference (B) based on the formula for average genetic homozygosity used in population genetics (Long & Gillespie, 1990).

MATERIALS AND METHODS

The data set (shown in Table 1) contains 20 genes of *S. mansoni*. Gene sequences were extracted from either GenBank or the EMBL sequence data bases using the accession numbers shown. They represent almost a complete set of gene sequence data that is currently available for this organism. Many of them code for proteases, which introduces a potential source of bias into the data. However, they were

Table 1. *Schistosoma mansoni* gene sequence data set

Gene*	Accession number	L†	GC§	N_c¶	GC_{3s}**	B††	Code‡‡	References
SCMCALPAIN	M74233	2277	0·36	39·64	0·15	0·303	1	Karcz *et al.* 1991
SCMEGFRA	M86396	5154	0·39	46·97	0·24	0·157	2	Shoemaker *et al.* 1992*a*
SCMGLUPER	M86510	510	0·41	61·00	0·40	0·137	3	Williams *et al.* 1991
SCMHEXK	L04480	1356	0·41	50·08	0·30	0·127	4	GenBank
SCMHSP70X	L02415	1914	0·47	43·03	0·47	0·190	5	Neumann *et al.* 1992
SCMMYH	L01634	5823	0·33	33·47	0·11	0·418	6	Weston *et al.* 1993
SCMNPK	M80542	594	0·35	46·15	0·24	0·236	7	Cao *et al.* 1992
SCMTPIPR	M83294	762	0·38	43·19	0·19	0·273	8	Shoemaker *et al.* 1992*b*
SCMCPROT	J03946	795	0·49	57·90	0·39	0·087	9	Newport *et al.* 1988
SCMCTSB	M21309	1020	0·41	51·20	0·28	0·160	10	Klinkert *et al.* 1989
SCMFERA	X15898	522	0·40	61·00	0·37	0·157	11	Dietzel *et al.* 1992
SCMHMGCOB	M27294	2844	0·36	48·02	0·27	0·164	12	Rajkovic *et al.* 1989
SCMHSP86	J04017	1329‡	0·43	58·10	0·45	0·067	13	GenBank
SCMIMP23A	M34453	654	0·43	61·00	0·41	0·123	14	Wright *et al.* 1990
SCMPYMA1	M35499	2598	0·31	29·46	0·05	0·564	15	Laclette *et al.* 1991
SCMPRSM	M17423	1059	0·35	47·21	0·24	0·224	16	Davis *et al.* 1987
SCMSM24A	M67506	555	0·32	35·63	0·11	0·419	17	Francis & Bickle, 1992
SCMSMC74	J05410	2028	0·39	49·14	0·28	0·146	18	Stein *et al.* 1990
S79195	M76387	1356	0·34	31·99	0·08	0·478	19	Duvaux-Miret *et al.* 1991
SCMHGPRT	X07883	693	0·39	52·63	0·31	0·202	20	Craig *et al.* 1988

* Gene, data base name.
† L, length in base pairs (‡ no initiation codon).
§ GC, gene G+C content as a fraction.
¶ N_c, effective number of codons.
** GC_{3s}, G+C content at silent third positions.
†† B, standardized measure of bias.
‡‡ Code, code used in the figures.

chosen for study because they are all greater than 500 bp long and are characterized by a single, major open reading frame (which in 19 out of the 20 is defined by an initiation and a stop codon). Although a large number of partial gene sequences are available for *S. mansoni* these were excluded from the analysis presented here because the partial sequence may not be typically representative of the complete gene sequence from which it was derived. Such variations result in bias measurements which do not reflect the entire gene sequence.

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 Genome Information Service.

Heterogeneity in codon usage between genes was investigated by metric multidimensional scaling (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 2-dimensional graph, where the difference between the entities on the graph represents their differences in codon usage. A 2-dimensional ordination was used, with the euclidean distance measure (Faith, Minchin & Belbin, 1987).

A computer program CODONS (Lloyd & Sharp, 1992*b*) 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 one 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).

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 within a gene sequence.

RESULTS

Codon usage was examined in 20 genes of *S. mansoni* containing 11287 codons (Table 1). Analysis of the total codon usage summed over all the genes is shown in Table 2. A χ^2 goodness-of-fit test was used to identify codons that were either under or over-represented in the data set. Twenty-five codons were

Table 2. Codon usage in *Schistosoma mansoni*

Amino acid	Codon	N*	F_{1000}†	FR_S‡	Amino acid	Codon	N*	F_{1000}†	FR_S‡
Gly	GGG	45	3·99	0·07	Trp	TGG	95	8·42	1·00
Gly	GGA	186	16·48	0·29	End	TGA	7	0·62	0·33
Gly	GGT	359	31·81	0·56	Cys	TGT	197	17·45	0·78
Gly	GGC	55	4·87	0·09	Cys	TGC	56	4·96	0·22
Glu	GAG	182	16·12	0·20	End	TAG	5	0·44	0·24
Glu	GAA	745	66·01	0·80	End	TAA	9	0·80	0·43
Asp	GAT	515	45·63	0·80	Tyr	TAT	239	21·17	0·71
Asp	GAC	126	11·16	0·20	Tyr	TAC	100	8·86	0·29
Val	GTG	136	12·05	0·20	Leu	TTG	195	17·28	0·18
Val	GTA	165	14·62	0·24	Leu	TTA	450	39·87	0·42
Val	GTT	299	26·49	0·43	Phe	TTT	250	22·15	0·59
Val	GTC	94	8·33	0·14	Phe	TTC	171	15·15	0·41
Ala	GCG	48	4·25	0·07	Ser	TCG	55	4·87	0·07
Ala	GCA	229	20·29	0·32	Ser	TCA	237	21·00	0·30
Ala	GCT	366	32·43	0·52	Ser	TCT	159	5·94	0·09
Ala	GCC	64	5·67	0·09	Ser	TCC	67	5·94	0·09
Arg	AGG	29	2·57	0·05	Arg	CGG	21	1·86	0·03
Arg	AGA	120	10·63	0·19	Arg	CGA	109	9·66	0·17
Ser	AGT	214	18·96	0·27	Arg	CGT	308	27·29	0·49
Ser	AGC	48	4·25	0·06	Arg	CGC	41	3·63	0·07
Lys	AAG	199	17·63	0·25	Gln	CAG	109	9·66	0·22
Lys	AAA	607	53·78	0·75	Gln	CAA	391	34·64	0·78
Asn	AAT	429	38·01	0·76	His	CAT	204	18·07	0·76
Asn	AAC	137	12·14	0·24	His	CAC	63	5·58	0·24
Met	ATG	279	24·72	1·00	Leu	CTG	85	7·53	0·08
Ile	ATA	181	16·04	0·29	Leu	CTA	108	9·57	0·10
Ile	ATT	336	29·77	0·53	Leu	CTT	185	16·39	0·17
Ile	ATC	112	9·92	0·18	Leu	CTC	56	4·96	0·05
Thr	ACG	62	5·49	0·10	Pro	CCG	43	3·81	0·11
Thr	ACA	270	23·92	0·44	Pro	CCA	199	17·63	0·50
Thr	ACT	220	19·49	0·36	Pro	CCT	125	11·07	0·31
Thr	ACC	61	5·40	0·10	Pro	CCC	30	2·66	0·08

* N, observed usage of codon.
† F_{1000}, codon usage as a frequency/1000 codons.
‡ FR_S, fraction of synonymous codon usage.

over-represented and 28 were under-represented ($P < 0·05$). All the over-represented codons contained either A or T at the third base position of the codon and most of the under-represented codons contained either G or C (Table 3). Twenty-four of the under-represented codons could be classified as low usage codons since their frequency of usage was extremely low (less than 10/1000 codons) and most of these contained G or C at the third base position (22/24). Most low-usage codons also contained G and/or C at two of the codon positions (21/24). Fifteen out of the 20 stop codons used also contained A at the third base position. The coding region of SCMGLUPER (glutathione peroxidase) contains a stop codon (TGA) that is in-frame and which probably codes for selenocysteine (Williams *et al.* 1991).

Metric multi-dimensional scaling was used to investigate the overall patterns of codon usage shown by the 20 genes of *S. mansoni*. In the first instance codon usage of each gene was expressed as a frequency/1000 codons and the data analysed by MMDS (Table 1 and Fig. 1). The scatter of the data points over the ordination shows that significant heterogeneity exists in the patterns of codon usage shown by the genes in the data set. Since the pattern of codon usage shown by a gene can result from a bias in the amino acid composition of the gene product, codon usage was also expressed as a frequency of synonymous codon usage and the data reanalysed by MMDS. This method of analysis removes the effect of the amino acid composition of the gene product from the measure of codon usage. The results obtained are shown in Fig. 2. A comparison of Figs 1 and 2 reveals that although the amino acid composition of the gene products does influence the pattern of codon usage shown by these genes, it is the pattern of usage of synonymous codons that has the greatest influence on codon usage. The scatter of the data points over the ordination shown in Fig. 2 therefore shows that large differences exist in the pattern of codon usage used by the 20 genes of *S. mansoni*.

Three measures of codon usage bias were calculated (GC_{3S}, N_c and B) and the results obtained are presented in Table 1. GC_{3S} values vary between 0·05 and 0·47; N_c varies between 29·46 and 61·00 and B

Table 3. Over and under-represented codons in biased genes of *Schistosoma mansoni*

Amino acid	Codon	SCM MYH	SCM PMYA1	SCM SM24A	S79195	Total*
Gly	GGG	U†	U	—	U	U
Gly	GGA	—	—	—	—	—
Gly	GGT	O	O	—	O	O
Gly	GGC	U	U	—	U	U
Glu	GAG	U	U	U	U	U
Glu	GAA	O	O	O	O	O
Asp	GAT	O	O	—	O	O
Asp	GAC	U	U	—	U	U
Val	GTG	U	—	—	U	U
Val	GTA	—	—	O	—	—
Val	GTT	O	—	—	O	O
Val	GTC	U	—	—	—	—
Ala	GCG	U	U	—	U	U
Ala	GCA	—	—	—	O	O
Ala	GCT	O	O	O	O	O
Ala	GCC	U	U	—	U	U
Arg	AGG	—	U	—	—	U
Arg	AGA	—	O	—	—	O
Ser	AGT	O	O	—	—	O
Ser	AGC	U	U	—	—	U
Lys	AAG	U	U	—	U	U
Lys	AAA	O	O	O	O	O
Asn	AAT	O	O	—	O	O
Asn	AAC	U	U	—	U	U
Ile	ATA	U	—	—	—	—
Ile	ATT	O	O	—	O	O
Ile	ATC	U	U	—	U	U
Thr	ACG	U	U	—	U	U
Thr	ACA	O	O	—	O	O
Thr	ACT	O	—	O	—	O
Thr	ACC	U	U	—	—	U
Cys	TGT	O	—	—	O	O
Cys	TGC	U	—	—	U	U
Tyr	TAT	O	O	—	O	O
Tyr	TAC	U	U	—	U	U
Leu	TTG	U	U	U	U	U
Leu	TTA	O	O	O	O	O
Phe	TTT	—	—	—	O	O
Phe	TTC	—	—	—	U	U
Ser	TCG	U	U	—	U	U
Ser	TCA	O	O	O	O	O
Ser	TCT	—	—	—	—	O
Ser	TCC	U	U	—	—	U
Arg	CGG	U	U	—	U	U
Arg	CGA	—	U	—	—	—
Arg	CGT	O	O	O	O	O
Arg	CGC	U	U	—	—	U
Gln	CAG	U	U	—	U	U
Gln	CAA	O	O	—	O	O
His	CAT	O	O	—	O	O
His	CAC	U	U	—	U	U
Leu	CTG	—	U	—	—	U
Leu	CTA	—	—	—	—	—
Leu	CTT	O	O	O	—	O
Leu	CTC	U	U	—	—	U
Pro	CCG	U	—	—	U	U
Pro	CCA	O	O	O	O	O
Pro	CCT	—	—	—	—	O
Pro	CCC	U	—	—	—	U

* Total codon analysis.

† U, Under-represented codon; O, over-represented codon.

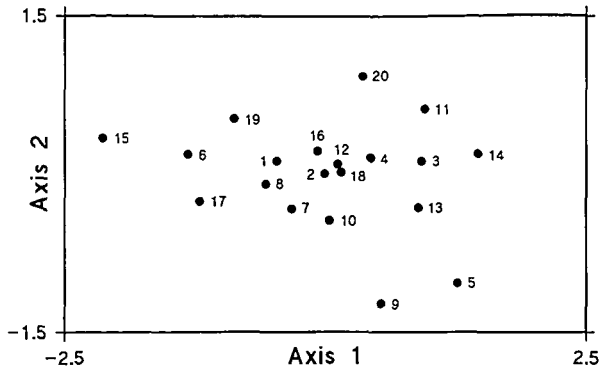

Fig. 1. Metric multidimensional scaling (Ellis *et al.* 1993) was used to investigate the pattern of codon usage shown by gene sequences of *Schistosoma mansoni*. For each gene sequence codon usage was expressed as a frequency/1000 codons (Table 2). A 2-dimensional ordination was used, with the euclidean distance measure (Faith *et al.* 1987). Gene codes are as in Table 1.

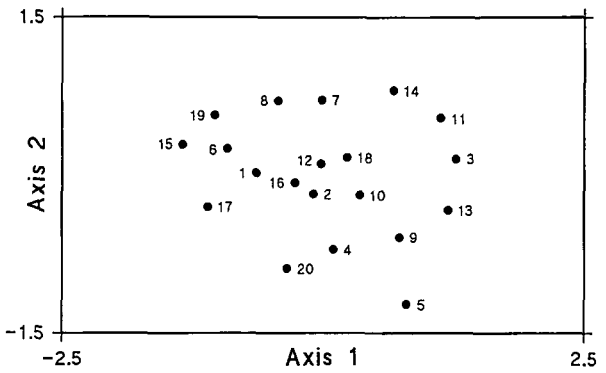

Fig. 2. Metric multidimensional scaling (Ellis *et al.* 1993) was used to investigate the pattern of codon usage shown by gene sequences of *Schistosoma mansoni*. For each gene sequence codon usage was expressed as a fraction of synonymous codon usage (Table 2). A 2-dimensional ordination was used, with the euclidean distance measure (Faith *et al.* 1987). Gene codes are as in Table 1.

values vary between 0·067 and 0·564. The variation in these values between different genes indicates substantial heterogeneity in the patterns of codon usage present in genes of this species.

A plot of N_c versus B for the 20 genes of *S. mansoni* is shown in Fig. 3. The plot shows that 4 genes are significantly biased according to both measures of bias (SCMPMYA1, S79195, SCMMYH and SCMSM24A). All of these genes have peripheral positions in both of the MMDS ordinations of Figs 1 and 2, indicating that their position is a result of codon usage bias.

A plot of GC_{3S} versus gene GC content is shown in Fig. 4. The GC_{3S} values vary between 0·05 and 0·5 whereas the gene GC content varies over a much narrower range (0·3–0·5). Four genes are identified as being highly biased by their GC_{3S} value and they are SCMPMYA1, S79195, SCMMYH and

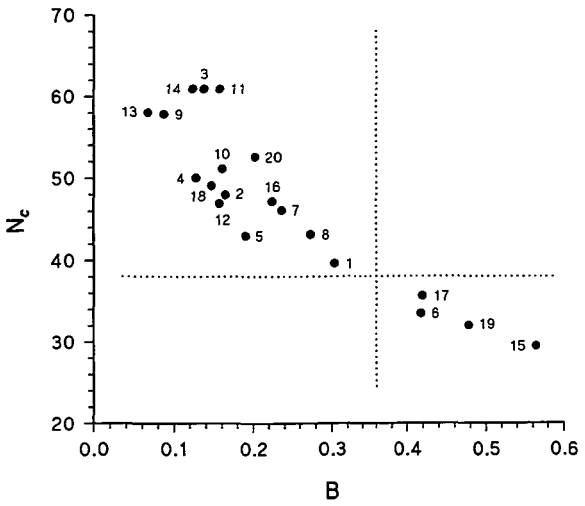

Fig. 3. 'Effective' number of codons used by the gene sequence (N_c) relative to the standardized synonymous codon bias (B). Gene sequences below the horizontal line are indicated by N_c as having higher codon usage bias than those above the line, while gene sequences to the right of the vertical line are indicated by B as having higher codon usage bias than those to the left of the line. Gene codes are as in Table 1.

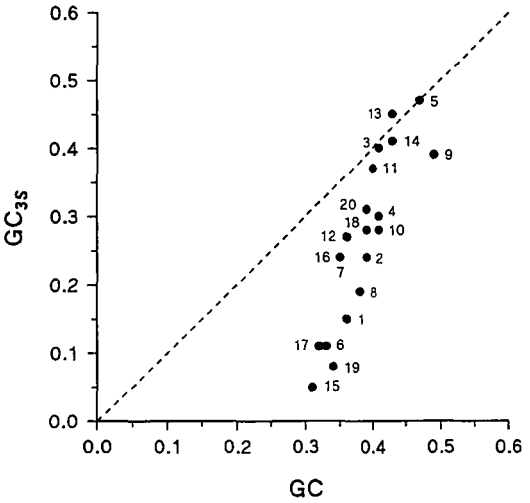

Fig. 4. Frequency of G+C at silent third positions of synonymous codons (excluding codons for Trp, Met and stop codons) (GC_{3S}) relative to G+C frequency of the whole gene sequence. The line is $GC_{3S} = GC$. Gene codes are as in Table 1.

SCMSM24A. At the other extreme of Fig. 4, SCMFERA, SCMG-LUPER, SCMIMP23A, SCMHSP86, SCMHSP-70X and SCMCPROT are all identified as having GC_{3S} and gene GC contents higher than the other genes in the data set. A comparison of the results from the analyses of Figs 3 and 4 with those ordinations presented in Figs 1 and 2 clearly shows a correlation between the patterns of codon usage, codon usage bias and gene GC content.

The patterns of codon usage bias detected were further studied using a χ^2-goodness-of-fit test in order to identify those codons that were either over- or under-represented in the gene sequences of

S. mansoni. In the first instance, a codon usage table was constructed from 5 gene sequences that were defined as unbiased in that these genes clustered together in the MMDS and had high N_c, GC_{3S} values and a low value of B (SCMFERA, SCMGLUPER, SCMIMP23A, SCMHSP86 and SCMCPROT). The χ^2-goodness-of-fit test revealed only 4 codons that were over-represented (GGT (Gly), GAT (Asp), GCA (Ala), CGT (Arg) and 3 codons that were under-represented (GGG (Gly), GAC (Asp) and CTA (Leu) $P < 0{\cdot}05$). In contrast, a codon usage table was compiled from each of the 4 genes that were identified as having a biased pattern of codon usage (SCMMYH, S79195, SCMPMYA1 and SCM-SM24A). A χ^2-goodness-of-fit test of these data revealed 22 codons that were over-represented and 30 codons that were under-represented (Table 3). All of the over-represented codons contained either A or T at the third base position of the codon, whereas all (except 2) codons that were under-represented contained either G or C at this position. The overall identity of the codons that were either over or under-represented were similar to those identified in the total data set (Table 3).

DISCUSSION

Codon usage and bias has been examined in 20 genes comprising 11287 codons from the trematode *S. mansoni*. Meadows & Simpson (1989) analysed the overall pattern of codon usage shown by 19 genes from two schistosome species (*S. mansoni* and *S. japonicum*). More recently, Wada *et al.* (1990) similarly presented a codon usage analysis of 21 genes of *S. mansoni*. However, as pointed out by Lloyd & Sharp (1992*a*), analyses such as these are flawed in several important aspects. In the first instance, the data set of Meadows & Simpson (1989) contained genes from two schistosome species. Secondly, and of more importance is the fact that both of these studies considered only the total codon usage summed over all the genes, and did not take into account any heterogeneity within the data set (Sharp & Devine, 1989; Lloyd & Sharp, 1992*a*). It is now well documented that the extent to which alternative synonymous codons are used by an organism is non-random, and that the pattern of codon usage varies not only between species but also between genes of the same species. The analysis presented here shows that this is also the case for genes of *S. mansoni*. Significant levels of codon usage bias were detected by 3 general indicators of codon usage bias (GC_{3S}, N_c and B). Like the sporozoan parasites (Ellis *et al.* 1994), the magnitude of the bias observed was dependent on the overall base composition of the genes analysed. Such a conclusion has also been drawn for a wide range of taxa (Bernardi & Bernardi, 1985). The identification of a large number of low usage codons, many of which contain G and/or C at 2 of the 3 base positions within a codon, provides further evidence in support of this concept. This is partially explained by the fact that since the base composition of schistosomes is only 34% GC (G. Hillyer, 1974, cited by Meadows & Simpson, 1989) dinucleotides containing G and/or C are generally under-represented in schistosome DNA.

Many of the schistosome genes revealed only low levels of bias and the large number of codons found to be either over or under-represented in the total codon analysis were not similarly reflected in these genes. The pattern of codon usage bias revealed by the total codon analysis could generally be explained by bias detected in just 4 genes of the data set (SCMMYH, SCMPMYA1, SCMCM24A and S79195). A variety of explanations may account for the bias detected. For example, both paramyosin (SCMPMYA1) and myosin heavy chain (SCMMYH) proteins contain characteristic amino acid repeats that explain some but not all of the bias detected in these 2 genes (compare their positions on the ordinations of Figs 1 and 2) (Weston *et al.* 1993; Laclette *et al.* 1991). However, although the amino acid composition of the gene product does influence the pattern of codon usage in these 2 genes it is the pattern of synonymous codon usage that has the largest effect on the codon usage observed.

In unicellular organisms genes that are highly expressed also show codon usage bias and therefore the discovery of bias in S79195 (the alpha tubulin gene which is normally expressed at high levels in a cell) is not that surprising (Andersson & Kurland, 1990; Osawa *et al.* 1992). It is possible that bias in *S. mansoni* may reflect the magnitude of gene expression; however, the available evidence does not support this concept. For example, during the isolation of the SCMPMYA1 and SCMCM24A from cDNA libraries only 0·026 (Laclette *et al.* 1991) and 0·034% (Francis & Bickle, 1992) of the total number of clones screened were positive implying that the mRNA from these genes is present at only low levels in the total mRNA pool of *S. mansoni*. In contrast SCMFERA (Dietzel *et al.* 1992), SCMEGFRA (Shoemaker *et al.* 1992*a*), SCMTPIPR (Shoemaker *et al.* 1992*b*), SCMPRSM (Davis, Nanduri & Watson, 1987), SCMGLUPER (Williams *et al.* 1991), SCMCPROT (Newport *et al.* 1988) and SCMCALPAIN (Andresen, Tom & Strand, 1991) transcripts constitute approximately 0·0008, 0·003, 0·006, 0·008, 0·2, 0·32 and 0·32% of the total mRNA pool respectively. Even though the representation of a cDNA within a cDNA library is an extremely inaccurate method for estimating mRNA levels there is no obvious correlation in this data set between the pattern of codon usage detected and the level of gene expression.

Finally, it has been shown that the genomes of eukaryotic organisms may comprise a mosaic of long

regions of different base composition ('isochores') and that the base composition of genes is related to their genomic context (Sharp & Lloyd, 1993). One possible explanation for the variation in gene GC content seen in the study presented here is that the genome of *S. mansoni*, like yeast, human, birds and monocoytledenous plants, has an isochore structure and that variations in isochore GC content are reflected in the GC content of the genes studied.

REFERENCES

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

ANDRESEN, K., TOM, T. D. & STRAND, M. (1991). Characterisation of cDNA clones encoding a novel calcium-activated neutral proteinase from *Schistosoma mansoni*. *Journal of Biological Chemistry* **266**, 15085–90.

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

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

CAO, M., AKRIDGE, R., WESTON, D., KEMP, W. M. & DOUGHTY, B. L. (1992). *Schistosoma mansoni*: cloning and sequencing of a gene for adenylate kinase. *Experimental Parasitology* **74**, 357–9.

CRAIG, S. P., MCKERROW, J. H., NEWPORT, G. R. & WANG, C. C. (1988). Analysis of cDNA encoding the hypoxanthine–guanine phosphoribosyltransferase (HGPRTase) of *Schistosoma mansoni*: a putative target for chemotherapy. *Nucleic Acids Research* **16**, 7087–101.

DAVIS, A. H., NANDURI, J. & WATSON, D. C. (1987). Cloning and gene expression of *Schistosoma mansoni* protease. *Journal of Biological Chemistry* **262**, 12851–5.

DEVEREUX, J., HAEBERLI, P. & SMITHIES, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 216–23.

DIETZEL, J., HIRZMANN, J., PREIS, D., SYMMONS, P. & KUNZ, W. (1992). Ferritins of *Schistosoma mansoni*: sequence comparison and expression in female and male worms. *Molecular and Biochemical Parasitology* **50**, 245–54.

DUVAUX-MIRET, O., BLANDINE, B., DISSOUS, C. & CAPRON, A. (1991). Molecular cloning and sequencing of the alpha tubulin gene from *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **49**, 337–40.

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.

ELLIS, J., MORRISON, D. A., AVERY, D. & JOHNSON, A. M. (1994). Codon usage and bias among individual genes of the coccidia and piroplasms. *Parasitology* (in the Press).

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.

FRANCIS, P. & BICKLE, Q. (1992). Cloning of a 21·1 kDa vaccine-dominant antigen gene of *Schistosoma mansoni* reveals an EF hand-like motif. *Molecular and Biochemical Parasitology* **50**, 215–24.

JOHNSTON, D. A., DIAS NETO, E., SIMPSON, A. J. G. & ROLLINSON, D. (1993). Opening the can of worms: molecular analysis of schistosome populations. *Parasitology Today* **9**, 286–91.

KARCZ, S. R., PODESTR, R. B., SIDDIQUI, A. A., DEKABAN, G. A., STREJAN, G. H. & CLARKE, M. W. (1991). Molecular cloning and sequence analysis of a calcium-activated neutral protease (calpain) from *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **49**, 333–6.

KLINKERT, M. Q., FELLEISEN, R., LINK, G., RUPPEL, A. & BECK, E. (1989). Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. *Molecular and Biochemical Parasitology* **33**, 113–22.

LACLETTE, J. P., LANDA, A., ARCOS, L., WILLMS, K., DAVIS, A. E. & SHOEMAKER, C. B. (1991). Paramyosin is the *Schistosoma mansoni* (Trematoda) homologue of antigen B from *Taenia solium* (Cestoda). *Molecular and Biochemical Parasitology* **44**, 287–95.

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

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

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

MEADOWS, H. M. & SIMPSON, A. J. G. (1989). Codon usage in *Schistosoma*. *Molecular and Biochemical Parasitology* **36**, 291–9.

MEHLHORN, H. & WALLDORF, V. (1988). Life cycles. Phylum Platyhelminthes. In: *Parasitology in Focus* (ed. Mehlhorn, H.), pp. 53–87. Berlin: Springer-Verlag.

NEUMANN, S., ZIV, E., LANTNER, F. & SCHECTER, I. (1992). Cloning and sequencing of an hsp70 gene of *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **56**, 357–60.

NEWPORT, G., MCKERROW, J. H., HEDSTROM, R., PETITT, M., MCGARRIGLE, L., BARR, P. J. & AGABIAN, N. (1988). Cloning of the proteinase that facilitates infection by schistosome parasites. *Journal of Biological Chemistry* **263**, 13179–84.

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

RAJKOWIC, A., SIMONSEN, J. N., DAVIS, R. E. & ROTTMAN, F.M. (1989). Molecular cloning and sequence analysis of 3-hydroxy-3-methylglutaryl-coenzyme A reductase from the human parasite *Schistosoma mansoni*. *Proceedings of the National Academy of Sciences, USA* **86**, 8217–21.

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.

SHARP, P. M. & LLOYD, A. T. (1993). Regional base composition along yeast chromosome III: evolution of chromosome promary structure. *Nucleic Acids Research* **21**, 179–183.

SHOEMAKER, C. B., RAMACHANDRAN, H., LANDA, A., DOS-REIS, M. G. & STEIN, L. D. (1992*a*). Alternative splicing of the *Schistosoma mansoni* gene encoding a homologue of epidermal growth factor receptor. *Molecular and Biochemical Parasitology* **53**, 17–32.

SHOEMAKER, C., GROSS, A., GEBREMICHAEL, A. & HARN, D. (1992*b*). cDNA cloning and functional expression of the *Schistosoma mansoni* protective antigen triose-phosphate isomerase. *Proceedings of the National Academy of Sciences, USA* **89**, 1842–6.

STEIN, L. D., HARN, D. A. & DAVID, J. R. (1990). A cloned ATP:guanidino kinase in the trematode *Schistosoma mansoni* has a novel duplicated structure. *Journal of Biological Chemistry* **265**, 6582–8.

WADA, K., AOTA, S., TSUCHIYA, R., ISHIBASHI, F., GOJOBORI, T. & IKEMURA, T. (1990). Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Research* **18**, 2367–8.

WESTON, D., SCHMITZ, J., KEMP, W. M. & KUNZ, W. (1993). Cloning and sequencing of a complete myosin heavy chain cDNA from *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **58**, 161–4.

WILLIAMS, D. L., PIERCE, R. J., COOKSON, E. & CAPRON, A. (1991). Molecular cloning and sequencing of glutathione peroxidase from *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **52**, 127–30.

WORLD HEALTH ORGANIZATION (1985). The control of schistosomiasis. Report of a WHO Expert Committee, Geneva. *WHO, Technical Report Series*, No. 728.

WORLD HEALTH ORGANIZATION (1990). Health Education in the Control of Schistosomiasis. Geneva: World Health Organization.

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

WRIGHT, M., HENKLE, K. & MITCHELL, G. (1990). An immunogenic M_r 23000 integral membrane protein of *Schistosoma mansoni* worms that closely resembles a human tumour-associated signal. *Journal of Immunology* **144**, 3195–200.