The Korean Journal of Parasitology
Vol. 42, No. 4. 159-167, December 2004

Codon usage and bias in mitochondrial genomes of parasitic platyhelminthes

Thanh Hoa LE[1),2)]*, Donald Peter McMANUS[1)] and David BLAIR[3)]

[1)]*Molecular Parasitology Laboratory, Australian Centre for International and Tropical Health and Nutrition, The Queensland Institute of Medical Research and The University of Queensland, Brisbane, Queensland 4029, Australia*
[2)]*Institute of Biotechnology (IBT), Hanoi, Vietnam*
[3)]*School of Tropical Biology, James Cook University, Townsville, Queensland 4811, Australia*

Abstract: Sequences of the complete protein-coding portions of the mitochondrial (mt) genome were analysed for 6 species of cestodes (including hydatid tapeworms and the pork tapeworm) and 5 species of trematodes (blood flukes and liver- and lung-flukes). A near-complete sequence was also available for an additional trematode (the blood fluke *Schistosoma malayensis*). All of these parasites belong to a large flatworm taxon named the Neodermata. Considerable variation was found in the base composition of the protein-coding genes among these neodermatans. This variation was reflected in statistically-significant differences in numbers of each inferred amino acid between many pairs of species. Both convergence and divergence in nucleotide, and hence amino acid, composition was noted among groups within the Neodermata. Considerable variation in skew (unequal representation of complementary bases on the same strand) was found among the species studied. A pattern is thus emerging of diversity in the mt genome in neodermatans that may cast light on evolution of mt genomes generally.

Key words: base composition, codon usage, mitochondrial genome organisation, mitochondrial genomes, Platyhelminthes, skew

INTRODUCTION

Mitochondrial (mt) genomes are an evolutionary paradox. There are many reasons why genomes should not have persisted in mitochondria — and yet they have indeed persisted (Saccone et al. 2002). Mitochondrial genomes also exhibit features not seen, or not as pronounced, in nuclear genomes. Among these are biases in base composition that must have an influence on the protein subunits for which they code. Studies on mitochondrial codon usage investigating this phenomenon have mainly focused on vertebrates.

We have an ongoing program of sequencing and characterising mitochondrial genomes from parasitic flatworms (Le et al., 2000a, 2000b). The major classes of parasitic flatworms, Trematoda, Monogenea and Cestoda, belong to a larger monophyletic taxon, the Neodermata. This taxon is distinct from other members of the Platyhelminthes (Littlewood and Bray 2001). Arising largely from our previous work, it is now clear that mt genomes of neodermatans resemble those of other metazoans in their organisation (Le et al., 2000a, 2000b). We have previously noted striking differences in base composition among these genomes

• Received 4 June 2004, accepted after revision 27 August 2004.
• This work was supported in part by Wellcome Trust (Ref: 068762) and by the National Health and Medical Research Council of Australia.
*Corresponding author (e-mail: im-ibt@hn.vnn.vn)

(Le et al., 2002b). Here, we report codon usage and associated phenomena for as many of these genomes as are currently available.

MATERIALS AND METHODS

Available to us for analysis were DNA sequences encompassing all protein-coding genes of the mitochondria of 11 species of neodermatans, 5 trematodes and 6 cestodes. For an additional trematode species, *Schistosoma malayensis*, we had sequence for most of the protein-coding genes. Further information on the taxa, GenBank accession numbers and reference sources, are in the footnote to Table 1.

Sequences were aligned using AssemblyLIGN v 1.9c and analysed by MacVector 6.5.3 package (Oxford Molecular Group). Pairwise comparisons of nucleotide and amino acid (aa) sequences of individual genes were undertaken using ClustalW as incorporated into the MacVector 6.5.3 package. Base composition and codon usage was calculated with MacVector 6.5.3, the DNA Strider program (Douglas 1995) and MEGA v2.1 (Kumar et al., 2001 - Arizona State University, Tempe, Arizona, USA).

Translations were done using the neodermatan mt genetic code most recently discussed in Blair et al. (1999), Nakao et al. (2000) and Telford et al. (2000). This code differs from the universal code in that TGA specifies tryptophan, AGA and AGG specify serine, ATA specifies isoleucine and AAA specifies asparagine. Initiation and termination codons will be discussed further below.

The program Tree-Puzzle v5 (Strimmer and von Haeseler, 1996) was used to explore base composition variation and amino-acid composition variation among the mt genomes examined. This program uses a chi-square test to determine whether the base composition of each sequence is identical to the average base composition of the whole alignment.

Mitochondrial genes of neodermatans are all encoded on the same strand. All calculations of base composition, skew etc used the strand reported in GenBank which is equivalent in sequence to the mRNAs of the various genes.

Skew is the unequal representation on a single strand of complementary bases such as G and C, something frequently reported from mt genomes (e.g. Saccone et al., 2002). Skew was estimated using the AT and GC-skew indices (Perna and Kocher, 1995) where:

AT skew = (A-T)/(A+T) and
GC skew = (G-C)/(G+C)

Values for the skew indices can range from -1 to +1. A value of zero indicates that A = T or G = C in frequency on the strand being analysed. A negative value for AT skew implies that T occurs more frequently than A, and so on.

RESULTS

Termination and initiation codons

The alignments of 3 gene-junctional blocks are presented in Fig. 1. These alignments comprise junctions between i) two protein-encoding genes (*atp6* and *nad2*; Fig. 1A); ii) a protein-encoding gene and a tRNA gene (*nad4* and *trnQ*; Fig. 1B); and iii) a tRNA and a protein-encoding gene (*trnG* and cox3; Fig. 1C). As seen in Fig. 1, the amino acid composition of the N-terminus of Atp6 (Fig. 1A) and Nad4 (Fig. 1B) as well as the C-terminus of Cox3 (Fig. 1C) are highly conserved. The C-terminus of Nad2 is less conserved across the range of species surveyed. The conservation of amino acid tracts strongly suggests that *atp6* and *nad4* can terminate at TAA and *cox3* and *nad2* may initiate at GTG. The stop codon for *atp6* in *Taenia crassiceps* is discussed below.

Variation in gene length

The general features of the 12 protein-encoding genes are presented in Table 1. Lengths of the individual genes are very similar among closely related species (Table 1). The most striking exceptions are *cox1* (encodes 510-609 aa) and *cox2* (191-219 aa). Most of the length difference in the former is due to an insertion of about 60 codons near the 5' end in *Schistosoma mansoni* and in the latter to an insertion of

Table 1. Characteristics of protein-encoding genes in the mitochondrial genomes of parasitic platyhelminthes[a)]

Genes	Class Trematoda						Class Cestoda					
	Sja	Sme	Sml	Smn	Fhe	Pwe	Tcr	Tso	Emu	EgrG1	EgrG4	Hdi
atp6	ATG 172	ATG 173	ATG 173	ATG 173	ATG 172	ATG 170	**GTG** 170	ATG 171	ATG 171	ATG 170	ATG 170	ATG 171
	TAA 519	TAA 522	TAG 522	TAG 522	TAG 519	TAG 513	TAA 513	TAA 516	TAG 516	TAG 513	TAG 513	TAG 516
cob	ATG 371	ATG 372	ATG 372	**GTG** 364	ATG 370	ATG 372	ATG 357	ATG 355	ATG 355	ATG 355	ATG 355	ATG 365
	TAG 1116	TAA 1119	TAA 1119	TAG 1095	TAG 1113	TAG 1119	TAA 1074	TAA 1068	TAA 1068	TAA 1068	TAA 1068	TAG 1098
cox1	**GTG** 547	ATG 549	ATG	ATG 609	ATG 510	ATG 511	ATG 537	ATG 539	ATG 535	ATG 535	**GTG** 526	**GTG** 521
	TAG 1644	TAA 1650	---	TAG 1830	TAG 1533	TAG 1536	TAG 1614	TAG 1620	TAG 1608	TAG 1608	TAG 1581	TAG 1566
cox2	ATG 202	ATG 219		ATG 197	ATG 200	ATG 199	ATG 194	ATG 193	**GTG** 193	**GTG** 193	**GTG** 193	ATG 191
	TAA 609	TAA 660		TAA 594	TAG 603	TAG 600	TAG 585	TAG 582	TAG 582	TAA 582	TAG 582	TAA 579
cox3	ATG 214	ATG 216	ATG 216	**GTG** 217	ATG 213	ATG 214	**GTG** 214	ATG 214	ATG 215	ATG 215	ATG 215	ATG 216
	TAG 645	TAG 651	TAA 651	TAG 654	TAG 642	TAG 645	TAG 645	TAG 645	TAG 648	TAG 648	TAG 648	TAG 651
nad1	ATG 296	ATG 295	ATG 297	**GTG** 296	**GTG** 300	ATG 296	ATG 297	ATG 297	ATG 297	**GTG** 297	ATG 297	ATG 296
	TAG 891	TAA 888	TAA 894	TAG 891	TAG 903	TAG 891	TAG 894	**T** 892	TAG 894	TAA 894	TAA 894	TAG 891
nad2	ATG 284	ATG 283	ATG 283	**GTG** 279	ATG 288	ATG 288	ATG 292	ATG 293	ATG 293	ATG 293	ATG 293	ATG 293
	TAG 855	TAA 852	TAG 852	TAA 840	TAG 867	TAA 867	TAG 879	TAA 882	TAG 882	TAG 882	TAG 882	TAG 882
nad3	ATG 117	ATG 120	ATG 120	ATG 120	ATG 118	ATG 118	**GTG** 115	ATG 115	ATG 115	ATG 115	ATG 115	ATG 115
	TAG 354	TAG 363	TAG 363	TAG 363	TAG 357	TAG 357	TAG 348	TAG 348	TAA 348	TAG 348	TAA 348	TAG 348
nad4	ATG 424	ATG 423	ATG 423	ATG 419	**GTG** 423	ATG 420	**GTG** 419	ATG 417	ATG 419	ATG 419	ATG 419	ATG 415
	TAG 1275	TAA 1272	TAA 1272	TAA 1260	TAA 1272	TAG 1263	TAG 1260	TAG 1254	TAG 1260	TAG 1260	TAA 1260	TAG 1248
nad4L	ATG 87	ATG 87	ATG 87	ATG 86	**GTG** 90	**GTG** 85	ATG 86	ATG 86	**GTG** 86	**GTG** 86	**GTG** 86	ATG 86
	TAA 264	TAA 264	TAA 264	TAA 261	TAG 273	TAG 258	TAG 261	TAA 261	TAG 261	TAA 261	TAA 261	TAG 261
nad5	ATG 528	**GTG** 530	---	ATG 527	**GTG** 522	**GTG** 527	ATG 522	ATG 522	ATG 524	ATG 523	ATG 523	ATG 524
	TAG 1587	TAA 1593	TAG	TAG 1584	TAG 1569	TAG 1584	TAA 1569	TAA 1569	TAA 1575	TAG 1572	TAA 1572	TAG 1575
nad6	ATG 152	**GTG** 153		ATG 149	ATG 150	ATG 150	ATG 150	ATG 150	ATG 151	ATG 151	ATG 151	ATG 152
	TAG 459	TAA 462		TAA 450	TAG 453	TAG 453	TAA 453	TAG 453	TAA 456	TAG 456	TAG 456	TAA 459

[a)]In each cell, the predicted initiation codon is shown at the upper left, the termination codon at the lower left, the number of amino acids at upper right and the number of nucleotides at the lower right. Start codons other than ATG & TAA and stop codons other than TAG are shown in bold and underlined. Dashes (---) indicate the gene has not been fully sequenced (*S. malayensis* only). Trematodes - Sja: *Schistosoma japonicum* (Schistosomatidae, GenBank accession AF215860); Sme: *S. mekongi* (Schistosomatidae, AF217449); Sml: *S. malayensis* (Schistosomatidae, AF295106); Smn: *S. mansoni* (Schistosomatidae, AF216698); Fhe: *Fasciola hepatica* (Fasciolidae, AF216697); Pwe: *Paragonimus westermani* (Paragonimidae, AF219379). Cestodes - Tcr: *Taenia crassiceps* (Taeniidae, AF216699); Tso: *Taenia solium* (Taeniidae, AB086256); Emu: *Echinococcus multilocularis* (Taeniidae, AB018440); Egr: *E. granulosus* (Taeniidae, G1: genotype 1 (sheep-dog strain), AF297617 and G4: genotype 4 (horse-dog strain), AF346403); Hdi: *Hymenolepis diminuta* (Hymenolepididae, AF314223).

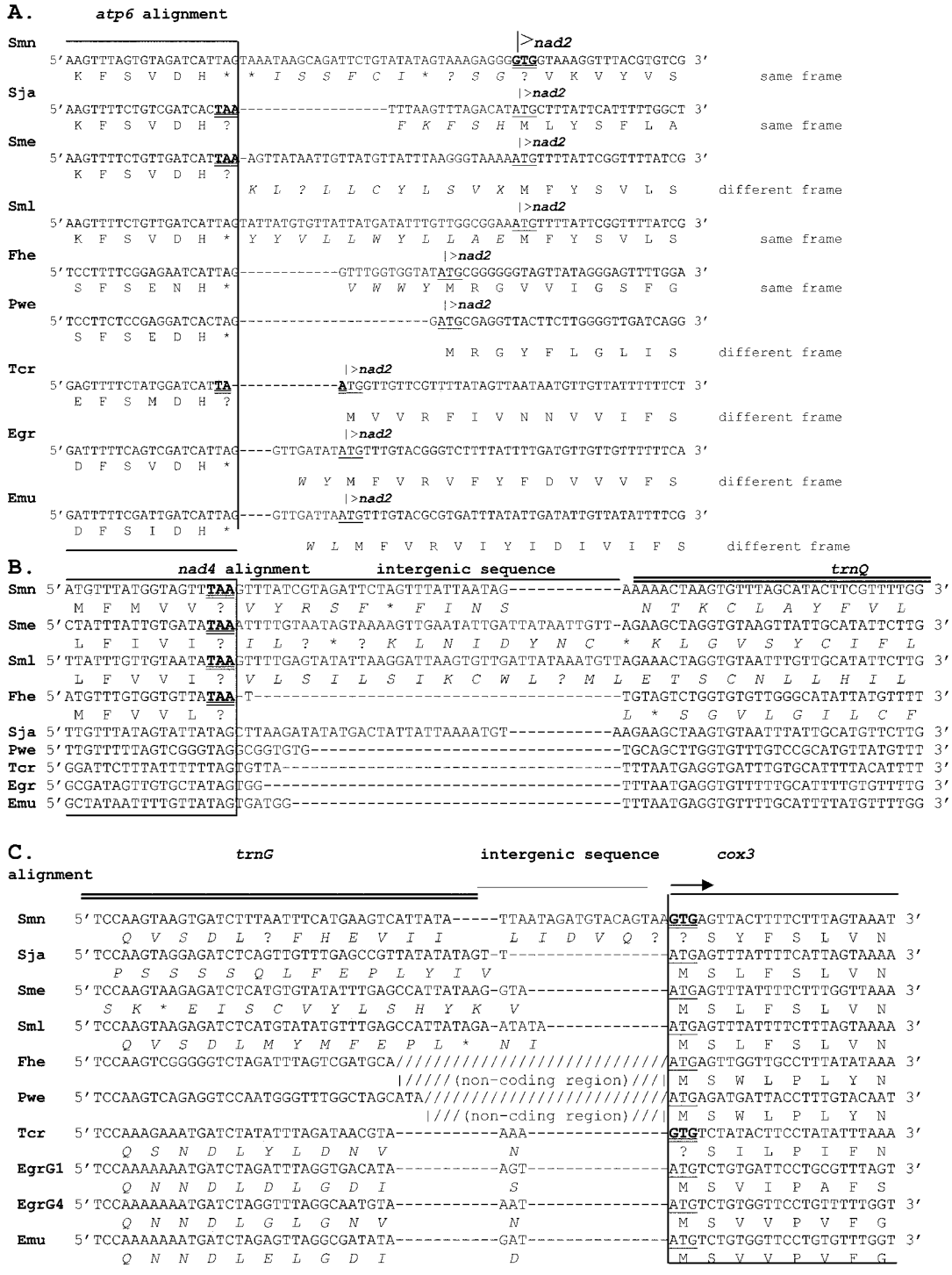

Fig. 1. Alignments of the junctions between pairs of mt genes from selected flatworms to demonstrate that TAA can act as a termination codon and GTG as an initiation codon. **A.** *atp6-nad2*; **B.** *nad4-trnQ* and **C.** *trnG-cox3* regions. Sequences encoding genes are highlighted. TAA and GTG codons are in bold and double-underlined. The typical initiation codon (ATG) is underlined; TAG (termination) is under-asterisked (*) where analysed. Dashes (-) indicate gaps inserted for alignment purposes. Supposed amino acid residues in tRNA and intergenic sequences are in italics. Slashes (/ /) show a long non-coding region encroaching between *trnG* and *cox3* in *F. hepatica* and *P. westermani*. See text for genetic code analysis. Names of species as for Table 1.

Table 2. Overall protein-coding and 3rd codon position (all codons and four-fold redundant (FFR) codons) base usage among parasitic platyhelminthes

Species	Base-composition[a]					Total	Total	Codon ending with[a]							
	T	C	A	G	T+A	bp	codon	T		C		A		G	
	%	%	%	%	%	usage[c]	No[d]	All	FFR	All	FFR	All	FFR	All	FFR
Trematodes															
S. mansoni	45.6	8.2	23.3	23.0	68.9	10344	3448	47.6	51.8	4.2	4.3	27.1	24.1	21.1	19.7
S. japonicum	48.3	8.0	23.0	20.7	71.3	10218	3406	54.8	64.3	3.4	3.7	23.9	19.6	17.9	12.5
S. mekongi	48.4	6.7	24.3	20.6	72.7	10296	3432	56.0	65.9	0.9	1.2	25.5	21.0	17.5	11.9
S. malayensis[b]	48.8	6.6	23.6	20.9	72.4	8145[b]	2714	56.1	67.9	1.0	1.0	25.2	20.7	17.6	10.4
F. hepatica	49.4	9.6	14.2	26.8	63.6	10104	3368	57.2	67.8	4.6	5.5	9.9	6.7	28.3	20.0
P. westermani	38.3	17.9	13.2	30.6	51.5	10086	3362	34.1	35.4	21.7	19.4	7.1	6.7	37.1	38.5
Cestodes															
T. solium	48.6	7.9	23.5	20.0	72.1	10092	3364	53.4	56.7	3.7	3.8	26.3	25.3	16.7	14.2
T. crassiceps	50.6	7.2	23.4	18.8	74.0	10095	3365	56.6	63.6	2.2	1.3	25.1	20.7	16.1	14.5
E. multilocularis	50.6	7.1	18.3	24.0	68.9	10098	3366	57.8	61.8	2.3	2.1	16.0	15.4	23.8	20.8
E. granulosus (G1)	49.9	7.6	16.8	25.7	66.7	10092	3364	56.3	58.7	3.4	4.0	12.6	11.1	27.7	26.2
E. granulosus (G4)	50.1	7.4	17.7	24.8	67.8	10065	3355	56.8	60.4	2.9	2.8	14.1	12.8	26.2	24.0
H. diminuta	47.4	9.5	23.7	19.4	71.1	10074	3358	50.7	52.1	6.0	5.1	26.7	25.8	16.6	16.9

[a]Excluding start and stop codons.
[b]Not all protein-encoding genes of *S. malayensis* are available for analysis.
[c]Bases overlapping between *nad4L* and *nad4* are counted twice.
[d]Includes start and stop codons.

about 20 codons near the 3' end in *Schistosoma mekongi*. Genes are particularly uniform in length among the cestodes (Table 1), perhaps partly due to the smaller phylogenetic range of cestodes sampled. Among the cestodes, most genes differ in length by 0-4 codons. However, *cob* in *Hymenolepis diminuta* is about 10 codons longer than in other cestodes and *cox1* in *H. diminuta* and the G4 genotype (horse-dog strain) of *Echinococcus granulosus* is 10 or more codons shorter than in other cestodes. The lengths of most genes in cestodes are similar to those of the corresponding genes in trematodes. However, *cob* in cestodes is shorter than that found in the Asian schistosomes (355-365 aa in cestodes, compared with 371-372 in Asian schistosomes and 364-372 in other trematodes). The *cox1* gene encoding 521-539 residues in cestodes is intermediate in length between that in the Asian schistosomes (547-549 aa) and other trematodes (510-511 aa for *Fasciola hepatica* and *Paragonimus westermani*, respectively). As mentioned above, *S. mansoni* has the longest *cox1*, encoding 609 residues.

Base composition and bias

Base compositions differ among the genomes examined (Table 2). When the DNA sequences of the concatenated protein-coding genes for all taxa were analysed using Tree-Puzzle, each differed significantly from the overall consensus. When *S. malayensis* (incomplete sequence available), *P. westermani*, *F. hepatica* and *Taenia crassiceps* (all with distinctive base compositions - Table 2) were omitted singly or together, the remaining sequences still differed significantly from the consensus.

Redundancy at the third codon position, especially at four-fold degenerate sites, means that sequences differing significantly in nucleotide frequencies may differ less when translated into amino acid sequences. This may also be expected because the greatest bias in base frequencies is usually observed at third codon positions (Saccone et al., 1999, 2002). However, Tree-Puzzle also rejected the hypothesis that inferred amino-acid composition is uniform across all species. Indeed, all species failed the test when all were included in a single test. Omission of single species,

Table 3. Numbers of each amino acid present in polypeptides encoded by mitochondrial genes of parasitic flatworms[a)]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Total[b)]
Sja	85	95	87	64	367	234	48	330	56	493	111	119	81	27	58	348	92	351	104	232	3382
Sme	78	102	82	68	342	234	55	352	66	496	117	123	82	24	55	354	87	348	100	243	3408
Smn	84	102	66	74	327	249	53	304	60	536	115	97	75	31	64	379	67	418	108	215	3424
Fhe	116	121	70	75	372	300	51	166	44	562	88	70	95	26	64	333	80	434	118	159	3344
Pwe	151	110	66	81	339	320	60	124	50	577	77	70	91	28	71	360	84	417	111	149	3336
EgrG1	80	149	80	65	398	243	50	199	42	507	82	97	69	25	50	338	89	464	97	216	3340
EgrG4	79	140	82	66	407	228	52	207	44	499	87	100	71	25	53	333	91	455	97	215	3331
Emu	82	148	76	63	418	236	49	221	43	499	82	106	71	24	51	344	88	436	93	212	3342
Tso	81	138	84	68	413	193	56	314	48	494	86	126	71	23	49	371	95	343	90	197	3340
Tcr	63	135	84	57	433	186	52	302	49	514	93	153	72	21	48	358	93	325	86	217	3341
Hdi	100	131	67	73	417	190	53	297	47	507	83	136	80	22	52	365	110	298	88	217	3333

[a)]Start codons omitted. *Schistosoma malayensis* was omitted because a full sequence was not available. Names of species as for Table 1.

[b)]Total numbers lie in a narrow range (3331-3424), so it was not felt necessary to also tabulate percentages of each.

or groups of species, did not yield any cases where remaining species passed the test.

Numbers of each amino acid encoded by each mt genome are shown in Table 3. A 2x20 contingency table showed that many pairs of species differed significantly (5% level) from each other in amino-acid composition. Exceptions were comparisons among the three *Schistosoma* species and comparisons among the three species / genotypes of *Echinococcus*. The two *Taenia* species did not differ significantly from each other or from *H. diminuta*, a surprising finding given that they differed significantly from taeniids in the genus *Echinococcus*. Interestingly, *S. mansoni* (a trematode) did not differ significantly from *Taenia solium* (a cestode). This is not the only apparent example of convergence among our neodermatans. *F. hepatica* and *P. westermani*, two trematodes that are not closely related to one another (Cribb et al., 2001), did not differ significantly. It is clear that divergences and convergences are frequent at various taxonomic levels, including between the Trematoda and Cestoda.

Despite the fact that codon usage generally differs significantly among species, the common pattern of bias, with high frequencies of T, means that the same amino acids tend to be most commonly observed in all taxa (Table 3). Leucine is the most common amino acid in every case, followed by serine, valine and phenylalanine (not necessarily in that order). The least common amino acids are glutamine, histidine, lysine and arginine (Table 3), regardless of the base composition of the genome.

Skew in base composition clearly exists among the neodermatan mt genomes (Table 2, Fig. 2). For example, among the neodermatans, T is the most common base, often reaching a frequency of around 50% (with the notable exception of *P. westermani*). By contrast, A is usually present at around half this frequency. Table 2 shows the overall base composition of protein-coding genes of the available neodermatans, and also the percentage of each base occurring in the third codon position overall and the third codon position of fourfold redundant codons (FFR). In most neodermatans, the percentage of codons ending with C is no greater than about half the representation of C in the protein-coding genes overall. Particularly low values are noted for *S. mekongi* (0.9%) and *S. malayensis* (1% - but note that not all genes were available for analysis in this species) (Table 2). Only in *P. westermani* is C more represented in third codon positions (21.7%) than in protein-coding genes overall (17.9%). This species also has an unusually high representation of G in third codon positions. Usage of T in third codon positions is usually greater than its overall representation in protein-coding genes (Table 2). The only exception is *P. westermani* which has by far the lowest overall percentage of T. Similarly, A is under-represented in

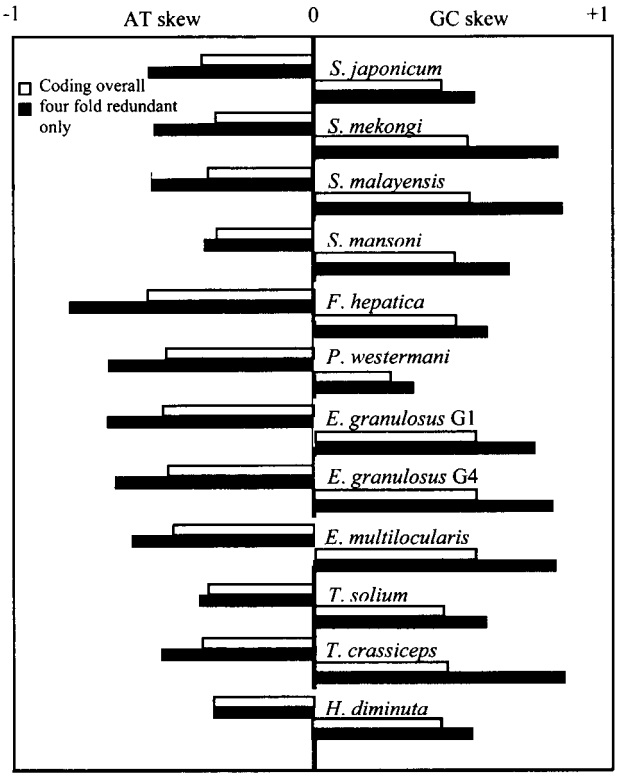

Fig. 2. Histogram showing AT-skew and GC-skew in concatenated protein-coding genes overall (minus start and stop codons) and at four-fold redundant (FFR) sites only.

third codon positions in *F. hepatica* and *P. westermani* (Table 2).

Among neodermatans, skew values vary considerably. AT-skew is least in *S. mansoni* and greatest in *F. hepatica* (Fig. 2). GC-skew is least in *P. westermani*, reflecting the relatively high occurrence of C. The greatest GC-skews are seen in S. *malayensis*, *S. mekongi* and the *Echinococcus* species.

DISCUSSION

A first requirement in characterising genes is to determine where they start and end. ATG and TAG are regarded as the typical mt stop and start codons respectively in neodermatans. However, we have presented evidence (Blair et al., 1999; Le et al., 2000a) that TAA can also act as a stop codon and GTG as an alternative start codon in neodermatans (as reported for other metazoans: Wolstenholme, 1992). In this study, we have confirmed that these two codons (GTG and TAA) can act respectively to initiate and terminate a gene.

Using a similar approach, we have inferred that the initiation codon of *cox1* in the cestode *H. diminuta* is GTT (Le et al., 2002b) (Table 1) thus agreeing with von Nickisch-Rosenegk et al. (2001). In the latter paper, the initiation codon for *nad4* is stated to be ATT. However, an in-frame ATG is situated 6 codons further upstream and we regard this as the correct start codon.

Pairs of genes may overlap, leading to interpretive difficulties. For example, in *T. crassiceps*, the last A of the stop codon (TAA) of *atp6* is shared with the ATG start codon in the following *nad2* gene (Fig. 1A). The former could be interpreted as a truncated stop codon or as an actual overlap between two genes. Abbreviated stop codons (T or TA) are known from some metazoans (Wolstenholme, 1992). There is only one example of this among the neodermatans. In *T. solium*, *nad1* ends with a T (Nakao, unpublished). Von Nickish-Rosenegk et al. (2001) suggested that *cox1* in *H. diminuta* is terminated with such a codon (T). However, our later analysis including sequences from several other cestodes indicate that a normal stop codon (TAG) is present here in cestodes but this requires overlap with the downstream *trnT* (Le et al., 2002a).

Tree-Puzzle always rejected the null hypothesis of equal base or amino-acid composition. A partial explanation for this might be that Tree-Puzzle requires a minimum of four species before it will run the analysis, but we did not have four sufficiently close relatives to include. When only four taeniid cestodes, *E. granulosus* G1 (sheep-dog strain) and G4 (horse-dog strain) genotypes, *E. multilocularis* and *T. solium* were included, all failed the test, although marginally in the case of the *E. granulosus* G4 genotype and *E. multilocularis*.

Associated with codon bias is the phenomenon of skew (unequal representation on a single strand of complementary bases). A well-developed theory is available to explain this situation, at least in the case of mammals (Saccone et al., 2002). The asymmetric nature of mt replication means that one of the strands

remains in a single-stranded state for relatively long periods. During this time it is prone to particular mutational changes, specifically a reduction in C and A on that strand and a corresponding increase in G and T. Skew is likely to be most pronounced at third codon positions, and especially at four-fold degenerate sites, where any mutational change is synonymous and not subject to selection pressure. Although little is known about the mode of mt replication in phyla other than vertebrates, an echinoderm and a few insects, base composition bias and strong skew are observed in many phyla (Saccone et al., 1999) including flatworms, suggesting that similar mechanisms may operate. As predicted by the theory outlined above, base composition bias and skew are most evident in third codon positions among neodermatans. The pattern observed among neodermatans is similar to that seen in vertebrates, with four-fold redundant sites showing the most extreme skew (Fig. 2). The main difference is that, among vertebrates, GC-skew has a negative value and AT-skew a positive value (Perna and Kocher, 1995). Nematodes and at least some molluscs exhibit negative AT-skew values and positive GC-skew values, as in neodermatans (Perna and Kocher, 1995). The sign of the skew value reflects only the strand being investigated: each strand will have the same value for each skew statistic, but with the opposite sign (Perna and Kocher, 1995).

Reyes et al. (1999) found that, among mammals, bias and skew were greatest in the regions of the mt genomes where the heavy strand was exposed as single-stranded for the longest time during replication. If the mode of replication in neodermatans is similar to that in mammals, we might expect the same pattern. Given that the locations of the origin(s) of replication are unknown in neodermatan mt genomes, and that they probably differ among taxa (as evidenced by the differing location of long non-coding regions (Le et al., 2002a), we did not feel able to explore this in detail. However, marked differences in skew and bias were noted among genes in a single neodermatan genome (data not shown) as reported by Reyes et al. (1999) for mammals. Analysis of sequences from additional neodermatan taxa may make it possible to infer the nature of the replication process.

Deviations in base composition (and amino-acid composition) among species will violate a basic assumption implicit in many algorithms used to infer phylogenies and can lead to construction of incorrect topologies (Foster and Hickey, 1999). Despite the considerable differences among neodermatan species in nucleotide sequences, and corresponding differences in amino-acid sequences, phylogenetic trees (not shown) inferred from either class of data recover the topology expected from the known systematic relationships among these taxa. This should not be taken as evidence that base composition differences will not affect tree topologies when sequences from a wider array of neodermatan taxa are added. All but two of the species included here fall into two relatively narrow but well-separated systematic groups (trematodes of the family Schistosomatidae and cestodes of the order Cyclophyllidea). Base composition differences would presumably have to be extreme to fail to recover these two groups. The two remaining trematodes, *P. westermani* and *F. hepatica*, which have similar base compositions, form a group elsewhere in the tree.

Previous studies have revealed that mt genomes of neodermatans are similar in most respects to those of other bilateral metazoans (Le et al., 2000a). However, these genomes vary considerably in the location(s) of long non-coding regions presumed to have a role in replication, and in the length and structure of these regions (Le et al., 2000a). There is also variation in gene order which, in African schistosomes, is very different from that of all other taxa sequenced to date (Le et al., 2001). Here, we have shown that there can be convergence as well as divergence in nucleotide, and hence amino acid, composition among taxa. A pattern is thus emerging of diversity in the mt genome in neodermatans that may cast light on evolution of mt genomes generally. Certainly, there is a need of data from additional neodermatan taxa, and from flatworms generally. Once these patterns are better understood, mt genome data can be put to practical use in evolutionary and population/species-level studies of important parasites such as the schisto-

somes (Le et al., 2000b).

REFERENCES

Blair D, Le TH, Després L, McManus D (1999) Mitochondrial genes of *Schistosoma mansoni*. *Parasitology* **119:** 303-313.

Cribb TH, Bray RA, Littlewood DTJ, Pichelin SP, Herniou EA (2001) The Digenea. *In* Interrelationships of the Platyhelminthes, Littlewood DTJ, Bray RA (eds). pp168-185, Taylor & Francis, London, UK.

Douglas SE (1995) DNA Strider. An inexpensive sequence analysis package for the Macintosh. *Mol Biotechnol* **3:** 37-45.

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* **48:** 284-290.

Le TH, Blair D, Agatsuma T, et al. (2000a) Phylogenies inferred from mitochondrial gene orders — a cautionary tale from the parasitic flatworms. *Mol Biol Evol* **17:** 1123-1125.

Le TH, Blair D, McManus DP (2002a) Mitochondrial genomes of parasitic flatworms. *Trends Parasitol* **18:** 206-213.

Le TH, Blair D, McManus DP (2000b) Mitochondrial genomes of human helminths and their use as markers in population genetics and phylogeny. *Acta Trop* **77:** 243-256.

Le TH, Humair PF, Blair D, Agatsuma T, Littlewood DT, McManus DP (2001) Mitochondrial gene content, arrangement and composition compared in African and Asian schistosomes. *Mol Biochem Parasitol* **117:** 61-71.

Le TH, Pearson MS, Blair D, Dai N, Zhang LH, McManus DP (2002b) Complete mitochondrial genomes confirm the distinctiveness of the horse-dog and sheep-dog strains of *Echinococcus granulosus*. *Parasitology* **124:** 97-112.

Littlewood DTJ, Bray RA (2001) Interrelationships of the Platyhelminthes. Taylor & Francis, London, UK.

Nakao M, Sako Y, Yokoyama N, Fukunaga M, Ito A (2000) Mitochondrial genetic code in cestodes. *Mol Biochem Parasitol* **111:** 415-424.

Perna NT, Kocher TD (1995) Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J Mol Evol* **41:** 353-358.

Reyes A, Gissi C, Pesole G, Saccone C (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* **15:** 957-966.

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* **238:** 195-209.

Saccone C, Gissi C, Reyes A, Larizza A, Sbisa E, Pesole G (2002) Mitochondrial DNA in Metazoa: degree of freedom in a frozen event. *Gene* **286:** 3-12.

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol* **13:** 964-969.

Telford MJ, Herniou EA, Russell RB, Littlewood DT (2000) Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proc Natl Acad Sci USA* **97:** 11359-11364.

von Nickisch-Rosenegk M, Brown WM, Boore JL (2001) Complete sequence of the mitochondrial genome of the tapeworm *Hymenolepis diminuta*: gene arrangements indicate that platyhelminths are eutrochozoans. *Mol Biol Evol* **18:** 721-730.

Wolstenholme DR (1992) Animal mitochondrial DNA, structure and evolution. *Int Rev Cytol* **141:** 173-216.