Genomics **74,** 197–210 (2001)
doi:10.1006/geno.2001.6531, available online at http://www.idealibrary.com on IDEAL®

# Influence of Intercodon and Base Frequencies on Codon Usage in Filarial Parasites

A. Fadiel,*,†,[1] S. Lithwick,* M. Q. Wanas,† and A. Jamie Cuticchia*

*Bioinformatics Supercomputing Centre, The Hospital for Sick Children, Toronto, Ontario M5G 1Z8, Canada; and
†Laboratory of Parasitology, Department of Zoology, Faculty of Science, Al-Azhar University, Cairo, Egypt

Received November 15, 2000; accepted February 26, 2001; published online May 7, 2001

**Base frequency, codon usage, and intercodon identity were analyzed in five filarial parasite species representing five *Onchocercidae* genera. *Wucheria bancrofti, Brugia malayi, Onchocerca volvulus, Acanthocheilonema viteae,* and *Dirofilaria immitis* gene sequences were downloaded from NCBI, and analysis was performed using locally designed computer programs and other freely available applications. A clear sequence bias was observed among the nematode species examined. At the nucleotide level, AT basepairs were present in gene sequences at higher frequencies than GC. In addition, codons ending in A or T were used proportionately more than those with G or C in the third-codon position. In addition, the amino acids used most often corresponded to codons ending in AT basepairs. Intercodon base proportion was biased in that A was found most often at N4, second only to T in certain specific cases. Since all of these sequence biases were observed in a relatively consistent fashion among all of the organisms studied, we conclude that sequence bias is a genetic characteristic, which is associated with multiple filarial genera.** © 2001 Academic Press

## INTRODUCTION

The invasion of filarial larvae into the human body stimulates an immunogenic response, which causes subsequent pathological complications varying from minor to major filariasis. In addition, the presence of invading larvae can damage vital organs such as the eye, leading to blindness and death if not treated properly. Filarial infection is prevalent in both the Third World and many cosmopolitan centers. Current estimates suggest that approximately 150 million people, equivalent to 2.45% of the world population (U.S. Bureau of the Census 2001, http://www.census.gov/cgi-bin/ipc/popclockw), are infected with filarial parasites (Peters and Gilles, 1995; Ogunrinade *et al.,* 1999; Basanez *et al.,* 2000), and 1 billion more are at risk (Johnston *et al.,* 1999).

According to the genome hypothesis, genes in a single genome and those in phylogenetically related species use similar patterns of codon usage (Levin and Whittome, 2000). However, most codons have a synonymous alternative except for those encoding methionine (AUG) and tryptophan (UGG), suggesting instead that variability in codon selection may exist. Therefore, the genetic code is degenerate, and synonymous codons are used with unequal frequencies in different genes (McClellan, 2000). Heterogeneity of codon usage has been found in the genes of eukaryotes such as *Saccharomyces cerevisiae* (Sharp and Cowe, 1991), *Arabidopsis thaliana* (Mathe *et al.,* 1999), *Drosophila melanogaster* (Moriyama and Powell, 1998), *Caenorhabditis elegans* (Stenico *et al.,* 1994), and *Schistosoma mansoni* (Ellis *et al.,* 1995). In *C. elegans* a codon usage bias for AT basepairs is present at the wobble position, and biased sequences are associated with high levels of gene expression (Stenico *et al.,* 1994). In addition, data indicate that codon usage bias is dependent upon the overall base composition of the genes analyzed, which defines the types of codons present in the nucleotide sequences. Similar studies have also been carried out in the *Brugia* genus, identifying a bias for A or T at the third-codon position and a composition dependency (Hammond, 1994). Furthermore, sequence patterns in *Sc. mansoni* have been studied and are characterized by similar AT biases in codon usage (Musto *et al.,* 1998). Also, codon usage patterns in the genome of *Onchocerca volvulus* have recently been investigated and are subject to bias (Unnasch and Williams, 2000). In contrast, relatively low levels of codon usage bias have been detected in sporozoan parasites. It is speculated that the codon usage patterns detected in these parasitic protozoa are the result of directional genomic mutation pressure (Ellis *et al.,* 1994). In addition, serine protease inhibitors (serpins) have recently been investigated in *Brugia malayi* in an effort to identify candidate vaccine proteins (Zang *et al.,* 1999).

Similarly, shared nonrandom dinucleotide patterns

[1] To whom correspondence should be addressed at Bioinformatics Supercomputing Centre, The Hospital for Sick Children, Suite 1300, 180 Dundas Street West, Toronto, M5G 1Z8 Ontario, Canada. Telephone: (416) 813-8428. Fax: (416) 813-6110. E-mail: afadiel@bioinfo.sickkids.on.ca.



0888-7543/01 $35.00
Copyright © 2001 by Academic Press
All rights of reproduction in any form reserved.

have been identified in DNA sequences corresponding to phylogenetically related species Russell *et al.*, 1976; Karlin and Ladunga, 1994). These sequence signatures, known as "general design" patterns (Russell *et al.*, 1976), are generated as a result of evolutionary pressure, methylation patterns, and mechanisms of DNA replication and repair. Furthermore, context-dependent mutation patterns, which influence the general composition of the genomic sequence, impact upon these dinucleotide regions (Russell *et al.*, 1976; Karlin *et al.*, 1997; Karlin and Mrazek, 1997). As a result, it is anticipated that intercodon dinucleotide pairs, namely, the wobble position nucleotide of one codon and the first position base in the following codon, would have an effect on codon choice (de Amicis and Marchetti, 2000).

Base composition has been found to influence both codon usage and gene function (Seetharaman and Srinivasan, 1995; Karlin and Mrazek, 2000). GC-rich genes tend to be of greater transcriptional and mitogenic significance than AT-rich genes, consistent with GC → AT mutational drift in methylated genomic regions. Moreover, in murine rodents coding exons tend to be rich in GC basepairs, primarily at the third-codon position, compared to introns (Hughes and Yeager, 1997). Third-base GC retention also identifies critical amino acids within individual proteins, as indicated by nonrandom patterns of codon variation between gene homologues and also by differential sequelae of site-directed mutagenesis. Amino acids corresponding to codons with G or C basepairs at the third position thus appear more tightly linked to cell function and survival than those encoded by codons with A or T at this site (Epstein *et al.*, 2000).

Sequences corresponding to highly expressed genes tend to make use of single optimal codons in the translation of specific amino acids (Sharp and Devine, 1989). In addition, genes possessing a codon usage pattern that varies from the mean but is similar to that of ribosomal protein genes are expressed at high levels (Karlin and Mrazek, 2000). This correlation between codon usage and gene expression (Sharp *et al.*, 1993) can be useful in the characterization of specific genes with respect to expression level and function (Mathe *et al.*, 1999). Moreover, this relationship is also useful in the identification of coding and noncoding regions (McLachlan *et al.*, 1984). Furthermore, it has been suggested that rare codons can be used for the regulation of specialized gene expression in bacteria (Saier, 1995; Sharp and Li, 1986). In addition, similar atypical codon usage among phylogenetically related organisms can be used to infer that genes have been acquired through horizontal transfer (Groisman *et al.*, 1992; Medigue *et al.*, 1991).

Codon usage analyses require the application of the codon adaptation index (CAI), a simple and effective measure of synonymous codon usage bias (Sharp and Li, 1987). The index uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for each gene is calculated from the overall codon usage frequency of that gene. Specifically, the index assesses the extent to which selection has been effective in molding the pattern of codon usage. The CAI has also been used to identify a gene family found in multiple organisms, which is associated with low codon usage bias (Bulmer, 1990). In addition, $N_c$, which represents the number of unique codons used per gene, can be used to investigate the codon usage patterns present within specific genes (Wright, 1990). This value represents the extent of codon preference used by a given gene and is plotted as a function of GC content at each codon position.

In some cases, the accuracy of synonymous codon usage statistics has been shown to decrease with decreasing gene length (Comeron and Aguade, 1998). Such a result was observed in the measurement of the scaled $\chi^2$ value, which estimates the departure of a given sequence from equal synonymous codon usage. A similar gene length dependence was observed with respect to the codon bias index (CBI) and the intrinsic codon bias index (ICDI). However, CAI and $N_c$ measurements, both relevant to this analysis, were not dependent upon gene length (Comeron and Aguade, 1998).

Correlations have often been made between codon usage and gene expression; furthermore, the amount of sequencing information currently available is unparalleled. Therefore, it is an opportune time to investigate gene expression levels and sequence usage in filarial nematodes. Studying codon and intercodon usage patterns will better our understanding of filarial evolution with respect to genetics (McWeeney and Valdes, 1999). In addition, *in silico* analyses of codon usage and base composition will provide valuable information for use in molecular investigations (Chen and Cheng, 1999). Moreover, the prediction of gene expression levels will clarify the mechanisms by which these parasites adapt to their host environment. In addition, this information will be invaluable in the design of primers for PCR and in the determination of exon boundaries (McInerney, 1998). Clearly, investigations into parasitic codon usage will benefit both the scientific and the medical communities.

## MATERIALS AND METHODS

Patterns of codon usage were investigated in five filarial parasitic species, namely *Wucheria bancrofti*, *B. malayi*, *O. volvulus*, *Acanthocheilonema viteae*, and *Dirofilaria immitis*. The organisms belong to five separate genera and are all members of the family Onchocercidae. Codon analysis was performed on data sets consisting of 240,816 bp making up 161 genes, coding for 80,272 amino acids. Complete coding regions were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/entrez/) and then filtered to exclude partial and redundant sequences. Of the 161 gene sequences analyzed, 10 were derived from *W. bancrofti*, 58 from *B. malayi*, 60 from *O. volvulus*, 5 from *Ac. viteae*, and 28 from *Di. immitis*.

Base composition and intercodon frequencies were investigated using freely available codon analysis programs. In particular, GCUA (McInerney, 1998) and ADE-4 (Perriere *et al.*, 1996) were used to

determine the base, codon, and amino acid number per gene. The ADE-4 package can be used through the Web (http://pbil.univlyon1.fr/mva/coa.html). In addition, GCUA was used to perform statistical analyses, namely, a relative synonymous codon usage analysis (RSCUA), as described by Sharp and Li (1986), and a factorial correspondence analysis (FCA), as described by Greenacre (1984). Furthermore, the program provided a measurement of the CAI (Sharp and Li 1987) and the number of unique codons used per gene ($N_c$) (Wright, 1990). To analyze intercodon frequencies, we designed a tool written in Perl, which recognizes DNA sequences saved in FASTA format. The intercodon base frequency per gene is measured as a function of the nucleotide present at the wobble position. The analytical output is then provided in text format to facilitate further analysis and interpretation using other programs. Amino acids tryptophan and methionine were ignored in all analyses; they are both encoded by a single codon, eliminating the possibility of synonymous codon interactions.

Base composition, intercodon frequency, and codon analyses were performed on an SGI Supercomputer with 4 GB of RAM and 1 TB of HD capacity. Results were then transferred to an IBM-compatible PC microcomputer equipped with 160 MB of RAM and a 20-GB hard drive running Microsoft Windows 98. Further statistical and graphical analyses were performed using SPSS and MS Excel, respectively.

## RESULTS

Genes corresponding to the genomes of five parasite species, *W. bancrofti*, *B. malayi*, *O. volvulus*, *Ac. viteae*, and *Di. immitis*, were analyzed with respect to base and codon usage, as well as amino acid composition. Selection biases were observed at all levels from DNA to amino acids among all organisms being studied.

Gene sequences were analyzed with respect to specific mononucleotide and dinucleotide usage. Among all five organisms studied, A and T were present at significantly higher proportions than G and C (Table 1). In addition, A was present at higher proportions than T, and G was significantly more common than C. Therefore, this indicates that among all of these organisms there is a bias for AT basepairs over GC nucleotides and a further bias for A over T and G over C.

Gene sequences were then analyzed with respect to base usage at each of the specific codon positions. In many cases, the third position, known as the wobble site, is that which varies among synonymous codons. Therefore, studies dealt mainly with base usage at this specific site. Among all organisms studied, A and T were found more often at this wobble position than G and C, with A more prevalent than T and G more prevalent than C (Fig. 1). Therefore, a codon-specific bias for AT over GC also appears to be present in the filarial coding sequences.

An analysis of the number of unique codons used per gene ($N_c$) was then performed, and results were plotted as a function of the GC frequency at the third-codon position (GC3). The $N_c$ value, defined as $2 + GC3 + (29/(GC3 + (1 - GC3)^2))$, is a measure of the codon usage specificity in each organism (Wright, 1990). Among all of the organisms studied, plots conformed to a reverse-parabolic distribution covering the entire data range (Fig. 2). Therefore, since the $N_c$ reached a maximum at high GC3 frequency instead of remaining linearly correlated, this indicates that there is a bias against the usage of codons with G or C at the wobble position when the GC3 frequency is high.

Most often, multiple synonymous codons are used to encode identical amino acids. If conditions were fully random, each synonymous trinucleotide would be used at an equal frequency. However, unequal synonymous codon usage would indicate codon bias. As a result, to evaluate codon composition, relative synonymous codon usage (RSCU) values per amino acid were examined for each of the five genetic data sets (Table 2). Values near 1 would indicate random codon usage, while values significantly greater or less than this threshold value would identify biased sequence usage. Codons used most often were rich in A or T at the wobble position. However, these codons possessed equal or lower AT frequency at positions 1 and 2. Specifically, in *W. bancrofti* the codon GCA possessed an RSCU value of 1.74, while the codon AUA possessed an RSCU of 0.80. A similar pattern among applicable codons was observed in all of the other organisms analyzed. Therefore, this indicates that the AT sequence bias was specific to the third-codon position and that a secondary bias might be present at the first two codon positions.

Recently, a clear influence of intercodon frequency on codon usage was described in plant genes (de Amicis and Marchetti, 2000). However, very little attention has been paid to investigate this influence in parasite genes. To fill this gap, base frequency at the intercodon site was analyzed as a function of the base at the wobble position for each filarial parasite. In general, GC basepairs were found less often at N4 than AT nucleotides. Furthermore, among the data sets corresponding to *W. bancrofti*, *O. volvulus*, *Ac. viteae*, and *Di. immitis*, A was present most often at the intercodon position in the presence of A, C, or G at the wobble site. However, A was much less common at N4 in the presence of T at N3 (Fig. 3). With respect to *B. malayi*, although A was found most often at N4 while C or G was present at N3, T was most common at N4 with A at N3. Among all species except for *Ac. viteae*, T was the most common nucleotide at N4 in the presence of T at N3. Intercodon results were found to be statistically significant ($P < 0.05$) through the use of a two-tailed unpaired heteroscedastic Student's $T$ test (Fig. 4).

An amino acid analysis was then performed for each of the sequence data sets. Specifically, each gene collection was analyzed with respect to the mean proportion of specific amino acids per gene. A general trend in amino acid usage was observed for *W. bancrofti*, *B. malayi*, and *O. volvulus*, with genes possessing proportionately high levels of leucine, lysine, isoleucine, and arginine and low levels of asparagine. With respect to *Ac. viteae*, a more gradual spectrum of amino acid usage was observed, with higher levels of glycine and valine. *Di. immitis* also deviated from the general pattern, possessing high levels of valine and glutamine.

**TABLE 1**

**Analysis of the Mean Nucleotide Proportions in Genes Corresponding to *W. bancrofti*, *B. malayi*, *O. volvulus*, *Ac. viteae*, and *Di. immitis***

| Accession No. | Gene name | Length (AA) | Length (bp) | Base proportion per gene | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | A | T | C | G | W (AT) | S (GC) |
| *W. bancrofti* | | | | | | | | | |
| gi\|9625063 | Cuticulin-1 mRNA | 432 | 1296 | 0.318 | 0.330 | 0.180 | 0.172 | 0.648 | 0.352 |
| gi\|7673687 | Cuticular endochitinase mRNA | 544 | 1632 | 0.324 | 0.263 | 0.230 | 0.183 | 0.587 | 0.413 |
| gi\|7673685 | Heat-shock protein-70 mRNA | 723 | 2169 | 0.289 | 0.279 | 0.237 | 0.195 | 0.568 | 0.432 |
| gi\|6708155 | Actin gene | 896 | 2688 | 0.257 | 0.362 | 0.195 | 0.187 | 0.618 | 0.382 |
| gi\|3599466 | Abundant larval transcript-2 protein (alt-2) mRNA | 128 | 384 | 0.323 | 0.232 | 0.276 | 0.169 | 0.555 | 0.445 |
| gi\|5882243 | GTP-binding protein mRNA | 257 | 771 | 0.305 | 0.293 | 0.209 | 0.193 | 0.598 | 0.402 |
| gi\|4324679 | Vespin allergen antigen homologue (VAH) mRNA | 281 | 843 | 0.332 | 0.275 | 0.210 | 0.183 | 0.607 | 0.393 |
| gi\|2789663 | MIF mRNA | 182 | 546 | 0.289 | 0.333 | 0.212 | 0.165 | 0.623 | 0.377 |
| gi\|3152687 | Antigen WB14 mRNA | 235 | 705 | 0.352 | 0.281 | 0.186 | 0.182 | 0.633 | 0.367 |
| gi\|162578 | Ribosomal protein S13 gene | 152 | 456 | 0.298 | 0.246 | 0.232 | 0.224 | 0.544 | 0.456 |
| *B. malayi* | | | | | | | | | |
| gi\|4406215 | TGF-$\beta$ homologue mRNA | 411 | 1233 | 0.275 | 0.234 | 0.247 | 0.243 | 0.509 | 0.491 |
| gi\|10086322 | Transcription factor DP1 mRNA | 381 | 1143 | 0.318 | 0.269 | 0.217 | 0.195 | 0.588 | 0.412 |
| gi\|9957282 | FABPDK mRNA | 212 | 636 | 0.355 | 0.278 | 0.193 | 0.173 | 0.634 | 0.366 |
| gi\|9717248 | mif-2 mRNA | 144 | 432 | 0.326 | 0.313 | 0.190 | 0.171 | 0.639 | 0.361 |
| gi\|1850558 | mif-1 mRNA | 189 | 567 | 0.314 | 0.323 | 0.210 | 0.153 | 0.637 | 0.363 |
| gi\|1814369 | alt-2 mRNA | 180 | 540 | 0.319 | 0.261 | 0.235 | 0.185 | 0.580 | 0.420 |
| gi\|7596931 | VAH mRNA | 292 | 876 | 0.361 | 0.269 | 0.195 | 0.175 | 0.630 | 0.370 |
| gi\|7159327 | Galectin mRNA | 280 | 840 | 0.329 | 0.264 | 0.240 | 0.167 | 0.593 | 0.407 |
| gi\|2305209 | Serpin precursor mRNA | 453 | 1359 | 0.347 | 0.291 | 0.185 | 0.177 | 0.638 | 0.362 |
| gi\|6646877 | Thioredoxin peroxidase mRNA | 213 | 639 | 0.277 | 0.275 | 0.241 | 0.207 | 0.552 | 0.448 |
| gi\|1518124 | Small heat-shock protein mRNA | 228 | 684 | 0.361 | 0.289 | 0.165 | 0.184 | 0.651 | 0.349 |
| gi\|6434852 | FKBP12 mRNA | 115 | 345 | 0.304 | 0.249 | 0.261 | 0.186 | 0.554 | 0.446 |
| gi\|5759312 | cpi-2 gene | 518 | 1554 | 0.330 | 0.357 | 0.184 | 0.129 | 0.687 | 0.313 |
| gi\|5759310 | cpi-1 gene | 341 | 1023 | 0.360 | 0.297 | 0.160 | 0.183 | 0.657 | 0.343 |
| gi\|5616167 | her-1 gene | 756 | 2268 | 0.340 | 0.340 | 0.165 | 0.154 | 0.680 | 0.320 |
| gi\|5107085 | Small zinc finger-like protein mRNA | 102 | 306 | 0.291 | 0.255 | 0.216 | 0.239 | 0.546 | 0.454 |
| gi\|3721893 | BrJTB mRNA | 327 | 981 | 0.243 | 0.212 | 0.277 | 0.268 | 0.455 | 0.545 |
| gi\|4102826 | FKBP13 mRNA | 177 | 531 | 0.311 | 0.262 | 0.249 | 0.179 | 0.573 | 0.427 |
| gi\|3264825 | 24-kDa secreted protein mRNA | 363 | 1089 | 0.336 | 0.263 | 0.193 | 0.208 | 0.599 | 0.401 |
| gi\|4097222 | $\gamma$-Glutamyl transpeptidase precursor mRNA | 563 | 1689 | 0.299 | 0.314 | 0.207 | 0.179 | 0.613 | 0.387 |
| gi\|1813697 | Cytidine deaminase mRNA | 157 | 471 | 0.346 | 0.278 | 0.208 | 0.168 | 0.624 | 0.376 |
| gi\|1813691 | Ribosomal protein P2 mRNA | 151 | 453 | 0.305 | 0.285 | 0.232 | 0.179 | 0.589 | 0.411 |
| gi\|1813689 | Ribosomal protein L44 mRNA | 144 | 432 | 0.361 | 0.236 | 0.238 | 0.164 | 0.597 | 0.403 |
| gi\|1813685 | bm-alt-3 mRNA | 139 | 417 | 0.341 | 0.379 | 0.173 | 0.108 | 0.719 | 0.281 |
| gi\|1813683 | Putative RNA-binding protein mRNA | 184 | 552 | 0.230 | 0.328 | 0.245 | 0.197 | 0.558 | 0.442 |
| gi\|1813681 | Cystatin mRNA | 160 | 480 | 0.392 | 0.246 | 0.188 | 0.175 | 0.638 | 0.363 |
| gi\|1813679 | Tumor protein homologue mRNA | 219 | 657 | 0.321 | 0.272 | 0.256 | 0.151 | 0.594 | 0.406 |
| gi\|1754683 | Ribosomal protein S23 mRNA | 173 | 519 | 0.316 | 0.268 | 0.245 | 0.171 | 0.584 | 0.416 |
| gi\|1480790 | Glia maturation factor gene | 763 | 2289 | 0.280 | 0.359 | 0.191 | 0.170 | 0.639 | 0.361 |
| gi\|1322371 | alt-1 mRNA | 168 | 504 | 0.339 | 0.258 | 0.228 | 0.175 | 0.597 | 0.403 |
| gi\|1480460 | Cyclophilin Bmcyp-2 mRNA | 216 | 648 | 0.326 | 0.313 | 0.213 | 0.148 | 0.639 | 0.361 |
| gi\|392787 | Intermediate filament protein mRNA | 532 | 1596 | 0.348 | 0.241 | 0.226 | 0.185 | 0.589 | 0.411 |
| gi\|563236 | Cuticular collagen Bmcol-2 mRNA | 358 | 1074 | 0.302 | 0.223 | 0.257 | 0.218 | 0.525 | 0.475 |
| gi\|460245 | a2 (IV) basement membrane collagen gene | 4102 | 12306 | 0.322 | 0.321 | 0.193 | 0.164 | 0.643 | 0.357 |
| gi\|984561 | Peptidylprolyl isomerase mRNA | 960 | 2880 | 0.355 | 0.235 | 0.241 | 0.169 | 0.591 | 0.409 |
| gi\|1002821 | Bm-tpx-1 mRNA | 295 | 885 | 0.285 | 0.324 | 0.207 | 0.184 | 0.609 | 0.391 |
| gi\|1679785 | Paramyosin mRNA | 784 | 2352 | 0.367 | 0.223 | 0.218 | 0.192 | 0.591 | 0.409 |
| gi\|156069 | Hsp70 gene | 1492 | 4476 | 0.275 | 0.313 | 0.224 | 0.188 | 0.588 | 0.412 |
| gi\|510683 | Vinculin mRNA | 1096 | 3288 | 0.309 | 0.243 | 0.233 | 0.216 | 0.551 | 0.449 |
| gi\|2642218 | Bmshp3a and Bmshp3 genes | 3335 | 10005 | 0.337 | 0.352 | 0.148 | 0.162 | 0.690 | 0.310 |
| gi\|2454545 | Bm-tgh-1 gene | 1015 | 3045 | 0.329 | 0.334 | 0.163 | 0.174 | 0.664 | 0.336 |
| gi\|2384712 | Cystatin-type cysteine proteinase inhibitor mRNA | 212 | 636 | 0.333 | 0.314 | 0.209 | 0.143 | 0.648 | 0.352 |
| gi\|2347060 | Bm-tpx-2 mRNA | 214 | 642 | 0.307 | 0.330 | 0.202 | 0.160 | 0.637 | 0.363 |
| gi\|2199569 | Her-1 mRNA | 259 | 777 | 0.347 | 0.296 | 0.181 | 0.175 | 0.644 | 0.356 |
| gi\|2190975 | bmmif gene | 298 | 894 | 0.308 | 0.332 | 0.210 | 0.150 | 0.640 | 0.360 |
| gi\|433399 | BmSERPIN mRNA | 401 | 1203 | 0.328 | 0.293 | 0.194 | 0.185 | 0.621 | 0.379 |

**TABLE 1**—*Continued*

| Accession No. | Gene name | Length (AA) | Length (bp) | Base proportion per gene | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | A | T | C | G | W (AT) | S (GC) |
| gi\|1562571 | bm20 mRNA | 177 | 531 | 0.382 | 0.235 | 0.190 | 0.192 | 0.618 | 0.382 |
| gi\|1518126 | 60S ribosomal protein mRNA | 167 | 501 | 0.279 | 0.263 | 0.273 | 0.184 | 0.543 | 0.457 |
| gi\|1305492 | Microfilarial sheath protein SHP3a mRNA | 202 | 606 | 0.304 | 0.323 | 0.158 | 0.215 | 0.627 | 0.373 |
| gi\|619942 | BmNDK mRNA | 197 | 591 | 0.352 | 0.264 | 0.208 | 0.176 | 0.616 | 0.384 |
| gi\|1155359 | shp3a mRNA | 217 | 651 | 0.329 | 0.217 | 0.207 | 0.247 | 0.545 | 0.455 |
| gi\|156086 | Myosin heavy chain gene | 3744 | 11232 | 0.335 | 0.295 | 0.195 | 0.175 | 0.630 | 0.370 |
| gi\|863005 | ORF1 mRNA | 560 | 1680 | 0.295 | 0.310 | 0.182 | 0.214 | 0.605 | 0.395 |
| gi\|603210 | BmG3PD mRNA | 431 | 1293 | 0.281 | 0.299 | 0.221 | 0.200 | 0.579 | 0.421 |
| gi\|453471 | γ-glutamyltransferase gene | 580 | 1740 | 0.264 | 0.302 | 0.229 | 0.205 | 0.566 | 0.434 |
| gi\|156071 | Filarial antigen mRNA | 459 | 1377 | 0.413 | 0.212 | 0.219 | 0.155 | 0.625 | 0.375 |
| gi\|156063 | Chitinase mRNA | 548 | 1644 | 0.339 | 0.262 | 0.230 | 0.170 | 0.600 | 0.400 |
| gi\|156052 | 63-kDa antigen mRNA | 564 | 1692 | 0.300 | 0.285 | 0.241 | 0.174 | 0.585 | 0.415 |
| gi\|475860 | Dg2 gene for D34 immunodominant antigen | 1534 | 4602 | 0.309 | 0.275 | 0.230 | 0.186 | 0.584 | 0.416 |
| gi\|7159325 | Galectin mRNA | 281 | 843 | 0.331 | 0.260 | 0.234 | 0.176 | 0.591 | 0.409 |
| gi\|7159289 | Intermediate filament protein mRNA | 552 | 1656 | 0.339 | 0.248 | 0.229 | 0.184 | 0.586 | 0.414 |
| gi\|7159287 | Putative gut-associated protein mRNA | 212 | 636 | 0.277 | 0.347 | 0.197 | 0.179 | 0.624 | 0.376 |
| gi\|6646875 | Thioredoxin peroxidase mRNA | 242 | 726 | 0.322 | 0.300 | 0.201 | 0.176 | 0.623 | 0.377 |
| gi\|6635936 | Asparaginase mRNA | 644 | 1932 | 0.319 | 0.312 | 0.202 | 0.167 | 0.631 | 0.369 |
| gi\|3273481 | Transglutaminase precursor mRNA | 590 | 1770 | 0.341 | 0.295 | 0.209 | 0.155 | 0.636 | 0.364 |
| gi\|4102824 | FKBP13 mRNA | 165 | 495 | 0.305 | 0.295 | 0.222 | 0.178 | 0.600 | 0.400 |
| gi\|3253096 | P22U mRNA | 321 | 963 | 0.333 | 0.334 | 0.178 | 0.155 | 0.668 | 0.332 |
| gi\|4115902 | Calreticulin precursor mRNA | 504 | 1512 | 0.370 | 0.273 | 0.212 | 0.145 | 0.643 | 0.357 |
| gi\|1480462 | Dicyp-2 mRNA | 206 | 618 | 0.330 | 0.290 | 0.228 | 0.152 | 0.620 | 0.380 |
| gi\|1373004 | Pepsin inhibitor precursor mRNA | 260 | 780 | 0.378 | 0.272 | 0.171 | 0.179 | 0.650 | 0.350 |
| gi\|1144144 | Larval 20/22-kDa protein mRNA | 192 | 576 | 0.361 | 0.283 | 0.179 | 0.177 | 0.644 | 0.356 |
| gi\|3057039 | Dicyp-3 mRNA | 562 | 1686 | 0.372 | 0.256 | 0.200 | 0.172 | 0.628 | 0.372 |
| gi\|2347118 | Thioredoxin peroxidase mRNA | 230 | 690 | 0.306 | 0.307 | 0.194 | 0.193 | 0.613 | 0.387 |
| gi\|3046902 | β-Tubulin mRNA | 441 | 1323 | 0.284 | 0.284 | 0.234 | 0.197 | 0.568 | 0.432 |
| gi\|2598121 | 1-cys peroxidoxin mRNA | 245 | 735 | 0.317 | 0.302 | 0.200 | 0.181 | 0.619 | 0.381 |
| gi\|2245507 | Venom allergen antigen 5-like protein mRNA | 264 | 792 | 0.354 | 0.311 | 0.197 | 0.139 | 0.664 | 0.336 |
| gi\|2209363 | Cytosolic Cu-Zn superoxide dismutase mRNA | 153 | 459 | 0.255 | 0.322 | 0.237 | 0.185 | 0.577 | 0.423 |
| gi\|2149474 | Glutathione peroxidase mRNA | 272 | 816 | 0.359 | 0.304 | 0.156 | 0.181 | 0.663 | 0.337 |
| gi\|1206024 | Low-molecular-weight heat shock protein p27 mRNA | 264 | 792 | 0.337 | 0.264 | 0.194 | 0.205 | 0.601 | 0.399 |
| gi\|905396 | Aspartyl protease inhibitor homologue mRNA | 232 | 696 | 0.361 | 0.274 | 0.174 | 0.191 | 0.635 | 0.365 |
| gi\|555946 | Extracellular superoxide dismutase mRNA | 197 | 591 | 0.301 | 0.347 | 0.181 | 0.171 | 0.648 | 0.352 |
| gi\|541629 | Glutathione *S*-transferase mRNA | 228 | 684 | 0.313 | 0.294 | 0.194 | 0.199 | 0.607 | 0.393 |
| gi\|452448 | Glutathione lipid hydroperoxidase mRNA | 264 | 792 | 0.343 | 0.312 | 0.158 | 0.187 | 0.655 | 0.345 |

Therefore, it appears as if the bias in base and codon usage does extend to the amino acid level.

Genes encoding proteins found to have either biased or unbiased amino acid levels were identified and characterized. Genes subject to significant amino acid bias encoded proteins such as Hsp70, abundant larval transcript peptides (Alt-1, Alt-2), and chitinase, all expressed at high levels during parasite development (Karlin and Brocchieri, 1998; Gregory *et al.*, 2000; Wu *et al.*, 1996). Genes with amino acid proportions similar to the mean also corresponded to highly expressed proteins, namely thioredoxin peroxidase, which is present throughout the parasite life cycle (Lu *et al.*, 1998). Therefore, results do not show that codon usage is associated with gene expression level.

The CAI was then calculated for each sequence. This important statistical factor provides a measure of gene expression levels relative to those of a specific control species, in this case *Dr. melanogaster*. The value is defined as the mean of the RSCU values corresponding to each codon used in the gene, divided by the maximum potential CAI that could correspond to a gene of identical amino acid composition (Sharp and Li, 1987). Therefore, based upon knowledge of the relative expression levels of each codon, values closest to 1 signify high levels of expression, while those approaching 0 are associated with lower levels of expression. The CAI values corresponding to each of the genes from the data set ranged from 0 to 0.2, indicating that none is expressed at a relatively high level.

## DISCUSSION

Nucleotide usage patterns play an integral role in the characterization of novel gene sequences. Primar-

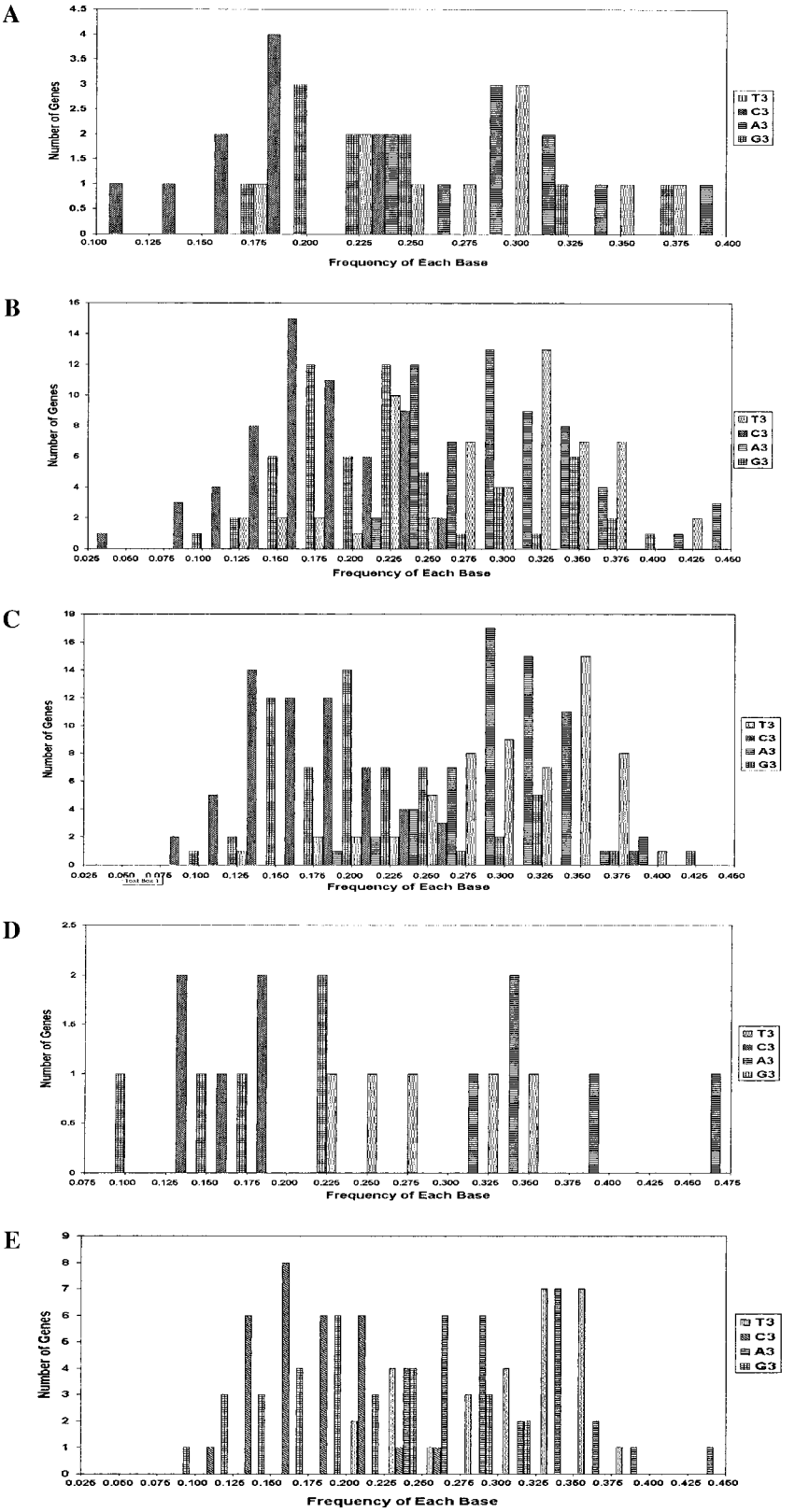


**FIG. 1.** Relative proportion of bases at the third-codon position in the genes of five common nematode species. The *X*-axis represents the frequency of respective bases, and the *Y*-axis indicates the number of genes with each base at the given frequency. (**A**) *W. bancrofti,* (**B**) *B. malayi,* (**C**) *O. volvulus,* (**D**) *Ac. viteae,* and (**E**) *Di. immitis.*

ily, correlations have been made between codon usage bias and levels of gene expression (Sharp *et al.,* 1993). Furthermore, horizontal gene transfer events among similar species have often been identified through the use of codon usage patterns (Groisman *et al.,* 1992; Medigue *et al.,* 1991). In addition, patterns of nucleo-

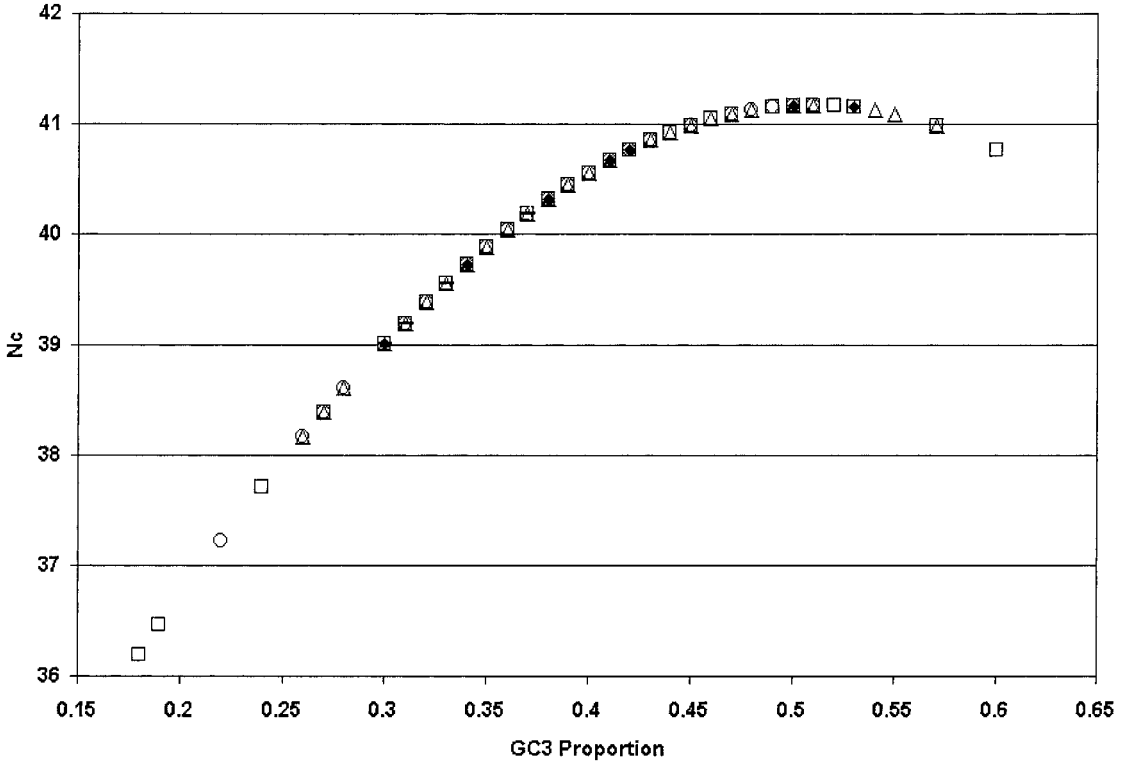


**FIG. 2.** Plot of the mean number of unique codons used per gene ($N_c$) as a function of GC3 proportion among specific nematode species. (◆) *W. bancrofti,* (□) *B. malayi,* (△) *O. volvulus,* (—) *Ac. viteae,* and (○) *Di. immitis.*

tide usage have been used within the medical community to select antigenic proteins for use as vaccines in the treatment of parasitic infection (Gregory *et al.*, 2000). Clearly, as large amounts of sequencing data become available, patterns of nucleotide usage will be of great importance in the definition and functional investigation of coding regions.

Base composition analyses revealed that the filarial gene sequences were AT-rich. Specifically, A was the most common nucleotide, while C was the least prevalent. A similar pattern has also been observed in *Taenia* sp., *Entamoeba histolytica,* and *Plasmodium falciparum* (Waterkeyn *et al.*, 1998; Ghosh *et al.*, 2000; Musto *et al.*, 1999), in which genomes are highly AT-rich. In addition, AT basepairs were more common at the third-codon position than GC basepairs in all of the organisms studied. This sequence bias could have been due to the base composition of the data set. Since the gene sequences were all AT-rich, through simple probability, codons ending in A or T would be more likely to arise than those ending in G or C. However, A was found to be more common than T and G more common than C. Therefore, this indicates that in addition to simple base composition, some other factor was selecting for codons with A at the third-codon position and against codons with C at this site. Such a factor would be able to generate a definite sequence bias over time through progressive natural selection.

Plots of $N_c$ as a function of GC3 were reverse-parabolic, indicating that $N_c$ increased as a function of GC3 but then reached a maximum. This limit suggests that at low GC3 frequency, both AT3 and GC3 codons were being used equally, while under high GC3 conditions, AT3 codons were being chosen more often. Therefore, selection was likely taking place, driven by both the genomic base composition and evolutionary pressure. This is in contrast to results that have been observed for *Taenia,* in which low GC3 proportions are associated with all $N_c$ values (Waterkeyn *et al.*, 1998).

Among all of the organisms considered, codons with A or T at the third position were found in significantly higher proportions than those with G or C at this site. However, the base composition at codon sites 1 and 2 was not significantly biased for AT basepairs. Therefore, it seems likely that the wobble position (N3) is the main site that is subject to an AT bias. Whether there is also a bias for GC basepairs at positions 1 and 2 was not considered; such a condition has never been observed in previous analyses with other similar organisms such as *Taenia sp., E. histolytica,* and *P. falciparum* (Waterkeyn *et al.*, 1998; Ghosh *et al.*, 2000; Musto *et al.*, 1999). Therefore, a bias for AT basepairs at N3 appears to exist among the five filarial parasites.

*W. bancrofti, B. malayi,* and *O. volvulus* possessed genes encoding proteins consisting of high proportions of lysine, leucine, arginine, and isoleucine. In contrast, *Ac. viteae* genes encoded proteins rich in threonine, glutamate, and glycine. *Di. immitis* also varied from

**TABLE 2**

**Analysis of the Mean Codon Frequency and Relative Synonymous Codon Usage (RSCU) in Genes Corresponding to *W. bancrofti*, *B. malayi*, *O. volvulus*, *Ac. viteae*, and *Di. immitis***

| Organism | AA | Codon | *N* | RSCU | AA | Codon | *N* | RSCU |
|---|---|---|---|---|---|---|---|---|
| *W. bancrofti* | Phe | UUU | 214 | 1.32 | Ser | UCU | 92 | 1.14 |
| | | UUC | 110 | 0.68 | | UCC | 72 | 0.89 |
| | Leu | UUA | 143 | 1.29 | | UCA | 110 | 1.36 |
| | | UUG | 185 | 1.66 | | UCG | 72 | 0.89 |
| | Tyr | UAU | 131 | 1.22 | | AGU | 76 | 0.94 |
| | | UAC | 84 | 0.78 | | AGC | 64 | 0.79 |
| | ter | UAA | 114 | 0.00 | Cys | UGU | 112 | 1.07 |
| | ter | UAG | 46 | 0.00 | | UGC | 97 | 0.93 |
| | Leu | CUU | 103 | 0.93 | ter | UGA | 134 | 0.00 |
| | | CUC | 50 | 0.45 | Trp | UGG | 115 | 1.00 |
| | | CUA | 75 | 0.67 | Pro | CCU | 45 | 1.00 |
| | | CUG | 111 | 1.00 | | CCC | 32 | 0.71 |
| | His | CAU | 107 | 1.13 | | CCA | 61 | 1.36 |
| | | CAC | 83 | 0.87 | | CCG | 42 | 0.93 |
| | Gln | CAA | 126 | 1.18 | Arg | CGU | 66 | 0.89 |
| | | CAG | 87 | 0.82 | | CGC | 47 | 0.64 |
| | Ile | AUU | 199 | 1.39 | | CGA | 70 | 0.95 |
| | | AUC | 115 | 0.8 | | CGG | 60 | 0.81 |
| | | AUA | 115 | 0.8 | Arg | AGA | 122 | 1.65 |
| | Met | AUG | 147 | 1.00 | | AGG | 78 | 1.06 |
| | Asn | AAU | 148 | 1.15 | Thr | ACU | 68 | 1.08 |
| | | AAC | 110 | 0.85 | | ACC | 51 | 0.81 |
| | Lys | AAA | 251 | 1.29 | | ACA | 87 | 1.38 |
| | | AAG | 138 | 0.71 | | ACG | 47 | 0.74 |
| | Val | GUU | 132 | 1.25 | Ala | GCU | 62 | 1.04 |
| | | GUC | 69 | 0.65 | | GCC | 30 | 0.50 |
| | | GUA | 93 | 0.88 | | GCA | 104 | 1.74 |
| | | GUG | 128 | 1.21 | | GCG | 43 | 0.72 |
| | Asp | GAU | 107 | 1.20 | Gly | GGU | 98 | 1.49 |
| | | GAC | 71 | 0.80 | | GGC | 44 | 0.67 |
| | Glu | GAA | 149 | 1.35 | | GGA | 91 | 1.38 |
| | | GAG | 71 | 0.65 | | GGG | 30 | 0.46 |
| *B. malayi* | Phe | UUU | 1670 | 1.36 | Ser | UCU | 452 | 0.93 |
| | | UUC | 777 | 0.64 | | UCC | 353 | 0.73 |
| | Leu | UUA | 982 | 1.63 | | UCA | 743 | 1.53 |
| | | UUG | 846 | 1.40 | | UCG | 344 | 0.71 |
| | Tyr | UAU | 932 | 1.33 | | AGU | 536 | 1.11 |
| | | UAC | 472 | 0.67 | | AGC | 478 | 0.99 |
| | ter | UAA | 957 | 0.00 | Cys | UGU | 640 | 1.02 |
| | ter | UAG | 378 | 0.00 | | UGC | 613 | 0.98 |
| | Leu | CUU | 598 | 0.99 | ter | UGA | 931 | 0.00 |
| | | CUC | 280 | 0.46 | Trp | UGG | 646 | 1.00 |
| | | CUA | 379 | 0.63 | Pro | CCU | 260 | 1.03 |
| | | CUG | 528 | 0.88 | | CCC | 105 | 0.42 |
| | His | CAU | 630 | 1.25 | | CCA | 428 | 1.69 |
| | | CAC | 377 | 0.75 | | CCG | 218 | 0.86 |
| | Gln | CAA | 996 | 1.24 | Arg | CGU | 291 | 0.78 |
| | | CAG | 610 | 0.76 | | CGC | 176 | 0.47 |
| | Ile | AUU | 1356 | 1.45 | | CGA | 405 | 1.08 |
| | | AUC | 587 | 0.63 | | CGG | 218 | 0.58 |
| | | AUA | 862 | 0.92 | | AGA | 688 | 1.83 |
| | Met | AUG | 888 | 1.00 | | AGG | 472 | 1.26 |
| | Asn | AAU | 1328 | 1.31 | Thr | ACU | 481 | 1.07 |
| | | AAC | 693 | 0.69 | | ACC | 351 | 0.78 |
| | Lys | AAA | 1979 | 1.35 | | ACA | 641 | 1.42 |
| | | AAG | 956 | 0.65 | | ACG | 330 | 0.73 |
| | Val | GUU | 679 | 1.41 | Ala | GCU | 430 | 1.23 |
| | | GUC | 283 | 0.59 | | GCC | 213 | 0.61 |
| | | GUA | 481 | 1.00 | | GCA | 531 | 1.51 |
| | | GUG | 479 | 1.00 | | GCG | 228 | 0.65 |
| | Asp | GAU | 692 | 1.36 | Gly | GGU | 466 | 1.26 |
| | | GAC | 322 | 0.64 | | GGC | 225 | 0.61 |
| | Glu | GAA | 1009 | 1.43 | | GGA | 587 | 1.59 |
| | | GAG | 406 | 0.57 | | GGG | 202 | 0.55 |

**TABLE 2**—*Continued*

| Organism | AA | Codon | *N* | RSCU | AA | Codon | *N* | RSCU |
|---|---|---|---|---|---|---|---|---|
| *O. volvulus* | Phe | UUU | 1175 | 1.27 | Ser | UCU | 448 | 1.05 |
| | | UUC | 675 | 0.73 | | UCC | 285 | 0.67 |
| | Leu | UUA | 703 | 1.47 | | UCA | 641 | 1.5 |
| | | UUG | 803 | 1.68 | | UCG | 437 | 1.02 |
| | Tyr | UAU | 771 | 1.27 | | AGU | 367 | 0.86 |
| | | UAC | 443 | 0.73 | | AGC | 381 | 0.89 |
| | ter | UAA | 609 | 0.00 | Cys | UGU | 463 | 1.03 |
| | ter | UAG | 254 | 0.00 | | UGC | 440 | 0.97 |
| | Leu | CUU | 482 | 1.01 | ter | UGA | 734 | 0.00 |
| | | CUC | 214 | 0.45 | Trp | UGG | 492 | 1.00 |
| | | CUA | 283 | 0.59 | Pro | CCU | 184 | 0.90 |
| | | CUG | 380 | 0.80 | | CCC | 93 | 0.45 |
| | His | CAU | 589 | 1.39 | | CCA | 348 | 1.69 |
| | | CAC | 260 | 0.61 | | CCG | 197 | 0.96 |
| | Gln | CAA | 810 | 1.31 | Arg | CGU | 333 | 1.07 |
| | | CAG | 425 | 0.69 | | CGC | 170 | 0.54 |
| | Ile | AUU | 1150 | 1.40 | | CGA | 407 | 1.3 |
| | | AUC | 620 | 0.75 | | CGG | 198 | 0.63 |
| | | AUA | 703 | 0.85 | | AGA | 517 | 1.66 |
| | Met | AUG | 840 | 1.00 | | AGG | 249 | 0.80 |
| | Asn | AAU | 1259 | 1.34 | Thr | ACU | 388 | 0.96 |
| | | AAC | 627 | 0.66 | | ACC | 313 | 0.77 |
| | Lys | AAA | 1817 | 1.39 | | ACA | 614 | 1.52 |
| | | AAG | 792 | 0.61 | | ACG | 304 | 0.75 |
| | Val | GUU | 585 | 1.39 | Ala | GCU | 377 | 1.17 |
| | | GUC | 260 | 0.62 | | GCC | 179 | 0.56 |
| | | GUA | 397 | 0.94 | | GCA | 492 | 1.53 |
| | | GUG | 445 | 1.06 | | GCG | 236 | 0.74 |
| | Asp | GAU | 811 | 1.49 | Gly | GGU | 432 | 1.30 |
| | | GAC | 281 | 0.51 | | GGC | 235 | 0.71 |
| | Glu | GAA | 1168 | 1.53 | | GGA | 519 | 1.57 |
| | | GAG | 362 | 0.47 | | GGG | 139 | 0.42 |
| *Ac. viteae* | Phe | UUU | 73 | 1.26 | Ser | UCU | 25 | 0.68 |
| | | UUC | 43 | 0.74 | | UCC | 20 | 0.54 |
| | Leu | UUA | 47 | 1.59 | | UCA | 61 | 1.66 |
| | | UUG | 39 | 1.32 | | UCG | 41 | 1.11 |
| | Tyr | UAU | 77 | 1.48 | | AGU | 42 | 1.14 |
| | | UAC | 27 | 0.52 | | AGC | 32 | 0.87 |
| | ter | UAA | 25 | 0.00 | Cys | UGU | 35 | 1.01 |
| | ter | UAG | 5 | 0.00 | | UGC | 34 | 0.99 |
| | Leu | CUU | 49 | 1.66 | ter | UGA | 37 | 0.00 |
| | | CUC | 8 | 0.27 | Trp | UGG | 39 | 1.00 |
| | | CUA | 15 | 0.51 | Pro | CCU | 18 | 0.69 |
| | | CUG | 19 | 0.64 | | CCC | 12 | 0.46 |
| | His | CAU | 37 | 1.23 | | CCA | 39 | 1.49 |
| | | CAC | 23 | 0.77 | | CCG | 36 | 1.37 |
| | Gln | CAA | 65 | 1.55 | Arg | CGU | 22 | 0.87 |
| | | CAG | 19 | 0.45 | | CGC | 13 | 0.52 |
| | Ile | AUU | 74 | 1.45 | | CGA | 41 | 1.63 |
| | | AUC | 32 | 0.63 | | CGG | 13 | 0.52 |
| | | AUA | 47 | 0.92 | | AGA | 40 | 1.59 |
| | Met | AUG | 56 | 1.00 | | AGG | 22 | 0.87 |
| | Asn | AAU | 91 | 1.42 | Thr | ACU | 53 | 0.97 |
| | | AAC | 37 | 0.58 | | ACC | 40 | 0.73 |
| | Lys | AAA | 168 | 1.62 | | ACA | 81 | 1.49 |
| | | AAG | 40 | 0.38 | | ACG | 44 | 0.81 |
| | Val | GUU | 47 | 1.59 | Ala | GCU | 41 | 1.1 |
| | | GUC | 22 | 0.75 | | GCC | 24 | 0.64 |
| | | GUA | 29 | 0.98 | | GCA | 62 | 1.66 |
| | | GUG | 20 | 0.68 | | GCG | 22 | 0.59 |
| | Asp | GAU | 103 | 1.57 | Gly | GGU | 63 | 1.22 |
| | | GAC | 28 | 0.43 | | GGC | 43 | 0.83 |
| | Glu | GAA | 146 | 1.49 | | GGA | 85 | 1.65 |
| | | GAG | 50 | 0.51 | | GGG | 15 | 0.29 |

**TABLE 2**—*Continued*

| Organism | AA | Codon | *N* | RSCU | AA | Codon | *N* | RSCU |
|---|---|---|---|---|---|---|---|---|
| *Di. immitis* | Phe | UUU | 552 | 1.28 | Ser | UCU | 169 | 0.98 |
| | | UUC | 309 | 0.72 | | UCC | 137 | 0.80 |
| | Leu | UUA | 312 | 1.44 | | UCA | 273 | 1.59 |
| | | UUG | 355 | 1.64 | | UCG | 147 | 0.85 |
| | Tyr | UAU | 387 | 1.4 | Ser | AGU | 162 | 0.94 |
| | | UAC | 167 | 0.6 | | AGC | 144 | 0.84 |
| | ter | UAA | 297 | 0.00 | Cys | UGU | 232 | 1.13 |
| | ter | UAG | 134 | 0.00 | | UGC | 179 | 0.87 |
| | Leu | CUU | 218 | 1.01 | ter | UGA | 359 | 0.00 |
| | | CUC | 102 | 0.47 | Trp | UGG | 280 | 1.00 |
| | | CUA | 145 | 0.67 | Pro | CCU | 82 | 0.83 |
| | | CUG | 164 | 0.76 | | CCC | 43 | 0.44 |
| | His | CAU | 269 | 1.27 | | CCA | 178 | 1.81 |
| | | CAC | 154 | 0.73 | | CCG | 90 | 0.92 |
| | Gln | CAA | 361 | 1.36 | Arg | CGU | 152 | 1.07 |
| | | CAG | 170 | 0.64 | | CGC | 71 | 0.50 |
| | Ile | AUU | 520 | 1.32 | | CGA | 156 | 1.10 |
| | | AUC | 292 | 0.74 | | CGG | 93 | 0.65 |
| | | AUA | 370 | 0.94 | Arg | AGA | 241 | 1.69 |
| | Met | AUG | 393 | 1.00 | | AGG | 141 | 0.99 |
| | Asn | AAU | 461 | 1.31 | Thr | ACU | 159 | 0.92 |
| | | AAC | 245 | 0.69 | | ACC | 124 | 0.71 |
| | Lys | AAA | 775 | 1.39 | | ACA | 273 | 1.57 |
| | | AAG | 344 | 0.61 | | ACG | 139 | 0.80 |
| | Val | GUU | 282 | 1.41 | Ala | GCU | 134 | 0.99 |
| | | GUC | 122 | 0.61 | | GCC | 81 | 0.60 |
| | | GUA | 177 | 0.89 | | GCA | 224 | 1.65 |
| | | GUG | 217 | 1.09 | | GCG | 105 | 0.77 |
| | Asp | GAU | 314 | 1.42 | Gly | GGU | 169 | 1.19 |
| | | GAC | 129 | 0.58 | | GGC | 128 | 0.90 |
| | Glu | GAA | 450 | 1.43 | | GGA | 183 | 1.29 |
| | | GAG | 179 | 0.57 | | GGG | 89 | 0.63 |

*Note.* The RSCU value is a proportionality factor, which has been corrected for gene length. Therefore, it is this value that may be considered in a more general fashion.

this general pattern, with genes encoding proteins rich in valine and glutamine. This makes intuitive sense in that many of the implicated amino acids are encoded by codons rich in A and T, namely lysine (AAA), leucine (CUU), isoleucine (AUU), and valine (GUU). Therefore, a complex sequence bias is clearly present at the amino acid level.

Base frequency at the intercodon position was examined in the filarial data set. A was not found to be common at N4 in the presence of T at N3. This agrees with results from multiple plant species, including *Ar. thaliana* and *Brassica napus,* which indicate that T3pA4 is very rare (de Amicis and Marchetti, 2000). In addition, T was found to be the predominant N4 nucleotide in the presence of T at N3 in all species except for *Ac. viteae.* This high frequency might have been the result of selection for T at N4. Although the presence of A at the third-codon position was favored in most cases, the presence of T at N4 might confer a selective advantage upon the use of codons ending in T.

Genes that encoded proteins that were biased with respect to certain amino acids were highly expressed. One such polypeptide, heat-shock protein-70 (Hsp70), which acts as a chaperone in protein folding and transport, is expressed at high levels under conditions of cell stress in organisms ranging from archaea to eukaryotes (Karlin and Brocchieri, 1998). In addition, high-level expression is associated with abundant larval transcript proteins (Alt-1, Alt-2) (Gregory *et al.,* 2000). Alt-1 mRNA and alt-2 mRNA account for greater than 1% of the total mRNA present at the third larval stage of *B. malayi* development (Gregory *et al.,* 2000). Furthermore, the gene encoding chitinase protein, which is involved in postinfective parasite ecdysis in *Ac. viteae* and *O. volvulus,* is highly expressed (Wu *et al.,* 1996). Interestingly, genes encoding unbiased proteins were also associated with high levels of expression. In particular, the gene encoding thioredoxin peroxidase, vital to the neutralization of oxygen radicals that result from aerobic metabolism, is highly expressed in *O. volvulus* larvae (Lu *et al.,* 1998). These results conflict with previous studies that have shown that less biased sequences are associated with low levels of expression (Stenico *et al.,* 1994). Therefore, within this data set, levels of gene expression do not appear to be related to sequence bias.

In summary, gene sequences deriving from five filarial parasites were found to possess an AT nucleotide bias, which impacted upon codon and amino acid usage, as well as base usage at intercodon sites. These biases were present in sequences from all five parasites, suggesting that the patterns might be con-

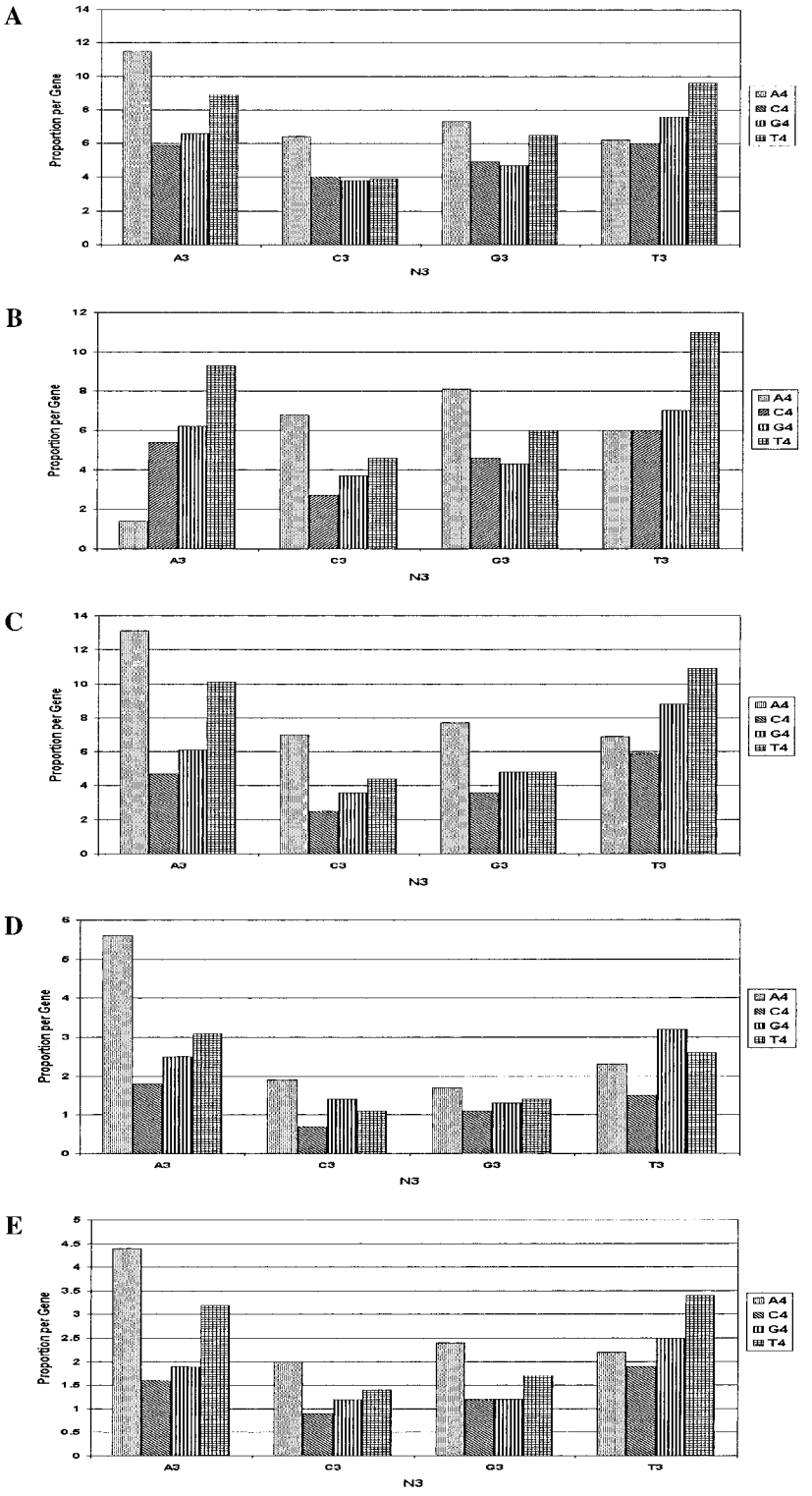


**FIG. 3.** Plot of the base frequency at the intercodon position (N4) as a function of the base at N3. The various N3 bases are plotted on the *X*-axis, and the proportion of each N4 nucleotide in the presence of each possible nucleotide at N3 is plotted on the *Y*-axis. (**A**) *W. bancrofti*, (**B**) *B. malayi*, (**C**) *O. volvulus*, (**D**) *Ac. viteae*, and (**E**) *Di. immitis*.

served among all members of the common phylogenetic family. However, it must be noted that the genomes of the filarial parasites examined in this research study have not yet been completely sequenced. Once this goal is reached, permitting all open reading frames to be defined, extensive codon usage investigations and expression analyses will be required for the functional characterization of the genome sequence. Furthermore, future work will undoubtedly define whether AT sequence bias is a dis-

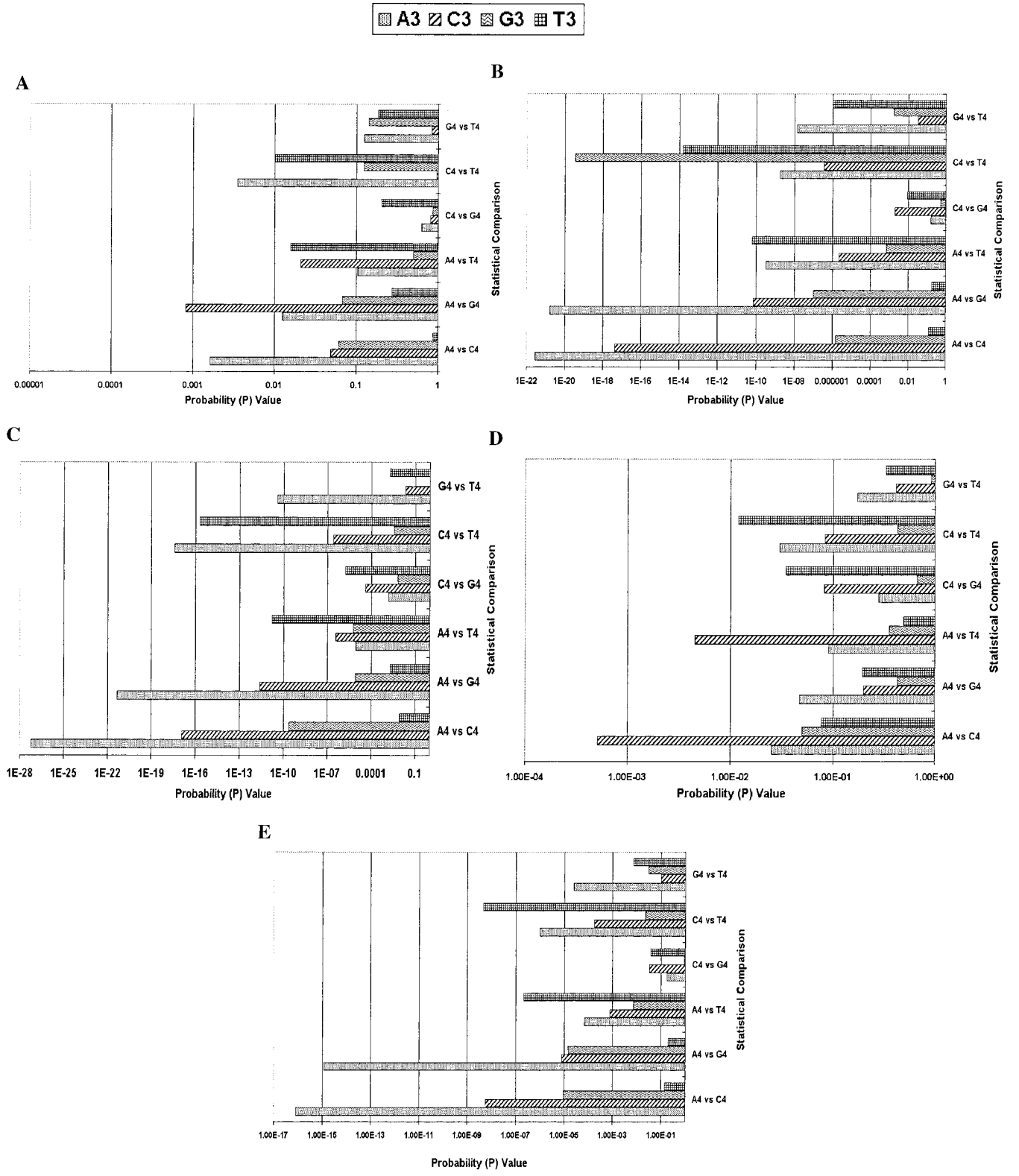


**FIG. 4.** Logarithmic plot comparing the statistical significance of associations established between intercodon (N4) and wobble position (N3) nucleotides. (**A**) *W. bancrofti,* (**B**) *B. malayi,* (**C**) *O. volvulus,* (**D**) *Ac. viteae,* and (**E**) *Di. immitis.*

crete characteristic associated with all phylogenetically related parasites or whether it is simply an evolutionary artifact that has been maintained in only a small number of species.

## REFERENCES

Basanez, M. G., Yarzabal, L., Frontado, H. L., and Villamizar, N. J. (2000). *Onchocerca–Simulium* complexes in Venezuela: Can hu-

man onchocerciasis spread outside its present endemic areas? *Parasitology* **120**(Pt. 2): 143–160.

Breitschopf, K., Haendeler, J., Malchow, P., Zeiher, A. M., and Dimmeler, S. (2000). Posttranslational modification of Bcl-2 facilitates its proteosome-dependent degradation: Molecular characterization of the involved signaling pathway. *Mol. Cell. Biol.* **20**(5): 1886–1896.

Bulmer, M. (1990). The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res.* **18**(10): 2869–2873.

Chen, N., and Cheng, Q. (1999). Codon usage in *Plasmodium vivax* nuclear genes. *Int. J. Parasitol.* **29**(3): 445–449.

Comeron, J. M., and Aguade, M. (1998). An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.* **47**(3): 268–274.

de Amicis, F., and Marchetti, S. (2000). Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Res.* **28**(17): 3339–3345.

Ellis, J., Morrison D. A., and Kalinna, B. (1995). Comparison of the patterns of codon usage and bias between *Brugia, Echinococcus, Onchocerca* and *Schistosoma* species. *Parasitol. Res.* **81**(5): 388–393.

Ellis, J. T., Morrison D. A., Avery D., and Johnson, A. M. (1994). Codon usage and bias among individual genes of the coccidia and piroplasms. *Parasitology* **109**(Pt. 3): 265–272.

Epstein, R. J., Lin, K., and Tan, T. W. (2000). A functional significance for codon third bases. *Gene* **245**(2): 291–298.

Ghosh, T. C., Gupta, S. K., and Majumdar, S. (2000). Studies on codon usage in *Entamoeba histolytica. Int. J. Parasitol.* **30:** 715–722.

Gomez-Escobar, N., Lewis, E., and Maizels, R. M. (1998). A novel member of the transforming growth factor-beta (TGF-beta) superfamily from the filarial nematodes *Brugia malayi* and *B. pahangi. Exp. Parasitol.* **88**(3): 200–209.

Greenacre, M. J. (1984). "Theory and Applications of Correspondence Analysis," Academic Press, London.

Gregory, W. F., Atmadja, A. K., Allen, J. E., and Maizels, R. M. (2000). The abundant larval transcript-1 and -2 genes of *Brugia malayi* encode stage-specific candidate vaccine antigens for filariasis. *Infect. Immun.* **68**(7): 4174–4179.

Groisman, E. A., Saier, M. H., Jr., and Ochman, H. (1992). Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the *Salmonella* genome. *EMBO J.* **11**(4): 1309–1316.

Hammond, M. P. (1994). Codon usage and gene organization in *Brugia. Parasitol. Res.* **80:** 173–175.

Hughes, A. L., and Yeager, M. (1997). Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45**(2): 125–130.

Johnston, D. A., Blaxter, M. L., Degrave, W. M., Foster, J., Ivens, A. C., and Melville, S. E. (1999). Genomics and the biology of parasites. *BioEssays* **21:** 131–147.

Karlin, S., and Brocchieri, L. (1998). Heat shock protein 70 family: Multiple sequence comparisons, function, and evolution. *J. Mol. Evol.* **47:** 565–577.

Karlin, S., Campbell, A. M., and Mrazek, J. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179:** 3899–3913.

Karlin, S., and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* **91**(26): 12832–12836.

Karlin, S., and Mrazek, J. (1997). Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **94**(19): 10227–10232.

Karlin S., and Mrazek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* **182**(18): 5238–5250.

Kurland, C. G. (1991). Codon bias and gene expression. *FEBS Lett.* **285**(2): 165–169.

Levin, D. B., and Whittome, B. (2000). Codon usage in nucleopolyhedroviruses. *J. Gen. Virol.* **81**(Pt. 9): 2313–2325.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P., and Darnell, J. (1995). "Molecular Cell Biology," 3rd ed., Scientific American Books, New York.

Lu, W., Egerton, G. L., Bianco, A. E., and Williams, S. A. (1998). Thioredoxin peroxidase from *Onchocerca volvulus:* A major hydrogen peroxide detoxifying enzyme in filarial parasites. *Mol. Biochem. Parasitol.* **91**(2): 221–235.

Mathe, C., Peresetsky, A., Dehais, P., Van Montagu, M., and Rouze, P. (1999). Classification of *Arabidopsis thaliana* gene sequences: Clustering of coding sequences into two groups according to codon usage improves gene prediction. *J. Mol. Biol.* **285**(5): 1977–1991.

McClellan, D. A. (2000). The codon-degeneracy model of molecular evolution. *J. Mol. Evol.* **50**(2): 131–140.

McInerney, J. O. (1998). GCUA: General codon usage analysis. *Bioinformatics* **14:** 372–373.

McInerney, J. O. (1997). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. *Microb. Comp. Genomics* **2**(1): 1–10.

McLachlan, A. D., Staden, R., and Boswell, D. R. (1984). A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res.* **12**(24): 9567–9575.

McWeeney, S. K., and Valdes, A. M. (1999). Codon usage bias and base composition in MHC genes in humans and common chimpanzees. *Immunogenetics* **49**(4): 272–279.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**(4): 851–856.

Moriyama, E. N., and Powell, J. R. (1998). Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cereviseae* and *Escherichia coli. Nucleic Acids Res.* **26**(13): 3188–3193.

Musto, H., Romero, H., and Rodriguez-Maseda, H. (1998). Heterogeneity in codon usage in the flatworm *Schistosoma mansoni. J. Mol. Evol.* **46**(2): 159–167.

Musto, H., Romero, H., Zavala, A., Jabbari, K., and Bernardi, G. (1999). Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum:* Compositional constraints and translational selection. *J. Mol. Evol.* **49:** 27–35.

Nakamura, T., Suyama, A., and Wada, A. (1991). Two types of linkage between codon usage and gene-expression levels. *FEBS Lett.* **289**(1): 123–125.

Ogunrinade, A., Boakye, D., Merriweather, A., and Unnasch, T. R. (1999). Distribution of the blinding and nonblinding strains of *Onchocerca volvulus* in Nigeria. *J. Infect. Dis.* **179**(6): 1577–1579.

Piacentini, M., Rodolfo, C., Farrace, M. G., and Autuori, F. (2000). "Tissue" transglutaminase in animal development. *Int. J. Dev. Biol.* **44**(6 Spec. No): 655–662.

Perriere, G., Lobry, J. R., and Thioulouse, J. (1996). Correspondence discriminant analysis: A multivariate method for comparing classes of protein and nucleic acid sequences. *Comput. Appl. Biosci.* **12**(6): 519–524.

Peters, W., and Gilles, H. M. (1995). "Color Atlas of Tropical Medicine and Parasitology," 4th ed., Mosby-Wolfe, London.

Pogonka, T., Oberlander, U., Marti, T., and Lucius, R. (1999). *Acanthocheilonema viteae:* Characterization of a molt-associated excretory/secretory 18-kDa protein. *Exp. Parasitol.* **93**(2): 73–81.

Rao, U. R., Salinas, G., Mehta, K., and Klei, T. R. (2000). Identification and localization of glutathione *S*-transferase as a potential target enzyme in *Brugia* species. *Parasitol. Res.* **86**(11): 908–915.

Russell, G. J., Walker, P. M. B., Elton, R. A., and Subak-Sharpe, J. H. (1976). Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J. Mol. Biol.* **108**(1): 1–23.

Saier, M. H., Jr. (1995). Differential codon usage: A safeguard

against inappropriate expression of specialized genes? *FEBS Lett.* **362**(1): 1–4.

Seetharaman, J., and Srinivasan, R. (1995). Analysis of codon usage: Positional preference in various organisms. *Ind. J. Biochem. Biophys.* **32**(3): 156–160.

Sharp P. M., and Cowe, E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae. Yeast* **7**(7): 657–678.

Sharp P. M., and Devine, K. M. (1989). Codon usage and gene expression level in highly expressed genes do 'prefer' optimal codons. *Nucleic Acids Res.* **17**(13): 5029–5039.

Sharp, P. M., and Li, W. (1987). The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**(3): 1281–1295.

Sharp, P. M., and Li, W. H. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res.* **14**(19): 7737–7749.

Sharp, P. M., and Matassi, G. (1994). Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **4**(6): 851–860.

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.

Stenico M., Lloyd A. T., and Sharp, P. M. (1994). Codon usage in *Caenorhabditis elegans:* Delineation of translational selection and mutational biases. *Nucleic Acids Res.* **22**(13): 2437–2446.

Unnasch, T. R., and Williams, S. A. (2000). The genomes of *Onchocerca volvulus. Int. J. Parasitol.* **30**(4): 543–552.

Waterkeyn, J. G., Gauci, C., Cowman, A. F., and Lightowlers, M. W. (1998). Codon usage in *Taenia* species. *Exp. Parasitol.* **88:** 76–78.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* **87**(1): 23–29.

Wu, Y., Adam, R., Williams, S. A., and Bianco, A. E. (1996). Chitinase genes expressed by infective larvae of the filarial nematodes, *Acanthocheilonema viteae* and *Onchocerca volvulus. Mol. Biochem. Parasitol.* **75**(2): 207–219.

Zang, X., Yazdanbakhsh, M., Jiang, H., Kanost, M. R., and Maizels, R. M. (1999). A novel serpin expressed by blood-borne microfilariae of the parasitic nematode *Brugia malayi* inhibits human neutrophil serine proteinases. *Blood* **94**(4): 1418–1428.