Volume 16 Number 17 1988 Nucleic Acids Research

Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity

Paul M.Sharp*, Elizabeth Cowe, Desmond G.Higgins, Denis C.Shields, Kenneth H.Wolfe and Frank Wright

Department of Genetics, Trinity College, Dublin 2, Ireland

Received May 21, 1988; Revised and Accepted July 26, 1988

ABSTRACT

The genetic code is degenerate, but alternative synonymous codons are generally not used with equal frequency. Since the pioneering work of Grantham's group (1,2) it has been apparent that genes from one species often share similarities in codon frequency; under the "genome hypothesis" (1,2) there is a species-specific pattern to codon usage.

However, it has become clear that in most species there are also considerable differences among genes (3-7). Multivariate analyses have revealed that in each species so far examined there is a single major trend in codon usage among genes, usually from highly biased to more nearly even usage of synonymous codons. Thus, to represent the codon usage pattern of an organism it is not sufficient to sum over all genes (8), as this conceals the underlying heterogeneity. Rather, it is necessary to describe the trend among genes seen in that species. We illustrate these trends for six species where codon usage has been examined in detail, by presenting the pooled codon usage for the 10% of genes at either end of the major trend (Table 1).

Closely-related organisms have similar patterns of codon usage, and so the six species in Table 1 are representative of wider groups. For example, with respect to codon usage, Salmonella typhimurium closely resembles E.coli (9), while all mammalian species so far examined (principally mouse, rat and cow) largely resemble humans (4,8).

CAUSES OF WITHIN-SPECIES DIVERSITY

Biased codon usage may result from a combination of several factors, viz. biases in the pattern of mutation, (translational) selection among synonymous codons, or selection against particular structures in DNA. Within-species heterogeneity in codon usage has been most clearly elucidated in E.coli; the major trend is from a strong bias towards a particular subset of codons in highly expressed genes to more even codon usage in lowly expressed genes (3,4,7). The heavily favoured codons in highly expressed E.coli genes are those best recognised by the most abundant tRNA species (3,4), and it seems clear that selection mediated by the translation process can occur among alternative synonymous codons (10,11). In contrast, most of the deviation from equal synonym use in the lowly expressed genes is likely to reflect nonrandom patterns of mutation (7,12). Then the pattern of bias in a particular gene reflects a mutation-selection balance at a point determined by the strength of translational selection on that gene (7,9,12).

Similar observations have been made for S.cerevisiae (4,5,12,13). In B.subtilis (14) and S.pombe (15) there are similar trends among genes, but there is less information about tRNA abundances. The pattern of codon

Table 1. Codon usage diversity within six species.

		E.coli		B.subtilis		S.cerevisiae		S.pombe		Drosophila		Human		
		high	low	high	low	high	low	high	low	high	low	G+C	A+T	
Phe	UUU	0.34	1.33	0.70	1.48	0.19	1.38	0.44	1.28	0.12	0.86	0.27	1.20	UUU
	UUC	1.66	0.67	1.30	0.52	1.81	0.62	1.56	0.72	1.88	1.14	1.73	0.80	UUC
Leu	UUA	0.06	1.24	2.71	0.66	0.49	1.49	0.28	1.79	0.03	0.62	0.05	0.99	UUA
	UUG	0.07	0.87	0.00	1.03	5.34	1.48	2.16	0.80	0.69	1.05	0.31	1.01	UUG
Leu	CUU	0.13	0.72	2.13	1.24	0.02	0.73	2.44	1.55	0.25	0.80	0.20	1.26	CUU
	CUC	0.17	0.65	0.00	0.93	0.00	0.51	1.13	0.31	0.72	0.90	1.42	0.80	CUC
	CUA	0.04	0.31	1.16	0.34	0.15	0.95	0.00	0.87	0.06	0.60	0.15	0.57	CUA
	CUG	5.54	2.20	0.00	1.80	0.02	0.84	0.00	0.68	4.25	2.04	3.88	1.38	CUG
Ile	AUU	0.48	1.38	0.91	1.38	1.26	1.29	1.53	1.77	0.74	1.27	0.45	1.60	AUU
	AUC	2.51	1.12	1.96	1.14	1.74	0.66	1.47	0.59	2.26	0.95	2.43	0.76	AUC
	AUA	0.01	0.50	0.13	0.48	0.00	1.05	0.00	0.64	0.00	0.78	0.12	0.64	AUA
Met	AUG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	AUG
Val	GUU	2.41	1.09	1.88	0.83	2.07	1.13	1.61	2.04	0.56	0.74	0.09	1.32	GUU
	GUC	0.08	0.99	0.25	1.49	1.91	0.76	2.39	0.65	1.59	0.93	1.03	0.69	GUC
	GUA	1.12	0.63	1.38	0.76	0.00	1.18	0.00	1.06	0.06	0.53	0.11	0.80	GUA
	GUG	0.40	1.29	0.50	0.92	0.02	0.93	0.00	0.24	1.79	1.80	2.78	1.19	GUG
Ser	UCU	2.81	0.78	3.45	0.77	3.26	1.56	3.14	1.33	0.87	0.55	0.45	1.63	UCU
	UCC	2.07	0.60	0.00	0.81	2.42	0.81	2.57	0.52	2.74	1.41	2.09	0.80	UCC
	UCA	0.06	0.95	1.50	1.29	0.08	1.30	0.00	1.56	0.04	0.84	0.26	1.23	UCA
	UCG	0.00	1.04	0.00	0.94	0.02	0.66	0.00	0.67	1.17	1.30	0.68	0.13	UCG
Pro	CCU	0.15	0.75	2.29	0.99	0.21	1.17	2.00	1.21	0.42	0.43	0.58	1.50	CCU
	CCC	0.02	0.68	0.00	0.27	0.02	0.75	2.00	0.83	2.73	1.02	2.02	0.83	CCC
	CCA	0.42	1.03	1.14	1.08	3.77	1.38	0.00	1.51	0.62	1.04	0.36	1.57	CCA
	CCG	3.41	1.54	0.57	1.66	0.00	0.70	0.00	0.45	0.23	1.51	1.04	0.10	CCG
Thr	ACU	1.87	0.76	2.21	0.39	1.83	1.23	1.89	1.52	0.65	0.70	0.36	1.45	ACU
	ACC	1.91	1.29	0.00	0.98	2.15	0.78	2.11	1.04	3.04	1.58	2.37	0.92	ACC
	ACA	0.10	0.68	1.38	1.64	0.00	1.38	0.00	1.04	0.10	0.77	0.36	1.45	ACA
	ACG	0.12	1.28	0.41	0.98	0.01	0.60	0.00	0.40	0.21	0.95	0.92	0.18	ACG
Ala	GCU	2.02	0.61	2.94	0.78	3.09	1.07	2.30	1.79	0.95	0.91	0.45	1.59	GCU
	GCC	0.18	1.18	0.08	1.14	0.89	0.76	1.49	0.50	2.82	1.93	2.38	0.92	GCC
	GCA	1.09	0.79	0.60	1.19	0.03	1.49	0.21	1.14	0.09	0.59	0.36	1.38	GCA
	GCG	0.71	1.42	0.38	0.89	0.00	0.68	0.00	0.57	0.14	0.57	0.82	0.11	GCG

Relative Synonymous Codon Usage (RSCU; Ref.5) values are presented for two groups of genes from each of six species: Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens.
(An RSCU value is the observed number of codons divided by the number expected if all codons for that amino acid were used equally.)

Table 1 (cont.)

	E.coli		B.subtilis		S.cerevisiae		S.pombe		Drosophila		Human		
	high	low	high	low	high	low	high	low	high	low	G+C	A+T	
Tyr UAU	0.38	1.28	0.50	1.29	0.06	1.13	0.48	1.24	0.23	0.96	0.34	1.17	UAU
UAC	1.63	0.72	1.50	0.71	1.94	0.87	1.52	0.76	1.77	1.04	1.66	0.83	UAC
ter UAA	--	--	--	--	--	--	--	--	--	--	--	--	UAA
UAG	--	--	--	--	--	--	--	--	--	--	--	--	UAG
His CAU	0.45	1.21	2.00	1.28	0.32	1.16	0.56	1.44	0.29	0.86	0.30	1.28	CAU
CAC	1.55	0.79	0.00	0.72	1.68	0.84	1.44	0.56	1.71	1.14	1.70	0.72	CAC
Gln CAA	0.12	0.76	1.71	0.88	1.98	1.10	1.85	1.67	0.03	0.88	0.21	0.98	CAA
CAG	1.88	1.24	0.29	1.13	0.02	0.90	0.15	0.33	1.97	1.12	1.79	1.02	CAG
Asn AAU	0.02	1.12	0.47	1.21	0.06	1.28	0.30	1.41	0.13	1.13	0.33	1.20	AAU
AAC	1.98	0.88	1.53	0.79	1.94	0.72	1.70	0.59	1.87	0.87	1.67	0.80	AAC
Lys AAA	1.63	1.50	1.83	1.47	0.16	1.24	0.10	1.27	0.06	0.81	0.34	1.17	AAA
AAG	0.37	0.50	0.17	0.53	1.84	0.76	1.90	0.73	1.94	1.19	1.66	0.83	AAG
Asp GAU	0.51	1.43	0.53	1.16	0.70	1.38	0.78	1.56	0.90	1.10	0.36	1.29	GAU
GAC	1.49	0.57	1.47	0.84	1.30	0.62	1.22	0.44	1.10	0.90	1.64	0.71	GAC
Glu GAA	1.64	1.28	1.40	1.27	1.98	1.29	0.69	1.20	0.19	0.73	0.26	1.33	GAA
GAG	0.36	0.72	0.60	0.73	0.02	0.71	1.31	0.80	1.81	1.27	1.74	0.67	GAG
Cys UGU	0.60	0.94	0.00	0.94	1.80	1.10	0.14	1.56	0.07	0.71	0.42	1.09	UGU
UGC	1.40	1.06	2.00	1.06	0.20	0.90	1.86	0.44	1.93	1.29	1.58	0.91	UGC
ter UGA	--	--	--	--	--	--	--	--	--	--	--	--	UGA
Trp UGG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	UGG
Arg CGU	4.47	1.71	3.11	0.54	0.63	0.64	5.17	1.89	2.65	0.69	0.38	0.64	CGU
CGC	1.53	2.41	1.78	1.21	0.00	0.39	0.83	0.26	3.07	1.55	2.72	0.36	CGC
CGA	0.00	0.52	0.00	0.74	0.00	0.65	0.00	0.86	0.07	1.12	0.31	0.81	CGA
CGG	0.00	0.80	0.00	0.81	0.00	0.34	0.00	0.43	0.00	1.12	1.53	0.51	CGG
Ser AGU	0.13	1.01	0.45	0.56	0.06	0.97	0.14	1.48	0.04	0.89	0.31	1.26	AGU
AGC	0.93	1.62	0.60	1.63	0.16	0.70	0.14	0.44	1.13	1.01	2.22	0.94	AGC
Arg AGA	0.00	0.37	1.11	2.02	5.37	2.51	0.00	1.71	0.00	0.56	0.22	2.40	AGA
AGG	0.00	0.19	0.00	0.67	0.00	1.47	0.00	0.86	0.21	0.95	0.84	1.28	AGG
Gly GGU	2.27	1.29	1.38	0.54	3.92	1.32	3.36	1.87	1.34	0.91	0.34	0.84	GGU
GGC	1.68	1.31	0.97	1.30	0.06	0.92	0.59	0.27	1.66	1.65	2.32	0.76	GGC
GGA	0.00	0.64	1.66	1.24	0.00	1.22	0.05	1.60	0.99	0.98	0.29	1.79	GGA
GGG	0.04	0.76	0.00	0.92	0.02	0.55	0.00	0.27	0.00	0.46	1.05	0.61	GGG

For each species, genes have been ranked according to their position along the major intraspecific trend in codon bias (see text). The highest 10% and the lowest 10% of genes have been drawn from: 165 E.coli genes (7), 76 B.subtilis genes (8,14), 154 S.cerevisiae genes (5,8), 40 S.pombe genes (15), 84 D.melanogaster genes (16) and 290 human genes (8). The sample size for S.pombe is rather small, but the codon frequencies appear to be reliable (15). Full gene listings are available from the authors.

frequencies in lowly expressed genes from B.subtilis is most strongly indicative of mutational bias (14).

Recently, we have reported evidence of selection among synonymous codons in the multicellular organism D.melanogaster (16). In contrast, among human genes the major variation is in G+C content associated with the local base composition around the gene (6). This variation has not been attributed to translational selection, and is most easily explained in terms of variation in mutation biases among chromosomal regions.

CODON BIAS RANKINGS

For E.coli, B.subtilis, S.cerevisiae and S.pombe codon bias in a gene is measured by the Codon Adaptation Index (CAI). A species-specific reference set of very highly expressed genes is used to assess the relative fitness of each synonymous codon, and the CAI for a gene is then calculated as the geometric mean of the fitness values for each codon in that gene. (For a full description, see Ref.17.)

Since the biological basis of codon frequencies in Drosophila is not yet so firmly established (for example, there may be more than one optimal set of codons, depending on the tissue of gene expression) we have simply estimated codon bias as the deviation from equal synonym use, by a "chi-square" scaled by gene length (16); this index is very highly correlated with the major trend among genes. Finally, human genes are ranked by G+C content at silent positions, since this is the major source of variation among genes (4,6).

FORTRAN 77 programs to calculate these indices are available (on IBM-type floppy disks) from the authors on request.

ACKNOWLEDGEMENT: This is a publication from the Irish National Centre for Bioinformatics. This work was supported in part by a grant from the European Community Biotechnology Action Programme.

*To whom correspondence should be addressed

REFERENCES

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Nucleic Acids Res. 8, r49-r62.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. 9, r43-r74.
3. Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. 10, 7055-7074.
4. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
5. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) Nucleic Acids Res. 14, 5125-5143.
6. Aota, S. and Ikemura, T. (1986) Nucleic Acids Res. 14, 6345-6355.
7. Sharp, P.M. and Li, W-H. (1986) Nucleic Acids Res. 14, 7737-7749.
8. Aota, S., Gojobori, T., Ishibashi, F., Maruyama, T. and Ikemura, T. (1988) Nucleic Acids Res. 16, r315-r402.
9. Ikemura, T. (1985) in Molecular Evolution and Population Genetics, Aoki, K. and Ohta, T. Eds., pp. 385-406, Springer-Verlag, Berlin.

10. de Boer, H. and Kastelein, R.A. (1986) in From gene to protein; steps dictating the maximal level of gene expression, Reznikoff, W.S. and Gold, L. Eds., pp. 225-283, Butterworths, Mass.
11. Kurland, C.G. (1987) Trends Biochem. Sci. 12, 126-128.
12. Bulmer, M. (1988) J. Evol. Biol. 1, 15-26.
13. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
14. Shields, D.C. and Sharp, P.M. (1987) Nucleic Acids Res. 15, 8023-8040.
15. Sharp, P.M. and Wright, F. (1988) in preparation
16. Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) Mol. Biol. Evol. 5 (in press)
17. Sharp, P.M. and Li, W-H. (1987) Nucleic Acids Res. 15, 1281-1295.