Analysis of codon usage patterns using graphical methods

F. Wright

Codon usage patterns The genetic code maps the 61 sense codons in protein-coding DNA to the choice of the 20 amino acids in the resultant protein sequence. Eighteen of the amino acids are coded for by more than one codon (Table 1). Three amino acids have six synonymous codons (e.g. Leucine with UUU, UUC, CUU, CUC, CUA, and CUG codons), five have four, one has three, and nine have two. Two amino acids (Methionine and Tryptophan) are each encoded by a single codon.

When molecular sequence data started to accumulate nearly 20 years ago, it was noted that DNA sequences that code for proteins did not randomly choose among codons that specify the same amino acid. Indeed the observed departure from uniform codon usage for each amino acid was a common feature. These observed patterns in synonymous codon usage were not simple and varied among genes within a genome, and among genomes. For *E. coli* and yeast genes, the main evolutionary force varying among genes was considered to be natural selection to optimise protein production (translational selection, acting on highly expressed genes), whereas with human genes, it was thought to be variation among chromosomal regions in the mutation process (biased mutation, resulting in base composition differences). It is only recently that sufficient DNA sequence data has accumulated to allow large scale studies of codon usage patterns in higher plants.

Studies of synonymous codon usage reveal information about the molecular evolution of individual genes and can provide data to train methods (genome-specific gene recognition algorithms) that detect protein coding regions in uncharacterised genomic DNA. Knowledge of codon usage patterns can also be utilised to design DNA primers and to detect horizontal transfer events.

Statistical analysis of genomic data and bioinformatics Earlier studies of synonymous codon usage were typically based on a sample of 100 genes from a genome. The ongoing eukaryotic genome projects are producing large volumes of sequence data and thus surveys of codon usage require to be automated. This will involve the automation of the extraction of protein coding DNA from the primary databases and of the subsequent statistical analysis of thousands of genes. In this article, some preliminary analysis of syn-

Phe	UUU	2	0.31	Ser	UCU	5	1.67	Tyr	UAU	0	0.00	Cys	UGU	0	0.00
	UUC	11	1.69		UCC	6	2.00		UAC	6	2.00		UGC	0	0.00
Leu	UUA	0	0.00		UCA	0	0.00	TER	UAA	1	3.00	TER	UGA	0	0.00
	UUG	23	2.56		UCG	1	0.33	TER	UAG	0	0.00	Trp	UGG	0	0.00
	CUU	13	1.44	Pro	CCU	5	1.67	His	CAU	1	0.33	Arg	CGU	14	3.50
	CUC	11	1.22		CCC	0	0.00		CAC	5	1.67		CGC	0	0.00
	CUA	2	0.22		CCA	1	0.33	Gln	CAA	5	0.28		CGA	0	0.00
	CUG	5	0.56		CCG	6	2.00		CAG	31	1.72		CGG	0	0.00
Ile	AUU	14	1.00	Thr	ACU	17	1.62	Asn	AAU	2	0.33	Ser	AGU	3	1.00
	AUC	28	2.00		ACC	18	1.71		AAC	10	1.67		AGC	3	1.00
	AUA	0	0.00		ACA	3	0.29	Lys	AAA	9	0.43	Arg	AGA	5	1.25
Met	AUG	6	1.00		ACG	4	0.38		AAG	33	1.57		AGG	5	1.25
Val	GUU	5	0.83	Ala	GCU	4	0.89	Asp	GAU	13	0.72	Gly	GGU	14	1.56
	GUC	7	1.17		GCC	13	2.89		GAC	23	1.28		GGC	5	0.56
	GUA	1	0.17		GCA	0	0.00	Glu	GAA	8	0.53		GGA	16	1.78
	GUG	11	1.83		GCG	1	0.22		GAG	22	1.47		GGG	1	0.11

Table 1 The codon usage table for an *A. thaliana* gene (Polyubiquitin; accession number L05361), consisting of 458 codons. Integer values denote usage of each codon; real numbers represent 'Relative Synonymous Codon Usage' (RSCU) values which are obtained by dividing the observed occurrence by the expected (e.g. the CUC Leucine codon occurs 1.22 times more often than the expected 9.0 assuming uniform usage of Leucine codons). The location of this gene on Figure 1 is at Nc=39.8 and GC3s=0.64.

		EXT-1		EXT-2	
Phe	UUU	0.53	(73)	1.26	(227)
	UUC*	1.47	(205)	0.74	(132)
Leu	UUA	0.27	(20)	0.88	(140)
	UUG	1.21	(90)	1.44	(229)
	CUU	1.55	(115)	1.63	(259)
	CUC*	2.39	(177)	0.65	(104)
	CUA	0.23	(17)	0.74	(118)
	CUG	0.35	(26)	0.65	(104)
Ile	AUU	0.88	(92)	1.28	(205)
	AUC*	2.00	(209)	0.69	(111)
	AUA	0.12	(12)	1.02	(164)
Met	AUG	1.00	(151)	1.00	(315)
Val	GUU	1.42	(153)	1.85	(320)
	GUC*	1.46	(158)	0.41	(71)
	GUA	0.29	(31)	0.70	(122)
	GUG	0.83	(90)	1.04	(180)
Tyr	UAU	0.36	(32)	1.32	(170)
	UAC*	1.64	(146)	0.68	(88)
TER	UAA	1.18	(11)	1.29	(12)
TER	UAG	0.32	(3)	0.43	(4)
His	CAU	0.65	(40)	1.52	(157)
	CAC*	1.35	(84)	0.48	(50)
Gln	CAA	0.69	(60)	1.03	(200)
	CAG*	1.31	(113)	0.97	(187)
Asn	AAU	0.36	(39)	1.23	(310)
	AAC*	1.64	(176)	0.77	(194)
Lys	AAA	0.59	(113)	0.92	(387)
	AAG*	1.41	(273)	1.08	(455)
Asp	GAU	0.93	(142)	1.51	(478)
	GAC*	1.07	(164)	0.49	(156)
Glu	GAA	0.71	(124)	1.02	(427)
	GAG*	1.29	(225)	0.98	(409)

		EXT-1		EXT-2	
Ser	UCU	1.58	(122)	1.87	(305)
	UCC*	1.41	(109)	0.39	(64)
	UCA	0.88	(68)	1.31	(214)
	UCG	0.53	(41)	0.50	(82)
Pro	CCU	1.35	(88)	1.85	(251)
	CCC*	0.89	(58)	0.33	(45)
	CCA	1.20	(78)	1.26	(171)
	CCG	0.57	(37)	0.57	(77)
Thr	ACU	1.30	(122)	1.58	(200)
	ACC*	1.56	(146)	0.52	(66)
	ACA	0.73	(68)	1.37	(173)
	ACG	0.41	(38)	0.52	(66)
Ala	GCU	1.75	(235)	1.89	(335)
	GCC*	1.42	(191)	0.47	(83)
	GCA	0.41	(55)	1.15	(203)
	GCG	0.42	(56)	0.49	(87)
Cys	UGU	0.90	(53)	1.25	(106)
	UGC*	1.10	(65)	0.75	(64)
TER	UGA	1.50	(14)	1.29	(12)
Trp	UGG	1.00	(55)	1.00	(126)
Arg	CGU*	2.21	(106)	0.64	(63)
	CGC*	0.88	(42)	0.12	(12)
	CGA	0.31	(15)	0.77	(76)
	CGG	0.17	(8)	0.58	(57)
Ser	AGU	0.47	(36)	1.09	(178)
	AGC	1.13	(87)	0.83	(135)
Arg	AGA	1.35	(65)	2.32	(228)
	AGG	1.08	(52)	1.57	(154)
Gly	GGU*	1.67	(333)	1.36	(245)
	GGC*	0.71	(141)	0.50	(89)
	GGA	1.44	(287)	1.38	(247)
	GGG	0.19	(38)	0.76	(137)

Table 2 Extremes (denoted EXT-1 and EXT-2) of the main trend from CA based on pooled codon usage from 28 genes (6273 and 10836 codons, respectively). Integers and real numbers are used as in Table 1. Preferred codons in the EXT-1 group are marked with an asterisk.

onymous codon usage among protein-coding genes in *Arabidopsis thaliana* will be discussed, using the 560 well-annotated genes analysed by Mathé[1] *et al.* This will serve as a pilot study of a higher plant genome, after which we intend to investigate the software requirements and statistical methods for the automated analysis of larger datasets.

Statistical methods can be split into two main types: (1) graphical methods that display the trends in the data, and (2) methods that explicitly test the significance of possible factors (e.g. base composition, protein expression level). The former approach is discussed here.

Graphical methods I: correspondence analysis An appropriate graphical multivariate method for count or proportion data is correspondence analysis (CA). Codon usage data from a sample of G genes can be arranged as a two-way contingency table, with G rows and 61 columns. CA can be used to extract the trends in this dataset, using either the raw counts (containing amino acid usage as well as synonymous codon usage information) or counts corrected for amino acid usage, i.e. relative synonymous codon usage (RSCU) values (see Table 1 for an example of RSCU values for a gene). Here we concentrate on trends among the genes. The CA of the RSCU values of the 560 *A. thaliana* genes produced a main trend that explained 11.2% of the variation. To illustrate this trend in synonymous codon usage, we have pooled codon usage tables of 28 genes that lie at either end of the trend (see Table 2). Initial inspection reveals that genes at

one end of the trend (EXT-1) tend to use synonymous codons ending in C, and that many of these genes are known to be expressed in large amounts. Similar CA results were obtained by Chiapello[2] *et al.*

The CA analysis produces other trends in the data. For brevity, we have only shown the main trend. Other multivariate methods can be applied to codon usage data. If distinct clusters are expected, then cluster analysis is an appropriate method. Cluster analysis was used by Mathé[1] *et al.* to partition the 560 genes in this study into two groups which, as might be expected, have similar codon usage to the patterns of EXT-1 and EXT-2, respectively.

Graphical methods II: the effective number of codons used in a gene Another approach to exploring patterns in synonymous codon usage among genes is to quantify, for each gene, the extent of the departure from uniform usage within each amino acid class. A commonly-used measure, the "effective number of codons used in a gene", N_c, was developed by Wright[3]. This produces a number, for each individual gene, that lies between 20 (when only one codon is used for each amino acid) and 61 (when all codons are uniformly used). N_c is based on the "effective number of alleles" (N_a) statistic from Population Genetics theory and is approximately the sum, over all amino acids, of N_a. This intuitive measure is essentially independent of gene length and a recent comparative simulation study[4] has shown it to be the best overall estimator of absolute synonymous codon usage bias. Nc can be plotted against factors (e.g. G+C content in the third codon position, GC3s) to investigate patterns of codon usage (this is analogous to plotting residuals against possible additional factors when carrying out a multiple regression analysis). Figure 1 shows the plot for the 560 *A. thaliana* genes in this study. The plot contains a reference line, labelled GC(ref), that shows the expected position of genes whose codon usage is only determined by variation in GC3s. This GC(ref) line is an approximate upper limit for the value of N_c. For example, if GC3s is zero, then only codons ending in A and T will be used, thus restricting the number of codons used to 30 out of the 61 sense codons. The position of genes lying well below the GC(ref) curve reveals that GC3s composition only explains some of the variation in *A. thaliana* codon usage.

Software for analysing synonymous codon usage patterns While correspondence analysis can be carried out on count data by most commercial statistics packages, CodonW (by John Peden; available from http://www.molbiol.ox.ac.uk/cu/Readme.html), a public domain program, will read in a set of protein-coding DNA sequences in FASTA format and carry out a range of analyses, including CA and the calculation of N_c and GC3s for each gene.

Production of codon usage data from DNA sequence data The preparation of error-free protein-coding DNA sequences as input to a survey of codon usage analysis can be time consuming. In particular, it is important to exclude coding regions that have been predicted by gene recognition algorithms, rather than by direct experiment. The Codon Usage Database[5] is a useful source of pre-calculated codon usage data, although it is not completely up-to-date with the primary nucleotide sequence database. Ideally, the permanent storage of codon usage data should be avoided and instead the codon usage table should be automatically calculated "on the fly" from each database entry. We are investigating the automatic extraction of such data from molecular sequence databases with the help of our colleagues at Bioinformatics and I.T. Research at SCRI.

Trends in *Arabidopsis thaliana* codon usage The amount of the synonymous codon usage variation

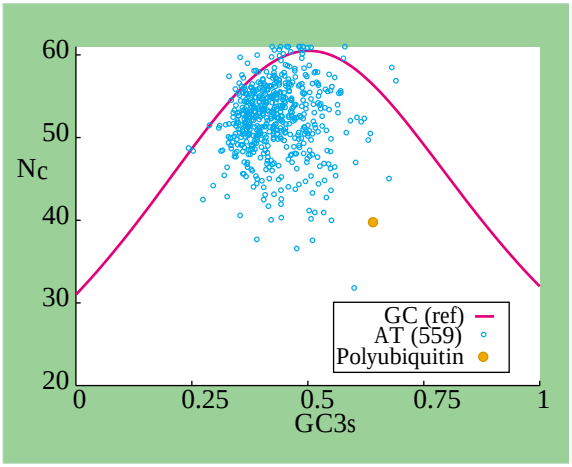

Figure 1 A plot for the 560 *A. thaliana* genes in the study.

explained by the main trend (11.2%) in this study of 560 *A. thaliana* genes is low compared to similar CA analyses[6] of human (32.5%), yeast (38.7%) and *E. coli* (30.5%). The main trend is partly due to variation in G+C (mainly C) composition among genes. The range of variation in base composition is not as extreme as found in Human genes, but is more extensive than found in yeast and *E. coli*. However, many of the high G+C genes are known to be highly expressed[2]. This confounding of G+C variation and protein expression levels complicates the interpretation of the analysis. Chiapello[2] *et al.* concluded that translation selection was the major factor, on the basis that intron and synonymous sites differed in G+C composition (if biased mutation was the main factor, we would expect that intron and synonymous sites would have similar base composition). Further analysis of this dataset may reveal other factors influencing codon usage. In particular, the chromosomal location of *A. thaliana* genes is a factor excluded from our analysis.

There are currently 6297 protein-coding regions in Nakamura's Codon Usage Database and more in the primary nucleotide sequence database. We intend to extend the analysis to this larger dataset once we have automated the analysis and improved the statistical analysis.

Future statistical work N_C is the best overall measure of absolute synonymous codon usage bias[4]. However, N_C did not perform well for very short genes (e.g. less than 200 codons long) where there are some amino acids that are unused, and we plan to produce an improved estimator of N_C.

Rather than use graphical methods to explore the data, we could use a Generalized Linear Model (GLM) to fit a model to the codon usage table. Such a model could be used to fit factors like G+C composition, chromosomal location, tissue-type, and gene expression level (if available). GLMs have not been applied to large scale surveys of codon usage, but have the potential to aid our understanding of the relative importance of factors. In particular, the interaction of mutation bias and translational selection in genomes with skewed base composition, e.g. *Dictyostelium discoideum* (mean genomic G+C approx. 25%) and *Streptomyces* spp. (mean genomic G+C 75%), is of particular interest.

References

1 Mathé, C., Peresetsky, A., Dehais, P., Van Montagu, M. & Rouze, P. (1999). *Journal of Molecular Biology* 285, 1977-1991.

2 Chiapello, H., Lisacek, F., Caboche, M. & Henaut, A. (1998). *Gene* 209, GC1-GC38.

3 Wright, F. (1990). The effective number of codons used in a gene. *Gene* 87, 23-29.

4 Comeron, J.M. & Aguade, M. (1998). *Journal of Molecular Evolution* 47, 268-274.

5 Nakamura, Y., Gojobori, T. & Ikemura, T. (1999) Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic Acids Research* 27, 292-292.

6 Wright, F. (unpublished).