Volume 14 Number 13 1986 Nucleic Acids Research

Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes

Paul M.Sharp*, Therese M.F.Tuohy and Krzysztof R.Mosurski[1]

Departments of Genetics and [1]Statistics, Trinity College, Dublin 2, Ireland

Received 4 April 1986; Accepted 23 May 1986

ABSTRACT

Codon usage data has been compiled for 110 yeast genes. Cluster analysis on relative synonymous codon usage revealed two distinct groups of genes. One group corresponds to highly expressed genes, and has much more extreme synonymous codon preference. The pattern of codon usage observed is consistent with that expected if a need to match abundant tRNAs, and intermediacy of tRNA-mRNA interaction energies are important selective constraints. Thus codon usage in the highly expressed group shows a higher correlation with tRNA abundance, a greater degree of third base pyrimidine bias, and a lesser tendency to the A+T richness which is characteristic of the yeast genome. The cluster analysis can be used to predict the likely level of gene expression of any gene, and identifies the pattern of codon usage likely to yield optimal gene expression in yeast.

INTRODUCTION

The usage of different synonymous codons is clearly not random in the majority of genes so far examined. It has been concluded that natural selection distinguishing between synonymous codons constrains the rate of nucleotide substitution (1,2), and that the rate varies somewhat between genes (3). This constraint presumably reflects differences in translational efficiency of different codons. It would be expected then that levels of expression of heterologous genes would be influenced by the degree of correspondence between the pattern of codon usage in the introduced gene and the preferred profile in the host organism. Early reports suggest that this may indeed be true (4,5). Thus it is of great interest to detail the precise preferred pattern of codon usage which might yield optimal expression of heterologous genes in species of biotechnological importance.

Nucleotide sequence data are now available for many genes from a wide variety of organisms (6,7). However, determination of the precise pattern of codon usage, and its possible causative factors, has been carried out for large data sets from very few species. The outstanding exception is Escherichia coli (8-10), while the availability of the entire DNA sequences of several coliphages has enabled the investigation of total genomic codon usage (11,12). Smaller compilations have been made for Bacillus (13,14), for yeast (15,16), and for several multicellular eukaryotes, including Drosophila (17), chicken and man (18). From these compilations it has become clear that, in genes from the same taxa, there are broad similarities in direction of codon bias. This had led Grantham and co-workers to formulate the "genome hypothesis" of codon preference (19). However, there are also clear differences between genes from the same species. In E.coli genes thought to be highly and lowly expressed differ in their extent of codon bias, with the bias being more extreme in highly expressed genes (8,9).

From the first available yeast gene sequences a pattern of strong codon bias, most prominent in highly expressed genes, was established (20). It has been reported that a compilation of about 40 genes confirms this pattern, and suggests that tRNA abundance appears to be an important influence (18). Whether there is a causal link between codon usage and level of gene expression is as yet controversial. Here we compile codon usage data for 110 yeast (mainly Saccharomyces cerevisiae) genes. A cluster analysis, based only on pattern of synonymous codon preference, yields two distinct groups. Inspection reveals that one group contains almost all of the (and perhaps only) highly expressed genes. Thus from the pattern of codon usage in a yeast gene it appears that we can predict the likely level (high or low) of its expression. We also detail the patterns of codon preference in each group and discuss their possible basis.

DATA

The 110 yeast genes examined are detailed in Table 1. Unless otherwise indicated the genes were isolated from S.cerevisiae. Genes from S.carlsbergensis have been included

because it is not regarded as a separate species (21). A few genes are from plasmids (also indicated in Table 1). Sources of data were the GenBank (6) and EMBL (7) data libraries (when possible) or original publications -- all referenced in Table 1 and listed in the Appendix.

ANALYSES

To examine synonymous codon usage without the confounding influence of amino acid composition of different gene products, observed numbers of codons were converted to relative synonymous codon usage values, $RSCU_{ij}$:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}$$

-- where X_{ij} is the number of occurrences of the jth codon for the ith amino acid, which is encoded by n_i synonymous codons. (More simply RSCU is the observed number of occurrences divided by that expected if usage of synonymous codons was uniform). The values for UGG (Trp) and AUG (Met) are always 1.0, and excluding also termination codons, each gene is then characterized by 59 variables (of which 41 are independent).

Cluster analysis:

The 110 genes were subjected to cluster analysis, using Ward's method (22) and grouping genes on the basis of their 59 RSCU values. The Clustan 2 package, from the Computer Centre, James Cook University of North Queensland, Australia was run on the DEC 20-60 at Trinity College, Dublin. This method considers the N items of data (genes), discerns the two which are most similar, records the "distance" (difference in codon usage) between these two, then clusters them to form a new item of data (at the mid point of the distance between them), and thus reducing the total number of items to N-1. This algorithm is performed N-1 times until only one cluster remains. Thus all points must be progressively clustered, but a dendogram derived

from the distances at each clustering indicates whether any truly distinct homogeneous clusters have been formed (see below). This can be more rigorously tested by partitioning the total variation between all genes into components between and within clusters. The stability of clusters can be tested by a relocation procedure (RELOCATE in Clustan 2) which determines whether any genes are better fitted in a different cluster. To ascertain whether any genes are "outliers", i.e. do not truly belong in any cluster, a threshold can be imposed such that any gene greater than a certain distance from the nearest gene or cluster centre is discarded. Obviously, in general, the smaller the threshold value chosen, the greater the number of outliers produced. However, if the threshold distance is varied, a critical value can be selected within a range where the number of outliers produced does not vary. To assess the effect of using RSCU values, genes were clustered by the same method, but using percentage codon usage values (with no correction for amino acid usage).

Codon bias indices:

Several indices of codon usage bias were calculated for each gene individually:

(i) The extent of preference (codon bias index, CBI) for 22 particular codons identified by Bennetzen and Hall (20) as being strongly preferred in three highly expressed yeast genes. This index has been used previously for comparisons between yeast genes (23).

(ii) The degree of bias within all synonymous groups, estimated by a G2 statistic (measuring the deviation from random synonymous codon usage) scaled by division by two times the number of codons considered. This index, since it does not measure bias towards a particular subset of codons, could be used in comparisons between genes from different species.

(iii) The linear correlation between usage of each codon and the relative abundance in yeast of the relevant cognate tRNA species (data from Ikemura (16)). A correlation of 0.54 would result from equal use of synonymous codons, given the average amino acid composition of these yeast genes.

(iv) The degree of third base C/U bias yielding intermediate

codon-anticodon interaction strengths (24) (summarised as a P2 statistic (9)). A P2 value of 0.5 indicates no bias.

For each of these indices higher values indicate stronger bias.

RESULTS

The dendogram depicting the result of a cluster analysis of yeast genes, grouping those genes according to synonymous codon usage, is shown in Figure 1a. The genes fall into two clear groups, such that the differences (horizontal distances in the dendogram) between genes within groups are very small compared to the difference between the two groups. Of the total variation between genes 40% lies between these two clusters. Comparison with the result of the same cluster analysis applied to the 50 genes of bacteriophage T7 (data from ref.11), where no real grouping is apparent (Figure 1b), suggests that the two clusters of yeast genes are highly significant. The relocation procedure did not change the composition of these major clusters, but subsequent application of a threshold distance, in conjunction with the relocation, suggested 6 outliers.

Details of the genes grouped into the two major clusters, and those not clustered, are given in Table 1. Consideration of Table 1 shows that many of the yeast genes thought to be highly expressed have been clustered into group A, while few if any appear in group B. For example, of 16 ribosomal protein genes 14 appear in group A (the other two, one in group B and one an outlier, are conspicuously short). Alcohol dehydrogenase, enolase, glyceraldehyde-3-phosphate dehydrogenase, histone, elongation factor, pyruvate kinase, phosphoglycerate kinase and glutamate dehydrogenase genes are all known to be highly expressed, and are clustered in group A. Several of the outliers are very short genes (Table 1c) and the 'peculiarity' of the patterns of relative synonymous codon usage in these cases is probably due to small numbers of occurrences of many amino acids.

Clustering on percentage codon usage, rather than RSCU values, yields an essentially similar dendogram, except six of the genes in group A (see Table 1 for details) are then

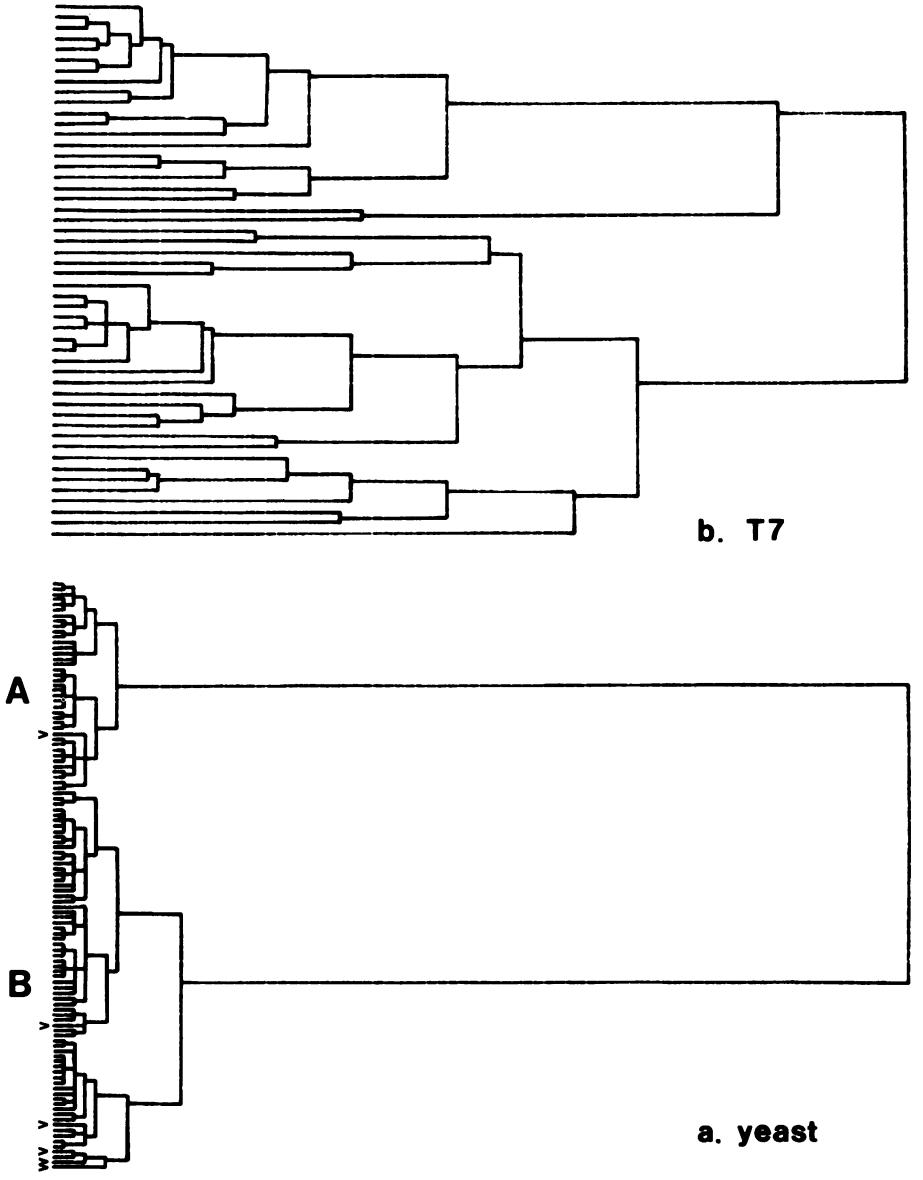

Figure 1.
Cluster analysis dendograms of RSCU values for (a) 110 yeast genes, (b) 50 T7 genes. The horizontal length of branches represents the distance between two groups when clustered. In each case all branch lengths are scaled relative to the distance between the last two clusters. In (a) A and B refer to the groups of genes in Table 1, and the six outliers are indicated>.

Table 1a. Details of yeast genes clustered in group A.

	Gene / product	codons	CBI	G2	P2	tRNA	Ref.
	Ribosomal protein L16	175	0.83	0.70	0.80	0.79	(1)
*1	Ribosomal protein L17a	138	0.79	0.68	0.72	0.63	(2)
	Ribosomal protein L25	138	0.86	0.72	0.82	0.52	(3)
	Ribosomal protein L29	150	0.79	0.73	0.83	0.66	(4)
	Ribosomal protein L34	114	0.84	0.75	0.79	0.57	(5)
	Ribosomal protein 13	388	0.89	0.78	0.86	0.70	(7)
	Ribosomal protein 28	187	0.89	0.86	0.89	0.61	(8)
	Ribosomal protein 51a	137	0.87	0.85	0.86	0.68	(9)
	Ribosomal protein 59	138	0.88	0.85	0.79	0.63	(10)
*1	Ribosomal protein S10	238	0.94	0.86	0.86	0.67	(11)
	Ribosomal protein S16a	145	0.88	0.80	0.78	0.78	(12)
	Ribosomal protein S24	131	0.86	0.77	0.67	0.81	(13)
*4	Actin	376	0.82	0.66	0.80	0.83	(15)
	ADH 1	349	0.91	0.76	0.79	0.74	(16)
	ADR 2	349	0.71	0.45	0.73	0.76	(17)
*4	iso-1-cytochrome C	110	0.47	0.37	0.63	0.51	(27)
	enolase A	438	0.93	0.82	0.85	0.78	(29)
	enolase B	438	0.96	0.85	0.86	0.75	(30)
	GA-3-PDH 1	331	0.99	0.86	0.86	0.81	(34)
	GA-3-PDH 3	331	0.94	0.75	0.81	0.78	(35)
	Histone 2A1	133	0.78	0.78	0.77	0.62	(38)
	Histone 2A2	133	0.68	0.72	0.71	0.62	(39)
	Histone 2B1	132	0.77	0.64	0.77	0.62	(40)
	Histone 2B2	132	0.71	0.58	0.63	0.65	(41)
	Histone 3	137	0.77	0.72	0.68	0.61	(42)
	Histone 4	104	0.82	0.81	0.84	0.68	(43)
*4	Heat shock protein 90	710	0.66	0.41	0.68	0.64	(46)
	Leu 2	365	0.60	0.44	0.71	0.78	(48)
	Pyruvate kinase	500	0.95	0.79	0.87	0.77	(53)
*4	Ubiquitin	382	0.50	0.36	0.68	0.71	(62)
	PGK	417	0.91	0.75	0.85	0.75	(67)
	TEF 1 Elong. factor 1a	459	0.93	0.78	0.83	0.73	(69)
	TPI	249	0.90	0.75	0.84	0.78	(77)
	Ribosomal protein 29	156	0.83	0.72	0.83	0.55	(86)
*4	Porin	284	0.50	0.35	0.65	0.74	(88)
	GDH 1	455	0.75	0.52	0.78	0.79	(96)
	Ribosomal protein 51B	137	0.83	0.78	0.83	0.68	(102)
*4	HXK 2	486	0.73	0.55	0.72	0.84	(109)
	average indices :		0.81	0.69	0.78	0.70	

codons : length of gene (including termination codon)
CBI : codon bias index of Bennetzen & Hall (23).
G2 : overall codon bias statistic (see text).
P2 : measure of third base pyrimidine bias (4).
tRNA : linear correlation of codon usage with tRNA abundance.

Table 1b. Details of yeast genes clustered in group B.

	Gene / product	codons	CBI	G2	P2	tRNA	Ref.
	Ribosomal protein S33	68	0.63	0.68	0.79	0.31	(14)
	Arg 4	464	0.32	0.20	0.61	0.79	(18)
	ATP 2	313	0.50	0.44	0.53	0.78	(19)
	B-tubulin	458	0.41	0.21	0.64	0.74	(20)
	CBP 2	631	0.09	0.07	0.54	0.49	(21)
	CPA 2	1119	0.30	0.20	0.53	0.78	(23)
	Citrate synthetase	481	0.30	0.28	0.55	0.74	(24)
*3	Cup 1 X	247	0.03	0.19	0.34	0.46	(25)
	iso-2-cytochrome C	114	0.16	0.31	0.65	0.58	(28)
	Gal 1	529	0.20	0.12	0.46	0.67	(31)
	Gal 4	882	0.04	0.06	0.44	0.63	(32)
	Gal 7	185	0.20	0.30	0.48	0.75	(33)
	GCN 4	282	0.30	0.23	0.49	0.64	(36)
	Gal 10	446	0.14	0.15	0.43	0.73	(37)
	His 1	298	0.23	0.20	0.53	0.67	(44)
	His 4	800	0.37	0.21	0.59	0.63	(45)
	Invertase	533	0.43	0.22	0.61	0.76	(47)
	Mat al	176	-0.04	0.17	0.48	0.51	(50)
	Mes 1	752	0.32	0.27	0.56	0.76	(51)
	Pho 5	468	0.56	0.33	0.68	0.71	(52)
*2	PKT 1	317	-0.03	0.12	0.48	0.53	(54)
	PPR 1	905	-0.01	0.10	0.42	0.69	(55)
	Ras 1	310	0.17	0.13	0.44	0.81	(56)
	Trp 1	225	0.05	0.18	0.46	0.76	(57)
	Trp 2	529	0.17	0.15	0.48	0.69	(58)
	Trp 3	485	0.22	0.12	0.59	0.71	(59)
	Trp 5	708	0.45	0.27	0.65	0.76	(60)
	Tuf M	438	0.41	0.31	0.61	0.64	(61)
*1	Mel 1	472	0.23	0.15	0.54	0.78	(63)
	Ura 3	268	0.21	0.23	0.51	0.71	(65)
	Mating factor alpha	166	0.34	0.41	0.62	0.39	(66)
	Cytochrome C oxidase 4	156	0.36	0.36	0.66	0.78	(68)
	PPR 2	129	0.30	0.35	0.59	0.53	(70)
	Car 1	334	0.34	0.20	0.63	0.78	(71)
	Pho 3	468	0.47	0.26	0.66	0.67	(72)
	Rad 6	173	0.21	0.31	0.41	0.53	(73)
*2	2u plasmid - able	424	-0.06	0.14	0.47	0.50	(74)
*2	2u plasmid - baker	374	-0.05	0.11	0.38	0.48	(75)
*2	2u plasmid - charlie	297	0.06	0.15	0.49	0.55	(76)
	Mat A2	120	-0.04	0.23	0.45	0.25	(78)
	Mat a2	211	-0.03	0.16	0.40	0.34	(79)
	Ras 2	323	0.22	0.14	0.58	0.73	(80)
	Ade 4	511	0.28	0.23	0.59	0.72	(81)
	Ade 8	215	0.10	0.10	0.51	0.56	(82)
	CBP 1	655	0.11	0.09	0.46	0.57	(83)
	CBP 6	163	0.00	0.30	0.38	0.57	(84)
	CDC 8	217	0.12	0.12	0.43	0.56	(85)
	Ilv 2	688	0.36	0.24	0.67	0.64	(87)
	Outer membrane prot. 70	618	0.32	0.23	0.55	0.59	(89)

Cytochrome C oxidase 5	154	0.21	0.28	0.59	0.65	(90)
Rad 10	196	0.06	0.22	0.47	0.53	(91)
Ilv 1	577	0.43	0.30	0.66	0.78	(92)
Rad 2	976	0.04	0.08	0.44	0.57	(93)
Rad 3	779	0.10	0.16	0.41	0.59	(94)
Spt 2	334	0.02	0.13	0.44	0.53	(95)
Cpa 1	412	0.28	0.19	0.56	0.78	(97)
Mn SOD	234	0.34	0.25	0.63	0.60	(98)
Rad 52	505	0.09	0.12	0.46	0.58	(99)
Rad 1	973	0.01	0.09	0.33	0.61	(100)
Put 2	576	0.17	0.12	0.48	0.73	(101)
Gal 80	436	0.08	0.11	0.49	0.68	(103)
UCCR 14	128	0.31	0.31	0.56	0.55	(104)
SIR 2	563	0.08	0.10	0.44	0.68	(105)
SIR 3	979	0.01	0.11	0.40	0.57	(106)
Cytochrome C oxidase 6	149	0.31	0.37	0.42	0.46	(107)
CDC 28	299	0.19	0.17	0.54	0.66	(108)
average indices :		0.20	0.21	0.52	0.63	

Table 1c. Details of yeast genes not clustered ("outliers").

Gene / product	codons	CBI	G2	P2	tRNA	Ref.
Ribosomal protein L46	52	0.93	0.99	1.00	0.13	(6)
*3 Cen 3	53	-0.06	0.40	0.36	0.11	(22)
Cup 1 Cu chelatin	62	0.11	0.42	0.62	0.12	(26)
Mat A1	149	0.00	0.29	0.32	0.38	(49)
UCCR	148	0.23	0.18	0.55	0.29	(64)
YP2	207	0.25	0.24	0.67	0.47	(110)

Groups A and B are defined in Figure 1.

*1 S.carlsbergensis, *2 plasmid borne gene, *3 Unidentified open reading frame, *4 genes excluded from Group A when clustered on % codon usage values.

For references to original sequence papers, see Appendix.

clustered into group B. Since several of these genes would be expected to be highly expressed, e.g. actin and ubiquitin, the clustering based on RSCU values is to be preferred. The difference between these two results is due to the effect of amino acid composition, which is successfully removed by use of relative usage within synonymous groups of codons.

Total codon usage data for each of the two major clusters, summed over genes, and excluding outliers, are presented in

Table 2. Relative synonymous codon usage in highly and lowly expressed genes from yeast, and highly expressed genes from E.coli.

		Y-L	Y-H	E-H			Y-L	Y-H	E-H			Y-L	Y-H	E-H			Y-L	Y-H	E-H
Phe	UUU	1.17	0.42	0.46	Ser	UCU	1.87	3.17	2.57	Tyr	UAU	1.05	0.26	0.39	Cys	UGU	1.41	1.78	0.67
	UUC	0.83	1.58	1.54		UCC	0.93	2.17	1.91		UAC	0.95	1.74	1.61		UGC	0.59	0.22	1.33
Leu	UUA	1.63	0.80	0.11		UCA	1.17	0.23	0.20	*	UAA	1.34	2.61	--	*	UGA	1.06	0.24	--
	UUG	1.95	4.50	0.11		UCG	0.50	0.09	0.04	*	UAG	0.60	0.16	--	Trp	UGG	1.00	1.00	1.00
Leu	CUU	0.67	0.13	0.22	Pro	CCU	1.18	0.50	0.23	His	CAU	1.29	0.52	0.45	Arg	CGU	1.06	0.64	4.39
	CUC	0.33	0.02	0.20		CCC	0.62	0.03	0.04		CAC	0.71	1.48	1.55		CGC	0.30	0.01	1.56
	CUA	0.80	0.42	0.04		CCA	1.81	3.44	0.44	Gln	CAA	1.42	1.89	0.22		CGA	0.34	0.00	0.02
	CUG	0.62	0.13	5.33		CCG	0.39	0.03	3.29		CAG	0.58	0.11	1.78		CGG	0.18	0.00	0.02
Ile	AUU	1.52	1.36	0.47	Thr	ACU	1.43	1.94	1.80	Asn	AAU	1.12	0.28	0.10	Ser	AGU	0.93	0.17	0.22
	AUC	0.79	1.58	2.53		ACC	0.91	1.78	1.87		AAC	0.88	1.72	1.90		AGC	0.60	0.16	1.05
	AUA	0.70	0.06	0.01		ACA	1.17	0.22	0.14	Lys	AAA	1.11	0.38	1.60	Arg	AGA	3.00	5.20	0.02
Met	AUG	1.00	1.00	1.00		ACG	0.48	0.06	0.18		AAG	0.89	1.62	0.40		AGG	1.12	0.14	0.00
Val	GUU	1.61	2.18	2.24	Ala	GCU	1.57	2.72	1.88	Asp	GAU	1.31	0.84	0.61	Gly	GGU	2.23	3.80	2.28
	GUC	0.92	1.65	0.15		GCC	0.98	1.13	0.25		GAC	0.69	1.16	1.39		GGC	0.72	0.15	1.65
	GUA	0.75	0.04	1.11		GCA	1.10	0.12	1.10	Glu	GAA	1.43	1.83	1.59		GGA	0.67	0.02	0.02
	GUG	0.72	0.13	0.50		GCG	0.35	0.04	0.80		GAG	0.57	0.17	0.41		GGG	0.38	0.03	0.04

Y : yeast, E : E.coli, H : highly, L : lowly expressed.
Total number of codons : Y-L 28415; Y-H 10172; E-H 6240.

E-H data : rpsU rpsJ rpsL rpsT rpsA rpsB rpsO rpsG rpmB rpmG rpmH rplK rplJ rplA rplL rplQ rplC lpp ompA ompC ompF recA dnaK tufA tufB tsf fusA (all sequences from GenBank, Ref.6)

Table 2. Generally it can be seen that a similar direction of bias occurs in both clusters, but that the bias is much more extreme in the group of highly expressed genes. The various coefficients of bias for individual genes are shown in Table 1. Genes in the highly expressed group have significantly higher bias as assessed by both the CBI and G2 statistics. Those genes also have significantly higher P2 values than genes in group B, where the average value is not significantly greater than 0.5 (no bias). On average, codon usage in the highly expressed genes is more highly correlated with tRNA abundance, but the difference between the groups is small.

DISCUSSION

Cluster analysis of yeast genes by synonymous codon usage clearly yields two relatively homogeneous groups with different patterns of codon preference. With a few possible exceptions, the genes in one of these groups are those known or expected to be highly expressed. Thus genes in yeast can be divided into two groups, of biological significance, on the basis of a purely statistical analysis of synonymous codon usage. If this is confirmed by further experimental data, then it will be possible to predict the likely level of expression of any yeast gene given only the nucleotide sequence of the coding region.

Yeast genes in the highly expressed group have a distinctly more extreme pattern of codon bias. The pattern of synonymous codon preference in highly expressed genes in E.coli can be compared to that in yeast (Table 2). The degree of bias is similar, but the codons preferred in each species are quite different for Leu, Cys, Gln, Arg, Lys and Pro. For Ala and Val the second most favoured codon differs. Thus a gene with the pattern of codon usage optimal for expression in E.coli should not be as highly expressed in yeast as a gene with optimal yeast codons. A preliminary report confirms this (5). Optimizing the expression of heterologous genes in yeast is of great potential interest. While several factors will influence the level of gene expression (25), the presence of codons other than those identified (in Table 2) as being strongly preferred in highly expressed genes may well reduce expression below optimal

Table 3. G+C content of yeast genes.

Genes	G+C content	
	3rd position	Total
A (high)	0.45 (0.50)	0.44 (0.48)
B (low)	0.38 (0.50)	0.41 (0.46)
Average protein	(0.51)	(0.48)

Values that would arise from uniform synonymous codon usage are given in parentheses. Average protein composition from Ref. 28.

levels. Thus the data presented here point the way for in vitro mutagenesis or complete de novo gene synthesis to optimize codon usage to yield maximal gene expression.

It is useful to divide possible influences on codon usage into two types. First, fundamental properties of the DNA molecule in which the gene is embedded, which may not have any direct effect on the gene product encoded, may (nevertheless) influence codon usage. Thus it might be expected that total G+C content would influence the choice of base in degenerate positions of codons (26). Also higher order DNA structures, the simplest being dinucleotides, are often nonrandom in frequency (12,27), and to an extent that it is unlikely to be simply a result of nonrandom codon usage. Second, the interaction of mRNA and tRNA molecules in the translation process may lead to differences in codon fitness, reflecting the direct action of natural selection on codon usage. This would include the influence of tRNA abundance (18), and the hypothesized advantage of intermediate bond strengths between tRNA and mRNA (24), thought to yield third base pyrimidine bias. Within an evolutionary framework, the degree of codon bias in any one gene presumably reflects a balance between selection for use of optimal codons, and synonymous mutations (probably only mildly deleterious) tending to drive towards random codon usage. The point of equilibrium will depend on the strength of selection for optimal codons or, conversely, the extent to which a synonymous mutation yielding a less optimal codon is

deleterious. While codon biases in the first category might apply equally to all genes in a genome, those in the second category should be stronger for genes where high expression is important. Consequently the degree of bias is expected to be higher in genes where high expression is necessary. The 110 genes examined here have a low G+C content (41.4%), as would be expected in yeast. From the base composition of the two clusters of genes (Table 3), it can be seen that the highly expressed group have a higher G+C content, due largely to a difference between the two groups in G+C at the third position of codons. More precisely there is an increase in C, at the expense of A, in the highly expressed group. This is most easily interpreted as the result of selection for particular codons in those genes overcoming the intrinsic mutational bias (to A and T) of the yeast genome. Both the correlation of codon usage with tRNA abundance and the strength of third base pyrimidine bias are stronger in the yeast genes thought to be highly expressed, as had been found in an examination of 83 E.coli genes (9). This difference in pattern between highly and lowly expressed genes explains some but not all of the within species heterogeneity.

There is experimental evidence for the effect of codon usage on gene expression mediated by tRNA abundance (29,30), but the true relevance of the observation of third base pyrimidine bias is still open to question. Translation of the codons UUU and AAU is more error prone than for UUC and AAC, respectively (31). However, this does not appear to be simply due to the G+C content of the codon-anticodon interaction, as no difference in efficiency of translation of poly(U) and poly(UG) has been detected (30). It should be noted that the pattern of codon usage predicted, if intermediate codon-anticodon interaction energies are optimal, is certainly not universally observed. For example, while genes in bacteriophage T7 do show third base pyrimidine bias, highly and lowly expressed genes do not appear to differ in degree of bias (11). Also, a preliminary survey of codon usage in Bacillus genes suggests that this bias may be absent (13).

Grantham and colleagues have used correspondence analysis to group genes according to codon usage (8). Cluster analysis is

an alternative statistical procedure which has desirable properties in certain circumstances. Importantly cluster analysis is perhaps easier to conceptualise. When clear groupings exist, as in the case of these yeast genes, the cluster analysis output in the form of a dendogram (as in Figure 1a) is also easily interpretable. Note that a similar analysis applied to a set of E.coli genes is not so clear, giving results intermediate between the two dendograms shown in Figure 1.

CONCLUSIONS

This analysis of 110 yeast genes has confirmed the pattern of codon usage associated with those genes which are highly expressed (20). Cluster analysis seems to clearly differentiate between these and other genes. Identification of the patterns of yeast codon usage may prove useful in the design of oligonucleotide probes (33), in deducing whether open reading frames in yeast DNA are likely to be protein coding (34), determining the probable level of expression of genes (both heterologous and from yeast) in yeast, and indicating the codons to use in synthetic genes to be expressed in yeast.

ACKNOWLEDGEMENTS

We are grateful to B.A.Cantwell, R.Grantham, M.Gouy, D.G.Higgins, D.C.Shields, G.M.Butler, and particularly David J. McConnell for discussion. Thanks also to P.F.O'Brien and P.Pamilo for assistance.

*To whom correspondence should be addressed

REFERENCES

1. Li, W-H., Gojobori, T. and Nei, M. (1981) Nature 292, 237-239.
2. Miyata, T. and Hayashida, H. (1981) Proc. Natl. Acad. Sci. USA 78, 5739-5743.
3. Li, W-H., Wu, C-I. and Luo, C-C. (1985) Mol. Biol. Evol. 2, 150-174.
4. Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nucl. Acids Res. 12, 6663-6671.
5. Jones, M., Koski, R., Richards, R., Stabinsky, Z., Ferguson, B., Stabinsky, Y. and Alton, K. (1984) Int. Conf. Yeast Genetics, Edinburgh (abstract).
6. GenBank, Genetic Sequence Data Bank, Bolt Beranek and Newman, Cambridge, MA.

7. E.M.B.L. Nucleic Acid Sequence Data Library, E.M.B.L., Heidelberg.
8. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucl. Acids Res. 9, r43-r74.
9. Gouy, M. and Gautier, C. (1982) Nucl. Acids Res. 10, 7055-7074.
10. Ikemura, T. (1981) J. Mol. Biol. 151, 389-409.
11. Sharp, P.M., Rogers, M.S. and McConnell, D.J. (1985) J. Mol. Evol. 21, 150-160.
12. Grantham, R., Greenland, T., Louail, S., Mouchiroud, D., Prato, J.L., Gouy, M. and Gautier, C. (1985) Bull. Inst. Pasteur 83, 95-148.
13. McConnell, D.J., Cantwell, B.A., Devine, K.D., Forage, A.J., Laoide, B.M., O'Kane, C., Ollington, J.F. and Sharp, P.M. (1986) Annals N.Y. Acad. Sci. (in press)
14. Ogasawara, N. (1985) Gene 40, 145-150.
15. Guthrie, C. and Abelson, J. (1982) in The Molecular Biology of the Yeast Saccharomyces, Strathern, J.N., Jones, E.W. and Broach, J.R. Eds., pp. 487-528, Cold Spring Harbor.
16. Ikemura, T. (1982) J. Mol. Biol. 158, 573-597.
17. Ashburner, M., Bodmer, M. and Lemeunier, F. (1984) Develop. Genet. 4, 295-312.
18. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
19. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Nucl. Acids Res. 8, r49-r62.
20. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
21. Yarrow, D. (1984) in The Yeasts a taxonomic study, Kreger-van Rij, N.J.W. Ed., 3rd edn. pp.379-395, Elsevier, Amsterdam.
22. Ward, J.H.,Jr. (1963) J. Amer. Stat. Ass. 58, 236.
23. Kammerer, B., Guyonvarch, A. and Hubert, J.C. (1984) J. Mol. Biol. 180, 239-250.
24. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
25. Mellor, J., Dobson, M.J., Roberts, N.A., Kingsman, A.J. and Kingsman, S.M. (1985) Gene 33, 215-226.
26. Bibb, M.J., Findlay, P.R. and Johnson, M.W. (1984) Gene 30, 157-166.
27. Nussinov, R. (1984) Nucleic Acids Res. 12, 1749-1763.
28. Dayhoff, M.O. and Hunt, L.T. (1972) in Atlas of Protein Sequences, Dayhoff, M.O. Ed., Vol.5, p.D-355, N.B.R.F.
29. Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) J. Mol. Biol. 180, 549-576.
30. Pedersen, S. (1984) EMBO J. 3, 2895-2898.
31. Johnston, T.C., Borgia, P.T. and Parker, J. (1984) Mol. Gen. Genet. 195, 459-465.
32. Andersson, S.G.E., Buckingham, R.H. and Kurland, C.J. (1984) EMBO J. 3, 91-94.
33. Lathe, R. (1985) J. Mol. Biol. 183, 1-12.
34. Staden, R. (1984) Nucleic Acids Res. 12, 551-567.

APPENDIX References for yeast gene sequences (see Table 1).

1. Teem, J.L., Abovich, N., Kaufer, N.F., Schwindinger, W.F., Warner, J.R., Levy, A., Woolford, J., Leer, R.J., van Raamsdonk-Duin, M.M.C., Mager, W.H., Planta, R.J., Schultz,

L., Friesen, J.D., Fried, H. and Rosbash, M. (1984) Nucl. Acids Res. 12, 8295-8312.
2-3. Leer, R.J., van Raamsdonk-Duin, M.M.C., Hagendoorn, M.J.M., Mager, W.H. and Planta, R.J. (1984) Nucl. Acids Res. 12, 6685-6700.
4. Kaufer, N.F., Fried, H.M., Schwindinger, W.F., Jasin, M. and Warner, J.R. (1983) Nucl. Acids Res. 11, 3123-3135.
5. Schaap, P.J., Molenaar, C.M.T., Mager, W.H. and Planta, R.J. (1984) Current Genetics 9, 47-52.
6. Leer, R.J., van Raamsdonk-Duin, M.M.C., Kraakman, P., Mager, W.H. and Planta, R.J. (1985) Nucl. Acids Res. 13, 701-709.
7. Schultz, L.D. and Friesen, J.D. (1983) J. Bact. 155, 8-14.
8. Molenaar, C.M.T., Woudt, L.P., Jansen A.E.M., Mager, W.H., Planta, R.J., Donovan, D.M. and Pearson, N.J. (1984) Nucl. Acids Res. 12, 7345-7358.
9. Teem, J.L. and Rosbash, M. (1983) Proc. Natl. Acad. Sci. USA 80, 4403-4407.
10. see 1.
11. Leer, R.J., van Raamsdonk-Duin, M.M.C., Molenaar, C.M.T., Cohen, L.H., Mager, W.H. and Planta, R.J. (1982) Nucl. Acids Res. 10, 5869-5878.
12. see 8.
13. see 6.
14. Leer, R.J. et al (1983) Nucl. Acids Res. 11, 7759-7768.
15. Gallwitz, D. and Sures, I. (1980) Proc. Natl. Acad. Sci. USA 77, 2546-2550.
16. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3018-3025.
17. Russell, D.W., Smith, M., Williamson, V.M. and Young, E.T. (1983) J. Biol. Chem. 258, 2674-2682.
18. Beacham, I.R., Schweitzer, B.W., Warrick, H.M. and Carbon, J. (1984) Gene 29, 271-279.
19. Saltzgaber-Muller, J., Kunapuli, S.P. and Douglas, M.G. (1983) J. Biol. Chem. 258, 11465-11470.
20. Neff, N.F., Thomas, J.H., Grisafi, P. and Botstein, D. (1983) Cell 33, 211-219.
21. McGraw, P. and Tzagoloff, A. (1983) J. Biol. Chem. 258, 9459-9468.
22. Fitzgerald-Hayes, M., Clarke, L. and Carbon, J. (1982) Cell 29, 235-244.
23. Lusty, C.J., Widgren, E.E., Broglie, K.E. and Nyunoya, H. (1983) J. Biol. Chem. 258, 14466-14472.
24. Suissa, M., Suda, K. and Schatz, G. (1984) EMBO J. 3, 1773-1781.
25-26. Karin, M., Najarian, R., Haslinger, A., Valenzuela, P., Welch, J. and Fogel, S. (1984) Proc. Natl. Acad. Sci. USA 81, 337-341.
27. Smith, M., Leung, D.W., Gillam, S., Astell, C.R., Montgomery, D.L. and Hall, B.D. (1979) Cell 16, 753-761.
28. Montgomery, D.L., Leung, D.W., Smith, M., Shalit, P., Faye, G. and Hall, B.D. (1980) Proc. Natl. Acad. Sci. USA 77, 541-545.
29-30. Holland, M., Holland, J.P., Thill, G.P. and Jackson, K.A. (1981) J. Biol. Chem. 256, 1385-1395.
31. Citron, B.A. and Donelson, J.E. (1984) J. Bact. 158, 269-278.

32. Laughon, A. and Gesteland, R.F. (1984) Mol. Cell. Biol. 4, 260-267.
33. see 31.
34. Holland, J.P. and Holland, M.J. (1979) J. Biol. Chem. 254: 9839-9845.
35. Holland, J.P., Labieniec, L., Swimmer, C. and Holland, M.J. (1983) J. Biol. Chem. 258: 5291-5299.
36. Thireos, G., Penn, M.D. and Greer, H. (1984) Proc. Natl. Acad. Sci. USA 81, 5096-5100.
Hinnebusch, A. (1984) Proc. Natl. Acad. Sci. USA 81, 6442-6446.
37. see 31.
38-39. Choe, J., Kolodrubetz, D. and Grunstein, M. (1982) Proc. Natl. Acad. Sci. USA 79, 1484-1487.
40-41. Wallis, J.W., Hereford, L. and Grunstein, M. (1980) Cell 22, 799-805.
42-3. Smith, M.M. and Andresson, O.S. (1983) J. Mol. Biol. 169, 663-690.
44. Hinnebusch, A.G. and Fink, G.R. (1983) J. Biol. Chem. 258, 5238-5247.
45. Donahue, T.F., Farabaugh, P.J. and Fink, G.R. (1982) Gene 18, 47-59.
46. Farrelly, F.W. and Finkelstein, D.B. (1984) J. Biol. Chem 259, 5745-5751.
47. Carlson, M., Taussig, R., Kustu, S. and Botstein, D. (1983) Mol. Cell. Biol. 3, 439-447.
48. Andreadis, A., Hsu, Y-P., Hermodson, M., Kohlhaw,G. and Schimmel, P. (1984) J. Biol. Chem. 259, 8059-8062.
49-50. Astell, C.R., Ahlstron-Jonasson, L., Smith, M., Tatchell, K., Nasmyth, K. and Hall, B.D. (1981) Cell 27,15-23.
51. Walter, P., Gangloff, J., Bonnet, J., Boulanger, Y., Ebel, J.P. and Fasiolo, F. (1983) Proc. Natl. Acad. Sci. USA 80, 2437-2441.
52. Arima, K., Oshima, T., Kubota, I., Nakamura, N., Mizunaga, T. and Toh-E, A. (1983) Nucl. Acids Res. 11, 1657-1672.
53. Burke, R.L., Tekamp-Olson, P. and Najarian, R. (1983) J. Biol. Chem. 258, 2193-2201.
54. Bostian, K.A., Elliot, Q., Bussey, H., Burn, V., Smith, A. and Tipper, D.J. (1984) Cell 36, 741-751.
55. Kammarer, B., Guyonvarch, A. and Hubert, J.C. (1984) J. Mol. Biol. 180, 239-250.
56. Dhar, R., Nieto, A., Koller, R., DeFeo-Jones, D. and Scolnick, E.M. (1984) Nucl. Acids Res. 12, 3611-3618.
57. Tschumper, G. and Carbon, J. (1980) Gene 10, 157-166.
58-59. Zalkin, H., Paluh, J.L., van Cleemput, M., Moye, W.S. and Yanofsky, C. (1984) J. Biol. Chem. 259, 3985-3992.
60. Zalkin, H. and Yanofsky, C. (1982) J. Biol. Chem. 257, 1491-1500.
61. Nagata, S., Tsunetsugu-Yokota, Y., Naito, A. and Kaziro, Y. (1983) Proc. Natl. Acad. Sci. USA 80, 6192-6196.
62. Ozkaynak, E., Finley, D. and Varshavsky, A. (1984) Nature 312, 663-666.
63. Summer-Smith, M., Bozzato, R.P., Skipper, N., Davies, R.W. and Hopper, J.E. (1985) Gene 36, 333-340.
64. van Loon, A.P.G.M., de Groot, R.J., de Haan, M., Dekker, A. and Grivell, L.A. (1984) EMBO J. 3, 1039-1043.

65. Rose, M., Grisafi, P. and Botstein, D. (1984) Gene 29, 113-124.
66. Kurjan, J. and Herskowitz, I. (1982) Cell 30, 933-943.
67. Hitzeman, R.A., Hagie, F.E., Hayflick, J.S., Chen, C.Y., Seeburg, P.H. and Derynck, R. (1982) Nucl. Acids Res. 10, 7791-7808.
68. Maarse, A.C., van Loon, A.P.G.M., Riezman, H., Gregor, I., Schatz, G. and Grivell, L.A. (1984) EMBO J. 3, 2831-2837.
69. Nagata, S., Nagashima, K., Tsunetsugu-Yokota, Y., Fujimura, K., Miyazaki, M. and Kaziro, Y. (1984) EMBO J. 3, 1825-1830.
70. Hubert, J-C., Guyonvarch, A., Kammerer, B., Exinger, F., Liljelund, P. and Lacroute, F. (1983) EMBO J. 2, 2071-2073.
71. Sumrada, R.A. and Cooper, T.G. (1984) J. Bact. 160, 1078-1087.
72. Bajwa, W., Meyhack, B., Rudolph, H., Schweingruber, A-M. and Hinnen, A. (1984) Nucl. Acids Res. 12, 7721-7739.
73. Reynolds, P. Weber, S. and Prakash, L. (1985) Proc. Natl. Acad. Sci. USA 82, 168-172.
74-76. Hartley, J.L. and Donelson, J.E. (1980) Nature 286, 860-865.
77. Alber, T. and Kawasaki, G. (1982) J. Mol. Appl. Genet. 1, 419-434.
78-79. see 49.
80. see 56.
81. Mantsala, P. and Zalkin, H. (1984) J. Biol. Chem. 259, 8478-8484.
82. White, J.H., Lusnak, K. and Fogel, S. (1985) Nature 315, 350-352.
83. Dieckmann, C.L., Homison, G. and Tzagoloff, A. (1984) J. Biol. Chem. 259, 4732-4738.
84. Dieckmann, C.L. and Tzagoloff, A. (1985) J. Biol. Chem. 260, 1513-1520.
85. Birkenmeyer, L.G., Hill, J.C. and Dumas, L.B. (1984) Mol. Cell. Biol. 4, 583-590.
86. Mitra, G. and Warner, J.R. (1984) J. Biol. Chem. 259, 9218-9224.
87. Falco, S.C., Dumas, K.S. and Livak, K.J. (1985) Nucl.Acids Res. 13, 4011-4027.
88. Mihara, K. and Sato, R. (1985) EMBO J. 4, 769-774.
89. Hase, T., Riezman, H., Suda, K. and Schatz, G. (1983) EMBO J. 2, 2169-2172.
90. Seraphin, B., Simon, M. and Faye, G. (1985) Current Genetics 9, 435-439.
91. Weiss, W.A. and Friedberg, E.C. (1985) EMBO J. 4, 1575-1582.
92. Kielland-Brandt, M.C., Holmberg, S., Petersen, J.G.L. and Nillson-Tillgren, T. (1984) Carlsberg Res. Comm. 49, 567-575.
93. Nicolet, C.M., Chenevert, J.M. and Friedberg, E.C. (1985) Gene 36, 225-234.
94. Naumovski, L., Chu, G., Berg, P. and Friedberg, E.C. (1985) Mol. Cell. Biol. 5, 17-26.
95. Roeder, G.S., Beard, C., Smith, M. and Keranen, S. (1985) Mol. Cell. Biol. 5, 1543-1553.

96. Moye, W.S., Amuro, N., Rao, J.K.M. and Zalkin, H. (1985) J. Biol. Chem. 260, 8502-8508.
97. Nyunoya, H. and Lusty, C.J. (1984) J. Biol. Chem. 259, 9790-9798.
98. Marres, C.A.M., van Loon, A.P.G.M., Oudshoorn, P., van Steeg, H., Grivell, L.A. and Slater, E.C. (1985) Eur. J. Biochem. 147, 153-161.
99. Adzuma, K., Ogawa, T. and Ogawa, H. (1984) Mol. Cell. Biol. 4, 2735-2644.
100. Yang, E. and Friedberg, E.C. (1984) Mol. Cell. Biol. 4, 2161-2169.
101. Krzywicki, K.A. and Brandriss, M.C. (1984) Mol. Cell. Biol. 4, 2837-2842.
102. Abovich, N. and Rosbash, M. (1984) Mol. Cell. Biol. 4, 1871-1879.
103. Nogi, Y. and Fukasawa, T. (1984) Nucl. Acids Res. 12, 9287-9298.
104. de Haan, M., van Loon, A.P.G.M., Kreike, J., Vaessen, R.T.M.J. and Grivell, L.A. (1984) Eur. J. Biochem. 138, 169-177.
105-106. Shore, D., Squire, M. and Nasmyth, K.A. (1984) EMBO J. 3, 2817-2823.
107. Wright, R.M., Ko, C., Cumsky, M.G. and Poyton, R.O. (1984) J. Biol. Chem. 259, 15401-15407.
108. Lorincz, A.T. and Reed, S.I. (1984) Nature 307, 183-185.
109. Frohlich, K-U., Entian, K-D. and Mecke, D. (1985) Gene 36, 105-111.
110. Gallwitz, D., Donath, C. and Sander, C. (1983) Nature 306, 704-707.