

# Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*

Andrew T.Lloyd and Paul M.Sharp\*

Department of Genetics, Trinity College, Dublin 2, Ireland

Received August 6, 1992; Revised and Accepted September 9, 1992

## ABSTRACT

**Codon usage in a sample of 28 genes from the pathogenic yeast *Candida albicans* has been analysed using multivariate statistical analysis. A major trend among genes, correlated with gene expression level, was identified. We have focussed on the extent and nature of divergence between *C.albicans* and the closely related yeast *Saccharomyces cerevisiae*. It was recently suggested that significant differences exist between the subsets of preferred codons in these two species [Brown *et al.* (1991) Nucleic Acids Res. 19, 4293]. Overall, the genes of *C.albicans* are more A + T-rich, reflecting the lower genomic G + C content of that species, and presumably resulting from a different pattern of mutational bias. However, in both species highly expressed genes preferentially use the same subset of 'optimal' codons. A suggestion that the low frequency of NCG codons in both yeast species results from selection against the presence of codons that are potentially highly mutable is discounted. Codon usage in *C.albicans*, as in other unicellular species, can be interpreted as the result of a balance between the processes of mutational bias and translational selection. Codon usage in two related *Candida* species, *C.maltosa* and *C.tropicalis*, is briefly discussed.**

## INTRODUCTION

In many genes from many species, the extent to which alternative synonymous codons are used has been seen to be nonrandom: the pattern of codon usage varies among species, and among genes from the same genome (1). Characterizing the pattern of codon usage in a particular organism may be of interest for several reasons. Knowledge of this pattern may be of practical application, for example in the design of oligonucleotide probes and primers for this species. Elucidation of the biological bases of codon usage in a particular species may add to a general understanding of its molecular biology and evolution, such as the mutational biases present in the genome, and the extent to which different genes are under selection for efficient translation. On a broader scale, since the principal determinants of codon usage are quite well understood in some species, it is of interest

to investigate the generality of these factors. In particular, it is becoming possible to examine to what extent, and how, codon usage patterns have diverged among related species.

Among eukaryotes the species for which codon usage has been best characterized, and is best understood, is the budding yeast *Saccharomyces cerevisiae* (2). In *S.cerevisiae*, highly expressed genes have synonymous codon usage strongly biased towards those codons that are efficiently translated by the most abundant tRNA species (3–5). Genes with low expression, by contrast, have generally more uniform codon usage which appears to be largely determined by mutation biases (5–7). From an evolutionary perspective, the overall codon usage pattern in *S.cerevisiae* appears to be highly co-adapted with the relative abundance of its tRNAs (4,8), and the pattern of codon usage in any particular gene reflects a balance between mutation and selection (6,9). The most closely related species in which codon usage has been examined in detail is the ascomycete *Aspergillus (Emericella) nidulans* (10), where codon usage appears to be determined by similar principles to that in *S.cerevisiae*, but where the particular codons which are preferred for many amino acids differ. Since *S.cerevisiae* and *A.nidulans* are quite divergent, it is interesting to examine a more closely related species.

*Candida albicans* is a dimorphic fungus, growing alternatively as a yeast and a hyphal mass. It is part of the normal flora of human mucous membranes, but is of increasing interest because of its particular pathogenicity in immuno-compromised AIDS patients. Research has been carried out into developmental processes, such as the control mechanisms of the switch to the hyphal growth phase in which its invasive properties are most manifest (11,12), and into the causes of susceptibility and resistance to various drugs (13,14). *C.albicans* is quite closely related to *S.cerevisiae*: they are perhaps one third/one half as divergent from each other as from *A.nidulans* (15,16). However, the overall genomic base composition of *C.albicans* (G+C = 35%) is more A+T-rich than that of *S.cerevisiae* (G+C = 40%) (17).

In a recent survey (18), Brown *et al.* have suggested that some of the 'preferred' codons in *C.albicans* differ from those in *S.cerevisiae*. However, their analysis may be flawed in two respects. First, data for only 11 genes were available. Second, and more important, Brown *et al.* considered only the total codon

\* To whom correspondence should be addressed

Table 1. *Candida albicans* gene sequence dataset.

Gene	Product	L	GC <sub>3S</sub>	N <sub>c</sub>	CAI	Acc. #	Reference
<i>TEF1</i>	elongation factor 1- $\alpha$	458	0.35	26.1	0.77	M29934	JBa 172:2036
<i>TEF2</i>	elongation factor 1- $\alpha$	458	0.36	26.2	0.77	M29935	JBa 172:2036
<i>ACT1</i>	actin	376	0.26	27.3	0.73	X16377	NAR 17:9488
<i>PMA1</i>	plasma membrane ATPase	895	0.31	28.0	0.69	M74075	JBa 173:6826
<i>CYP1</i>	peptidyl-propyl isomerase	162	0.34	31.4	0.60	M60628	Gene 96:189
<i>CEF3</i>	elongation factor 3	1049	0.29	28.3	0.64	Z11484	Unpublished
<i>TUB2</i>	$\beta$ -tubulin	449	0.19	28.8	0.45	M19398	Gene 63:53
<i>PRA1</i>	aspartyl proteinase	380	0.37	37.6	0.34	X13669	NAR 17:1779
<i>PRA2</i>	aspartyl proteinase	398	0.16	30.3	0.30	M83663	Unpublished
<i>PRA3</i>	aspartyl proteinase	391	0.20	32.3	0.35	X56867	JMVM 29:129
<i>CMD1</i>	calmodulin	149	0.24	40.3	0.27	M61128	Gene 106:43
<i>RBP1</i>	rapamycin binding protein	124	0.07	27.5	0.35	M84759	Gene 113:125
<i>ERG16</i>	14- $\alpha$ -demethylase	528	0.11	28.1	0.30	X13296	NAR 17:804
<i>PHR1</i>	glycolipid-anchored membrane protein	551	0.31	38.0	0.29	M90812	Unpublished
<i>BMR</i>	benomyl resistance	564	0.25	35.3	0.24	X53823	MGG 227:318
<i>CHS1</i>	chitin synthase	776	0.32	42.4	0.17	X52420	MM 4:197
<i>KRE1</i>	$\beta$ -glucan synthesis	130*	0.25	51.1	0.13	M81588	JBa 173:6859
<i>CAG1</i>	G-protein $\alpha$ subunit	429	0.28	42.5	0.22	M88113	MCB 12:1977
<i>ARF1</i>	ADP ribosylation factor	179	0.40	47.1	0.17	M54910	Unpublished
<i>URA3</i>	orotidine-decarboxylase	270	0.18	37.5	0.19	X14198	Unpublished
<i>TSA1</i>	thymidylate synthase	315	0.20	41.7	0.18	J04230	JBa 171:1372
<i>ERK1</i>	protein kinase	417	0.10	33.1	0.20	M76585	Unpublished
<i>DFR1</i>	dihydrofolate reductase	192	0.27	44.5	0.17	-	+
<i>CHS2</i>	chitin synthase	1009	0.17	38.8	0.18	M82937	MM 6:497
<i>CLN1</i>	cyclin	646*	0.11	35.9	0.18	M76587	Unpublished
<i>SOR2</i>		471	0.28	47.1	0.15	-	+
<i>ORF</i>	mating type interference	346	0.38	46.0	0.14	M83991	Unpublished
<i>ZNF1</i>	zinc-finger protein	388	0.29	50.3	0.14	M76586	Unpublished

L, length in amino acid residues (\* incomplete sequence); G+C, gene G+C content; GC<sub>3S</sub>, G+C content at silent third positions; N<sub>c</sub>, effective number of codons; CAI, Codon Adaptation Index; Acc. #, GenBank/EMBL/DDBJ accession number. The following reference abbreviations are used: JBa, *J. Bacteriol.*; JMVM, *J. Med. Vet. Mycol.*; MCB, *Mol. Cell. Biol.*; MGG, *Mol. Gen. Genet.*; MM, *Mol. Microbiol.*; NAR, *Nucleic Acids Res.*  
+ codon usage taken from Ref.14.

usage over all 11 genes, and did not take account of any heterogeneity within the data set; this approach can lead to serious misconceptions in the conclusions drawn (19). Here, we conduct a more thorough analysis of codon usage in *C.albicans*, exploiting a larger data set, examining the substantial heterogeneity among genes, and investigating the pattern of divergence from *S.cerevisiae* in the context of the heterogeneity seen within each species.

## MATERIALS AND METHODS

A data set consisting of 28 genes from *C.albicans* is shown in Table 1. Codon usage data were calculated for protein coding gene sequences extracted from the GenBank/EMBL/DDBJ DNA sequence data library (GenBank release 72) using the ACNUC retrieval system (20). In addition, codon usage data were obtained directly from the literature (14) for two sequences not yet incorporated into the database.

Codon usage differences among genes were investigated using correspondence analysis (21), the multivariate statistical technique most often used in the analysis of codon usage data (see, for example, Refs.10,19,22). This method can identify major trends in the dataset as a series of orthogonal axes in an n-dimensional hyperspace. The first axis explains the highest (and subsequent axes a diminishing) proportion of the variation in the data. Correspondence analysis was implemented using the Cornell Ecology Group program 'DECORANA', written in FORTRAN by M.O.Hill.

A number of indices of codon usage bias were calculated for each gene:

GC<sub>3S</sub>: the frequency of G+C at *silent* (i.e., synonymously variable) third positions of sense codons (i.e., excluding Trp, Met and stop codons).

N<sub>c</sub>: the 'effective number of codons' used in a gene (23). This is a measure of general nonuniformity of codon usage: for a gene with extreme codon usage bias (where only one codon is used for each amino acid) the N<sub>c</sub> value equals 20, while for a gene with random codon usage the N<sub>c</sub> value is 61. (For details of the calculation, see Ref.23).

CAI: the Codon Adaptation Index (24) is the geometric mean of the 'relative adaptedness' (*w*) values of all the codons in a gene, where *w* is defined from each codon's frequency in a reference set of highly expressed genes from the appropriate species. In this case, the codon usage of highly expressed genes from *S.cerevisiae* (24) was used.

A computer program (CODONS; Ref.25) to calculate codon usage and these indices is available on request.

## RESULTS AND DISCUSSION

### Codon usage in *Candida albicans*

Codon usage has been examined in a sample of 28 *C.albicans* genes, encoding a variety of different products expressed at a wide range of levels (Table 1). The total codon usage for these genes (Table 2) reveals a general bias towards U- and A-ending codons, as expected in an A+T-rich genome. However, the range

**Table 2.** Codon usage in *Candida albicans*.

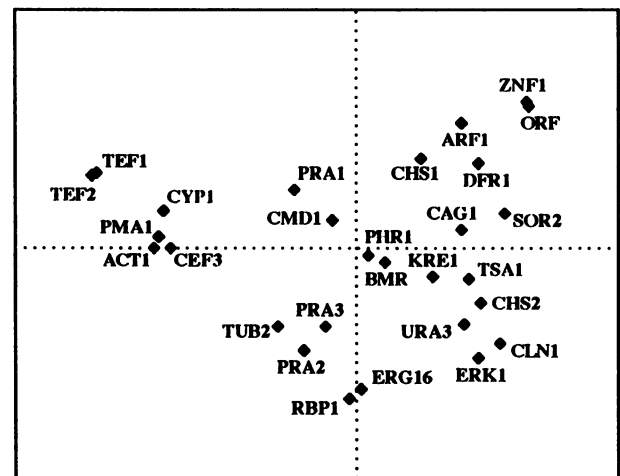
N RSCU (S.c.)		N RSCU (S.c.)		N RSCU (S.c.)		N RSCU (S.c.)	
Phe	UUU 345 1.17 (1.08)	Ser	UCU 322 2.00 (1.83)	Tyr	UAU 300 1.32 (1.02)	Cys	UGU 149 1.86 (1.34)
	UUC 245 0.83 (0.92)		UCC 144 0.90 (1.09)		UAC 153 0.68 (0.98)		UGC 1 0.14 (0.66)
Leu	UUA 430 2.44 (1.60)		UCA 252 1.57 (1.16)	ter	UAA 15 1.80 (1.56)	ter	UGA 2 0.24 (0.84)
	UUG 467 2.65 (2.08)		UCG 45 0.28 (0.51)	ter	UAG 8 0.96 (0.60)	Trp	UGG 143 1.00 (1.00)
Leu	CUU 96 0.54 (0.66)	Pro	CCU 134 0.94 (1.18)	His	CAU 167 1.44 (1.20)	Arg	CGU 75 0.97 (0.99)
	CUC 19 0.11 (0.28)		CCC 40 0.28 (0.54)		CAC 65 0.56 (0.80)		CGC 3 0.04 (0.29)
	CUA 20 0.11 (0.79)		CCA 376 2.63 (1.88)	Gln	CAA 474 1.75 (1.46)		CGA 33 0.42 (0.30)
	CUG 26 0.15 (0.59)		CCG 22 0.15 (0.39)		CAG 67 0.25 (0.54)		CGG 8 0.10 (0.17)
Ile	AUU 557 2.00 (1.47)	Thr	ACU 427 2.10 (1.50)	Asn	AAU 437 1.28 (1.11)	Ser	AGU 153 0.95 (0.84)
	AUC 180 0.65 (0.89)		ACC 204 1.00 (0.97)		AAC 244 0.72 (0.89)		AGC 49 0.30 (0.57)
	AUA 98 0.35 (0.63)		ACA 158 0.78 (1.06)	Lys	AAA 592 1.44 (1.05)	Arg	AGA 329 4.24 (3.20)
Met	AUG 261 1.00 (1.00)		ACG 26 0.13 (0.47)		AAG 230 0.56 (0.95)		AGG 18 0.23 (1.05)
Val	GUU 490 2.42 (1.73)	Ala	GCU 487 2.38 (1.73)	Asp	GAU 571 1.58 (1.25)	Gly	GGU 592 2.84 (2.35)
	GUC 170 0.84 (0.96)		GCC 176 0.86 (0.97)		GAC 152 0.42 (0.75)		GGC 49 0.24 (0.65)
	GUA 70 0.35 (0.66)		GCA 138 0.68 (0.95)	Glu	GAA 659 1.82 (1.46)		GGA 127 0.61 (0.64)
	GUG 79 0.39 (0.65)		GCG 16 0.08 (0.35)		GAG 64 0.18 (0.54)		GGG 66 0.31 (0.37)

N, number of codons; RSCU, relative synonymous codon usage; S.c., *Saccharomyces cerevisiae* RSCU values (from Ref.2).

in the values of two indicators of general codon usage bias, namely the G+C content at silent sites ( $GC_{3S}$ , which varies from 0.10 to 0.40) and the effective number of codons used in the gene ( $N_c$ , which varies from 26.1 to 51.1), indicate that there is substantial heterogeneity of codon usage patterns among these genes (Table 1). Thus, the overall pattern of codon usage seen in Table 2 may have some general practical utility in the design of oligonucleotide probes and primers, but is not very useful in investigating the biological bases of codon usage in this species.

Codon usage patterns are complex, and so to elucidate the pattern of differences among genes it is necessary to use multivariate statistical analyses. Correspondence analysis of the usage of 59 codons (i.e., excluding Met, Trp and stop codons) across the 28 *C.albicans* genes yields a first axis (the most important trend among genes) that accounts for a high proportion (40%) of the variation in the dataset. The second axis accounts for a further 15% of the variation, and the remaining (fifty seven) axes are each responsible for relatively trivial amounts. Thus, there is one major, and perhaps one secondary, trend in the dataset: the position of each gene on these first two axes is shown in Fig. 1. In Table 1 the genes are presented in order of their appearance on axis 1. Position on axis 1 is highly correlated with  $N_c$ , the measure of general codon usage bias (correlation coefficient,  $r=0.80$ ). Genes at the top of the table appear to the left in Fig. 1, and have low  $N_c$  values indicating strong codon usage bias. The second axis, on the other hand, is highly correlated with  $GC_{3S}$  ( $r=0.81$ ): genes towards the top of Fig. 1 are more G+C-rich at silent sites.

The first axis of the correspondence analysis appears to have differentiated genes according to their expression level. Thus, the genes at one extreme of the first axis (towards the left in Fig. 1), which have the most highly biased codon usage, are those which are known or expected to be expressed at the highest levels; e.g., they include genes encoding elongation factors and actin, which are highly abundant proteins. Genes at the other extreme of the first axis encode protein kinase or other regulatory products, generally expressed at low levels. In between, there are genes encoding products generally expressed at moderate



**Figure 1.** Correspondence analysis of codon usage in 28 *Candida albicans* genes. Each point is a gene plotted at its coordinates on the first two axes produced by the analysis (axis 1 is horizontal)

levels. It is not clear whether there is any property of the genes systematically related to variation in G+C-content at silent sites, to cause the dispersion on the second axis produced by correspondence analysis. It has been suggested that there may be regional differences in G+C content around the *S.cerevisiae* genome (26), and it would be expected that if such an effect exists it would also be found in *C.albicans*; however, to our knowledge, there is not yet any direct evidence for this phenomenon.

Where highly and lowly expressed genes clearly differ in their codon usage, a comparison of these patterns should indicate which codons are translationally 'optimal'. Codon usage for groups of 6 genes drawn from each end of the major trend, representing high (*TEF1*, *TEF2*, *ACT1*, *PMA1*, *CYP1* and *CEF3*) and low (*ZNF1*, *ORF*, *SOR2*, *CLN1*, *CHS2* and *DFR1*) expression levels, is indicated in Table 3. Twenty one codons are significantly more common (as assessed by  $2 \times 2$   $\chi^2$  contingency tables,  $p < 0.01$ )

in the highly expressed genes. Thus, 'optimal' codons can be identified for 17 of the 18 amino acids encoded by more than one triplet. The exception is Cys, where UGU may be optimal, but it occurs at such high frequencies in the 'low' dataset that the difference between the 'high' and 'low' groups is not significant in the current data set.

Recently, Brown *et al.* (18) designated 24 codons as 'preferred' in *C.albicans*. However, they examined the total codon usage across 11 genes, including both highly and lowly expressed genes; the pattern of codon usage they observed was similar to that in Table 2. That such an approach can lead to erroneous conclusions, particularly in an A+T-rich genome, was previously found in an analysis of codon usage in *Dictyostelium discoideum* (19). The 24 codons identified by Brown *et al.* include the singleton codons for Met (AUG) and Trp (UGG), which are necessarily the 'optimal' codons for those amino acids. Among the remaining 22 codons are 7 (UUU, UUA, UAU, CAU, AAU, AAA and GAU) which are clearly not 'optimal' (in the sense of being used more often in highly expressed genes than in lowly expressed genes, see Table 3), but which occur at high frequencies in the total codon usage data (Table 2) because of the general A+T-richness of the *C.albicans* genome. Because of their approach, Brown *et al.* failed to identify 7 C- or G-ending codons (UUC, CAC, AUC, ACC, AAG, GUC and GAC) which are used at significantly higher frequencies in the more highly expressed genes (Table 3). Note that, because of the overall A+T-richness of the *C.albicans* genome, even in the highly expressed genes two of these 'optimal' codons (AAG and GAC) do not become the most commonly used.

### CpG dinucleotide frequencies

Drouin has recently noted the low frequency of CpG dinucleotides in the actin genes of several fungal species including *C.albicans* and *S.cerevisiae*, and has speculated that it results from a 'positive selection for codons that do not have highly mutable CpG dinucleotides' (27). However, while the CpG dinucleotide is frequently methylated (and consequently highly mutable) in vertebrate genomes, the fraction of methylated cytosine in fungi

is generally very low (28). Furthermore, it has been pointed out (in a different context) that selection against codons based on their potential to mutate to deleterious codons is most unlikely to be effective (Ref.29, p.187).

NCG (and CGN) codons are generally rare in *C.albicans* and *S.cerevisiae* (Table 2), and are especially unusual in highly expressed genes, such as those encoding actin (Table 3; see also Ref.2). The two factors discussed above as shaping *C.albicans* codon usage in general, can also be invoked to explain the low occurrence of CpG-containing codons in particular. First, CpG is rare because of the strong mutational bias to A+T-richness. This is supported by the observation that the *C.albicans* genome, which is more A+T-rich than the *S.cerevisiae* genome, has even fewer NCG codons (Table 2). Second, in *S.cerevisiae* (and probably also *C.albicans*) NCG codons are not recognized by the major tRNA species for the relevant amino acids (at least for Ser, Thr and Ala, and presumably also for Pro; Ref.30), and nor are the CGN codons for Arg, and so CpG-containing codons are selected against on this basis. This is correlated with the exaggerated rarity of NCG and CGN codons in more highly expressed genes (Table 3).

### Comparison of codon usage in *C.albicans* and *S.cerevisiae*

Brown *et al.* have suggested that there are significant differences between the subsets of strongly preferred codons in *C.albicans* and *S.cerevisiae* (18). However, they compared overall codon usage in all 11 *C.albicans* genes then available, with codon usage in highly expressed *S.cerevisiae* genes; the latter data (from Ref.31) had been taken from the most extremely biased 10% of 154 genes.

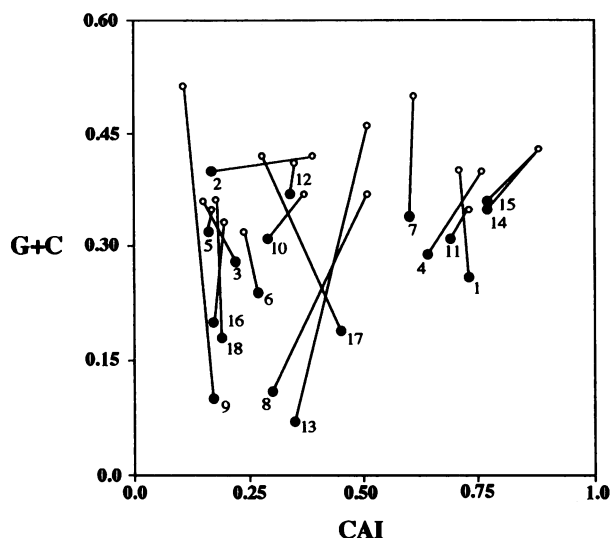
With respect to total codon usage (Table 2), the two yeasts differ primarily with respect to G+C content at silent sites (the overall GC<sub>3S</sub> values are 0.25 and 0.37, for *C.albicans* and *S.cerevisiae*, respectively), reflecting the greater A+T-richness of the *C.albicans* genome. However, the set of 21 optimal codons identified above for *C.albicans* (see Table 3) are also those identified for *S.cerevisiae* (2). The single difference between the sets of optimal codons for the two species pertains to UGU, which

Table 3. Codon usage in high and low bias *Candida albicans* genes.

	High N	Low N	High RSCU	Low RSCU		High N	Low N	High RSCU	Low RSCU		High N	Low N	High RSCU	Low RSCU				
Phe	UUU	33	0.49	101	1.59	Ser	UCU*	121	3.69	61	1.24	Tyr	UAU	17	0.45	100	1.59	
	UUC*	101	1.51	26	0.41		UCC*	53	1.61	26	0.53		UAC*	58	1.55	26	0.41	
Leu	UUA	63	1.44	119	2.86	Leu	UCA	13	0.40	96	1.96	ter	UAA	6	3.00	1	1.00	
	UUG*	191	4.37	71	1.70		UCG	1	0.03	21	0.43		UAG	0	0.00	2	2.00	
Leu	CUU	8	0.18	28	0.67	Pro	CCU	4	0.11	60	1.31	His	CAU	19	0.59	67	1.84	
	CUC	0	0.00	7	0.17		CCC	2	0.05	24	0.52		CAC*	45	1.41	6	0.16	
	CUA	0	0.00	10	0.24		CCA*	143	3.84	88	1.92		Gln	CAA*	98	2.00	169	1.59
	CUG	0	0.00	15	0.36		CCG	0	0.00	11	0.24		CAG	0	0.00	44	0.41	
Ile	AUU*	172	2.20	118	1.86	Thr	ACU*	132	2.43	86	1.67	Asn	AAU	41	0.63	167	1.49	
	AUC*	63	0.80	25	0.39		ACC*	82	1.51	32	0.62		AAC*	89	1.37	57	0.51	
	AUA	0	0.00	47	0.74		ACA	2	0.04	73	1.42		AAA	162	1.22	132	1.53	
Met	AUG	81	1.00	56	1.00	ACG	1	0.02	15	0.29	Lys	AAG*	103	0.78	41	0.47		
Val	GUU*	187	2.67	62	1.70	Ala	GCU*	213	3.06	54	1.59	Asp	GAU	129	1.28	121	1.78	
	GUC*	85	1.21	23	0.63		GCC	65	0.94	26	0.76		GAC*	73	0.72	15	0.22	
	GUA	2	0.03	30	0.82		GCA	0	0.00	48	1.41		GAA*	237	1.98	153	1.75	
	GUG	6	0.09	31	0.85		GCG	0	0.00	8	0.24		GAG	3	0.03	22	0.25	
											Gly	GGU*	262	3.95	70	1.90		
												GGC	1	0.02	18	0.49		
												GGA	1	0.02	37	1.01		
												GGG	1	0.02	22	0.60		

N, number of codons; RSCU, relative synonymous codon usage; \* 'optimal' codons.

can be seen to be optimal in *S.cerevisiae*, and is probably also optimal in *C.albicans*, although it has not been designated as such here for reasons discussed above. Thus, with respect to the codons which appear to be preferred by natural selection for translational efficiency, there is no evidence that the two species have diverged. Nevertheless, codon usage in highly expressed genes is more A+T-rich in *C.albicans* than in *S.cerevisiae*.



**Figure 2.** Comparison of codon usage bias in *C.albicans* and *S.cerevisiae*. Axes: CAI, Codon Adaptation Index. G+C, G+C content at silent third positions of codons. The lines connect points for homologous genes from *C.albicans* (filled circles) and *S.cerevisiae* (open circles). The genes are: 1, *C.albicans ACT1* (*S.cerevisiae* homologue: *ACT1*, GenBank/EMBL accession number L00026, proteins have 95% amino acid identity); 2, *ARF1* (*ARF2*, M35158, 78%); 3, *CAG1* (*GPA1*, M15867, 66%); 4, *CEF3* (*YEF3*, J05583, 78%); 5, *CHS1* (*CHS2*, M23865, 56%); 6, *CMD1* (*CMD1*, M14760, 62%); 7, *CYP1* (*CPH*, M30513, 81%); 8, *ERG16* (*14DM*, M21483, 65%); 9, *ERK1* (*FUS3*, M31132, 62%); 10, *PHR1* (*GAS1*, X53424, 55%); 11, *PMA1* (*PMA1*, X03534, 83%); 12, *PRA1* (*PEP4*, M13358, 71%); 13, *RBP1* (*RBP1*, M63892, 61%); 14, *TEF1* (*TEF1*, X01638, 90%); 15, *TEF2* (*TEF1*, X01638, 90%); 16, *TSA1* (*TMP1*, J02706, 76%); 17, *TUB2* (*TUB2*, J01384, 83%); 18, *URA3* (*URA3*, K02207, 72%).

Among low expression genes, the overall degree of bias as measured by  $N_c$  is greater in *C.albicans* (Table 1) than in *S.cerevisiae* (see Fig. 1 in Ref.2). This difference could be due to the relatively small sample of genes in the *C.albicans* data set. However, comparison of the RSCU values for the lowly expressed genes of *C.albicans* (Table 3) with those of *S.cerevisiae* (see Table 4 in Ref.2, or Table 1 in Ref.31) reveals that the greater degree of nonuniformity of codon usage in *C.albicans* is due to a strong bias toward A+T-rich codons in these genes. This is to be expected. In the presence of only weak selection for optimal codons, genes have a codon usage pattern shaped by mutational biases—the genomic G+C contents of the two yeasts suggest that the mutational bias is more extreme in *C.albicans*. (See also Wright's discussion of the relationship between  $N_c$  and G+C content at silent sites; Ref.23.)

A more direct comparison of codon usage in the two yeasts can be made for 18 pairs of homologous genes which have been sequenced in both species. The codon adaptation index, CAI, is a species-specific measure of codon usage bias defined for *S.cerevisiae* taking into account the 'adaptedness' of synonymous codons derived from their relative usage in those genes where codon selection is strongest, i.e., highly expressed genes (24). Since the optimal codons in *S.cerevisiae* and *C.albicans* do not appear to have diverged, it may be appropriate to calculate CAI values for the *C.albicans* genes using the *S.cerevisiae* adaptedness values. CAI values for the pairs of homologues are very highly correlated ( $r = 0.92$ ) (Fig. 2), suggesting that expression levels (and the consequent strengths of selection) are similar in the two species. However, G+C content at silent sites ( $GC_{3S}$ ) is quite different between some homologues—without exception, the *C.albicans* gene is more A+T-rich at silent sites (Fig. 2). It might be expected that the largest shifts in G+C content would be seen in the lowly expressed genes with low CAI values (as seen in the divergence between *Escherichia coli* and *Serratia marcescens*; Ref.32). However, while only genes with low CAI values have very large differences in G+C, there are nevertheless some other low bias genes with only small G+C shifts (Fig. 2). It is not clear why these genes which are apparently under little selective constraint at silent sites (as evidenced by their low CAI values) have not diverged more in G+C content.

**Table 4.** *Candida tropicalis* gene sequences.

Gene	product	L	$GC_{3S}$	$N_c$	CAI	Acc. #	Reference
<b>Peroxisomal genes</b>							
<i>ICL</i>	isocitrate lyase	550	0.54	26.1	0.77	D00703	JBio 107:262
<i>CAT</i>	catalase	485	0.53	29.5	0.67	X13978	NAR 17:3600
<i>POX4</i>	acyl-CoA oxidase II	709	0.54	28.3	0.68	M12160	PNAS 83:1232
<i>POX5</i>	acyl-CoA oxidase I	662	0.57	28.8	0.61	M12161	PNAS 83:1232
<i>HDE</i>	trifunctional enzyme	906	0.52	30.2	0.68	X57854	Gene 105:129
<i>POX2</i>	acyl-CoA oxidase	724	0.53	29.3	0.61	M18259	Gene 58:37
<i>AOX</i>	acyl-CoA oxidase	709	0.52	30.1	0.58	M16193	Gene 51:119
<b>Other genes</b>							
<i>14DM</i>	14- $\alpha$ -demethylase	95*	0.25	28.5	0.63	M17595	BBRC 146:1311
<i>ATP2</i>	ATPase $\beta$ subunit	511	0.20	28.6	0.54	X54875	unpublished
<i>ORF</i>	(unknown)	1088	0.21	29.4	0.48	M64984	unpublished
<i>CPR</i>	NADPH cytochrome P450 reductase	680	0.23	31.6	0.38	M35199	JBC 265:16428
<i>P450alk2</i>	cytochrome P450	543	0.28	34.0	0.33	M63258	Gene 106:51
<i>ACP</i>	acid protease	394	0.32	36.8	0.34	X61438	FEBS 286:181
<i>P450alk1</i>	cytochrome P450	522	0.25	35.6	0.29	M63258	Gene 106:51

See Table 1 for explanation of symbols. The following additional reference abbreviations are used: BBRC, *Biochem. Biophys. Res. Comm.*; FEBS, *FEBS Lett.*; JBC, *J. Biol. Chem.*; JBio, *J. Biochem.*; PNAS, *Proc. Natl. Acad. Sci., USA*.

Table 5. *Candida maltosa* gene sequences.

Gene	Product	L	GC <sub>3S</sub>	N <sub>c</sub>	CAI	Acc. #	Reference
<i>POX4</i>	acyl-CoA oxidase	709	0.30	27.2	0.65	X06721	NAR 16:365
<i>POX18Cm</i>	alkane-induced protein	127	0.32	36.8	0.45	M61102	Gene 106:61
<i>FAHR</i>	formaldehyde resistance	381	0.25	31.7	0.52	M58332	unpublished
<i>P450alk</i>	cytochrome P450	523	0.25	30.0	0.47	D00481	ABC 53:2217
<i>P450cm1</i>	cytochrome P450	523	0.23	29.4	0.47	X51931	unpublished
<i>P450cm2</i>	cytochrome P450	538	0.26	31.2	0.40	X51932	unpublished
<i>ADE1</i>	succinyl-carboxamide synthase	290	0.26	33.8	0.43	M58322	Gene 107:161
<i>LEU2</i>	isopropyl-malate synthase	373	0.25	33.7	0.43	X05459	CG 11:451
<i>AL11</i>	alkane assimilation	276	0.33	35.8	0.37	M61102	Gene 106:61
<i>CYHR</i>	cyclohexamide resistance	552	0.28	37.4	0.26	M64932	unpublished
<i>HIS5</i>	histidinol NH <sub>2</sub> transferase	389	0.26	39.2	0.24	X17310	CG 16:261
<i>ORF</i>	(unknown)	140*	0.27	41.3	0.24	X17310	CG 16:261

See Table 1 for explanation of symbols. The following additional reference abbreviations are used: ABC, *Agric. Biol. Chem.*; CG, *Curr. Genet.*

### Other *Candida* species

We have also examined codon usage in *C. maltosa* and *C. tropicalis*, the two other *Candida* species for which the most sequence data are currently available. There are 14 *C. tropicalis* genes (Table 4), totalling 8578 codons, and 12 *C. maltosa* genes (Table 5), totalling 4821 codons. While molecular phylogenetic analyses have revealed that species that have been placed in the genus *Candida* do not form a monophyletic group, small subunit ribosomal RNA comparisons indicate that *C. tropicalis* is much more closely related to *C. albicans* than to other yeasts (15). The genomic G+C contents of *C. albicans* and *C. tropicalis* are similar (17).

We subjected the *C. tropicalis* and *C. maltosa* genes to correspondence analysis together with the *C. albicans* genes, to investigate whether codon usage in all three species follow the same trends. The results are not shown, but the genes in Tables 4 and 5 are presented (as with *C. albicans*) in order of appearance on the first axis. Among the *C. tropicalis* genes there are seven encoding peroxisomal proteins, which are highly expressed (32,33). These genes have a highly biased codon usage, with a similar position on axis 1 to the highly expressed *C. albicans* genes, and with high CAI values (Table 4). However, there is a tendency to a higher G+C content at silent sites in these peroxisomal genes (Table 4), and they are differentiated from the *C. albicans* genes on axis 2 of the correspondence analysis. Among the seven other *C. tropicalis* genes, that encoding 14- $\alpha$ -demethylase also has high codon usage bias (like its homologue from *S. cerevisiae*; Ref.2), and lies close to the highly expressed *C. albicans* genes on both axes of the correspondence analysis. Codon usage in the other six genes is moderately biased, and they lie among the central cluster of *C. albicans* genes. There are no *C. tropicalis* genes in the current data set with codon usage similar to the low extreme of the *C. albicans* trend, but this may well reflect the small number of genes so far sequenced for the former species.

The correspondence analysis results indicate that the *C. maltosa* genes all fall along the same trend as those from *C. albicans*. The most highly biased gene is *POX4* (another peroxisomal gene), with codon usage similar to the highly expressed *C. albicans* genes, as indicated by its CAI and GC<sub>3S</sub> values (Table 5). There is a trend across the other genes (including, first, *POX18*) toward progressively lower codon usage bias, though again there are no genes as extreme as the low group of *C. albicans* genes.

Thus, codon usage patterns in the three *Candida* species appear to be similar, with the exception of the increased G+C content of the *C. tropicalis* peroxisomal genes. This could reflect the location of expression of these genes, though we have noted that the peroxisomal genes from *C. maltosa* do not show the same effect. As more genes of different expression levels are sequenced it should be possible to determine whether there are any real differences among these species, or effects peculiar to genes expressed in the peroxisome.

### CONCLUSIONS

The patterns and biological bases of codon usage in *Saccharomyces cerevisiae* have previously been investigated in extensive detail, and appear to be well understood (2-7). *Candida albicans* is closely related to *S. cerevisiae*, but its codon usage has been reported to be somewhat divergent (18). However, in that report only a very small number of genes were examined, and there was no attempt to take account of the heterogeneity of codon usage patterns which are usually evident among genes, even when they are derived from a single species (31). We have found that there is some overall difference between the codon usage patterns in these two species, in that all *C. albicans* genes show an increased tendency to A+T-richness at silent sites. However, we conclude that the codons which are translationally optimal in the two species are identical. These results are easily interpretable in the light of the mutation-selection balance theory of codon usage (6,9): the pattern of mutation bias has diverged between the two species, but the direction of natural selection has not. Ultimately, if divergence of mutational biases continued, it might be expected that changes in the optimal codons would result (34), and this may explain why the more distantly related species *A. nidulans* has a more divergent pattern of codon usage, in which the optimal codons differ from those of the two yeasts (10).

In their recent compilation of codon usage from the GenBank database, Wada et al. (35) presented overall codon usage values for 27 *Candida* genes. Their values differ quite considerably from the total values presented here for *C. albicans* (Table 2). In principle, this could be due to the fact that Wada et al. included in their compilation genes from two different species (*C. albicans* and *C. tropicalis*). In practice, in this case this is probably not the explanation, since we have found above that (at least for the

three species *C.albicans*, *C.maltosa* and *C.tropicalis*) codon usage patterns are broadly similar. In fact, it appears that the compilation of Wada *et al.* may have been heavily biased by the presence of several *C.tropicalis* peroxisomal genes which are more G+C-rich than other genes from that species, and from *C.albicans*. We have also emphasized before (19,31), and above, the potentially misleading effects of examining only a total codon usage table, even when the genes are all drawn from one species.

## ACKNOWLEDGEMENTS

This is a paper from the Irish National Centre for Bioinformatics. We thank Peter Whittaker for discussion. This work was supported by grant SC/91/603 from EOLAS (The Irish Science and Technology Agency).

## REFERENCES

1. Aota, S.-i., Gojobori, T., Ishibashi, F., Maruyama, T. and Ikemura, T. (1988) *Nucleic Acids Res.*, **16**, r315-r402.
2. Sharp, P.M. and Cowe, E. (1991) *Yeast*, **7**, 657-678.
3. Bennetzen, J.L. and Hall, B.D. (1982) *J. Biol. Chem.*, **257**, 3026-3031.
4. Ikemura, T. (1982) *J. Mol. Biol.*, **158**, 573-597.
5. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) *Nucleic Acids Res.*, **14**, 5125-5143.
6. Bulmer, M. (1988) *J. Evol. Biol.*, **1**, 15-26.
7. Bulmer, M. (1990) *Nucleic Acids Res.*, **18**, 2869-2873.
8. Bulmer, M. (1987) *Nature*, **325**, 728-730.
9. Sharp, P.M. and Li, W.-H. (1986) *J. Mol. Evol.*, **24**, 28-38.
10. Lloyd, A.T. and Sharp, P.M. (1991) *Mol. Gen. Genet.*, **230**, 288-294.
11. Au-Young, J. and Robbins, P.W. (1990) *Mol. Microbiol.*, **4**, 197-207.
12. Saporito, S.M. and Sypherd, P.S. (1991) *Gene*, **106**, 43-49.
13. Smith, H.A., Allaudeen, H.S., Whitman, M.H., Koltin, Y. and Gorman, J.A. (1988) *Gene*, **63**, 53-63.
14. Fling, M.E., Kopf, J., Tamarkin, A., Gorman, J.A., Smith, H.A. and Koltin Y. (1991) *Mol. Gen. Genet.*, **227**, 318-329.
15. Hendriks, L., Goris, A., Van de Peer, Y., Neefs, J.-M., Vancanneyt, M., Kersters, K., Hennebert, G.L. and De Wachter, R. (1991) *J. Gen. Microbiol.*, **137**, 1223-1230.
16. Chen, M.-W., Anné, J., Volckaert, G., Huysmans, E., Vandenberghe, A. and De Wachter, R. (1984) *Nucleic Acids Res.*, **12**, 4881-4892.
17. Stenderup, A. and Leth Bak, A. (1968) *J. Gen. Microbiol.*, **52**, 231-236.
18. Brown, A.J.P., Bertram, G., Feldmann, P.J.F., Pegg, M.W. and Swoboda, R.K. (1991) *Nucleic Acids Res.*, **19**, 4298.
19. Sharp, P.M. and Devine, K.M. (1989) *Nucleic Acids Res.*, **17**, 5029-5039.
20. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and di Paola, G. (1985) *CABIOS*, **1**, 167-172.
21. Greenacre, M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, London.
22. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) *Nucleic Acids Res.*, **9**, r43-r74.
23. Wright, F. (1990) *Gene*, **87**, 23-29.
24. Sharp, P.M. and Li, W.-H. (1987) *Nucleic Acids Res.*, **15**, 1281-1295.
25. Lloyd, A.T. and Sharp, P.M. (1992) *J. Hered.*, **83**, 239-240.
26. Sueoka, N. (1992) *J. Mol. Evol.*, **34**, 95-114.
27. Drouin, G. (1991) *J. Mol. Evol.*, **33**, 237-240.
28. Antequera, F., Tamame, M., Villanueva, J.R. and Santos, T. (1984) *J. Biol. Chem.*, **259**, 8033-8036.
29. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
30. Ikemura, T. (1985) *Mol. Biol. Evol.*, **2**, 13-34.
31. Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. and Wright, F. (1988) *Nucleic Acid Res.*, **16**, 8207-8211. 32. Sharp, P.M. (1990) *Mol. Microbiol.*, **4**, 119-122.
32. Murray, W.W. and Rachubinski, R.A. (1987) *Gene*, **61**, 401-413.
33. Atomi, H., Ueda, M., Hishida, T., Teranishi, Y. and Tanaka, A. (1990) *J. Biochem.*, **107**, 262-266.
34. Shields, D.C. (1990) *J. Mol. Evol.*, **31**, 71-80.
35. Wada, K.-n., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T. (1992) *Nucleic Acids Res.*, **20**, 2111-2118.