Yeast
Yeast 2000; **16**: 1131–1145.

Research Article

Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*

Avril Coghlan and Kenneth H. Wolfe*
Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland

**Correspondence to:*
K. H. Wolfe, Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland.
E-mail: khwolfe@tcd.ie

Abstract

In 1982, Ikemura reported a strikingly unequal usage of different synonymous codons, in five *Saccharomyces cerevisiae* nuclear genes having high protein levels. To study this trend in detail, we examined data from three independent studies that used oligonucleotide arrays or SAGE to estimate mRNA concentrations for nearly all genes in the genome. Correlation coefficients were calculated for the relationship of mRNA concentration to four commonly used measures of synonymous codon usage bias: the codon adaptation index (CAI), the codon bias index (CBI), the frequency of optimal codons (F_{op}), and the effective number of codons ($\hat{N}_c$). mRNA concentration was best approximated as an exponential function of each of these four measures. Of the four, the CAI was the most strongly correlated with mRNA concentration ($r_s = 0.62 \pm 0.01$, $n = 2525$, $p < 10^{-17}$). When we controlled for CAI, mRNA concentration and protein length were negatively correlated (partial $r_s = -0.23 \pm 0.01$, $n = 4765$, $p < 10^{-17}$). This may result from selection to reduce the size of abundant proteins to minimize transcriptional and translational costs. When we controlled for mRNA concentration, protein length and CAI were positively correlated (partial $r_s = 0.16 \pm 0.01$, $n = 4765$, $p < 10^{-17}$). This may reflect more effective selection in longer genes against missense errors during translation. The correlation coefficients between the mRNA levels of individual genes, as measured by different investigators and methods, were low, in the range $r_s = 0.39$–0.68. Copyright © 2000 John Wiley & Sons, Ltd.

Keywords: ***Saccharomyces cerevisiae*; synonymous codon bias; codon adaptation index; codon bias index; frequency of optimal codons; effective number of codons; oligonucleotide arrays; serial analysis of gene expression**

Received: 11 February 2000
Accepted: 17 April 2000

Introduction

The genome hypothesis (Grantham *et al*., 1980) proposed that the genes in one species will all display the same usage bias with respect to synonymous codons. In *Saccharomyces cerevisiae*, an early study of nine genes (Bennetzen and Hall, 1982) revealed a strong bias toward usage of just 25 'preferred' codons in genes with high cytoplasmic mRNA levels. In contrast, genes with low mRNA levels showed very little synonymous codon usage bias. The preferred synonymous codons were discovered to be codons recognised by the most abundant cognate tRNA species (Bennetzen and Hall, 1982; Ikemura, 1982). Usage bias was proposed to result from translational selection, since using a codon that is translated via an abundant tRNA species was hypothesized to boost translational efficiency (Bennetzen and Hall, 1982; Ikemura, 1982). Subsequent analysis of about 40 genes (Ikemura, 1985) lent weight to the prediction that in *S. cerevisiae* there will be a strong correlation between the frequency of such translationally preferred codons in a gene and that gene's protein concentration. Theoretical models of protein production have corroborated this prediction (Solomovici *et al*., 1997; Xia, 1998). Indeed, an early cluster analysis of 110 *S. cerevisiae* genes based on their synonymous codon usage (Sharp *et al*., 1986) distinguished two groups: genes with high and moderate/low protein levels. In the decade since then, codon usage bias was often used as a predictor of expression levels (i.e., either protein levels or mRNA levels) in *S. cerevisiae*, although this is now largely obsolete due to the availability of data from direct studies of genome-wide transcript levels. Here, we retrospectively examine the relationship between codon bias and mRNA levels in yeast.

The release of the genome sequence (Goffeau *et al.*, 1996) and the development of new technologies have allowed synonymous codon bias and mRNA concentrations to be quantified for all *S. cerevisiae* genes, and protein concentrations to be measured for many. This has prompted three recent studies of the relationships between these variables in *S. cerevisiae*. Futcher *et al.* (1999) analysed their own protein concentration data from two-dimensional gel electrophoresis, and a combination of mRNA concentration data from Velculescu *et al.*'s (1997) serial analysis of gene expression (SAGE) experiment and from the high-density oligonucleotide array (HDA) experiment of Wodicka *et al.* (1997). Gygi *et al.* (1999) compared their own protein concentration data from two-dimensional gel electrophoresis to the same SAGE mRNA concentration data set (Velculescu *et al.*, 1997). Furthermore, in a third recent study, Pavesi (1999) also examined data from the same SAGE experiment (Velculescu *et al.*, 1997). A strong correlation between mRNA concentration and codon bias [the codon adaptation index (CAI) of Sharp and Li, 1987] was reported by Futcher *et al.* (1999) for 71 genes; Pavesi (1999) confirmed this for both CAI and CBI [the codon bias index (CBI) of Bennetzen and Hall, 1982] for 72 genes (no correlation coefficients given). Protein concentration was also reported to be strongly correlated with codon bias (CAI) ($r_s = 0.80$, $n = 71$, $p < 0.0001$) by Futcher *et al.* (1999). Although Gygi *et al.* (1999) previously asserted that codon bias (CBI) is not a predictor of either protein or mRNA levels, Futcher *et al.* (1999) pointed out that no statistics were used. A strong correlation between mRNA and protein concentrations ($r_s = 0.74$, $n = 71$, $p < 0.0001$) was reported by Futcher *et al.* (1999). In contrast, Gygi *et al.* (1999) concluded that mRNA and protein concentrations are not correlated ($r_p = 0.356$, $n = 73$), but their result's validity was questioned by Futcher *et al.* (1999) because inappropriate statistics were used. The correlation between mRNA and protein concentrations is high but not perfect, because protein concentration is determined not only by mRNA concentration but also, among other factors, by translational initiation and elongation rates and protein half-life (Futcher *et al.*, 1999; Gygi *et al.*, 1999; VanBogelen *et al.*, 1999). Thus, codon bias will show different correlations with mRNA and protein levels, since protein levels are affected by translational rates and protein half-life, and translation rates are themselves affected by codon bias.

We aimed to investigate the relationship in *S. cerevisiae* between the frequency of preferred codons in a gene and mRNA levels, using published data sets that include most genes in the genome. We used three independent mRNA concentration data sets (Cho *et al.*, 1998; Holstege *et al.*, 1998; Velculescu *et al.*, 1997) and four methods of measuring synonymous codon bias. One goal was to determine which codon bias measures were the best predictors of mRNA concentration, because this may be applicable to other organisms for which whole-genome transcription data are not available. A second aim was to examine the concurrence of mRNA concentration data from the three different whole-genome studies: the SAGE data of Velculescu *et al.* (1997), and the high-density oligonucleotide array (HDA) data of Cho *et al.* (1998) and Holstege *et al.* (1998). Our third aim was to investigate whether dependencies exist between codon bias, protein length and mRNA levels in *S. cerevisiae*. It has been hypothesized that protein length correlates with mRNA and protein levels in *S. cerevisiae* and *Drosophila melanogaster* (Moriyama and Powell, 1998), and that codon bias correlates with protein length in *S. cerevisiae*, *Escherichia coli*, *Caenorhabditis elegans*, *D. melanogaster* and *Arabidopsis thaliana* (Duret and Mouchiroud, 1999; Eyre-Walker, 1996; Moriyama and Powell, 1998).

Databases and methods

Synonymous codon usage bias measures

Methods to quantify synonymous codon usage bias either measure the deviation from equal use of synonymous codons ('H_0^*-based measures'), or measure the frequency of putative translationally optimal codons ('H_1-based measures') (Wright, 1990). Since H_1-based measures are more likely to detect codon bias caused by translational selection, we studied just one H_0^*-based measure, the effective number of codons ($\hat{N}_c$), and three H_1-based measures, CAI, CBI, and F_{op}, as described briefly below.

The effective number of codons ($\hat{N}_c$; Wright, 1990) measures deviation from equal use of synonymous codons. The $\hat{N}_c$ is the number of codons that, if used equally, would generate the level of codon bias observed. $\hat{N}_c$ takes values from 20.0 (maximum bias) to 61.0 (no bias). The CAI (Sharp and Li, 1987) is a measure of synonymous codon usage bias in the direction of the bias seen in a reference set of 24 *S. cerevisiae* genes having high protein levels. CAI takes values from 0.0 (no bias) to 1.0 (maximum bias). The CBI (Bennetzen and Hall, 1982) is a measure of the frequency of optimal codons. Its 22 optimal codons are those present in more than 85% of cases in the *S. cerevisiae ADH1, TDH2* and *TDH3* genes, and are complementary to the anticodons of the major *S. cerevisiae* tRNA species. CBI generally takes values from 0.0 (no bias) to 1.0 (maximum bias), although negative CBI values can occur if optimal codons occur less often in a gene than in a sequence having an equal use of the synonymous codons for each amino acid. The frequency of optimal codons (F_{op}; Ikemura, 1985) measures the frequency of optimal codons chosen based on tRNA anticodon sequences and isoacceptor tRNA concentrations. The set of 22 optimal codons used to calculate F_{op} is almost the same as that used for CBI, except that in F_{op} GCC (Ala) is excluded and GAC (Asp) is included (Sharp and Cowe, 1991). F_{op} takes values from 0.0 (no bias) to 1.0 (maximum bias).

Codon bias in each open reading frame (ORF) was calculated using a modified version of the FORTRAN 77 program CODONS (version 1.4); ftp://acer.gen.tcd.ie/pub/cod/ (Lloyd and Sharp, 1992a). We altered CODONS so that it could accept more ORFs, and added a subroutine so it could calculate CBI, as well as CAI, F_{op}, and $\hat{N}_c$.

Correspondence analysis of synonymous codon usage

Correspondence analysis is a multivariate statistical method often used to analyse synonymous codon usage. It identifies the main trends in data as a series of orthogonal axes in an *n*-dimensional hyperspace. The first axis explains the highest proportion of the variation in the data, and successive axes a decreasing proportion. The first axis in correspondence analysis of synonymous codon usage in *S. cerevisiae* is hypothesized to reflect a relationship between codon bias and protein or mRNA concentration (Lloyd and Sharp, 1992b; Sharp and Cowe, 1991). To compare the correlation between the first axis and mRNA levels to that between mRNA levels and CAI/CBI/F_{op}/$\hat{N}_c$, we carried out correspondence analysis of the synonymous codon usage of *S. cerevisiae*, estimated by relative synonymous codon usages (RSCUs). The RSCU of a codon is the observed frequency of the codon divided by the frequency expected if all the synonyms for that amino acid were used equally (Sharp *et al.*, 1986). Correspondence analysis on RSCUs was implemented using the program CodonW (version 1.4.2); ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z (Peden J. F., unpublished).

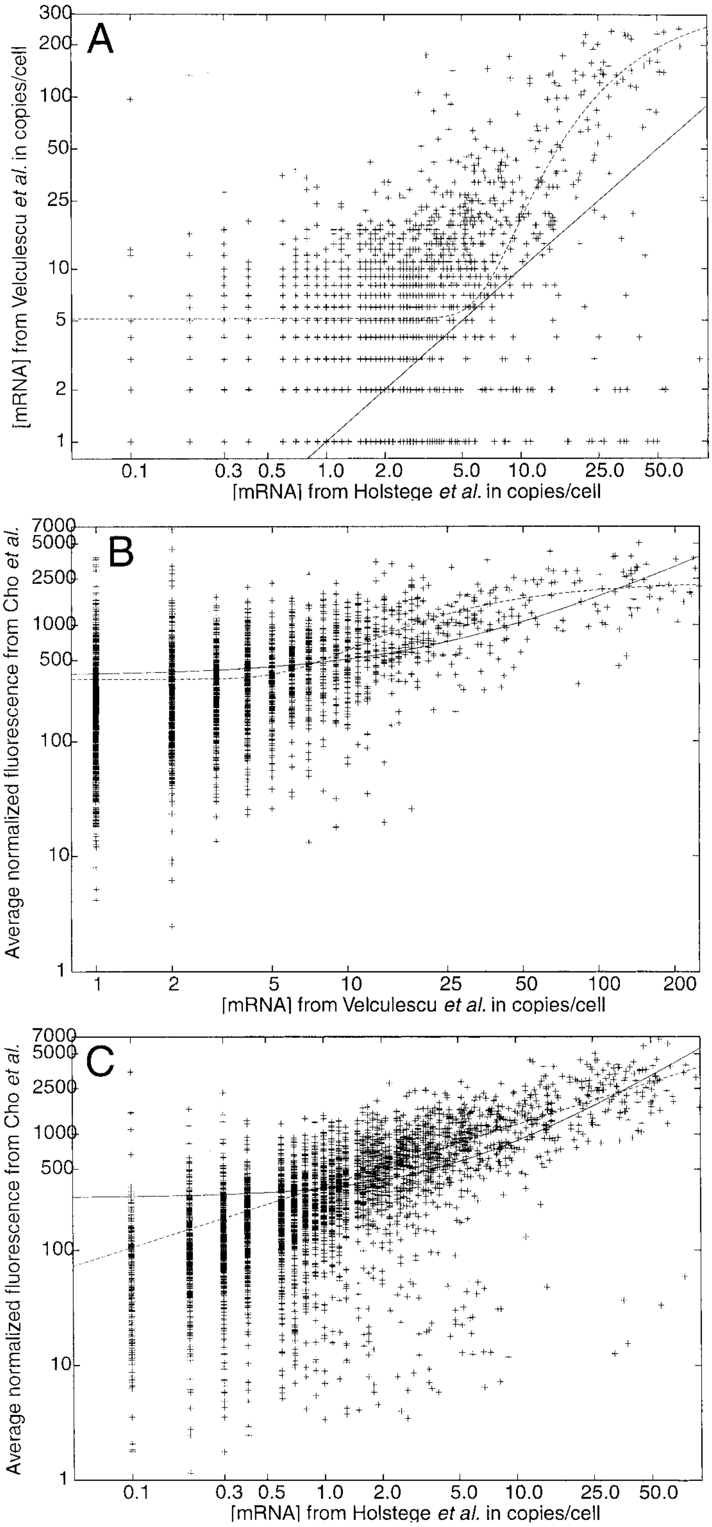
Sequence data

We obtained the *S. cerevisiae* nuclear genome sequence from ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/ (October 1998). This database included the coding sequence for 6217 current standard ORFs, with the introns removed. We excluded mitochondrial and plasmid ORFs, since their codon usage differs from that of nuclear genes.

mRNA concentration data

We studied genome-wide mRNA concentration data from three research groups: one group used the SAGE technique (Velculescu *et al.*, 1997), while the other two groups used HDAs (Cho *et al.*, 1998; Holstege *et al.*, 1998). Holstege *et al.* (1998) and Velculescu *et al.* (1997) estimated mRNA concentration in mRNA transcripts per cell. The third data set (Cho *et al.*, 1998) is in units of normalized fluorescence intensity; mRNA concentration is directly proportional to fluorescence intensity (Wodicka *et al.*, 1997).

The SAGE data (Velculescu *et al.*, 1997) cover 4665 ORFs, with mRNA concentration estimates from 0.3 to over 200.0 transcripts per cell. We obtained the data from ftp://genome-ftp.stanford.edu/pub/yeast/tables/ (December 1998). Velculescu *et al.* (1997) calculated mRNA concentration for each ORF by assuming 15 000 transcripts per cell in total. For each SAGE tag, we averaged the three tag concentration estimates from logarithmic phase, from S phase-arrested cells, and


A
300
200
100
50
25
10
5
2
1
[mRNA] from Velculescu et al. in copies/cell
0.1
0.3
0.5
1.0
2.0
5.0
10.0
25.0
50.0
[mRNA] from Holstege et al. in copies/cell
B
7000
5000
2500
1000
500
100
10
1
Average normalized fluorescence from Cho et al.
1
2
5
10
25
50
100
200
[mRNA] from Velculescu et al. in copies/cell
C
7000
5000
2500
1000
500
100
10
1
Average normalized fluorescence from Cho et al.
0.1
0.3
0.5
1.0
2.0
5.0
10.0
25.0
50.0
[mRNA] from Holstege et al. in copies/cell

from G_2/M phase-arrested cells. We then calculated the mRNA concentration for each ORF to be the sum of the concentrations of its corresponding tags, as did Velculescu *et al.* (1997). One unsettled issue is whether data from a non-unique SAGE tag can reliably be divided among its ORFs of origin (e.g. sharing data between the related genes *TDH2* and *TDH3* from a tag contained in both of them). We considered that including just data from unique tags for an ORF would introduce less error than (somehow) including data from both non-unique and unique tags. Thus, we discarded data from non-unique tags, and included data corresponding to unique type 1 SAGE tags (within the ORF) and unique type 2 SAGE tags (within the 3′-untranslated region). Our sample of the SAGE data included 5358 tags (from 3817 genes). In contrast, other analyses of the same SAGE data have included all the tags from ORFs with non-unique tags (Futcher *et al.*, 1999; Gygi *et al.*, 1999; Pavesi, 1999; Velculescu *et al.*, 1997). Our genome-wide correlation coefficients (Table 2) were little affected by excluding the inconclusive data from non-unique tags.

The data of Cho *et al.* (1998) were obtained from http://genomics.stanford.edu/yeast/cellcycle.html (October 1998). They estimated mRNA concentrations by using commercial oligonucleotide arrays: *S. cerevisiae* Ye6100 GeneChips from Affymetrix. They took measurements for each of 6218 ORFs in synchronized cells every 10 min during the mitotic cell cycle. For each ORF, the data consist of 17 values, which are in units of normalized fluorescence intensity and correspond to the time-points between 0 and 160 min after cell cycle reinitiation. We used both the average fluorescence intensity during the cell cycle and the peak fluorescence intensity in our analysis of Cho *et al.*'s data. They normalized their raw fluorescence intensity data to account for fluctuations in hybridization conditions during the cell cycle, assuming that total mRNA concentration in the cell remains constant throughout. These normalization calculations (described at the web-site above) perhaps reduced their data's accuracy for some ORFs. To avoid temperature-induced effects caused by arresting the cell cycle, we analysed data only from time-points more than 40 min past release from arrest (just the last 13 data points). We examined data only for probes that hybridize to a unique target ORF. Where two different probes were targeted to the two exons of a split gene, we took data for the second exon only, since the first exon is usually short in *S. cerevisiae*.

We obtained the data of Holstege *et al.* (1998) from http://www.wi.mit.edu/young/expression.html (January 1999). Using the same Affymetrix oligonucleotide arrays as Cho *et al.* (1998), they estimated mRNA concentration for 5460 ORFs.

Statistical analysis

We used DataDesk 5.0.1 and Excel 4.0 for statistical analysis.

Results

Concurrence between mRNA concentrations estimated using SAGE and two oligonucleotide array experiments

The three whole-genome mRNA concentration studies analysed here all employed similar yeast strains grown under similar conditions, so we compared the mRNA levels reported for genes in the different experiments, using log–log plots (Figure 1). Taking data from the two groups that used the same commercial oligonucleotide array, the correlation between the average fluorescence intensity during the cell cycle, as measured by Cho *et al.* (1998), and the mRNA concentration data of Holstege *et al.* (1998) was $r_s = 0.68 \pm 0.01$ ($n = 4360$, $p < 10^{-17}$). Taking data from the two groups that

Figure 1. Concurrence between three different mRNA concentration data sets. (A) SAGE data from Velculescu *et al.* (1997) plotted on a log–log scale against oligonucleotide array data from Holstege *et al.* (1998) for 3432 ORFs ($r_s = 0.54 \pm 0.01$, $p < 10^{-17}$). (B) Average oligonucleotide array data from Cho *et al.* (1998) plotted on a log–log scale against SAGE data from Velculescu *et al.* for 3094 ORFs ($r_s = 0.50 \pm 0.01$, $p < 10^{-17}$). (C) Average oligonucleotide array data from Cho *et al.* plotted on a log–log scale against oligonucleotide array data from Holstege *et al.* for 4360 ORFs ($r_s = 0.68 \pm 0.01$, $p < 10^{-17}$). The regression curves for the non-log transformed data (dashed lines) were: (A) $y \approx 5.14 + e^{(5.88 + (-32.26/x))}$; (B) $y \approx 344.95 + e^{(7.61 + (-20.42/x))}$; and (C) $y \approx 355.88(x)^{0.53}$. In (A) the solid line is $y = x$; in (B) the solid line is the linear best-fit to the non-log transformed data $y \approx 370.40 + 13.98x$; and in (C) the solid line is the linear best-fit to the non-log transformed data $y \approx 283.50 + 58.92x$

estimated mRNA concentrations during particular cell cycle stages, the correlations between the data of Cho *et al.* (1998; HDA method) and Velculescu *et al.* (1997; SAGE method) are $r_s = 0.41 \pm 0.02$ ($n = 3094$, $p < 10^{-17}$) for the G_2/M boundary, and $r_s = 0.39 \pm 0.02$ ($n = 3094$, $p < 10^{-17}$) for the S phase.

How many *S. cerevisiae* ORFs show the influence of translational selection on their synonymous codon usage?

Every yeast gene has some non-randomness in its codon usage, which can be quantified using any of the methods for measuring codon bias discussed here. For genes with low bias, it was necessary to distinguish between genuine 'translational' bias caused by natural selection, and genomic mutational biases or bias randomly caused by sampling. To make this distinction, we examined the codon bias seen in simulated sequences using an approach inspired by Sharp and Li (1987). They estimated the average CAI of an *E. coli* sequence that has genomic mutational bias but lacks translational codon bias (Sharp and Li, 1987). We wrote a C++ program to act as a random number generator (Press *et al.*, 1994) to produce sets of simulated sequences having the same lengths and amino acid compositions as the real *S. cerevisiae* ORFs. The simulated sequences were given a 38.0% silent-site G+C content (GC3s), to match the real genome.

We then calculated the average ($\overline{x}$) and standard deviation (s) of the CAI, CBI, F_{op} and $\hat{N}_c$ for the simulated sequences. We excluded from subsequent analysis any real *S. cerevisiae* ORFs with codon usage less biased than $\overline{x}+2s$ (or $\overline{x}-2s$ for $\hat{N}_c$). We assumed that translational selection did not influence codon usage in the ORFs discarded, since evidence that synonymous codon usage in very low-bias genes reflects selection for translationally non-optimal codons is scarce and controversial (Sharp and Li, 1986; Sharp *et al.*, 1993). When codon bias was measured using CAI, 48.1% of the 6217 putative ORFs had synonymous codon bias above this threshold, compared to 43.8% using CBI, 40.2% using F_{op}, and 41.6% using $\hat{N}_c$ (Table 1). Thus, 50–60% of *S. cerevisiae* ORFs appear to lack significant translational codon bias and were excluded from further study.

However, for three reasons we could have wrongly discarded or retained some ORFs. First, the simulated sequences had constant codon bias along each sequence with no context effects on codon usage, while real *S. cerevisiae* genes do have weak intragenic differences in codon bias (Bulmer, 1988), including context effects (Bulmer, 1990) and differences due to constraints on protein folding (Crombie *et al.*, 1992). Second, $\overline{x}$ was presumed to estimate the CAI/CBI/F_{op}/$\hat{N}_c$ seen in *S. cerevisiae* sequences having mutational bias (38.0% GC3s) but no translational codon usage bias. However, mutational biases vary around the *S. cerevisiae* genome (Bradnam *et al.*, 1999) and so $\overline{x}$ depends on genomic location. Third, genes in which translational selection has not influenced codon usage may have biased codon usage due to mRNA or DNA structural constraints.

CAI is the best predictor of mRNA concentration in *S. cerevisiae*

Of the three mRNA concentration data sets, the oligonucleotide array data of Holstege *et al.* (1998) were the most strongly correlated with each of the four codon bias measures (Table 2). In general, CAI showed the strongest correlation with mRNA levels within each experiment, with $\hat{N}_c$ the weakest and

Table 1. Average and standard deviation of the codon bias calculated for simulated sequences

Codon bias estimator	Average codon bias ($\overline{x}$)	Standard deviation of codon bias (s)	Cut-off[a]	Real yeast ORFs with translational codon bias[b]
CAI	0.107	0.017	0.141	2988
CBI	−0.011	0.053	0.096	2720
F_{op}	0.399	0.031	0.461	2584
$\hat{N}_c$	57.2	3.0	51.1	2499

[a] For $\hat{N}_c$ the cut-off is $\overline{x}-2s$; ORFs with $\hat{N}_c$ above this are considered to lack 'translational' codon bias. For CAI, CBI and F_{op}, the cut-off is $\overline{x}+2s$.
[b] For CAI, CBI and F_{op}, the number of real *S. cerevisiae* ORFs analysed was 6217. For $\hat{N}_c$, the number was 6212, since five ORFs were too short for the program CODONS to calculate $\hat{N}_c$.

Table 2. Spearman rank correlation coefficients (r_s) between the three mRNA concentration data sets and four different codon bias measures

Codon bias estimator	Array (HDA) data (from Holstege *et al.*, 1998)	Array (HDA) data (from Cho *et al.* 1998; cell cycle average)	Array (HDA) data (from Cho *et al.* 1998; cell cycle maximum)	SAGE data (from Velculescu *et al.* 1997)
CAI	0.62 ± 0.01 [2525]	0.52 ± 0.01 [2458]	0.53 ± 0.01 [2458]	0.48 ± 0.02 [2067]
CBI	0.62 ± 0.01 [2303]	0.51 ± 0.02 [2259]	0.52 ± 0.02 [2259]	0.45 ± 0.02 [1893]
F_{op}	0.61 ± 0.01 [2105]	0.50 ± 0.02 [2032]	0.52 ± 0.02 [2032]	0.46 ± 0.02 [1714]
$\hat{N}_c$	-0.58 ± 0.01 [2167]	0.50 ± 0.02 [2183]	-0.52 ± 0.02 [2183]	-0.43 ± 0.02 [1750]
Axis 1[a]	0.57 ± 0.01 [4914]	0.48 ± 0.01 [5074]	0.48 ± 0.01 [5074]	0.43 ± 0.02 [3806]

The number of ORFs analysed is given in brackets. For CAI, CBI, F_{op} and $\hat{N}_c$, only ORFs with 'translational' codon bias above the cut-offs in Table 1 were included, while all ORFs were included for the 'correspondence analysis axis 1' row. All r_s values are highly significant ($p < 10^{-17}$) due to the large sample sizes. Standard error estimates were calculated, assuming that r_s forms a Gaussian distribution. r_s was considered a more appropriate statistic than the Pearson product–moment correlation coefficient (r_p), because once ORFs lacking translational codon bias were discarded, the data did not form a bivariate Gaussian distribution, as is necessary for using r_p. Furthermore, the relationship between mRNA concentration and codon bias was more curvilinear than linear, while r_p tests for linear correlation. ORFs of less than 100 codons were included, since short sequence length causes almost no systematic error in either $\hat{N}_c$ or CAI (Comeron and Aguadé, 1998). When ORFs of less than 100 codons were excluded, r_s was generally reduced (e.g. for 2452 ORFs of greater than 99 codons, for CAI vs. the Holstege *et al.* (1998) data, $r_s = 0.60 \pm 0.01$).

[a] Axis 1 from correspondence analysis of RSCU values.

CBI and F_{op} intermediate. CAI's performance was significantly better than $\hat{N}_c$ (but not CBI or F_{op}) for the SAGE data and for the oligonucleotide array data of Holstege *et al.* (1998). For the oligonucleotide array data of Cho *et al.* (1998), there was no significant difference between correlations with the four different codon bias measures. From correspondence analysis on RSCUs, position on axis 1 did not correlate with mRNA levels as well as CAI, CBI or F_{op} or $\hat{N}_c$ (Table 2). This may be because our correspondence analysis included very low bias genes as we used no low-bias cut-off, unlike for CAI/CBI/ F_{op} /$\hat{N}_c$. Alternatively, factors other than mRNA concentration may also affect axis 1 in *S. cerevisiae*; for example, we found protein length to be weakly correlated with axis 1 ($r_s = 0.14 \pm 0.02$, $n = 3401$, $p < 10^{-15}$).

Table 3. The sums of squared residuals from the best-fit curves in Figure 2

Codon bias estimator			
CAI	**CBI**	$\mathbf{F_{op}}$	$\hat{N}_c$[a]
6.7×10^4	7.5×10^4	7.1×10^4	8.3×10^4
[2525]	[2303]	[2105]	[2167]

The numbers of ORFs are in brackets.

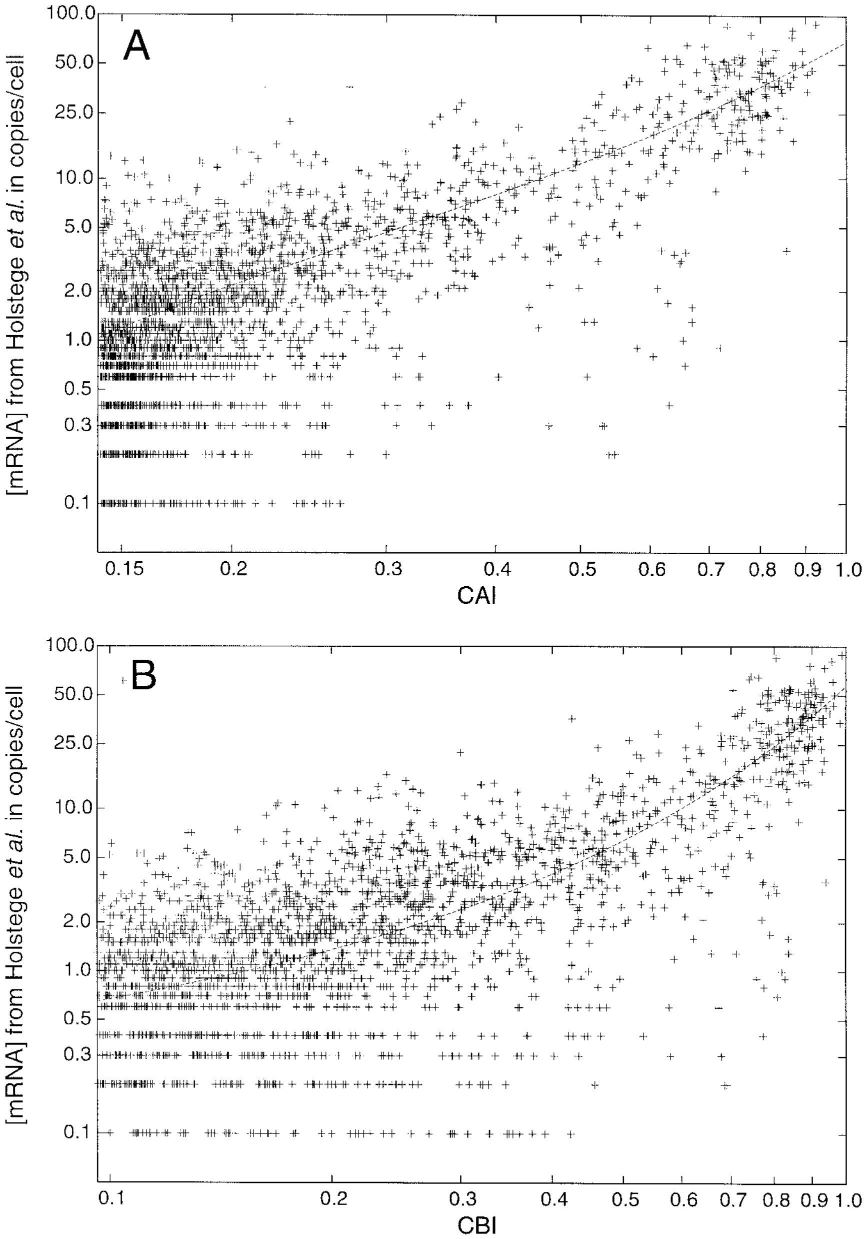
[a] $\hat{N}_c$ is expected to give a higher value for the sum of squared residuals than the three other codon bias estimators, since $\hat{N}_c$ took values 20.0–51.1, while the other three took values between approximately 0 and 1. As r_s indicated that $\hat{N}_c$ did not correlate as well with mRNA concentration as the other three (Table 2), this did not matter.

The experiment of Cho *et al.* (1998) studied mRNA concentration during the cell cycle and so contained information on both the average and the peak mRNA levels of each gene. It is not known which of these should have the stronger influence on codon bias. However, no significant differences were seen between the correlation coefficients calculated using the average and peak mRNA levels, with any of the four methods of calculating codon bias (Table 2).

For the four codon bias measures, the relationship between mRNA concentration and codon bias was best approximated by a curve of the form:

$$[mRNA] = c_1 + e^{c2 + c3 * x}$$

where $[mRNA]$ is the mRNA concentration, x is the codon bias, and c_{1-3} are constant coefficients. Log–log plots of mRNA levels (from Holstege *et al.*, 1998) vs. all four codon bias measures are shown in Figure 2. CAI produced the best regression curve (Table 3) and so is the best predictor of mRNA concentration, with F_{op} second best, and



CBI third best. There is almost no relationship between mRNA levels and codon bias (measured by any method) for mRNAs with fewer than about five transcripts per cell (Figure 2). This is to be expected because, first, there is little consistency among the different experiments for transcripts present at below five transcripts per cell (Figure 1), and second, selection for translationally optimal codons is hypothesized to be less effective in low-abundance mRNAs (Sharp and Li, 1986). For comparison with Figure 2A, log–log plots of CAI vs. the data of Velculescu *et al.* and Cho *et al.* are shown in Figures 3 and 4, respectively.

Protein length is negatively correlated with mRNA concentration in *S. cerevisiae* genes with the same level of codon bias

To take into account the dependencies of both protein length and mRNA concentration on CAI, we controlled for CAI (i.e. the effect of CAI was eliminated) by calculating the partial correlation

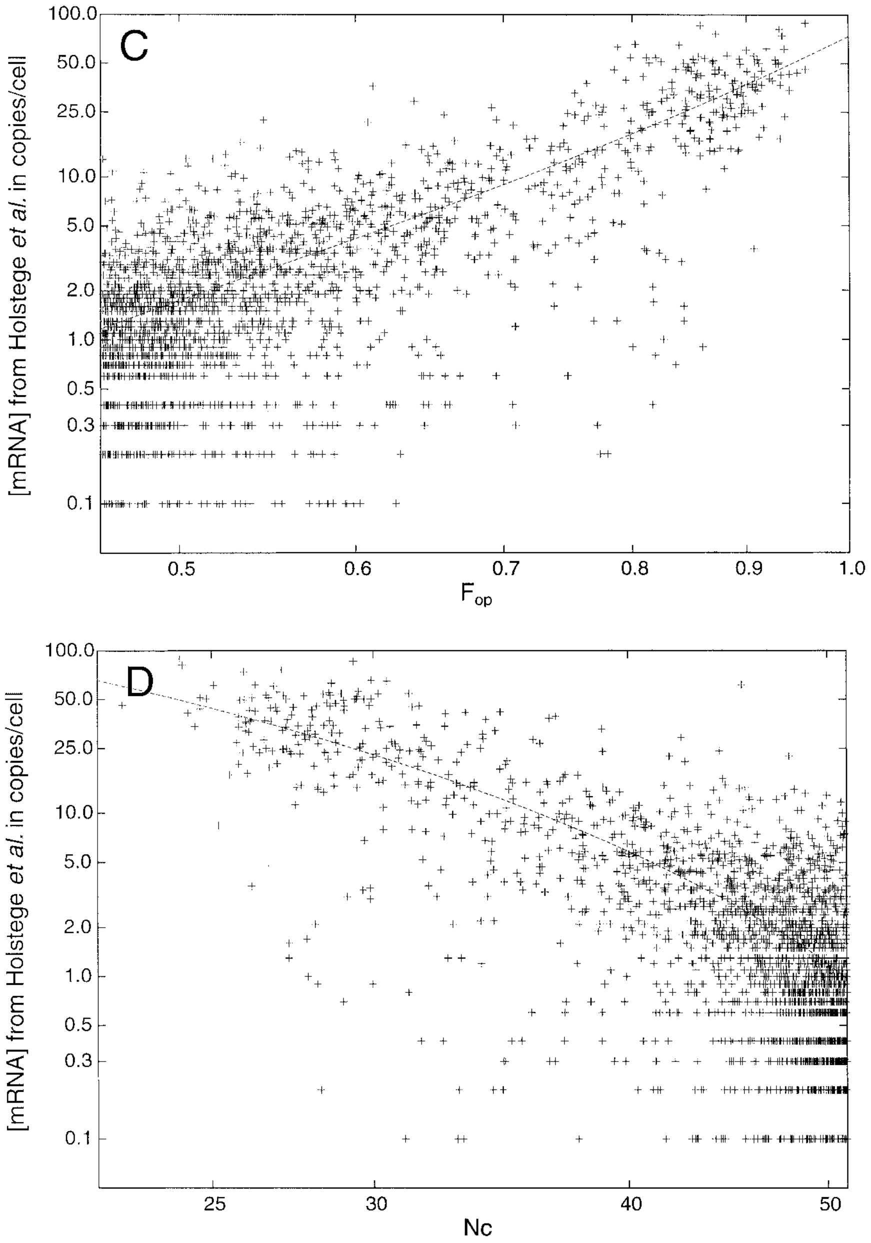


Figure 2. mRNA concentration estimates from Holstege *et al.* (1998), plotted on a log–log scale against (A) CAI for 2525 ORFs; (B) CBI for 2303 ORFs; (C) F$_{op}$ for 2105 ORFs; and (D) $\hat{N}_c$ for 2167 ORFs, for *S. cerevisiae* ORFs judged to have 'translational' codon bias. The regression curves for the non-log transformed data (dashed lines) are: (A) for CAI, $y \approx -5.24 + e^{(1.43+2.88x)}$; (B) for CBI, $y \approx -0.64 + e^{(-0.15+4.21x)}$; (C) for F$_{op}$, $y \approx -0.74 + e^{(-2.49+6.80x)}$; and (D) for $\hat{N}_c$, $y \approx -0.53 + e^{(7.05-0.13x)}$

coefficient of mRNA concentration and protein length, excluding CAI (Bailey, 1995). We excluded proteins of < 100 codons to avoid sampling effects in calculating the CAI. When we controlled for CAI in this way, mRNA concentration from Holstege *et al.* (1998) showed a weak negative partial correlation with protein length (partial $r_s = -0.23 \pm 0.01$, $n = 4765$, $p < 10^{-17}$).

In a similar way, CAI was found to be positively correlated with protein length in *S. cerevisiae* genes

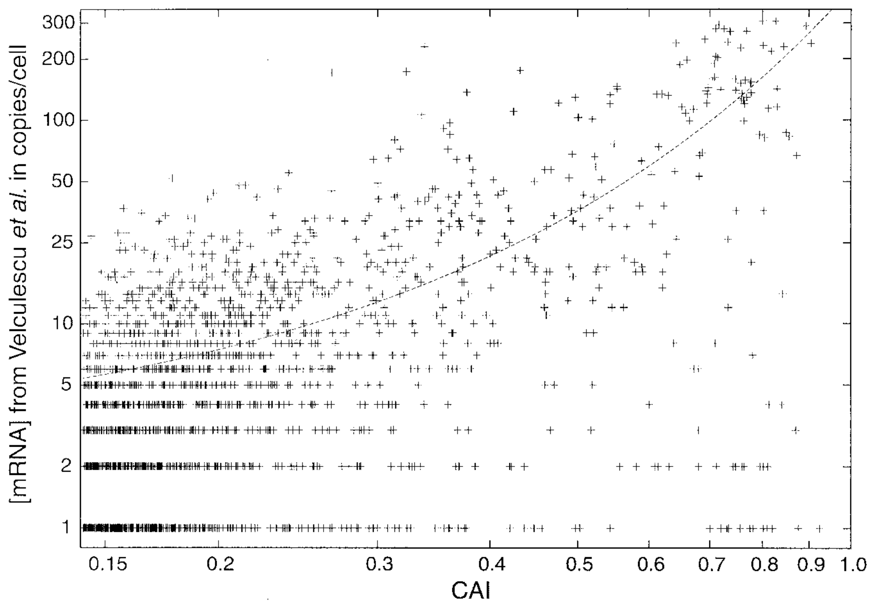


Figure 3. mRNA concentration estimates from Velculescu *et al.* (1997; SAGE method) plotted on a log–log scale against CAI for 2067 *S. cerevisiae* ORFs judged to have 'translational' codon bias. The ordinate is the average of three mRNA concentration estimates from logarithmic phase, from S phase-arrested cells, and from G_2/M phase-arrested cells. The regression curve for the non-log transformed data (dashed line) is $y \approx -0.75 + e^{(1.11+4.98x)}$

with the same mRNA concentration. For proteins of <100 codons, the partial correlation coefficient of CAI and protein length, excluding mRNA concentration, is $r_s = 0.16 \pm 0.01$, $n = 4765$, $p < 10^{-17}$.

Discussion

Concurrence between mRNA concentrations estimated using SAGE and oligonucleotide arrays

When the two data sets generated using oligonucleotide arrays (Cho *et al.*, 1998; Holstege *et al.*, 1998) were compared, the correlation was surprisingly low ($r_s = 0.68$), considering that both groups used the same commercial oligonucleotide array. This may indicate low precision in array mRNA concentration estimates. However, anomalous normalization of some data points by Cho *et al.* may have obscured a stronger correlation between their raw data and that of Holstege *et al.* When two cell cycle data sets from different techniques were compared, array data from Cho *et al.* were quite weakly correlated with the equivalent SAGE data from Velculescu *et al.* ($r_s = 0.41$ and 0.39, for G_2/M an S phase, respectively).

Why did the three mRNA concentration data sets not agree perfectly? First, SAGE and HDAs have different precision, accuracy, sensitivities, and resolving powers. Second, for some genes (we do not know how many) the three data sets probably disagreed due to minor strain, preparation and growth condition (e.g. cell density and growth medium) differences.

So how valid are our results on the relationship between codon bias and mRNA levels in *S. cerevisiae*? This depends on the quality (especially precision and accuracy) of the mRNA concentration data that we analysed. The quality of the data can be assessed by comparing the three data sets with each other (Figure 1), and from published studies of SAGE and HDA data quality. With respect to precision, in independent duplicate HDA experiments, Cho *et al.* (1998) and Holstege *et al.* (1998) found good precision in their mRNA concentration estimates. An earlier analysis of the precision of the HDA technique (Wodicka *et al.*, 1997), which also used *S. cerevisiae* Affymetrix Ye6100 arrays, concluded that 'a concentration for a given nucleic acid sequence can be assigned as *X*,

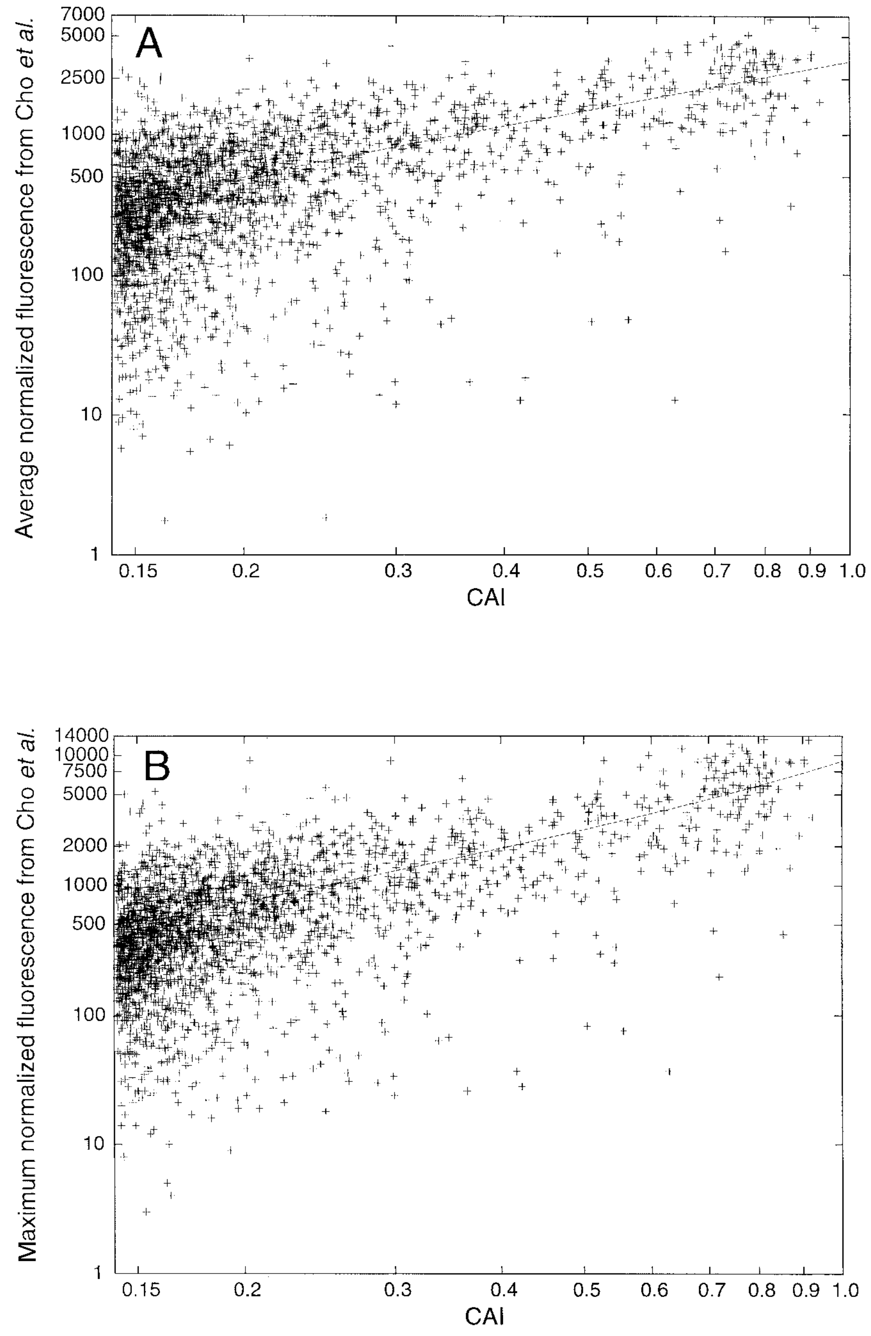


Figure 4. Average (A) and maximum (B) normalized fluorescence intensity measurements (a measure of mRNA concentration) from Cho *et al.* (1998) plotted on a log–log scale against CAI for 2458 *S. cerevisiae* ORFs judged to have 'translational' codon bias. The *y* axes in (A) and (B) are the average and maximum, respectively, of 13 fluorescence intensity measurements during the mitotic cell cycle. The regression curves for the non-log transformed data (dashed lines) are: (A) $y \approx -11{,}736.15 + e^{(9.36+0.26x)}$; and (B) $y \approx -2{,}398.69 + e^{(7.73+1.61x)}$

based on the observed fluorescence intensity, with a greater than 95% probability that the actual concentration is between $0.5X$ and $2X$'. Estimates of mRNA concentrations in duplicate array experiments varied more for some genes than for others; these genes were found to be very sensitive to growth conditions such as cell density and microenvironment (Lander, 1999; Wodicka *et al.*, 1997). There is so far no published in-depth study of the precision of SAGE, although it is likely that SAGE concentration estimates for low abundance mRNAs have low precision (Futcher *et al.*, 1999; Gygi *et al.*, 1999).

Accuracy of mRNA concentration estimates can best be judged by comparing data from different techniques, but no previous studies have compared SAGE results for a large number of genes to mRNA concentrations estimated by other methods (e.g. with oligonucleotide arrays, cDNA microarrays, or Northern blots). However, in several small studies the abundance of SAGE tags agreed well with relative mRNA concentration estimates from cDNA hybridization experiments (Velculescu *et al.*, 1995) or Northern blots (Velculescu *et al.*, 1997; Zhang *et al.*, 1997). HDA mRNA concentration estimates for 17 *S. cerevisiae* genes agreed well with Northern blots (Galitski *et al.*, 1999), but from unpublished data it has been speculated that arrays underestimate high mRNA concentrations and overestimate low mRNA concentrations (Anonymous, 1998; Futcher *et al.*, 1999). Thus, it is hypothesized that SAGE is more accurate than HDAs for abundant mRNAs (Futcher *et al.*, 1999). Indeed, the SAGE estimates of Velculescu *et al.* (1997) are higher than the HDA estimates of Holstege *et al.* (1998) for most abundant mRNAs (Figure 1A).

In *S. cerevisiae* mRNA concentration is correlated with codon bias

Why is this so? Two variables can be linearly correlated if one is the cause of the other, if they interact with each other, or if they are both effects of another third variable (Campbell, 1974). The reason why codon bias and mRNA concentration are correlated is an interesting case of cause and effect. *In a living cell*, codon bias may affect transcriptional rate and mRNA stability and so be a contributory cause of mRNA concentration (Xia, 1996). Messenger RNA concentration is without doubt a necessary and sufficient cause of protein concentration, and codon bias debatably affects translation rate and so probably is also a small contributory cause of protein concentration (Kurland, 1987; Sharp and Li, 1986; Sharp *et al.*, 1993; Xia, 1998). *Over evolution,* selection for protein concentration may have been a contributory cause of codon bias, via selection for translational efficiency, and of mRNA concentration (Sharp and Li, 1986; Sharp *et al.*, 1993). In addition selection for mRNA concentration may have been a contributory cause of codon bias, via selection for transcriptional efficiency and mRNA stability (Xia, 1996). That is, codon bias and mRNA concentration, codon bias and protein concentration, and mRNA concentration and protein concentration may all have interacted over time; these three interactions may all have contributed to the correlation between codon bias and mRNA concentration observed in this and other recent studies in *S. cerevisiae* (Futcher *et al.*, 1999; Pavesi, 1999) and in *C. elegans*, *D. melanogaster* and *A. thaliana* (Duret and Mouchiroud, 1999). How strongly are codon bias and mRNA concentration correlated when the effect of protein concentration is eliminated? This would be quantified by partial correlation, which unfortunately was not calculated in studies of the relationships between codon bias and mRNA and protein levels in *S. cerevisiae* (Futcher *et al.*, 1999; Gygi *et al.*, 1999).

CAI is the best predictor of mRNA concentration in *S. cerevisiae*

Of the four codon bias measures studied, the CAI was the best predictor of mRNA concentration (Table 3). It was not surprising that $\hat{N}_c$ was the most weakly correlated with mRNA concentration, since $\hat{N}_c$ is a H_0^*-based codon bias measure, while the CAI, CBI and F_{op} are H_1-based measures. Even though $\hat{N}_c$ is a H_0^*-based measure, it was still strongly correlated with mRNA concentration (Table 2). This is because GC3s in *S. cerevisiae* is reasonably close to 50% (it is in the range 35–45%) and so there is little influence from 'mutational' bias on codon usage (and so on $\hat{N}_c$) in *S. cerevisiae* (Wright, 1990). $\hat{N}_c$ may, however, be a poor predictor of mRNA concentration in species in which mutational bias is more pronounced (Wright, 1990).

Codon bias is often an imperfect measure of mRNA abundance due to differential gene regulation. For example, Holstege *et al.* (1998) detected relatively low mRNA levels for the high-bias gene *ENO1* (17.1 transcripts per cell; CAI = 0.87). *ENO2,* its homologue, had transcript levels more usual for their high codon bias (61.1 transcripts per cell; CAI=0.89). *ENO2* is dramatically induced by glucose (McAlister and Holland, 1982), while *ENO1* is glucose-repressed (Carmen *et al.*, 1995). Thus, under the glucose-rich growth conditions used by Holstege *et al.* (1998), *S. cerevisiae* transcribed abundant *ENO2* mRNA and little *ENO1* mRNA.

Why other estimators fall short of CAI is an interesting puzzle. There are three reasons. First, CAI assigns a relative translational 'adaptiveness' to each codon in the range 0.0 (non-optimal) to 1.0 (optimal), while CBI and F_{op} assign each codon a non-relative integer score of either 0 (non-optimal) or 1 (optimal) (Bennetzen and Hall, 1982; Ikemura, 1985; Sharp and Li, 1987). Thus, CBI and F_{op} are coarse-grained as compared to CAI. Second, simulations (Comeron and Aguadé, 1998) demonstrated that CAI has almost no systematic errors dependent on sequence length, and has low dispersion under different sequence length and codon bias conditions. CBI and F_{op} were not analysed but, given findings for other codon bias measures (Comeron and Aguadé, 1998), it would be worth investigating whether CBI and F_{op} have greater systematic errors and dispersion than CAI. Third, CAI quantifies the optimality of codons from their frequencies in a set of reference genes having high protein levels, so may be a good predictor of mRNA concentration by default.

CAI is an unreliable predictor of mRNA concentration, as the residuals for our regression curve were proportionately large (Figure 2); however, this was perhaps partly due to errors in the mRNA concentration estimates. CAI gives little insight into why its codons are preferred since, unlike CBI and F_{op}, its optimal codons were not chosen using physicochemical data. Another drawback of CAI is its inability to predict changes in mRNA or protein concentrations with changing physiological conditions. Indeed, equations constructed from physicochemical data may be a much more accurate way than CAI to predict protein concentrations; for example, from data such as codon usage, tRNA concentrations, ribosome binding site strengths, codon-anticodon binding energies, transcript lengths and protein half-lives. One possibility is that CBI or F_{op} could be updated using better *S. cerevisiae* empirical data (Percudani *et al.*, 1997) and theoretical models of protein production (Solomovici *et al.*, 1997; Xia, 1996, 1998). Such an improved translational codon bias measure might provide insights into the relative importance of contributory causes of protein concentration. In contrast, to detect transcriptional selection on codon usage (Xia, 1996), a completely new codon bias measure is needed. But will transcriptional codon bias turn out to be more or less correlated with mRNA concentration than is translational codon bias?

Protein length is negatively correlated with mRNA concentration in *S. cerevisiae* genes with the same level of codon bias

When we controlled for CAI, mRNA concentration showed a weak negative partial correlation with protein length (partial $r_s = -0.23 \pm 0.01$, $n = 4765$, $p < 10^{-17}$). This concurs with earlier evidence that genes in *D. simulans* are reduced in size compared to genes in its sister species *D. melanogaster* (Akashi, 1996). This was suggested to be due to more effective translational selection in *D. simulans* acting to reduce the size of abundant proteins, to minimize transcriptional and translational energy costs (Akashi, 1996). Moriyama and Powell (1998) hypothesized that this trend exists in *S. cerevisiae, D. melanogaster* and *E. coli*. Using very rough estimates of mRNA concentration level (categories low, moderate or high) from expressed sequence tags (ESTs), another study found no evidence for any such correlation in *A. thaliana*, and limited evidence for a positive correlation in *C. elegans* and *D. melanogaster* (Duret and Mouchiroud, 1999). It will be interesting to see if more precise concentration data confirm this.

CAI is positively correlated with protein length in *S. cerevisiae* genes with the same mRNA concentration

When we controlled for mRNA concentration, protein length showed a weak positive partial correlation with codon bias (CAI) (partial $r_s = 0.16 \pm 0.01$, $n = 4765$, $p < 10^{-17}$) in *S. cerevisiae*. It was necessary to control for mRNA levels to discover this, because mRNA concentration has a positive effect on CAI through one path (CAI ↑s as mRNA levels ↑) and a negative effect through another path (length ↓s as mRNA levels ↑, and CAI ↓s as length ↓s, so CAI ↓s as mRNA levels ↑). That is, mRNA concentration masks some of the CAI-length partial correlation that is visible when the effect of mRNA concentration is eliminated (Davis, 1985). It has been found that CAI and protein length are positively correlated in *S. cerevisiae* (Moriyama and Powell, 1998) and *E. coli* (Eyre-Walker, 1996; Moriyama and Powell, 1998) ribosomal protein genes; all ribosomal protein genes have approximately equal protein concentrations. It has been hypothesized that, since long proteins are energetically more expensive to produce, translational selection for codons which minimize missense errors during translation is more effective in long genes (Eyre-Walker, 1996).

In contrast to our results, negative correlations have been reported between CAI and protein length in *S. cerevisiae* and *D. melanogaster* (Moriyama and Powell, 1998; Powell and Moriyama, 1997), and between F_{av} ($\equiv F_{op}$) and protein length in *D. melanogaster*, *C. elegans* and *A. thaliana* (Duret and Mouchiroud, 1999). Why did Moriyama and Powell (1998) reach different conclusions than us for *S. cerevisiae*? First, they did not consider that dependencies of codon bias and protein length on mRNA concentration will distort the codon bias–protein length bivariate correlation. Second, they used Pearson product–moment correlation (r_p), which requires that both variables be Gaussian, otherwise p-values for r_p are meaningless (Bailey, 1995). Since they excluded proteins of <100 codons, their sample did not have a Gaussian distribution of lengths, and anyway in *S. cerevisiae* protein length is not Gaussian (Das *et al.*, 1997). Further, Pearson correlation measures linear correlation, not curvilinear correlation, but the relationship is clearly not linear. The results of Duret and Mouchiroud (1999) are also unconvincing. First, like Moriyama and Powell (1998), they calculated bivariate correlation, not partial correlation, and so failed to disentangle the mutual dependence of codon bias, protein length and mRNA levels. Second, their mRNA concentration data is very imprecise. Reanalysis of *Drosophila*, *Caenorhabditis* and *Arabidopsis* using better concentration data could be interesting.

Acknowledgements

We are grateful to A.T. Lloyd, D. Lynn and S. Stallone for helpful discussion. We would also like to thank two anonymous reviewers for their comments.

References

Akashi H. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144**: 1297–1307.

Anonymous. 1998. Getting hip to the chip. *Nature Genet* **18**: 195–197.

Bailey NTJ. 1995. *Statistical Methods in Biology*, 3rd edn. Cambridge University Press: Cambridge.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. *J Biol Chem* **257**: 3026–3031.

Bradnam KR, Seoighe C, Sharp PM, Wolfe KH. 1999. G+C content variation along and among *Saccharomyces cerevisiae* chromosomes. *Mol Biol Evol* **16**: 666–675.

Bulmer M. 1988. Codon usage and intragenic position. *J Theor Biol* **133**: 67–71.

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

Campbell SK. 1974. *Flaws and Fallacies in Statistical Thinking*, chapter 13. Prentice-Hall: Englewood Cliffs, NJ.

Carmen AA, Brindle PK, Park CS, Holland MJ. 1995. Transcriptional regulation by an upstream repression sequence from the yeast enolase gene *ENO1*. *Yeast* **11**: 1031–1043.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**: 65–73.

Comeron JM, Aguadé M. 1998. An evaluation of measures of synonymous codon usage bias. *J Mol Evol* **47**: 268–274.

Crombie T, Swaffield JC, Brown AJ. 1992. Protein folding within the cell is influenced by controlled rates of polypeptide elongation. *J Mol Biol* **228**: 7–12.

Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, Bienkowska J, Adams RM, Smith TF, Lindelien J. 1997. Biology's new Rosetta stone. *Nature* **385**: 29–30.

Davis JA. 1985. *The Logic of Causal Order*. Sage: Beverly Hills, CA.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila,* and *Arabidopsis*. *Proc Natl Acad Sci USA* **96**: 4482–4487.

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol* **13**: 864–872.

Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. 1999. A sampling of the yeast proteome. *Mol Cell Biol* **19**: 7357–7368.

Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. 1999. Ploidy regulation of gene expression. *Science* **285**: 251–254.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. *Science* **274**: 546, 563–547.

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* **8**: r49–r62.

Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**: 1720–1730.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**: 717–728.

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**: 573–597.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13–34.

Kurland CG. 1987. Strategies for efficiency and accuracy in gene expression 1. The major codon preference: a growth optimization strategy. *Trends Biochem Sci* **12**: 126–128.

Lander ES. 1999. Array of hope. *Nature Genet* **21**: 3–4.

Lloyd AT, Sharp PM. 1992a. CODONS: a microcomputer program for codon usage analysis. *J Hered* **83**: 239–240.

Lloyd AT, Sharp PM. 1992b. Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. *Nucleic Acids Res* **20**: 5289–5295.

McAlister L, Holland MJ. 1982. Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. *J Biol Chem* **257**: 7181–7188.

Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res* **26**: 3188–3193.

Pavesi A. 1999. Relationships between transcriptional and translational control of gene expression in *Saccharomyces cerevisiae*: a multiple regression analysis. *J Mol Evol* **48**: 133–141.

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol* **268**: 322–330.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci USA* **94**: 7784–7790.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1994. *Numerical Recipes in C*, 2nd edn. Cambridge University Press: Cambridge; p.275.

Sharp PM, Cowe E. 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7**: 657–678.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**: 28–38.

Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281–1295.

Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* **21**: 835–841.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**: 5125–5143.

Solomovici J, Lesnik T, Reiss C. 1997. Does *Escherichia coli* optimize the economics of the translation process? *J Theor Biol* **185**: 511–521.

VanBogelen RA, Greis KD, Blumenthal RM, Tani TH, Matthews RG. 1999. Mapping regulatory networks in microbial cells. *Trends Microbiol* **7**: 320–328.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* **270**: 484–487.

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW. 1997. Characterization of the yeast transcriptome. *Cell* **88**: 243–251.

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol* **15**: 1359–1367.

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

Xia X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* **144**: 1309–1320.

Xia X. 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics* **149**: 37–44.

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. 1997. Gene expression profiles in normal and cancer cells. *Science* **276**: 1268–1272.