J Mol Evol (2003) 56:473–484
DOI: 10.1007/s00239-002-2418-x

JOURNAL OF MOLECULAR EVOLUTION

Tracing Specific Synonymous Codon–Secondary Structure Correlations Through Evolution

Matej Orešič, Michael H. H. Dehn, Daniel Korenblum, David Shalloway

Department of Molecular Biology and Genetics, Cornell University, 265 Biotechnology Building, Ithaca, NY 14853, USA

Received: 24 May 2002 / Accepted: 28 November 2002

Abstract. We previously showed that GAU codons are preferred (relative to synonymous GAC codons) for encoding aspartates specifically at the N-termini of α-helices in human, but not in *E. coli*, proteins. To test if this difference reflected a general difference between eucaryotes and procaryotes, we now extended the analysis to include the proteins and coding sequences of mammals, vertebrates, *S. cerevisiae*, and plants. We found that the GAU-α-helix correlation is also strong in non-human mammalian and vertebrate proteins but is much weaker or insignificant in *S. cerevisiae* and plants. The vertebrate correlations are of sufficient strength to enhance α-helix *N*-terminus prediction. Additional results, including the observation that the correlation is significantly enhanced when proteins that are known to be correctly expressed in recombinant procaryotic systems are excluded, suggest that the correlation is induced at the level of protein translation and folding and not at the nucleic acid level. To the best of our knowledge, it is not explicable by the canonical picture of protein expression and folding, suggesting the existence of a novel evolutionary selection mechanism. One possible explanation is that some α-helix N-terminal GAU codons may facilitate correct co-translational folding in vertebrates.

Key words: Co-translational folding — Aspartate — Statistical database analysis — Folding shift

Correspondence to: D. Shalloway; *email:* dis2@cornell.edu

Introduction

Synonymous codon (SC) usage is generally assumed to passively reflect influences such as mutational biases and selection for translational rate and accuracy (Ikemura 1985; Sharp and Matassi 1994). However, the redundancy of the genetic code means that a gene sequence could embed biological information beyond the amino acid sequence. One possibility is that some excess coding capacity may be used to assist co-translational protein folding (Orešič and Shalloway 1998). A signature of such effects would be specific correlations between synonymous codon (SC) usage and protein structure. While several searches for these have been conducted (Thanaraj and Argos 1996a, b; Adzhubei et al. 1996; Brunak and Engelbrecht 1996; Orešič and Shalloway 1998; Tao and Dafu 1998; Gupta et al. 2000), in most cases the methods used were flawed (see Orešič and Shalloway 1998) and further investigation is required.

We previously used contingency table analysis to rigorously investigate the correlations between relative SC usage (RSCU) and protein secondary structure in human and *E. coli* proteins (Orešič and Shalloway 1998). SC choice could, in principle, affect secondary structure away from the codon position, so we compared SC choice with secondary structure at "offset" positions. The chi-square statistic χ^2 was used to look for violations of the null hypothesis, H_0, that RSCU and secondary structure are uncorrelated. Monte Carlo simulations were performed to compute confidence levels that accounted for the simultaneous consideration of multiple contingency tables. We

found a highly significant preference for *E. coli* Asn AAC codons downstream from β-sheet segments and for human Asp GAU codons at the N termini of α-helices. Additional tests excluded the possibilities that the correlations resulted from nucleotide base composition variations, nucleotide context effects, or correlations with intron–exon boundaries or the position of secondary structure elements within the protein (Orešič and Shalloway 1998).

The *E. coli* AAC-β correlation can be explained by selection for translational accuracy: In this species the Asn (non-wobble) AAC codon is translated an order-of-magnitude more accurately than the Asn (wobble) AAU, which has the highest known basal mistranslation frequency, 5×10^{-3} (Parker 1992). The correlation is strongest near inter-strand β-turns (Orešič and Shalloway 1998), and Asn is commonly used in type II β-turns to stabilize β-sheets (Creighton 1993). This suggests that the enhanced use of AAC codons downstream of β-sheet segments could result from selection against misreading in combination with an important role for Asn in β-sheet formation. This extends the demonstration by Akashi (1994), that more-accurately translated SCs can be selected during evolution specifically at functionally important locations, to demonstrate that selection can also occur at locations that are important for folding. The fact that the AAC-β-sheet segment correlation is not evident in human proteins is explained by the fact that Asn translation is two orders-of-magnitude more accurate in mammals (Harley et al. 1981), so selection against inaccuracy would not be important in this taxon.

On the other hand, selection for translational fidelity cannot explain the strong human Asp GAU-α-helix correlation: GAU is the wobble codon and binds less tightly to $tRNA^{Asp}$ than GAG (Singhal and Kopper 1981; Steinberg et al. 1995), so it is expected to have the higher mistranslation rate. Evidence that effects at the level of protein expression (in contrast to effects at the nucleotide level) were involved came from considering the sources of the protein crystals used for structure determination: If human proteins require structure-specific GAUs for correct expression and folding, they would not be correctly produced in recombinant expression systems that do not support the taxon-specific mechanism. We tested this hypothesis indirectly using the assumption that (for cost and efficiency) there is a strong *de facto* bias for experimentalists to prepare crystals of human proteins using protein expressed in procaryotic recombinant expression systems. Thus, the use of protein from human tissue to prepare a crystal may provide some evidence that it was difficult to express and fold it in a recombinant system. Therefore, if the GAU-α correlation is important for this, we would expect it to be stronger within the "native" subset of human proteins whose crystals had been prepared using natively-expressed protein. This unusual prediction was verified, suggesting that a novel protein-level effect is involved (Orešič and Shalloway 1998).

We do not know why the correlation is species-specific. Some experimental data suggests that eucaryotes and procaryotes use different protein folding mechanisms (Netzer and Hartl 1997; Netzer and Hartl 1998; Ellis and Hartl 1999). This could explain the specificity if the correlation were related to protein folding. In any case, examination of the correlation in multiple taxa is needed to evaluate this and other potential hypotheses. Towards this end we examined the GAU-α (and also the AAC-β) correlation in all taxa for which there is sufficient structural data for statistically significant analysis: mammals, vertebrates, *S. cerevisiae*, and plants.

Materials and Methods

Datasets

Non-homologous (≤25% sequence similarity) datasets of coding sequences (from GenBank) and matching secondary structures [from the Protein Data Bank (PDB) and Ras-Mol program (Sayle and Milner-White 1995)] were assembled as described (Orešič and Shalloway 1998). Since codon-structure correlations may be species-specific, and to avoid complications from interspecies differences in SC usages, data should ideally be pooled only within a single species. The only species having enough non-homologous structures in the PDB for statistically significant analysis are *E. coli* (31 proteins), *S. cerevisiae* (34 proteins, 20 natively expressed), and human (35 proteins, 17 native). However, the variation in RSCU between most mammalian and vertebrate species represented in the databank is small ($<4\%$ and $<8\%$, respectively), so we pooled the non-homologous data from mammals (71 proteins, 42 native) and vertebrates (85 proteins, 51 native) into datasets. Mammalian-excluding-human (36 proteins, 25 native) and vertebrate-excluding-human (54 proteins, 37 native) datasets were also tested. A dataset containing all non-homologous plant protein structures (48 proteins, 33 native) was also assembled. However, inter-plant species RSCU variations are large ($>45\%$), so the plant results must be considered with caution. Dataset proteins are listed in the Appendix.

p-*Value,* p_5-*Value, and* $P_5(p_5)$ *Overall Likelihood*

These were computed from contingency table χ^2 statistics by Monte Carlo analysis as previously described (Orešič and Shalloway 1998) using $N_{MC} = 200{,}000$ simulated datasets. Each simulated dataset had the experimental amino acid sequences and secondary structure maps, but had pseudo-random coding sequences generated according to the dataset RSCUs. Contingency tables and chi-square values χ^2_{MC} were calculated for each amino acid and offset for each Monte Carlo dataset. The *p*-value for a contingency table was calculated as the number of Monte Carlo datasets having χ^2_{MC} greater than the corresponding experimental χ^2. p_5-values were calculated similarly using contingency tables summed over five adjacent offsets (i.e., including two on each side of the indexed offset). $P_5(p_5)$ likelihoods were computed as the fraction of Monte Carlo datasets that had at least one of their p_5-values less than the experimental p_5.

As a control, p-values, p_5-values, and $P_5(p_5)$ likelihoods were computed using Wilks' G^2 statistic (Agresti 1990) in lieu of the χ^2 statistic. Almost identical values were obtained.

Statistical power. The power of the p_5[Asp(2)] statistic was computed as described in Agresti 1990 (p. 241) using a log-linear model to compute the maximum likelihood windowed Asp(2) contingency tables for a specified strength of Asp(2)-α-helix correlation. The correlation strength was parameterized by the windowed α:not-α (α:$\bar{\alpha}$) odds-ratio $\rho_5^{\alpha:\bar{\alpha}}$[Asp(2)]:

$$\rho_5^{\alpha:\bar{\alpha}}[\mathrm{Asp}(2)] \equiv \frac{n_5^{\mathrm{Asp},2}(\mathrm{GAU},\alpha)}{n_5(\mathrm{GAU},\bar{\alpha})} : \frac{n_5^{\mathrm{Asp},2}(\mathrm{GAC},\alpha)}{n_5^{\mathrm{Asp},2}(\mathrm{GAU},\alpha)} \tag{1}$$

where $n_5^{\mathrm{Asp},2}(c,s)$ is the number of events in the windowed Asp(2) contingency table for codon c and secondary structure s. The log-linear model maximum likelihood contingency table occupancy numbers were

$$\begin{aligned}
n(\mathrm{GAU},\alpha) &= n(\mathrm{GAU})n(\alpha)e^{-\lambda}/n \\
n(\mathrm{GAC},\alpha) &= n(\alpha) - n(\mathrm{GAU},\alpha) \\
n(\mathrm{GAU},\beta) &= [n(\mathrm{GAU}) - n(\mathrm{GAU},\alpha)]n(\beta)/[n(\beta) + n(o)] \\
n(\mathrm{GAC},\beta) &= [n(\mathrm{GAC}) - n(\mathrm{GAC},\alpha)]n(\beta)/[n(\beta) + n(o)] \\
n(\mathrm{GAU},o) &= [n(\mathrm{GAU}) - n(\mathrm{GAU},\alpha)]n(o)/[n(\beta) + n(o)] \\
n(\mathrm{GAC},o) &= [n(\mathrm{GAC}) - n(\mathrm{GAC},\alpha)]n(o)/[n(\beta) + n(o)],
\end{aligned} \tag{2}$$

where $n \equiv n_5^{\mathrm{Asp},2}$. The model is parameterized by the experimental contingency table margins $[n_5^{\mathrm{Asp},2}(\alpha), n_5^{\mathrm{Asp},2}(\beta), n_5^{\mathrm{Asp},2}(o)$, $n_5^{\mathrm{Asp},2}$(GAC), and $n_5^{\mathrm{Asp},2}$(GAU)] and λ, which provides an alternative measure of the correlation strength. Its relationship to $\rho_5^{\alpha:\bar{\alpha}}$ is obtained by substituting Eqs. (2) into Eq. (1). H_0 corresponds to $\lambda = 0$, $\rho_5^{\alpha:\bar{\alpha}} = 1$.

Statistical confidence of native-expression enhancement. One hundred subsets (each having the same number of members as the native-expression subset) were randomly selected from the human and vertebrate subsets. For each random subset we calculated the overall likelihood $P_5(p_5)$ for Asp($\Delta = 3$) (the furthest outlier in the native-expression human subset) and compared it to the corresponding $P_5(p_5)$ occurring in the experimental native-expression subset (0.0001 for human and 0.0006 for vertebrates). Only 4 out of 100 random human subsets had $P_5(p_5) < 0.0001$ and only 3 out of 100 random vertebrate subsets had $P_5(p_5) < 0.0006$.

GC/AU-adjusted, Chi-square analysis. The elements of the simulated contingency table for gene i, $n^{\mathrm{Asp},2}(i;c,s)$, were generated holding the number of α and not-α residues fixed while randomly assigning GACs and GAUs in proportion to the third-base GC/AU usage ratio of the gene. A log-linear adjustment ξ was included so that the expected numbers of GACs and GAUs for the entire dataset were equal to the observed numbers. That is, the gene-specific SC-assignment probabilities, $p(i;\mathrm{GAC})$ and $p(i;\mathrm{GAU})$, were fixed by

$$\frac{p(i;\mathrm{GAC})}{p(i;\mathrm{GAU})} = \xi\frac{n(i;\mathrm{GC}_3)}{n(i;\mathrm{AU}_3)}, \tag{3}$$

$$1 = p(i;\mathrm{GAU}) + p(i;\mathrm{GAC}), \tag{4}$$

$$n^{\mathrm{Asp},2}(\mathrm{GAC}) = \sum_i n^{\mathrm{Asp},2}(i)p(i;\mathrm{GAC}), \tag{5}$$

$$n^{\mathrm{Asp},2}(\mathrm{GAU}) = \sum_i n^{\mathrm{Asp},2}(i)p(i;\mathrm{GAU}), \tag{6}$$

where $n(i;\mathrm{GC}_3)$ and $n(i;\mathrm{AU}_3)$ are the numbers of GCs or AUs in the third-base positions of gene i, $n^{\mathrm{Asp},2}$(GAC) and $n^{\mathrm{Asp},2}$(GAU) are the experimental dataset contingency table margins, and $n^{\mathrm{Asp},2}(i)$ is the total number of events in the experimental contingency table of gene i. ξ was determined (equivalently) by either Eq. 5 or 6. The individual gene Monte Carlo tables were then summed to generate a total Monte Carlo dataset table. The $p^{\alpha:\bar{\alpha}}$-value for the experimental Asp(2) contingency table was then computed by comparing its χ^2 with the χ^2_{MC} of 1,000,000 Monte Carlo tables.

Results

The Prior Statistical Analysis: Development of the Pre-Experimental Hypothesis

The prior analysis, which identified the GAU-α correlation in the human dataset, has been described (Orešič and Shalloway 1998). To summarize, protein secondary structure was classified into three categories: α-helix, β-sheet segment, and "other." For each multi-codon amino acid, a contingency table that confronted these structural categories with RSCU was constructed (e.g., see Fig. 1 of Orešič and Shalloway 1998). Contingency tables were also constructed that compared the SC at position i with the structure at an offset position $i + \Delta$, with Δ covering the "structural range," $-10 \le \Delta \le +10$. Outlier analysis was performed using the χ^2 value of each table as a statistic to test H_0, the null hypothesis of no SC-structure correlations. The corresponding p-value gives the likelihood that an individual contingency table would have the observed or greater χ^2 if H_0 were true. Biologically significant correlations will generate clusters of outliers at adjacent offsets while random fluctuations will not be clustered. Therefore, to increase statistical power we summed contingency tables over five adjacent offsets and computed their "windowed" p_5-values by Monte Carlo simulation.

It is necessary to account for the fact that there are 378 (= 18 multi-codon amino acids × 21 offsets) quasi-independent p_5-values when computing overall confidence levels under H_0. For this purpose Monte Carlo simulation was used to determine the probability distribution under H_0 of the *smallest* p_5-value among *all* the 378 contingency tables. This distribution was then used in a single-sided test with the experimental p_5-values to compute the overall likelihood, $P_5(p_5)$, for each amino acid and offset. The smallest $P_5(p_5)$ for the dataset measures the likelihood that H_0 is true.

Analysis of the human and *E. coli* datasets (some data reprinted in Table 1 here) identified a cluster of outliers in the 2 × 3 (GAU:GAC × α:β:other) Asp contingency tables in the offset range $2 \le \Delta \le 5$ in the human, but not *E. coli,* dataset. (The cluster of *E. coli* Asn outliers discussed in the Introduction is also evident in Table 1.) $P_5(p_5)$ was <0.01 at offsets −2

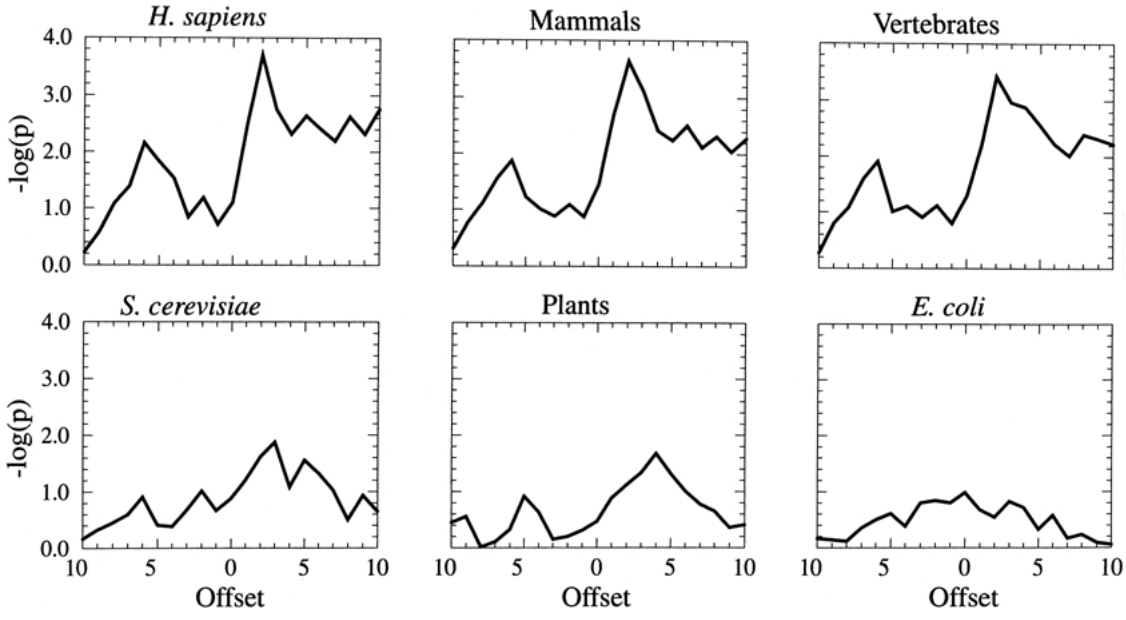

Fig. 1. Asp *p*-values as functions of offset Δ. High peaks (small *p*-values) are indicators of potential SC-structure correlations.

and −3 indicating that H_0 is violated at the 1% confidence level, even after accounting for the multiplicity of tests.

To identify the contrast that violated H_0, the 2×3 human Asp contingency tables were projected onto three 2×2 tables in which the structural categories were collapsed to either α or not-α ($\bar{\alpha}$), β and not-β ($\bar{\beta}$), or other and not-other ($\overline{\text{other}}$). Comparing the *p*-values of the projected tables showed that the violation results from a GAU-α-helix correlation, and examining the spatial variation of the GAC/GAU usage ratio showed that this preference is concentrated near α-helix *N*-termini.

Isochore effects (Bernardi 1995) can bias SC usage between genes (Karlin and Mrázek 1996) and could possibly induce an apparent SC-structure correlation *if* there were a complementary correlation between the structure of a protein and the GC/AU composition of its gene. We tested this possibility by testing for correlation between gene GC content and α-helix content. We also tested for correlations between GC content and the magnitude of the GAU-α-helix correlation. Both tests showed that isochore effects were not responsible for the SC-structure correlation.

Focused Versus Non-Focused Statistical Analyses

Two different statistical approaches can be used to identify the evolutionary conservation of the GAU-α-helix correlation: (1) Test the pre-experimental, focused null hypothesis $h_0^{\text{Asp},2}$ that the Asp at $\Delta = 2$ [Asp(2)] correlation does not exist in the non-human taxa. (2) Perform an unfocused analysis that ignores the prior study and replicates the full multiple contingency table test in the non-human taxa. These approaches are complementary: Since the focused analysis tests only one contingency table, it has stronger power to protect against type II errors (missing a violation of $h_0^{\text{Asp},2}$). Conversely, the unfocused analysis provides stronger protection again type I errors (falsely concluding that H_0 is violated). Both approaches were used so that conservative conclusions, protected against both types of errors, could be made. We first present the unfocused analysis.

Statistical Analysis Without a Pre-Experimental Hypothesis

The mammalian, mammalian-excluding-human, vertebrate, vertebrate-excluding-human, *S. cerevisiae,* and plant datasets were analyzed using the full contingency table analysis. The outliers having the smallest *p*-values are listed in Table 1 along with the previous human and *E. coli* results. We see that the human, mammalian, mammalian-excluding-human, vertebrate, and vertebrate-excluding-human datasets all contain clusters of Asp outliers in the offset range $2 \le \Delta \le 5$. The Asp contingency tables at $\Delta = 2$ or $\Delta = 3$ are the furthest outliers in all five datasets.The strong correlations found in the mammalian-excluding-human and vertebrate-excluding-human datasets demonstrate that the strong violations of H_0 in the mammalian and vertebrate datasets do not result just from the included human proteins. The spatial concentration of these outliers is visible in Fig. 1, which displays the *p*-values of Asp contingency tables as a function of offset within the structural range. The

Table 1. Contingency table outliers

Residue	Offset	−log(p)	−log(p_5)	$P_5(p_5)$	$P_5(p_5)$ [native]
			H. sapiens		
Asp	2	3.70	4.12	**0.008**	0.0006
Gly	−6	2.89	1.91	0.207	0.2078
Asp	10	2.76	1.19	0.612	0.3891
Asp	3	2.76	4.03	0.009	**0.0001**
Asp	5	2.64	3.10	0.021	0.0013
Asp	8	2.63	1.68	0.228	0.0749
			Mammals		
Asp	2	3.64	4.20	**0.006**	0.0007
Asp	3	3.12	3.94	0.007	**0.0005**
Glu	3	2.74	2.05	0.103	0.0927
Asp	1	2.68	3.86	0.007	0.0008
Gly	−6	2.53	1.51	0.328	0.3410
			Mammals-excluding-*H. sapiens*		
Asp	3	3.46	3.37	**0.010**	**0.0012**
Asp	4	3.17	2.98	0.013	0.0015
Asp	2	3.12	2.76	0.014	0.0015
Glu	4	2.65	1.49	0.347	0.3239
			Vertebrates		
Asp	2	3.43	3.57	0.008	0.0009
Asp	3	2.97	3.75	0.006	**0.0006**
Asp	4	2.89	3.88	**0.005**	0.0007
Asp	5	2.56	3.02	0.010	0.0013
			Vertebrates-excluding-*H. sapiens*		
Asp	3	3.60	3.90	**0.008**	**0.0011**
Asp	4	3.21	3.87	0.009	0.0013
Asp	2	2.97	2.94	0.012	0.0019
Glu	4	2.82	1.79	0.263	0.2955
Asp	5	2.67	2.15	0.056	0.0284
			S. cerevisiae		
Gly	−4	3.74	1.38	0.304	0.319
Glu	2	2.85	2.13	**0.142**	**0.082**
Cys	5	2.68	1.86	0.276	0.245
Asp	3	1.88	2.10	0.151	0.126
			Plants		
Gly	−8	2.73	1.47	0.329	0.297
Asn	3	2.61	1.74	**0.250**	**0.233**
Asp	4	1.68	1.52	0.301	0.267
			E. coli		
Asn	−6	5.34	2.94	**0.005**	
Asn	−5	3.86	2.92	0.007	
Asn	−7	3.69	2.85	0.007	
Asn	−3	3.20	2.04	0.042	
Asp	0	0.99	0.17	1.000	

All outliers with $\log(p) < -2.5$ are included, as well as the smallest Asp outlier for each dataset regardless of p-value. *H. sapiens* and *E. coli* values were previously reported in Orešič and Shalloway 1998. The smallest $P_5(p_5)$ and $P_5(p_5)$ [native] (in boldface) give confidence levels for violations of H_0 (accounting for the use of multiple statistics) in the full and native-expression datasets, respectively.

human, mammalian, and vertebrate Asp plots each have a large peak of very small (10^{-3}–10^{-4}) p-values within this range. Much smaller peaks are observed in the *S. cerevisiae* and plant datasets, and no peak is observed in the *E. coli* dataset.

To test the statistical significance of these outliers, their $P_5(p_5)$ overall likelihoods were computed as described above. The lowest $P_5(p_5)$ for *any* amino acid and offset in the human, mammalian, mammalian-excluding-human, vertebrate, and vertebrate-excluding-human datasets are all for Asp at $\Delta = 2$, 3, or 4 and are all ≤0.01 (Table 1 and Fig. 2), indicating strong violations of H_0. In contrast, there are no statistically significant outliers in the *S. cerevisiae* and plant datasets.

SC-structure correlations in the plant dataset could have been missed by this analysis because of the significant RSCU variations between the different included species. To at least partially correct for this, we also performed a Monte Carlo analysis in which the pseudo-random coding sequences were generated separately for proteins of each plant species using its

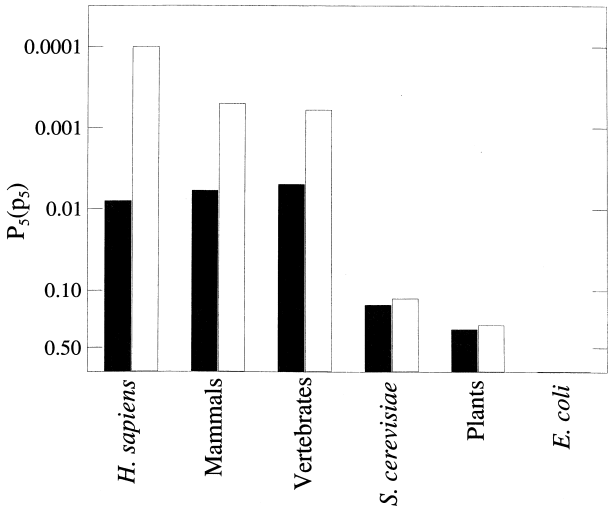

Fig. 2. Overall likelihoods P_5 (p_5) for the strongest Asp outliers within the structural range. Shaded and white bars indicate the P_5 (p_5) for the complete and native-expression datasets, respectively.

RSCU as reported in the Codon Usage Database (Nakamura et al. 1998). This analysis also found no additional far outliers (data not shown).

Following the procedures previously used to demonstrate that the violation of H_0 in the human Asp contingency tables resulted from a GAU-α-helix correlation (Orešič and Shalloway 1998), we compared the *p*-values of the projected 2×2 α:ᾱ, β:β̄, and other:other against the *p*-values of the full 2×3 table as functions of Δ. In all the eucaryotic datasets, the α:ᾱ *p*-values coincided with those of the full table *p*-values within the range $2 \leq \Delta \leq 5$ (data not shown), indicating that this contrast dominates the correlation.

To explore the spatial structure of the correlation, we aligned the N-termini of all α-helices in each dataset and plotted histograms of Asp GAC and GAU codon usage relative to the N-termini (Fig. 3). In the human, mammalian, and vertebrate datasets GAU usage is highly favored at the position of the "N_{cap}" (the residue just upstream of the first residue of the helix) and at the second residue in the helix. Much weaker, but similar, patterns are observed in *S. cerevisiae* and plants. No such bias is observed in *E. coli*.

Statistical Analysis with the Pre-Experimental Hypothesis

The unfocused analysis provides strong evidence for the GAU-α correlation in the human, mammalian, and vertebrate taxa, but it is necessary to test for differences in statistical power before concluding that the correlation is insignificant or much weaker in the other taxa (i.e., to control for type II errors). It is not possible to assess the statistical power of the P_5 test because H_0 represents an entire family of null statements, one for each amino acid and offset, each governed by an independent correlation-strength parameter (Miller 1981). Instead, to permit power comparisons, we tested the focused null hypothesis $h_0^{\mathrm{Asp},2}$, which specifically states that there are no Asp-α-helix SC-structure correlations at offset 2 (the furthest outlier in the prior human study). This was done by analyzing the collapsed 2×2 α:ᾱ windowed Asp(2) contingency tables using their $p_5^{\alpha:\bar{\alpha}}$-values as statistics. We emphasize that these single-table statistics are valid for the non-human taxa because the prior (human-only) analysis had already provided a basis for restricting focus.

These $p_5^{\alpha:\bar{\alpha}}$[Asp(2)]-values are listed in Table 2. Even more strongly than the unfocused P_5 test, they indicate that the correlation discovered in human proteins and sequences is also present in vertebrates (i.e., at the 0.0005 confidence level). The *S. cerevisiae* and plant GAU-α correlations are again much weaker, but this more powerful test provides some indication of a weak correlation (i.e., at the 0.05 confidence level). The strength of the correlations can be assessed by the deviation from 1 of the odds-ratio $p_5^{\alpha:\bar{\alpha}}$ of the GAU:GAC × α:ᾱ windowed Asp(2) contingency table. We see that the significant differences between the $p_5^{\alpha:\bar{\alpha}}$-values reflect large differences in the odds-ratios (Table 2).

The statistical power of this test was evaluated as described in Materials and Methods. The power curves for confidence level $\alpha = 0.05$ are plotted in Fig. 4. They display the probability that the $p_5^{\alpha:\bar{\alpha}}$-value will be <0.05 if there is an underlying correlation having strength $\rho_5^{\alpha:\bar{\alpha}}$ (or $1/\rho_5^{\alpha:\bar{\alpha}}$, since the test has this symmetry). Except for vertebrates-excluding-human (which has higher power because of the large dataset size), the power curves are similar. The *E. coli* and plant tests are at least as powerful as the human test, confirming that the observed correlation differences are not artifacts. Although the *S. cerevisiae* test is slightly weaker than the human test, the decrease in sensitivity to changes in $\rho_5^{\alpha:\bar{\alpha}}$ is only 0.014. This is $\sim 30 \times$ less than the difference in $\rho_5^{\alpha:\bar{\alpha}}$ between these datasets, and thus is insignificant. We conclude that the GAU-α correlation is much weaker in *S. cerevisiae* and plants than in vertebrates and humans.

Enhanced SC-Structure Correlations in Native-Expression Subsets

As discussed in the Introduction, if the GAU-α correlation were important for correct expression and folding, we would expect it to be stronger in the native-expression data subset. We tested this prediction for each of the eucaryotic datasets. The native-expression subsets contain approximately 50% (human), 60% (mammals, vertebrates, and *S. cerevisiae*), or 70% (plants) of the total number of proteins. As

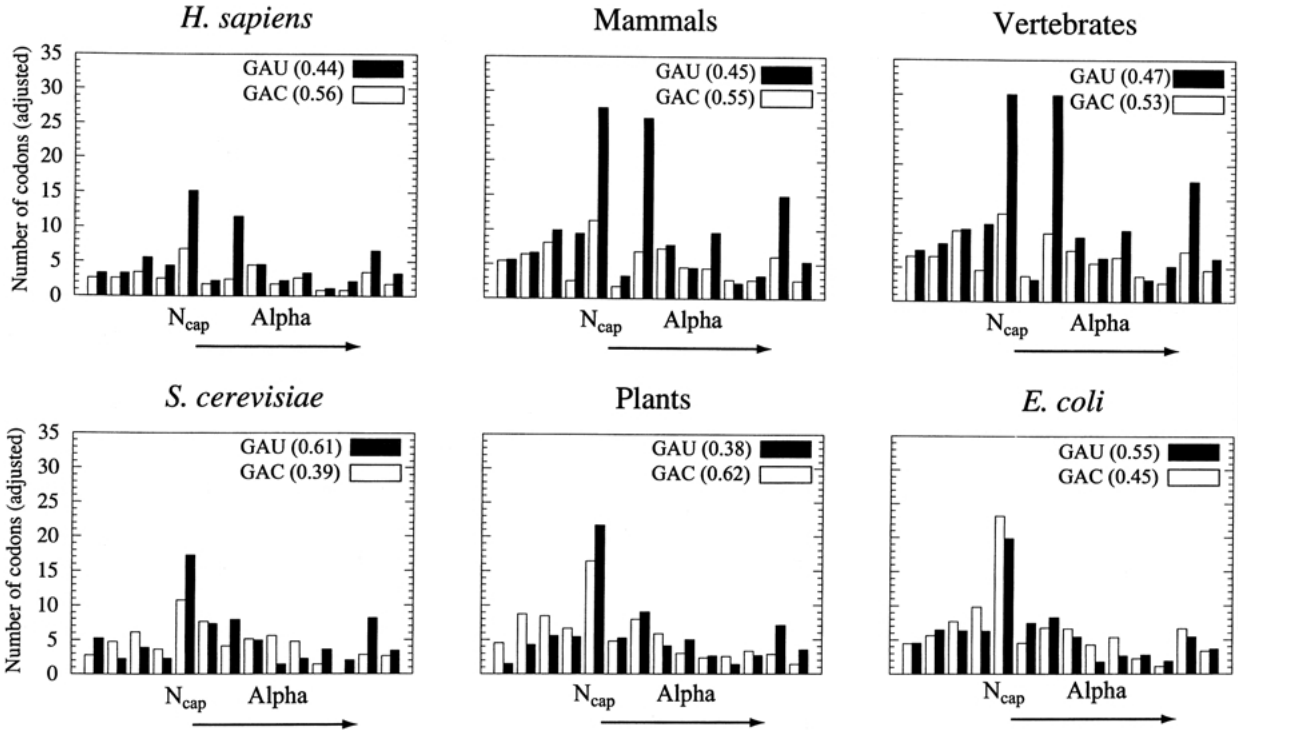

Fig. 3. Occurrences of Asp SCs near α-helix boundaries. Within each dataset, the N-termini of all α-helices were aligned and the numbers of Asp SCs at each position, *N*, were counted. To account for the different overall Asp RSCUs in each dataset, the adjusted numbers N'_{GAU} and N'_{GAC} are plotted. $N'_{GAU} = (N_{GAU} + N_{GAC}) \times \sigma/(1+\sigma)$ and $N'_{GAC} = (N_{GAU} + N_{GAC}) \times 1/(1+\sigma)$ where $\sigma = [(N_{GAU}/N_{GAU}\ (\text{total}))/(N_{GAC}/N_{GAC}\ (\text{total}))]$. $N'_{GAU} + N'_{GAC}$ at each position equals the total number of Asps at that position, and N'_{GAU}/N'_{GAC} equals the RSCU-adjusted SC usage ratio, σ. "N_{cap}" identifies the residue that is just upstream of the first residue of the helix.

Table 2. Statistical significance with pre-experimental hypothesis

	Asp (2)		Asp (3), native	
Dataset	$-\log(p_5^{\alpha:\bar{\alpha}})$	$\rho_5^{\alpha:\bar{\alpha}}$	$-\log(p_5^{\alpha:\bar{\alpha}})$	$\rho_5^{\alpha:\bar{\alpha}}$
H. sapiens	4.03	1.93	5.86	2.55
Vertebrates-excluding-*H. sapiens*	3.87	1.65	4.91	1.96
S. cerevisiae	1.79	1.19	2.35	1.21
Plants	1.51	1.12	1.68	1.09
E. coli	0.83	0.95	0.22	0.81

$p_5^{\alpha:\bar{\alpha}}$-values and $\rho_5^{\alpha:\bar{\alpha}}$ for the GAU:GAC × $\alpha:\bar{\alpha}$ contingency tables for the specified offsets and datasets are listed. The Asp(2) [or Asp(3)] $\alpha:\bar{\alpha}$ contrast was the strongest violator of H_0 in the prior study of the full (or native-expression) *H. sapiens* dataset. Thus, their $p_5^{\alpha:\bar{\alpha}}$-values are the best statistics for the focused tests of the non-human taxa.

shown in Table 1 and Fig. 2, this restriction enhances the significance of the violations of H_0 [i.e., reduces the values of the $P_5(p_5)$] by an order-of-magnitude in the human, mammalian, and vertebrate taxa. When only the natively-expressed proteins are included, the position of the furthest outlier in the native-expression human subset is shifted slightly to Asp(3), the odds-ratios for the human, mammalian, and vertebrate taxa strongly increase, and their p_5-values decrease to even more significant levels of $\sim 10^{-6}$–10^{-5} (Table 2). In contrast, only small changes are observed when the *S. cerevisiae* and plant datasets are restricted to natively-expressed proteins. (However, since fewer plant proteins are excluded in going from the total to the native-expression subset, the relative lack of change may not be meaningful in this case.) The net effect is that restriction to the native-expression subsets further increases the differences between the correlation significance and strength in vertebrates relative to *S. cerevisiae* and plants.

The GC contents of the human, human (native), vertebrate-excluding-human, and vertebrate-excluding-human (native) datasets were all 0.53 ± 0.01, so the expression system-dependent changes could not have been induced by differences in GC/AU usages. To check whether these enhancements could have

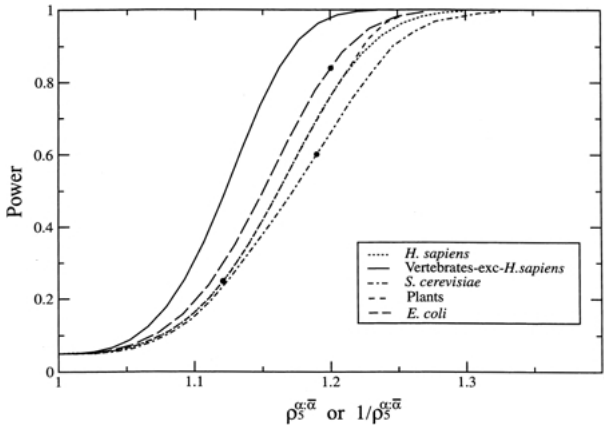

Fig. 4. Statistical power curves. Statistical power for detecting correlations in the windowed Asp(2) contingency tables were computed at the $\alpha = 0.05$ confidence level. The power is the probability that the dataset will have $p_5 < 0.05$ when the GAU-α correlation has strength $\rho_5^{\alpha:\bar{\alpha}}$. The experimental $\rho_5^{\alpha:\bar{\alpha}}$ for the *E. coli,* plant, and *S. cerevisiae* datasets are indicated with dots; the $\rho_5^{\alpha:\bar{\alpha}}$ for *H. sapiens* and vertebrates-excluding-*H. sapiens* are off-scale to the right.

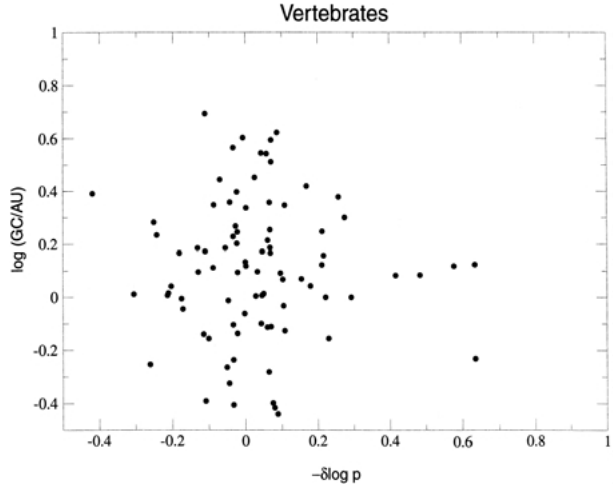

Fig. 5. Scatter plot contrasting $\log(GC/AU)_i$ with $-\delta \log p_i$ for the Asp(3) vertebrate contingency tables. GC/AU_i is the nucleotide content ratio for gene i and $-\delta \log p_i$ measures the decrease in the Asp ($\Delta = 3$) vertebrate p-value when protein/gene i is deleted from the dataset.

been random artifacts of subpartitioning, statistical confidence levels for the enhancements were determined by Monte Carlo analysis. These were 0.04 and 0.03 for the human and vertebrate enhancements, respectively. The confidence levels may be limited by the limited sizes of the datasets; studies with significantly larger datasets, when they become available, may improve the power of the analysis.

Potential Effects of Varying Gene-Specific GC/AU Usage

It is a priori possible that the SC-structure correlations could have been indirectly induced by the combined effects of an SC-isochore correlation and an isochore-structure correlation. We previously showed that there are no isochore-structure correlations for the human dataset (Orešič and Shalloway 1998), suggesting that this is not the case. To test whether isochore effects were responsible for the vertebrate correlation, we contrasted the GC/AU composition of each gene with the extent of its contribution to the observed Asp($\Delta = 3$) SC-structure correlation. The extent of contribution of protein i to the correlation was measured by its $\delta \log p_i$-value, the difference between the Asp(3) log p-value of the total vertebrate dataset and the log p-value with protein i excluded. If isochore effects had induced the observed correlation, we would expect that the genes having GC/AU content ratios $(GC/AU)_i$ farthest from 1 would have the largest values of $-\delta \log p_i$. However, there is no correlation between $|\log (GC/AU)_i|$ and $-\delta \log p_i$ (Fig. 5); the Spearman correlation coefficient is $r = -0.027$ ($p = 0.80$). Similar results were obtained with *S. cerevisiae;* $r = -0.013$ ($p = 0.94$).

In principle, another way to test for isochore effects would be to generate Monte Carlo contingency tables for each gene using the RSCU from the gene alone (rather than using the dataset RSCU as in the analyses above). However, the individual genes are not long enough for this. For example, the average gene length in the human dataset is ~180 codons, so there are many amino acids that are present only a few times within each gene. Thus, simulation with the individual gene RSCUs would, for many genes, force the simulated SC choices to be almost identical to the actual ones, and would severely reduce statistical power. Instead, we performed an alternative test in which the observed Asp(2) GAU:GAC × α:ᾱ contingency tables were compared against Monte Carlo tables that were pseudo-randomly generated for each gene separately using a GAC/GAU probability ratio that was proportional to the GC/AU nucleotide-usage ratio for third bases of the codons in each gene alone (see Materials and Methods). Thus, the Monte Carlo tables for genes having higher GC content tended to have proportionately higher fractions of GAC codons. $p_5^{\alpha:\bar{\alpha}}$-values computed in this way are displayed in Table 3. Including this bias only causes factor-of-two corrections that do not alter the conclusions that $h_0^{\mathrm{Asp},2}$ is significantly violated, that the violation is much stronger in the native-expressed dataset, and that it is not observed in *E. coli.*

Predictive Power

The Asp SC-α-helix correlation is, for vertebrate proteins, strong enough to contribute additional predictive power to secondary structure prediction. For example, the probability that an arbitrary residue

Table. 3. Effect of gene-specific GC/AU bias on significance levels

Dataset	$-\log[p^{\alpha:\bar{\alpha}}$ (Asp, 2)]	
	Uniform $\frac{p(GAC)}{p(GAU)}$	Gene-specific $\frac{p(GAC)}{p(GAU)}$
H. sapiens	3.4	3.0
H. sapiens (native)	4.1	4.0
H. sapiens (non-native)	0.27	0.24
E. coli	0.01	0.007

$p^{\alpha:\bar{\alpha}}$-values were computed for the χ^2 of the GAU:GAC $\times$ $\alpha:\bar{\alpha}$ Asp(2) contingency tables of the specified datasets by Monte Carlo analysis. The left column lists the $-\log(p^{\alpha:\bar{\alpha}})$ obtained when all the Asp codons were simulated according to the dataset GAC/GAU usage ratio. The right column lists the values obtained when the codons for each gene were generated in proportion to the GC/AU usage ratio of the third bases of the codons in that gene alone.

in the vertebrate database lies at a N_{cap} is $p(N_{cap}) = 0.028$ (Table 4). When the residue is an Asp, this is doubled to $p(N_{cap}|Asp) = 0.068$, and, when the codon is a GAU, it is increased an additional 30% to $p(N_{cap}|GAU) = 0.089$. The probability that a GAU-encoded residue lies within the region including N_{cap} plus two downstream positions is even higher: $p(N_{cap}\ldots N_2|GAU) = 0.187$. Similar results were observed with the mammalian and vertebrate subsets.

Discussion

We have extended a previous study (Orešič and Shalloway 1998), which discovered a surprising SC-secondary structure correlation in human (but not *E. coli*) coding sequences and proteins, to all other eucaryotic taxa having sufficient data for statistical analysis: mammals, vertebrates, *S. cerevisiae,* and plants. Using an unfocused statistical test we found significant evidence [$P_5(p_5) < 0.01$] that Asp GAU codons are preferred at the N-termini of α-helices in human, mammalian, and vertebrate proteins, and no evidence for this preference in *S. cerevisiae,* plant, and *E. coli* proteins. To protect against type II errors, we also tested the non-human taxa using a pre-experimental hypothesis that focused specifically on the Asp GAU-α-helix correlation at offset +2, the farthest human outlier. This provided very strong confirmation of the correlation in mammals and vertebrates ($p_5 < 0.0005$) and some indication ($p_5 < 0.05$) of a much weaker correlation in *S. cerevisiae* and plants. These differences did not result from differences in statistical power between the taxa, and two tests indicated that the correlations were not related to overall GC/AU usage ratios. The plant results must be interpreted with caution due to the unavoidable pooling of multiple species with different RSCUs, but their correspondence with the rigorous *S. cerevisiae* results provides some reassurance.

It is notable that the statistical significance of the human, mammalian, and vertebrate correlations is increased by at least an order-of-magnitude when recombinant-expressed proteins are excluded. This strongly suggests that the correlation depends on influences at the level of protein expression and/or folding. One possibility is that it is strongest in proteins that are highly expressed in human placenta and vertebrate tissue sources, a factor that would increase the likelihood that native rather than recombinant proteins would be purified for crystallization. This hypothesis must be tested experimentally, since (unlike some procaryotes, Sharp and Li 1987) vertebrate protein expression levels can vary widely between different cell types and cannot be determined from RSCUs alone. However, even if it were true, it could not account for the existence of the SC-structure correlation, since the GC and α-helical contents of the full and native-expression datasets were the same to within a few percent.

Mutational preferences are major determinants of species-specific RSCUs (Li 1997), but could only induce an SC-structure correlation if they were somehow coupled to protein structure. Dinucleotide mutational preferences that spanned codons could, in principle, do this, but we previously showed that this was not so for the human dataset. In fact, including the codon-spanning (''3–4'') dinucleotide preferences *increased* the statistical significance of the correlation (Orešič and Shalloway 1998). The possibilities that correlations with intron–exon boundaries or intragenic position might be involved were also eliminated. Thus, it is most likely that the SC-structure correlation results from a novel form of selection. The possibility that the selection is experimental (e.g., related to the types of proteins that are selected for structural studies) must be kept in mind. However, we do not know of any experimental selective pressure that could explain the observed structure-aligned correlations.

As previously suggested (Orešič and Shalloway 1998), an intriguing possibility that could explain both the correlation and its enhancement in the native-expression data subsets is that Asp SC choice plays a role in co-translational folding (Hardesty et al. 1999). Netzer and Hartl have suggested that co-translational folding is important in eucaryotes, but not in procaryotes (Netzer and Hartl 1997, 1998; Ellis and Hartl 1999). However, their experiments only contrasted mammals with *E. coli*, and there is no experimental basis for the generalization from mammals to all eucaryotes. If their ''folding shift hypothesis'' were modified to focus on vertebrates (and possibly some other higher eucaryotic taxa), it

would match the evolutionary pattern of the GAU-α-helix correlation.

One possible way for a SC to influence folding is pausing: Wobble codons (such as GAU) are generally translated more slowly (Thomas et al. 1988; Kato et al. 1990), and some evidence suggests that translational pauses caused by slowly translated SCs can be important for folding (Purvis et al. 1987; Crombie et al. 1992). Asp is commonly found at the N-termini of α-helices (Presta and Rose 1988; Richardson and Richardson 1988), and a GAU-induced pause might provide time for previously-translated residues to fold before α-helix synthesis. To speculate further, since α-helices are often found at the N-termini of folding domains, this might facilitate sequential domain folding. In principle, this hypothesis could be tested by looking for SC-folding domain correlations, with offsets measured relative to domain boundaries. Unfortunately, it is difficult to unambiguously define such boundaries at this time, but it may become possible in the future.

We cannot exclude the possibility that structure-specific SC selection also acts (albeit more weakly) on amino acids other than Asp, but that these effects are below the current threshold of statistical detectability. Thus, as the non-homologous taxon-specific structural datasets grow, it will be important to retest them for additional SC-structure correlations that may help edify the underlying mechanism. Larger datasets may also reveal if the correlation is concentrated in specific classes of genes/proteins. And, as data becomes available, tests on individual invertebrate and plant species can further elucidate the evolutionary emergence of the correlations.

New experimental tests may also be possible: Structural genomics projects that attempt automated expression of large numbers of eucaryotic proteins can provide well-controlled, direct information about individual vertebrate protein expression and folding in recombinant systems. This would permit a direct test of the relationship between the GAU-α correlation and recombinant expression/folding that would supersede the indirect (native-expression subset) analysis used here. Most interesting would be measurements of the effect of "silent" α-helix N-terminal GAU → GAC substitutions on the rate and extent of correct vertebrate protein folding. However, even effects that are strong enough for evolutionary selection may not be detectable in a direct experimental test.

Conversely, the GAU-α correlation, whatever its cause, may have practical applications: As shown in Table 4, the codon sequences possess more structural predictive power than the amino acid sequences alone, and algorithms can be developed that utilize this power for improved secondary structure prediction. In addition, it would be practically valuable for structural genomics if the SC-structure correlation, in combination with a secondary structure prediction algorithm, were able to increase the ability to predict the success of recombinant protein expression and folding.

Table. 4. Enhanced structure predictive power using the GAU-α-helix correlation

	N_{cap}	$N_{cap}...N_2$
$p\ (x)$	0.028	0.084
$p\ (x \mid \text{Asp})$	0.068	0.142
$p\ (x \mid \text{GAU})$	0.089	0.187

$p(x)$ is the probability that an amino acid is an α-helix N_{cap} (left column) or lies within the region including the N_{cap} and two downstream residues (right column). The second and third rows give the probabilities conditional upon the amino acid being an Asp or a GAU-encoded Asp, respectively.

Although it was not the focus of this investigation, we note that the preference for Asn AAC codons downstream from β-sheet segments, previously found in *E. coli* but not human proteins (Orešič and Shalloway 1998), was absent in the other eucaryotic taxa as well (Table 1). We have suggested that the *E. coli*-specific correlation is a consequence of the *E. coli*-specific high mistranslation rate of AAU (the other Asn SC) (Parker 1992) combined with the importance of Asn in inter-strand type II β-turns (Creighton 1993). While (as far as we are aware) AAU mistranslation rates have not yet been measured in eucaryotic taxa other than mammals, this hypothesis and the new data suggest that the mistranslation rates are relatively low in all eucaryotes. We predict that the AAC-β correlation will be an indicator of high AAU mistranslation rates in procaryotic species other than *E. coli*.

Acknowledgments. We thank Bruce Church, Jason Gans, and Chip Aquadro for helpful discussions, Rina Gendelman for help preparing the manuscript, the Intel Corporation for computing equipment. This study was supported in part by NSF Grant CCR-9988519 and by NIH Training Grant T32GM 08267 (to DK).

Appendix

Dataset Proteins

PDB and GenBank (in parentheses) identifiers are included for each protein. Proteins that were crystallized from material prepared using recombinant expression systems are marked with an asterisk. The human and *E. coli* datasets are described in (Orešič and Shalloway 1998).

Mammalian. In addition to human proteins, the following were included: 1atn (OCRNAASMA), 1bar* (BTFGFAR), 1bet (MUSNGFBA), 1btn

(MUSSPNA), 1cms* (BOVCHYMOB), 1dla (SSU46065), 1fkk (RABFKBP12A), 1lib* (MUSLBPA), 1mah (MMACHE), 1mdy* (MUSMYOD1A), 1mup (MUSMUPE), 1myg* (PIGMG), 1myp (S56200), 1nhm* (CHHMG1), 1phr (BOVPPTPPC), 1psp (SSPSPAA), 1rbb (S81740), 1rec* (BOVRECVR), 1rhd (BOVRHOD), 1rro* (RATOM), 1rtp (RATPALB), 1sxc (BOVCZSD), 1tag (BOVTRNAM), 1tfd (OCTRNFNM), 2cts (PIGCITSYN), 2ifb* (RATFABPX), 2ohx (HRSADHE), 2pld (BTPLCII), 2pnb* (BOVP85AA), 2tma (S78854), 3b5c (BOVCYTB01), 3dni (SYNDNASEI), 4gcr (BOVCRYGBA), 5gst* (RNGSTYBR), 6acn (PIGACON), and 6est (PIGELS1).

Vertebrate. In addition to all the mammalian proteins listed above, the following were used: 1atl (CRLPREHTD), 1ave (CHKAVID), 1cew (CHKCYS), 1gat* (CHKRERYF1), 1hst* (GGHI03), 1kba (BMKAPBUNG), 1ovt (GGCONR), 1srl* (CHKSRC), 1tar (CHKASPATM), 1tim (LSERABUB), 1tnx* (CHKTNC), 1xso* (XLXSODBG), 2ace (TCACER), and 2crt (NAU42585).

Vertebrate-excluding-human. This group also includes these vertebrate proteins that were excluded from the vertebrate dataset because of similarity to a human protein: 1bpd* (RNU38801), 1cid (RATCD4A), 1epi (MMEGF1), and 2ran (RATLC5).

S. cerevisiae. This group includes: 1apl* (YSCMATA), 1asz (SCAPSG), 1cca* (SCCCP1), 1cpy (YSCPRCCPY), 1csn* (SPU06929), 1d66* (YSCGAL4), 1eag (YSACPA), 1ebg (YSCENOA), 1fcb (SCCYTB2), 1gcb (YSCYCP1X), 1gky (YSCGUKI), 1lbt (CALIPASB), 1oya* (SCOYELE), 1plq* (SCPOL30), 1pxt (SCPOT1), 1pyd (SCPDC1A), 1pyi (SCPPR1), 1qpg (YSCPGK), 1sce* (YSPSUC1), 1sdy (YSCCUZNSD), 1tkb (SCTKL1), 1ukz* (YSCURA6), 1yat* (YSCFKB1), 1ygp* (SCPHOSG), 1ypi (YSCTPI), 1ypp (SCPPAG), 1ysa* (YSCGCN4), 1ytc* (YSCCYC7A), 2aky (YSCADK1), 2csm (YSCARO7A), 2cyp (SCCCP1), 2uce* (SCUBC4), 3pgm (SCGPM), and 3ypi* (YSCTPI).

Plant. This group includes: 1a6o* (ZMACK2), 1abr (APAAC1), 1afr* (RCCSACPD), 1ahc (MCAMC), 1air* (ERWPELC), 1amy (BLYAMY1), 1aok* (S82691), 1aoz (CSASOX), 1apx* (PSAPXI), 1aq0 (BLYGLB2), 1aun (TOBOLP), 1ba7 (SOYCIIPIB), 1bbg (AMBAMBT), 1bgp (BLYPRX5A), 1bhp (AF004018), 1bv1* (BVZ80104), 1bya (GMAMYB), 1cnv (CECONB), 1dhk* (PHVLECT), 1eno* (S60064), 1gnw* (ATHGLUGRFS), 1gox (SPIGLO), 1hpc (PEAGDC), 1hss (AB003682), 1jac (ARPJACD), 1jpc (GAAL2A), 1kbp (PVPAPHOSP), 1mzl (MZEPLTP), 1nar (VNNAN21NB), 1nls (CBLECTIN), 1pag (PAPAP), 1pcl (ERWPEL), 1plc (PNPCAMR), 1ppn (CPAPAP), 1smp* (SMMESM6), 1srd (SPICZD), 1thv (TDATHAU2), 1vok* (ATTFIIDB), 1who* (PPPHLPII), 1yge (SOYLOX), 1yve* (SOAHRI), 2aai* (RCRICIN), 2aak* (ATHUCP1B), 2baa (BLYCHI26A), 2ltn (PSLLECTIN), 2phl (PVPHASBR), 2wbc (S96732), and 9wga (WHTAGGTDD).

References

Adzhubei AA, Adzhubei IA, Krashenninnikov IA, Neidle S (1996) Non-random usage of "degenerate" codons is related to protein three-dimensional structure. FEBS Lett 399:78–82

Agresti A (1990) Categorical data analysis. J. Wiley & Sons, New York

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136:927–935

Bernardi G (1995) The human genome: organization and evolutionary history. Annu Rev Genet 29:445–476

Brunak S, Engelbrecht J (1996) Protein structure and the sequential structure of mRNA: α-helix and β-sheet signals at the nucleotide level. Proteins 25:237–252

Creighton TE (1993) Proteins: structures and molecular properties. WH Freeman & Company, New York

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

Ellis RJ, Hartl FU (1999) Principles of protein folding in the cellular environment. Curr Opin Struct Biol 9:102–110

Gupta SK, Majumdar S, Bhattacharya TK, Ghosh TC (2000) Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem Biophys Res Comm 269:692–696

Hardesty B, Tsalkova T, Kramer G (1999) Co-translational folding. Curr Opin Struct Biol 9:111–114

Harley CB, Pollard JB, Stanners JB, Goldstein S (1981) Model for messenger RNA translation during amino acid starvation applied to the calculation of protein synthetic error rates. J Biol Chem 256:10786–10793

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

Karlin S, Mrázek J (1996) What drives codon choices in human genes? J Mol Biol 262:459–472

Kato M, Nishikawa M, Uritani M, Miyazaki M, Takemura S (1990) The difference in the type of codon-anticodon basepairing at the ribosomal P-site is one of the determinants of the translation rate. J Biochem 107:242–247

Li W (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Miller Jr RG (1981) Simultaneous statistical inference, second edn. Springer-Verlag, New York

Nakamura Y, Gojobori T, Ikemura T (1998) Codon usage tabulated from the international DNA sequence databases. Nucl Acids Res 26:334

Netzer WJ, Hartl FU (1997) Recombination of protein domains facilitated by co-translational folding in eukaryotes. Nature 388:343–349

Netzer WJ, Hartl FU (1998) Protein folding in cytosol: chaperonin-dependent and independent mechanisms. Trends Biochem Sci 23:68–73

Orešič M, Shalloway D (1998) Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol 281:31–48

Parker J (1992) Variations in reading the genetic code. In: Hatfield DL, Lee BJ, Pirtle RM (eds) Transfer RNA in protein synthesis. CRC Press, Boca Raton, pp 191–267

Presta LG, Rose GD (1988) Helix signals in proteins. Science 140:1632–1641

Purvis IJ, Bettany AJE, Santiago TC, Coggins JR, Duncan K, Eason R, Brown AJP (1987) The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*—a hypothesis. J Mol Biol 193:413–417

Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of α helices. Science 240:1648–1652

Sayle RA, Milner-White EJ (1995) RasMol: biomolecular graphics for all. Trends Biochem Sci 20:374–376

Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res 15:1281–1295

Sharp PM, Matassi, G (1994) Codon usage and genome evolution. Curr Opin Genet Devel 4:851–860

Singhal RP, Kopper RA (1981) Effect of guanine to queuine modification on liver aspartate tRNA functions. Fed Proc Am Soc Expt Biol 40:1646

Steinberg S, Misch A, Sprinzl M (1995) Compilation of tRNA sequences and sequences of tRNA genes. Nucl Acids Res 21:3011–3015

Tao X, Dafu D (1998) The relationship between synonymous codon usage and protein structure. FEBS Lett 434:93–96

Thanaraj TA, Argos P (1996a) Ribosome-mediated translational pause and protein domain organization. Protein Sci 5:1594–1612

Thanaraj TA, Argos P (1996b) Protein secondary structural types are differentially coded on messenger RNA. Protein Sci 5:1973–1983

Thomas LK, Dix DB, Thompson RC (1988) Codon choice and gene expression: synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes *in vitro*. Proc Natl Acad Sci USA 85:4242–4246