Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 303 (2003) 177–185
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins

Dawn J. Brooks, Jacques R. Fresco*

Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

Received 20 June 2002; received in revised form 2 October 2002; accepted 20 November 2002
Received by F.G. Alvarez-Valin

Abstract

The possibility that RNY pattern bias in extant sequences is a remnant of more pronounced bias of this type in early ancestors was investigated. To this end, conserved residues (those residues for which the inferred ancestral and known descendant amino acids are identical) and non-conserved residues of ancient proteins dating to the Last Universal Ancestor were identified within six species: two archaea, two eubacteria and two eukaryotes. Bias within sequence elements encoding each subset of residues, conserved and non-conserved, was then determined. In all species, GNN bias is greater within conserved than non-conserved sequence elements, whereas ANN is not. This difference is statistically significant in all six species examined. Since the relative mutability of the GNN-encoded amino acids does not explain the greater bias in conserved sequences, it is concluded that early sequences probably possessed a strong GNN bias. It is suggested that this bias may be a consequence of the GNN codons being the first introduced into the genetic code. Although NNY bias is also greater within conserved sequence elements of the six species, that difference is statistically significant in only half of them. Therefore, the evidence for early NNY bias remains inconclusive. The findings of this study do not support the proposal of Diaz-Lazcoz et al. (J. Mol. Biol. 250 (1995) 123) that the codons of the TCN four-codon block were the first assigned to serine during the evolution of the genetic code.

Keywords: Codon usage bias; Serine codons; Last Universal Ancestor; Genetic code origins; Translational accuracy

1. Introduction

Pattern bias in coding sequences, i.e. a preference for specific nucleotides in distinct codon positions, was first reported by Shepherd (1981, 1983, 1990).[1] He proposed that the RNY pattern bias detected in coding sequences from a variety of sources ranging from viruses to eubacteria to eukaryotes is residual evidence of more extreme bias in the earliest coding sequences. The interpretation Shepherd placed on the data was influenced by the hypothesis that RNY codons would have best accommodated the presumed physicochemical constraints of primitive translation systems (Crick et al., 1976; Eigen and Schuster, 1978). In contrast, Wong and Cedergren (1986) argued that the observation of RNY coding sequence bias across such evolutionarily-divergent organisms is explained by selection. They proposed that RNY bias is a consequence of the adaptation of sequences to encoding protein. Specifically, they proposed that RNN bias reflects the amino acid composition of proteins and NNY bias reflects tRNA abundance.

In the present work, the origin of pattern bias in coding sequences was explored further by comparing bias within coding sequence elements corresponding to conserved residues (those amino acids inferred to be unchanged between the ancestral and descendant sequences) and non-conserved residues of ancient proteins. We reasoned that if pattern bias in contemporary coding sequences, especially in the first codon position, is inherited from early progenitor

Abbreviations: LUA, Last Universal Ancestor; N, any base; R, purine (G or A); Y, pyrimidine (C or T).

* Corresponding author. Tel.: +1-609-258-3927; fax: +1-609-258-2759.
E-mail address: jrfresco@princeton.edu (J.R. Fresco).

[1] Although codons function in the form of RNA sequences, many investigators have discussed them in terms of their DNA counterparts. We follow this practice for consistency with the work of those we cite.

0378-1119/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0378-1119(02)01176-9

sequences, it should be preserved best and therefore be greater in conserved elements of such sequences.

A previous analysis of codon usage in coding sequences corresponding to conserved and non-conserved residues of ancient proteins in three species reported that the TCN four-codon block of serine is systematically and disproportionately favored over the AGY two-codon block in conserved sequence elements (Diaz-Lazcoz et al., 1995). It was consequently proposed that during the establishment of the genetic code the TCN codon block was assigned to serine before the AGY block. We sought corroborative evidence for this proposal, as well as evidence for the possibility that either of the blocks coding for leucine or arginine, the other six-codon amino acids, were earlier additions to the genetic code.

2. Materials and methods

2.1. Choice of protein families and species

Protein families with members in species from the three primary lineages, eubacteria, archaea and eukaryotes, were chosen so that the study would include very ancient proteins, since these were the most likely extant proteins to have retained residual evidence of early sequence bias. Six species, including two eubacteria, *Aquifex aeolicus* and *Escherichia coli*, two archaea, *Aeropyrum pernix* and *Methanobacterium thermoautotrophicum*, and two eukaryotes, *Saccharomyces cerevisiae* and *Drosophila melanogaster*, were included in the analysis. These species include the three best characterized with respect to codon usage, *E. coli*, *S. cerevisiae*, and *D. melanogaster*, and represent a wide range of genomic GC content (Table 1), a factor known to influence codon usage (Ermolaeva, 2001).

In brief, the criteria for inclusion of a protein family in the analysis were that a member of the family be present in all six species and that the family members appear to have been vertically transmitted, rather than laterally transferred, over the course of evolution. For full details of the selection of the set of 59 protein families included in the study, see Brooks and Fresco (2002); the set included in this analysis is identical to the one described therein.

Table 1
GC content and codon counts in conserved and non-conserved sequence elements

Species[a]	GC content[b]	Conserved	Non-conserved
Aae	0.436	5255	14372
ENT	0.507	5195	15106
Ape	0.575	3953	15473
Mth	0.510	3631	14647
Sce	0.397	4123	19978
Dme	0.540	3651	21711

[a] Species abbreviations are: Aae, *A. aeolicus*; ENT, *E. coli*; Ape, *A. pernix*; Mth, *M. thermoautotrophicum*; Sce, *S. cerevisiae*; Dme, *D. melanogaster*.
[b] GC content is that of the entire set of coding sequences in each genome.

2.2. Identification of conserved residues within proteins

In order to identify conserved residues, maximum parsimony (MP) (Eck and Dayhoff, 1966) was used to partially reconstruct the ancestral protein sequences in the Last Universal Ancestor (LUA) that gave rise to each family of aligned descendants. The protein parsimony software 'protpars' included in the PHYLIP phylogenetic package (Felsenstein, 1993) was used to partially reconstruct ancestral sequences, assuming the phylogenetic tree indicated by small subunit rRNA data (Olsen et al., 1994). Because these ancient sequences have diverged to a great extent, only slightly more than a third (~37%) of the sites within the ancestral sequence could be reconstructed. Residues identical between the inferred ancestor and its descendant sequences were defined as conserved. At sequence positions for which no ancestral residue could be assigned, it was assumed that residues within none of the descendant sequences were conserved.

We wish to emphasize that MP was used in the present study to infer *partial* ancestral sequences, not complete ones; sites where the identity of the ancestral residue was ambiguous were left unassigned. Consequently, an ancestral residue was assigned only at sites for which a fairly strong inference regarding the identity of the residue (through MP) was possible. Using simulated sequence evolution, we found that the accuracy of assigned residues in the partially inferred ancestral sequences is ~80%. Thus, some conserved amino acids will be incorrectly identified as non-conserved, and vice versa. However, although error in inferring conserved and non-conserved residues will introduce noise into the data (i.e. reduce the degree of difference in bias we might hope to observe between the two classes of residue), it does not invalidate our findings.

2.3. Determination of codon usage and bias in conserved and non-conserved sequence elements

The coding sequence corresponding to each protein was collected from GenBank (Benson et al., 2002). Sequence elements corresponding to either conserved or non-conserved residues within the encoded proteins were then identified. For convenience, these are referred to as conserved and non-conserved sequence elements. However, it is important to note that it is the corresponding protein residue and not the codon itself that is defined as conserved or non-conserved; the codon may or may not be conserved at a sequence position in which the protein residue is conserved. Codon usage in conserved and non-conserved sequence elements was tabulated separately for each species. The number of codons in the two categories in each species is shown in Table 1. The frequencies of codons

with R, Y, G, and A in the first codon position, and Y and R in the third codon position, were then calculated. Separately, the frequencies of codons were normalized such that the frequencies of all synonymous codons associated with a particular amino acid would sum to one, allowing comparison of the relative usage of the different codons in conserved and non-conserved sequence elements.

2.4. Determination of statistical significance of data

Chi-squared tests for significance were performed using 2×2 contingency tables of frequencies of alternative sets of codons (i.e. GNN vs. non-GNN, NNY vs. NNR) within conserved and non-conserved sequence elements. To be considered statistically significant, the P value for any one statistical test was required to be <0.05.

3. Results

3.1. GNN, but not ANN, bias is greater within conserved sequence elements

RNN bias is greater within conserved than non-conserved sequence elements in all of the species but *E. coli* (Fig. 1A). However, whereas GNN bias is greater in conserved than in non-conserved sequence elements in all six species (Fig. 1B), ANN bias is consistently lower within conserved elements (Fig. 1C). The average frequency of GNN codons in conserved sequence elements is 43%, compared to 35% in non-conserved elements. The greater GNN bias in conserved sequence elements is statistically significant in all six species examined.

3.2. NNY usage is not significantly greater within conserved sequence elements

In contrast to early reports that NNY bias is a universal trait of coding sequences (Shepherd, 1981, 1983, 1990), we have found that NNR codons are preferred in both conserved and non-conserved sequence elements of three of the species studied, *A. aeolicus*, *A. pernix* and *M. thermoautotrophicum* (Fig. 1D). Nonetheless, the frequency of NNY codons is greater within conserved than non-conserved sequence elements in all six species (Fig. 1D). However, that difference is statistically significant in only half the species, i.e. *E. coli*, *A. pernix* and *M. thermoautotrophicum*.

3.3. Amino acid frequencies in conserved and non-conserved residues

Greater GNN bias in coding sequences corresponds to a greater frequency of the GNN-encoded amino acids as a set in conserved positions of the encoded proteins. It was therefore of interest to determine more specifically how

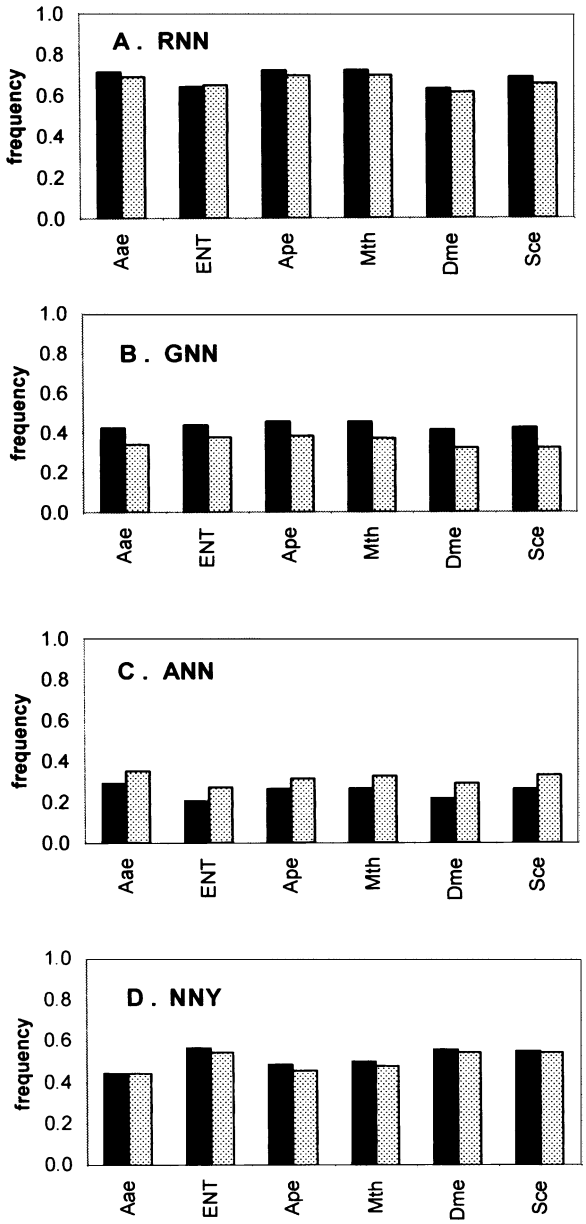

Fig. 1. Codon frequencies in conserved (black bars) and non-conserved (gray bars) sequence elements. Species abbreviations are as given in Table 1.

each of the amino acids in the set, glycine, alanine, valine, and aspartic and glutamic acids, differs in frequency between conserved and non-conserved residues. Glycine and valine were found to occur with greater frequency in conserved than in non-conserved residues in all six species, aspartic acid in four species, and alanine and glutamic acid in half the species (Table 2). On average, the frequencies of all five amino acids encoded by GNN are greater in conserved residues (Table 2, final column). Most notably, the average frequency of glycine within conserved sequence residues is more than double that within non-conserved residues.

Table 2
Ratio of amino acid frequencies in conserved and non-conserved residues

	Aae[a]	ENT	Ape	Mth	Dme	Sce	Average[b]
Val	1.12	1.19	1.03	1.25	1.32	1.32	1.21
Ala	1.31	0.88	0.97	1.18	0.94	1.06	1.06
Asp	1.27	1.01	1.18	0.89	1.14	0.97	1.08
Glu	0.88	1.01	0.99	0.97	1.15	1.15	1.03
Gly	2.12	1.89	1.96	2.00	1.99	2.17	2.02

[a] Species abbreviations are as given in Table 1.
[b] The final column gives the average over all six species.

3.4. Observed vs. predicted pattern bias in the third codon position

The greater frequency of NNY codons in conserved sequence elements could be a consequence of either greater codon usage bias, that is, greater bias in synonymous codon usage, or differences in amino acid composition, i.e. greater frequency of NNY-encoded amino acids, or both. To distinguish between these possibilities, the NNY usage predicted for a set of coding sequences in the absence of codon usage bias was calculated given the amino acid composition of the proteins they encode: a difference between NNY usage thus predicted and that observed is due to codon usage bias.

No species is predicted to display NNY bias in non-conserved sequence elements in the absence of codon usage bias (Table 3). However, *E. coli*, *D. melanogaster* and *S. cerevisiae* all display NNY bias in these elements. In addition, the observed frequency of NNY codons in non-conserved sequence elements is slightly higher than predicted in *M. thermoautotrophicum* (Table 3; ratio of Obs/Pred >1.0).

Conserved sequence elements of all species are also predicted to display NNR rather than NNY bias in the absence of codon usage bias. *Escherichia coli*, *D. melanogaster* and *S. cerevisiae* nevertheless display NNY bias in these sequence elements. Furthermore, the observed NNY bias is greater than predicted in conserved sequence elements in all species but *A. aeolicus*. Significantly, the discrepancy between observed and predicted NNY codon frequency is greater in conserved than in non-conserved sequence elements in all species but *A. aeolicus* (Table 3).

Table 3
Observed vs. predicted NNY codon frequencies

Species[a]	Non-conserved			Conserved		
	Pred[b]	Obs[c]	Obs/Pred	Pred	Obs	Obs/Pred
Aae	0.4459	0.4376	0.9814	0.4595	0.4441	0.9665
ENT	0.4701	0.5438	1.1568	0.4713	0.5675	1.2042
Ape	0.4569	0.4557	0.9974	0.4676	0.4849	1.0369
Mth	0.4716	0.4760	1.0093	0.4647	0.4968	1.0692
Dme	0.4877	0.5416	1.1106	0.4819	0.5564	1.1546
Sce	0.4698	0.5469	1.1151	0.4904	0.5492	1.1690

[a] Abbreviations are as given in Table 1.
[b] Predicted frequency based on the amino acid composition of the encoded protein, assuming no codon bias (i.e. equal usage of synonymous codons for a given amino acid).
[c] Observed frequency.

3.5. Changes in synonymous codon usage in conserved sequence elements show no consistent trend

We sought to determine whether the change in codon usage bias responsible for the greater frequency of NNY codons within conserved sequence elements could be explained by some consistent trend in the choice of synonymous codons within these elements. Specifically, it seemed possible that the same codons in all species, or the codons most preferred in genome-wide coding sequences of each species, would display the largest positive difference in frequency between conserved and non-conserved sequence elements. However, no such trend was observed. In fact, different synonymous codons show the greatest positive difference in frequency in different species (Table 4). Furthermore, there is a lack of correspondence between those synonymous codons that are favored in whole genome coding sequences and those that occur with greater relative frequency in conserved sequence elements. Therefore, greater NNY pattern bias within conserved sequence elements is due neither to greater usage of a uniform set of codons across species nor to greater usage of the synonymous codons preferred within genomic coding sequences.

3.6. Usage of two- and four-codon blocks for amino acids with six codons

None of the three amino acids with six codons, leucine, arginine or serine, displays a consistent preferential usage of either its two- or its four-codon block in conserved sequence elements of all six species (Fig. 2). For leucine, the TTR two-codon block is used preferentially in conserved sequence elements in *E. coli* and *D. melanogaster*, whereas the CTN four-codon block is used preferentially in the remaining four species (Fig. 2A). Interestingly, the same two species use the arginine CGN four-codon block preferentially in conserved elements, whereas the remaining species use the AGR block preferentially in these elements (Fig. 2B). In the case of serine, the AGY two-codon block is used preferentially in conserved sequence elements of all species but *A. pernix* (Fig. 2C).

4. Discussion

4.1. Early coding sequences may have been composed entirely of repeating GNN codons

The present analysis shows that GNN but not ANN bias is significantly greater in sequence elements encoding

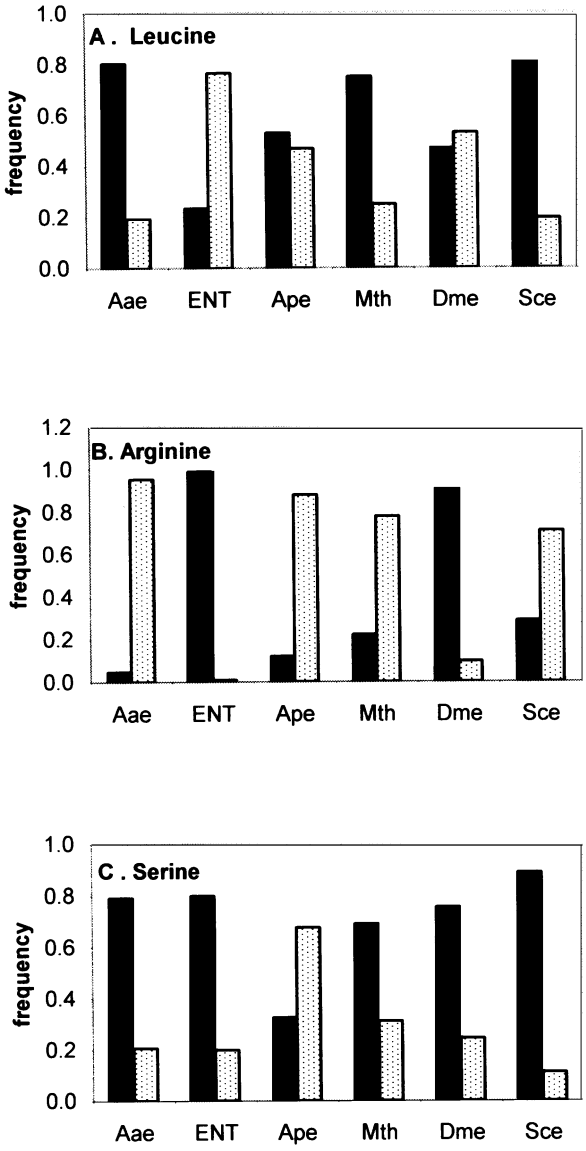

Fig. 2. Normalized frequency of four-codon (black bars) and two-codon (gray bars) blocks in conserved sequence elements. Species abbreviations are as given in Table 1.

conserved residues of ancient proteins. The GNN-encoded amino acids glycine and valine occur with greater frequency within conserved than non-conserved positions of the proteins themselves within all six species, and on average, the frequency of each of the other GNN-encoded amino acids is also greater in conserved positions. There are two conceivable explanations for this (Brooks and Fresco, 2002). Since conserved residues of contemporary proteins more closely reflect their ancestral sequences than do non-conserved residues, their composition might be enriched in those amino acids that were used more abundantly in the past. Thus, the GNN-encoded amino acids may have been more abundant within the ancestral proteins in the LUA than within modern-day proteins to which they gave rise. Alternatively, the GNN-encoded amino acids might be more likely than other amino acids to be conserved, and therefore they have become relatively enriched in conserved residues.

If one considers the empirically-determined relative mutability of the various amino acids (Dayhoff et al., 1978; Jones et al., 1992), it is clear that with the exception of glycine, the GNN-encoded amino acids are if anything less likely than others to be conserved. Although glycine does display a relatively low mutability, it is not so low as to explain the two-fold greater frequency within conserved sequence elements. Therefore, the most reasonable explanation for the greater frequency of the GNN-encoded amino acids in conserved sequence elements of ancient proteins is that these amino acids were more abundant within proteins of the LUA than those of modern species. This inference is supported elsewhere (Brooks et al., 2002).

Given that GNN-encoded amino acids have decreased in frequency between the LUA and today, it is possible that they have been decreasing in frequency since proteins first arose (Brooks et al., 2002). In fact, these amino acids may have been the first introduced into the genetic code, and the earliest coding sequences may have consisted entirely of GNN codons. Although the extreme GNN bias of the earliest coding sequences would have begun to erode as new codons and their assigned amino acids entered the genetic code, negative selection presumably prevented the rapid degradation of the early-established bias. Consequently, remnants of GNN bias could still be observed in the coding sequences of the LUA.

Ideas regarding the origin of the genetic code based entirely on theoretical considerations are consistent with the proposal that the GNN-encoded amino acids were the earliest additions (see Trifonov, 2000 for a review). Taking into account presumed restrictions on translation and replication of the earliest sequences in the prebiotic environment, Eigen and Schuster (1978) proposed that the first two sense codons were GGC and GCC, followed by GAC and GUC. It is reasonable to think that those amino acids believed to have been among the most abundant under prebiotic conditions, glycine, alanine, aspartate and valine (Miller, 1953, 1987), would have adopted the most abundant codons present in the earliest messages (Eigen and Schuster, 1978).

4.2. Cause of greater frequency of NNY codons in conserved sequence elements

In previous discussions of the cause of nucleotide bias in the third codon position, it was assumed that coding sequences universally display NNY bias (Shepherd, 1981; Wong and Cedergren, 1986; Jukes, 1996). However, we have found that three species, *A. aeolicus*, *A. pernix* and *M. thermoautotrophicum*, in fact display NNR bias (Table 3). Interestingly, both conserved and non-conserved sequence elements within all six species examined are predicted to display an NNR bias in the absence of codon usage bias

Table 4
Comparison of normalized codon frequencies[a] in conserved and non-conserved sequence elements and in whole genomes[b,c]

aa	Codon	Aae[d]		ENT		Dme		Sce		Ape		Mth	
		Cons[e]/non	Genome[f]	Cons/non	Genome	Cons/non	Genome	Cons/non	Genome	Cons/non	Genome	Cons/non	Genome
F	uuu	0.909	0.563	0.779	0.572	0.893	0.340	0.720	0.588	0.832	0.223	0.760	0.285
F	uuc	**1.095**	0.437	**1.149**	0.428	**1.048**	0.660	**1.241**	0.412	**1.025**	0.777	**1.080**	0.715
L	uua	0.894	0.167	0.935	0.135	1.109	0.047	1.045	0.278	0.188	0.042	0.767	0.044
L	uug	0.817	0.077	0.715	0.127	0.851	0.170	1.010	0.286	**1.168**	0.064	0.000	0.026
L	cuu	**1.080**	0.255	0.564	0.109	**1.161**	0.094	0.837	0.129	1.021	0.148	0.920	0.263
L	cuc	1.038	0.292	0.982	0.100	1.113	0.160	**1.178**	0.057	1.104	0.339	**1.188**	0.371
L	cua	0.842	0.073	**2.317**	0.039	0.690	0.084	0.964	0.141	0.822	0.159	0.820	0.049
L	cug	1.007	0.137	1.074	0.490	1.029	0.445	0.953	0.110	1.029	0.247	0.950	0.247
I	auu	0.849	0.238	0.952	0.501	0.940	0.324	0.928	0.464	1.015	0.154	0.924	0.154
I	auc	0.961	0.134	1.016	0.406	**1.099**	0.503	**1.116**	0.263	0.606	0.239	0.883	0.288
I	aua	**1.055**	0.628	**2.022**	0.093	0.819	0.173	0.931	0.273	**1.092**	0.606	**1.073**	0.558
M	aug	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
V	guu	0.986	0.380	1.084	0.277	0.715	0.215	**1.046**	0.392	0.998	0.241	1.003	0.306
V	guc	1.040	0.106	0.845	0.206	0.924	0.095	1.026	0.207	1.033	0.259	**1.103**	0.266
V	gua	0.880	0.321	**1.176**	0.161	0.831	0.120	0.854	0.210	0.846	0.167	0.957	0.132
V	gug	**1.261**	0.193	0.850	0.356	**1.168**	0.570	0.911	0.191	**1.038**	0.333	0.903	0.295
S	ucu	1.312	0.185	0.963	0.164	0.966	0.078	1.272	0.265	1.163	0.121	0.414	0.088
S	ucc	1.156	0.267	**1.546**	0.151	**1.254**	0.244	1.020	0.159	0.914	0.207	**1.331**	0.219
S	uca	0.856	0.148	1.241	0.135	0.867	0.088	0.782	0.211	0.385	0.093	0.979	0.341
S	ucg	**1.723**	0.073	1.187	0.143	1.126	0.211	**1.386**	0.097	0.510	0.144	1.211	0.045
S	agu	0.437	0.163	0.804	0.153	0.657	0.130	0.754	0.159	0.506	0.089	1.064	0.124
S	agc	0.762	0.165	0.530	0.254	0.820	0.249	0.219	0.109	**1.294**	0.344	0.882	0.184
P	ccu	0.833	0.267	**1.132**	0.168	0.739	0.120	0.843	0.310	0.876	0.279	0.739	0.213
P	ccc	**1.128**	0.422	1.088	0.119	0.918	0.339	1.010	0.155	**1.214**	0.375	1.140	0.371
P	cca	0.958	0.138	0.810	0.200	**1.191**	0.237	1.065	0.415	0.928	0.142	0.964	0.269
P	ccg	0.865	0.173	1.018	0.513	1.144	0.304	**1.083**	0.121	0.795	0.204	**1.183**	0.146
T	acu	0.644	0.232	0.938	0.182	0.613	0.160	0.983	0.345	**1.231**	0.191	0.589	0.093
T	acu	**1.208**	0.270	**1.126**	0.416	**1.346**	0.398	**1.346**	0.215	0.918	0.345	1.036	0.387
T	aca	0.848	0.208	0.622	0.150	0.695	0.188	0.631	0.303	0.838	0.227	**1.087**	0.426
T	acg	1.179	0.291	0.828	0.252	0.934	0.254	0.953	0.137	1.116	0.236	0.678	0.093
A	gcu	0.922	0.270	0.987	0.179	0.973	0.187	1.071	0.377	1.022	0.261	0.676	0.148
A	gcc	0.785	0.212	**1.097**	0.260	**1.188**	0.463	**1.078**	0.225	**1.066**	0.389	1.006	0.367
A	gca	0.920	0.287	0.993	0.223	0.862	0.165	0.807	0.289	0.902	0.143	1.068	0.395
A	gcg	**1.395**	0.232	0.970	0.338	0.573	0.185	0.772	0.109	0.915	0.207	**1.322**	0.090

Y	uau	**1.079**	0.186	**1.123**	0.579	0.883	0.345	0.905	0.561	0.892	0.379	**1.077**	0.298
Y	uac	0.986	0.814	0.911	0.421	**1.055**	0.655	**1.064**	0.439	**1.031**	0.621	0.969	0.702
H	cau	0.595	0.169	0.684	0.568	1.000	0.372	0.933	0.637	0.648	0.319	0.778	0.420
H	cac	**1.053**	0.831	**1.170**	0.432	1.000	0.628	**1.079**	0.363	**1.128**	0.681	**1.118**	0.580
Q	caa	**1.159**	0.358	0.695	0.339	0.538	0.284	0.948	0.693	0.000	0.163	0.000	0.068
Q	cag	0.926	0.642	**1.091**	0.661	**1.171**	0.716	**1.228**	0.307	**1.074**	0.837	**1.033**	0.932
N	aau	0.860	0.303	0.608	0.468	0.857	0.414	0.863	0.591	**1.099**	0.222	0.880	0.323
N	aac	**1.032**	0.697	**1.098**	0.532	**1.091**	0.586	**1.114**	0.409	0.985	0.778	**1.048**	0.677
K	aaa	**1.027**	0.481	**1.011**	0.752	0.754	0.272	0.918	0.578	0.920	0.220	0.851	0.382
K	aag	0.982	0.519	0.966	0.248	**1.074**	0.728	**1.060**	0.422	**1.012**	0.780	**1.077**	0.618
D	gau	0.722	0.377	**1.044**	0.626	**1.037**	0.502	0.999	0.651	0.856	0.326	0.863	0.513
D	gac	**1.126**	0.623	0.962	0.374	0.959	0.498	**1.001**	0.349	**1.042**	0.674	**1.119**	0.487
E	gaa	0.965	0.652	**1.023**	0.686	**1.203**	0.300	**1.038**	0.706	0.993	0.175	0.951	0.397
E	gag	**1.070**	0.348	0.944	0.314	0.916	0.700	0.861	0.294	**1.001**	0.825	**1.031**	0.603
C	ugu	1.000	0.494	0.862	0.451	**1.160**	0.261	**1.008**	0.630	0.993	0.344	**1.201**	0.381
C	ugc	1.000	0.506	**1.072**	0.549	0.943	0.739	0.973	0.370	**1.002**	0.656	0.904	0.619
W	ugg	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
R	cgu	0.526	0.031	1.035	0.377	**1.399**	0.163	1.075	0.146	**1.379**	0.049	**1.335**	0.069
R	cgc	**1.383**	0.024	0.985	0.377	0.947	0.352	1.044	0.059	1.126	0.088	1.298	0.055
R	cga	0.364	0.012	0.490	0.064	1.159	0.142	0.487	0.068	0.750	0.034	0.432	0.016
R	cgg	0.000	0.016	0.653	0.099	0.923	0.148	**1.740**	0.038	0.642	0.091	1.133	0.069
R	aga	0.954	0.389	0.881	0.051	0.671	0.088	1.013	0.480	0.818	0.146	0.974	0.184
R	agg	1.071	0.527	**1.469**	0.031	0.498	0.107	0.801	0.209	1.032	0.592	0.953	0.607
G	ggu	**1.239**	0.231	**1.151**	0.350	**1.165**	0.211	**1.172**	0.473	0.804	0.175	**1.118**	0.314
G	ggc	0.663	0.121	0.873	0.384	0.943	0.447	0.633	0.192	1.011	0.383	1.087	0.208
G	gga	1.045	0.500	0.760	0.118	0.927	0.274	0.603	0.216	1.078	0.141	0.927	0.261
G	ggg	0.728	0.148	0.917	0.149	1.070	0.069	0.817	0.119	**1.088**	0.302	0.799	0.217

[a] Codon frequencies in conserved and non-conserved sequence elements were normalized such that the frequencies of all synonymous codons encoding one amino acid sum to one.
[b] The frequency for the synonymous codon with the greatest enhancement in relative frequency within conserved sequence elements is highlighted in bold.
[c] The frequencies for the synonymous codons favored in whole genome sequences are underlined.
[d] Species abbreviations are as given in Table 1.
[e] Cons, normalized frequency in conserved sequence elements; non, normalized frequency in non-conserved sequence elements.
[f] Normalized frequency of each codon in whole-genome coding sequences.

(Table 3). This is due merely to the amino acid composition of the encoded proteins. Codon usage bias is responsible for NNY usage being greater than predicted within the non-conserved sequence elements of four of the species, and the conserved sequence elements of five of the species (Table 3; see the ratio of observed to predicted usage for each category). The discrepancy between the observed and predicted frequency of NNY codons is greater within conserved than non-conserved sequence elements of all species but *A. aeolicus*, indicating that the codon usage bias responsible for NNY bias is enhanced in conserved regions of these species.

The greater codon usage bias observed in conserved sequence elements of *Drosophila* has been ascribed to selection for increased translational accuracy (Akashi, 1994). However, it has been reported that codon bias is not influenced by selection for translational accuracy in either *E. coli* (Hartl et al., 1994) or *S. cerevisiae* (Percudani and Ottonello, 1999). Moreover, the report correlating selection of particular synonymous codons in *Drosophila* with the accuracy with which they are translated does not establish a causal link. Indeed, it is unknown which synonymous codons are most accurately translated in *Drosophila*. Even in the best characterized species, *E. coli*, there is a lack of systematic information regarding the relative translational accuracy of synonymous codons (Ulrich et al., 1991). Therefore, we see no basis for ascribing to selection for translational accuracy the greater frequency of NNY codons in conserved sequence elements of the six species examined here.

Diaz-Lazcoz et al. (1995) have reported greater codon bias in conserved sequence elements based on a comparison of codon usage in conserved and non-conserved sequence elements in 105 proteins common to *E. coli*, *Bacillus subtilis*, and *S. cerevisiae*. (It is worth noting that no strategy to eliminate horizontally-transferred genes from the set was described, although this limitation does not of itself invalidate their findings.) Only one amino acid, serine, displayed a statistically significant greater usage of the genome-wide preferred synonymous codon in conserved sequence elements of *E. coli* and *B. subtilis*. Nevertheless, these investigators reported that greater codon usage bias is a 'general feature' of conserved residues. In contrast to Akashi (1994), their explanation for greater bias in conserved sequence elements is that such positions have had longer to adapt to genome-wide codon usage than positions that have changed amino acid identity more recently. However, we found there to be a general lack of correspondence between those codons that are preferred in the whole genome and those that show the greatest positive difference in frequency in conserved sequence elements (Table 3). Therefore, the proposal that conserved positions show greater codon usage bias because they are the best adapted to genome-wide codon preferences is unsubstantiated.

The possibility that coding sequences historically displayed NNY bias is supported by the observation that the frequency of NNY codons is greater in conserved sequence elements even in species in which NNR codons are preferred. In those species that display NNR bias in their non-conserved sequence elements, the conserved elements may be lagging in their change in codon usage bias from a historical preference for NNY codons. In support of this idea, codons in conserved sequence elements have been shown to be substituted less frequently than those in non-conserved elements (Kafatos et al., 1977). However, codon usage is clearly influenced by complex factors, and it would be premature, based on our data, to infer that coding sequences have historically displayed an NNY bias.

4.3. Codon usage bias among amino acids with six codons

Codons within the TCN four-codon and AGY two-codon blocks of serine are separated by a minimum of two point mutations. Thus, unless a fortuitous double mutation occurs, interchanges between codons of either block require an intermediate change to a codon for a different amino acid (the most direct route being through codons for threonine or cysteine). Amino acid substitutions presumably have been accepted infrequently at conserved sites of proteins. Consequently, Diaz-Lazcoz et al. (1995) assumed that the ancestral usage of codons from either the TCN or AGY block would be well preserved at such sites. Accordingly, these investigators determined the usage of the two- and the four-codon blocks of serine in sequence elements encoding conserved positions of proteins presumed to date to the LUA.

The TCN codon block was preferred ($>72\%$) within these ancient sequence elements in all three species (Diaz-Lazcoz et al., 1995). Furthermore, the frequency of TCN codons in conserved elements was greater than within non-conserved sequence elements. These investigators interpreted their findings as evidence that the TCN codon block was favored in the ancestral sequence. They proposed that this preferential usage of the four-codon block was a consequence of this block being the first assigned to serine. Our findings are not consistent with this proposal, since one of the six species included in our study, *A. pernix*, displays a strong preference for the AGY block in conserved elements (Fig. 2C). (This species displays a less pronounced, yet clear preference for the AGY block in non-conserved sequence elements as well (data not shown).) In light of this evidence, the suggestion that the TCN codon block was assigned to serine before the AGY block during the establishment of the genetic code should be reassessed.

Acknowledgements

We wish to thank David Halitsky for suggesting that we examine pattern bias in sequences encoding ancient

conserved regions of proteins, and Ya'ir Aizenman (funded by NSF PECASE Grant MCB-0093399 to Mona Singh) for writing the program to identify conserved and non-conserved codons. D.J.B. was supported by a predoctoral traineeship from NIH grant 2T32GM07388-22 and then a traineeship from NSF grant DGE 9972930. The computational facility utilized for this work was obtained with funds provided to J.R.F. by the Department of Defense through MEDCOM at Fort Detrick, MD.

References

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

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L., 2002. GenBank. Nucleic Acids Res. 30, 17–20.

Brooks, D.J., Fresco, J.R., 2002. Increased frequency of cysteine, tyrosine and phenylalanine residues since the Last Universal Ancestor. Mol. Cell. Proteomics 1, 125–131.

Brooks, D.J., Fresco, J.R., Lesk, A.M., Singh, M., 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evol. 19, 1645–1655.

Crick, F.H.C., Brenner, S., Klug, A., Pieczenick, G., 1976. A speculation on the origin of protein synthesis. Orig. Life 7, 389–397.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. In: Dayhoff, M.O., (Ed.), Atlas of Protein Sequence and Structure, 5. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Diaz-Lazcoz, Y., Hénaut, A., Vigier, P., Risler, J.-L., 1995. Differential codon usage for conserved amino acids: evidence that the serine codons TCN were primordial. J. Mol. Biol. 250, 123–127.

Eck, R.V., Dayhoff, M.O., 1966. Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Silver Spring, MD.

Eigen, M., Schuster, P., 1978. The hypercycle: a principle of natural self-organization. Naturwissenschaften 65, 341–368.

Ermolaeva, M.D., 2001. Synonymous codon usage in bacteria. Issues Mol. Biol. 3, 91–97.

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package) version 35c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA.

Hartl, D.L., Moriyama, E.N., Sawyer, S.A., 1994. Selection intensity for codon bias. Genetics 138, 227–234.

Jones, D.R., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

Jukes, T.H., 1996. On the prevalence of certain codons ("RNY") in genes for proteins. J. Mol. Evol. 42, 377–381.

Kafatos, F.C., Efstratiadis, A., Forget, B.G., Weissman, S.M., 1977. Molecular evolution of human and rabbit beta-globin mRNAs. Proc. Natl. Acad. Sci. USA 74, 5618–5622.

Miller, S.L., 1953. Production of amino acids under possible primitive earth conditions. Science 117, 528–529.

Miller, S.L., 1987. Which organic compounds could have occurred on the prebiotic earth? Cold Spring Harbor Symp. Quant. Biol. 52, 17–27.

Olsen, G.J., Woese, C.R., Overbeek, R., 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. 176, 1–6.

Percudani, R., Ottonello, S., 1999. Selection at the wobble position of codons read by the same tRNA in Saccharomyces cerevisiae. Mol. Biol. Evol. 16, 1752–1762.

Shepherd, J.C.W., 1981. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J. Mol. Evol. 17, 94–102.

Shepherd, J.C., 1983. From primeval message to present-day gene. Cold Spring Harbor Symp. Quant. Biol. 47, 1099–1108.

Shepherd, J.C., 1990. Ancient patterns in nucleic acid sequences. Methods Enzymol. 183, 180–192.

Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene 261, 139–151.

Ulrich, A.K., Li, L.-Y., Parker, J., 1991. Codon usage, transfer RNA availability and mistranslation in amino acids starved bacteria. Biochim. Biophys. Acta 1089, 362–366.

Wong, J.T., Cedergren, R., 1986. Natural selection versus primitive gene structure as determinant of codon usage. Eur. J. Biochem. 159, 175–180.