J Mol Evol (2005) 61:597–607
DOI: 10.1007/s00239-004-0314-2

JOURNAL OF MOLECULAR EVOLUTION

Error Minimization and Coding Triplet/Binding Site Associations Are Independent Features of the Canonical Genetic Code

J. Gregory Caporaso,[1] Michael Yarus,[2] Rob Knight[3]

[1] Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, P.O. Box 6511, Mail Stop 8303, Aurora, CO 80045, USA
[2] Department of Molecular, Cellular and Developmental Biology, University of Colorado at Boulder, Campus Box 347, Boulder, CO 80309, USA
[3] Department of Chemistry and Biochemistry, University of Colorado at Boulder, Campus Box 215, Boulder, CO 80309, USA

Received: 4 November 2004 / Accepted: 5 May 2005 [*Reviewing Editor:* Dr. Stephen Freeland]

Abstract. The canonical genetic code has been reported both to be error minimizing and to show stereochemical associations between coding triplets and binding sites. In order to test whether these two properties are unexpectedly overlapping, we generated 200,000 randomized genetic codes using each of five randomization schemes, with and without randomization of stop codons. Comparison of the code error (difference in polar requirement for single-nucleotide codon interchanges) with the coding triplet concentrations in RNA binding sites for eight amino acids shows that these properties are independent and uncorrelated. Thus, one is not the result of the other, and error minimization and triplet associations probably arose independently during the history of the genetic code. We explicitly show that prior fixation of a stereochemical core is consistent with an effective later minimization of error.

Key words: Genetic code — Stereochemistry — Adaptation — Error minimization — Aptamers — SELEX

Correspondence to: Rob Knight; *email:* Rob@spot.Colorado.edu

Introduction

The canonical genetic code differs from random variants in two well-established ways. First, the pattern of codon assignments markedly reduces differences in hydrophobicity between amino acids whose codons differ by a single nucleotide and, thus, can potentially be interchanged by point misreading (Freeland and Hurst 1998). This pattern of codon assignments decreases the probability that a point mutation or translation error will inactivate the encoded protein and is, thus, considered error-minimizing (Haig and Hurst 1991). Such error-minimizing properties of the genetic code have been confirmed independently by many investigators (Haig and Hurst 1991; Freeland and Hurst 1998; Ardell 1998; Freeland et al. 2000; Zhu et al. 2003), suggesting that natural selection acted on the genetic code to reduce the effects of errors. Second, coding triplets (codons or anticodons) for an amino acid appear far more often than chance would predict at the active sites of RNA aptamers selected from random-sequence pools for their ability to bind that amino acid (Knight and Landweber 1998; Yarus 1998). These significant associations between coding triplets and amino acid binding sites, which we have recently reviewed (Yarus et al. 2005; Knight et al. 2003), suggest that stereochemical factors have influenced codon assignments. Several proposals for how information might be transmitted from primordial triplet/binding site interactions through either the codons or the anticodons by a number of mechanisms, including coding coenzyme handles (Szathmary 1993, 1999) and direct template recognition (Yarus 1998), have been suggested. Although selection for sequences with affinity for a ligand need not necessarily increase the chances of being able to catalyze reactions involving that ligand (Levy and Ellington 2001), the mounting evidence for statistical associations between triplets and binding sites (Knight and Landweber 2000; Yarus et al. 2005) requires some kind of explanation. It is possible that different amino acids entered the code via different mechanisms, which might be revealed according to whether they show associations with their cognate codons, anticodons, both, or neither (Szathmary 1999). A third theory, that the genetic code is structured according to relationships between amino acids within biosynthetic pathways due to its expansion from a simpler form (Wong 1975), has considerable intuitive appeal (Taylor and Coates 1989; Miseta 1989; Di Giulio 1991, 1998, 1999), although statistical support for particular suggestions about biosynthetic relationships within the code has been highly controversial (Amirnovin 1997; Di Giulio and Medugno 2000; Ronneberg et al. 2000; Di Giulio 2001; Yarus et al. 2005). Thus, although it may be possible to incorporate ideas about biosynthesis in future analyses, in the present work we focus on the adaptive and stereochemical hypotheses.

Both the error-minimizing (Sonneborn 1965; Alff-Steinberger 1969) and the stereochemical (Woese 1965; Yarus and Christian 1989) theories of the evolution of the genetic code have extensive histories, and predict that chemically similar amino acids will be assigned to chemically similar codons (as do essentially all other theories of the code's origin and evolution [Knight et al. 1999]). However, if the same features of amino acids were important both for function within proteins and for RNA–protein interactions, a random genetic code that was optimized for error minimization might automatically display relationships between coding triplets and amino acid binding sites within aptamers (and vice versa). In other words, the genetic code may appear optimal in one of these two respects as a side effect of being optimal in the other. A correlation between error-minimizing and triplet-association qualities would suggest that some chemical basis links these features. Because the link could be as simple as a mutual dependence on hydrophobicity, it is worthwhile to investigate a link between these features of the code.

To test whether error-minimizing codes necessarily produce associations between coding triplets and binding sites, we generated random genetic codes and scored them on both measures. As a control, we tested the extent to which codes that score highly on either measure do so because they share codon assignments with the canonical genetic code. Our set of sequences consists of previously published sequences (Knight et al. 2003), plus sequences subsequently isolated or currently in press from researchers in the Yarus lab.

This technique compares the canonical code to a distribution of possible codes. Any particular randomization method can be thought of as a hypothesis about which alternative codes could have been tested during the evolution of the canonical code. Since the mechanism by which the canonical genetic code evolved is uncertain, we used five different randomization algorithms, each with the ability either to preserve or to randomize the positions of stop codons, for a total of 10 distinct methods of randomizing the codes (and a total of 10^6 codes). These diverse randomization algorithms allow us to largely avoid making implicit assumptions about which properties of the canonical genetic code are fixed by the mechanism of its evolution and which are flexible. In addition, we can use numerical optimization techniques to evaluate whether a genetic code with a certain number of codon assignments fixed (for example, by stereochemistry) can still be optimized by evolution for error minimization, allowing the two processes to act separately in the genetic code's evolution. Consequently, we can test whether error minimization and triplet/binding site associations are independent features of the canonical genetic code.

Methods

In order to study the relationship between random genetic codes that minimize error and random genetic codes that contain coding triplet associations, we developed software that generates random genetic codes based on several randomization algorithms. Each random code is then scored on measures of both error minimization and coding triplet (codon or anticodon) association with amino acid binding sites within known RNA aptamers. The Pearson correlation coefficient is then used to test whether the same codes score well on each measure.

Random Genetic Code Generation

To account for uncertainty about the constraints on the form of the genetic code during evolution, we generated 200,000 random codes based on each of five different algorithms, each creating a different space of possible codes. We discuss each of our randomization algorithms in order, from the smallest to the largest set of possible codes.

Quartet Shuffler. In the Quartet Shuffler we assume that the arrangement of amino acids encoded in a quartet block is fixed among all possible genetic codes. We define a quartet block as an XYN block. During randomization, the intraquartet organization of the amino acids is held constant, while the codon assignments for each quartet are randomized. Under this mechanism of randomization, there are 16! (or 2.09×10^{13}) possible genetic codes. The Quartet Shuffler preserves the number of codons per amino acid and the quartet blocks of the canonical genetic code.

N-Block Shuffler. In the N-Block Shuffler the N-Blocks of the genetic code are held constant. We define an N-Block as a 1-, 2-, or 4-codon block in which all codons specify the same amino acid. The block structure is defined based on the canonical code, and all blocks of the same size are permuted among themselves. The N-Block Shuffler allows for a space of 4! × 14! × 8! (or 8.44×10^{16}) possible genetic codes. This randomization algorithm is designed based on the assumptions that the block structure of the genetic code is fixed, e.g., by constraints on tRNA specificity (Lagerkvist 1981). The N-Block Shuffler maintains the number of codons per amino acid and the block structure of the genetic code.

Amino Acid Shuffler. The Amino Acid Shuffler assumes that the groupings of codons which code for a single amino acid are fixed, but each codon group could code for any amino acid. This algorithm first determines the set of codons corresponding to each amino acid and then permutes the associations between amino acids and codon sets. This randomization algorithm yields a set of 21! (or 5.11×10^{19}) possible genetic codes. This method assumes that the structure of the genetic code is fixed but that amino acids could have been assigned to different codon blocks, and it has been widely used in other studies (Haig and Hurst 1991; Freeland and Hurst 1998; Ardell 1998; Freeland et al. 2000). The Amino Acid Shuffler maintains the same codon redundancy as the canonical genetic code, although the amino acid encoded by each codon set (and hence the number of codons for each amino acid) changes.

Codon Shuffler. The Codon Shuffler is designed based on the assumption that the important feature in determining possible genetic codes is the number of codons per amino acid. This randomization algorithm therefore uses the canonical genetic code as a template and randomizes the codon to amino acid relationships, maintaining the number of codons per amino acid. This gives $64!/6!^3 4!^5 3!^2 2!^9 = 2.3 \times 10^{69}$ possible genetic codes.

All Amino Acids Get At Least One Codon Shuffler. The All Amino Acids Get At Least One Codon (AAAGALOC) Shuffler makes the assumption that the only constraint on a code is that all amino acids are present. During randomization, a genetic code is built from the set of amino acids in the canonical genetic code in a manner that ensures that all amino acids are present. This is implemented by starting with a vector containing one of each amino acid, sampling amino acids for the remaining codons from the set of all possible amino acids, and permuting the assignments. Effectively, we are sampling directly from the space of codes with at least one of each amino acid. Without the constraint that each code must have at least one of each amino acid, there would be 21^{64} or 4.19×10^{84} codes generated by choosing 1 of the 21 possible symbols (20 amino acids plus stop) independently at each position. However, we must subtract from this figure all the codes with at most 20 distinct symbols. For a particular set of 20 symbols, there are 20^{64} or 1.84×10^{83} arrangements, leading to an equivalent number of possible codes, but there are 21 different ways to leave out a symbol, so we must subtract 21×20^{64} or 3.87×10^{84} codes, leaving 3.14×10^{83} codes. However, the set of codes lacking a particular symbol thus generated double-counts codes that lack more than one symbol: for example, codes lacking both D and E can be found in the codes lacking D and in the codes lacking E. Thus, we need to add back in all the codes that lack exactly two symbols, of which there are $(20 \times 19)/2 \times 19^{64}$, or 1.32×10^{84}, then subtract all the codes that lack exactly three symbols, add the codes that lack exactly four symbols, and so forth. Using the inclusion–exclusion principle, the formula for the total number of possible codes is

$$\sum_{i=1}^{21} \binom{21}{i} (-1)^{i+1} i^{64} \qquad (1)$$

Applying this formula yields 1.51×10^{84} possible codes. This method allows for the largest set of random variants.

Tests for Association Between Coding Triplets and Aptamer Binding Sites

Aptamer sequences were collated from the literature and from recent experiments in the Yarus lab (M. Illangasekare, I. Majerfeld, D. Puthenvedu, and M. Yarus, unpublished data). These sequences had all been isolated from random-sequence pools using selection-amplification or SELEX (Tuerk and Gold 1990; Ellington and Szostak 1990; Robertson and Joyce 1990) based on their ability to bind affinity columns derivatized with particular amino acids. Specific RNA aptamers now exist for eight amino acids: Arg (Connell and Yarus 1994; Geiger et al. 1996; Tao and Frankel 1996; Yang 1996), Ile (Majerfeld and Yarus 1998; Lozupone et al. 2003), Tyr (Manniróni et al. 2000), Gln (G. Tocchini-Valentini, pers. comm.), Phe (Illangasekare and Yarus 2002), and His, Trp, and Leu (I. Majerfeld and M. Yarus, unpublished data). The 43 characterized binding sites for these amino acids come from RNAs with 2791 total initially randomized nucleotides (for an average of 65 random nucleotides per sequence). The sequences and binding sites used are shown in Fig. 1.

To estimate the probability of observed associations between coding triplets and binding sites as conservatively as possible, we considered only those sequences for which direct experimental information about the binding sites was available. This information came from a mixture of chemical protection/modification/interference mapping, sequence conservation across isolates, and NMR. We considered only independently derived sequences in which the binding site occurred in backgrounds that shared no significant sequence identity. For example, the minimal isoleucine site has been isolated at least 63 times, and the minimal histidine site has been isolated at least 54 times. Consequently, only a small number of well-characterized examples are used for calculations in Table 1. Including the many other independent isolations of the same sites would increase the statistical significance of the results (i.e., diminish the probabilities) by many orders of magnitude.

Associations between coding triplets and binding sites were calculated as previously described (Knight and Landweber 1998). Specifically, we assigned each nucleotide to one of four categories—triplet and binding, triplet and not binding, not triplet but binding, or neither triplet nor binding—depending on whether it was in a coding triplet (either a codon or an anticodon for the cognate amino acid) and whether it was within the experimentally determined binding site sequence. We pooled these four counts across all the sequences for each amino acid and used the G test for independence (with the Williams correction) to test whether the binding site nucleotides were significantly more likely to form parts of codon or anticodon triplets than were non-binding-site nucleotides. The nonbinding parts of the aptamers thus act as internal controls both for the effects of nucleotide composition and for any other properties of the sequences that might affect the results. Counts are pooled across codons or anticodons and across all aptamers for that amino acid to reduce small-sample effects.

In order to compare results across multiple analyses (different amino acids, and codons and anticodons), we used Fisher's method for combining independent tests of a hypothesis. We combined the tests for codons and anticodons for each amino acid and combined the tests for all amino acids to get overall estimates of the probabilities bearing on the escaped triplet hypothesis (Table 1). The escaped triplet hypothesis suggests that triplets overrepresented in RNA binding sites for amino acids ultimately became part of the modern genetic code (Yarus et al. 2005).

Glutamine:

```
ucgauauuaaCACGGGUucuagcaaaagcucgugcugaccaucGGAUCAAGACGuguugcccgacaaggguggcguggg
acugGAUGUUCucugAUCGGGUaugcacucccuGGAgaaUCAAACGcgugcaugcgauguuugagacccgguggugg
```

Tyrosine:

```
ggcAGugaacucgugcgaucgugaaaacggggcaagaUGgccuuAcaGCGGUCAAUACGGGGGucauCAGAUAGGGAgGccuccuggu
gagcgacugagguucgccgcggAUUAUguuuugcgguuagAUcaggcaaCGGGUAAuaccgggucAGUCAGAuaGGGaggaucuacugcc
aagggcaguccCCCCUucgcugggggggguGGUUGUAGggcuaaaacaaaccaCGgGUGAUACGggggccauACCCuaggaaggcccugcuccc
```

Leucine:

```
ucUcucAacccCUAgcgUAguuuUgacUGcGAGAGgCAAAcgccacgguAgAaccGAaggGUAGgagggauuagcaug
```

Isoleucine:

```
ugggcuguucuccgcguUAUUGGGGuaccuccuccaacucugcugucgug
auccgucgcguUAUUGGGGuuccacuauaccagguucuuuugucugcgau
gauaaucugcguUAUUGGGGuguccucaucacuuuugcuccgcauugagg
uacacguUAUUGGGGuaucucuuacucgcuacaguagggcccucggga
uagugacaucgcuuucUAUUGGGGaucguacuuacgacgacgauuuacgagau
ggcaaguagguagcgucgaaaccgauuguguuccGGUAAUgaAAcgAguaaaagagcaagucgguuaga
```

Phenylalanine:

```
auuggaucgguaguaUUuAGGGUGAGAcacuucaugccuuuguugcaggcuggggugAAggcgcuacauggcgucUGAAA
gcGcGAGAaacggucacuagaauagugggccgUcAugcuaacgccucuuucgguguugGGGGAAUAUUggccaaucgagu
```

Histidine:

```
aAAGUGGGUUgAUGuAaGuAACAGGcgauaggcuuugcguuccaaauugcuaucuaacguuugcgcgcu
AAGUGGGGugAcGuAUGgAACAAcguuaguugcuuaggaacucucgguuggguguucgug
aaaaGAAGCGGGGguAAUGuuuuGuAACAacuuucuaaugagguaggagccucggauugcgugucgugu
ggcauaauacaAAGUGGAUGaGuuaGgAACAgguauuuaugcauggugggaguucggguacacgucg
agcugagauccgaugggaugauAAAGUGGGugaGGuGaagGgAACAgaucguccuaucgugacaagug
augacAAGAGGGuauAGuuaGGAACAGucauccugaugggagagungcuuacgugguccuaccau
ucuaauAGUGGGugacAUGgAAGgAACAguagacagagagaggagagguucccaacgcgaauaaaggcuu
aAAGGGGGAUGuaaGgAACAgcccgauaaugcgaaggacacccuuugagaaagcuaggugauagguuggg
acggcauAGAGGGauuguuAGuAACAgccacagauugagcagggaucgcaucguggauaggggguugucgc
auggcAAAGCGGGgggaaaGGAACAggccauuguuaggacuuaggcagauggcuggucgguugu
cuagucgggauuAguAAuggagAAGUGGGuguaAUGgugGuAACAauucgucaaaaucuuacuuagcug
uuaaccaggaguggaagugacgaAAGAGGGaauuguaAGgAACAgcgucacgauaacauacaggugaugu
```

Tryptophan:

```
guuaauaagaccucggaggaguuagggucaauucggcauaggcugcugAGGACCGuAaccagucgcuac
GACCgggacugguuuuccacaguuggcCGCUACcgacuagaagcgaaauucccgaacgcuguaggcu
aAGACCGUcgcgcaagguuuugugcgguaCGCUACUUcgaggcugugggcaucucggguguggccauga
cCGCUACuCgguuggggguuaucgAGGACCGggauagagcuaaggaugggguucguuguc
guaCGCUAagguagguguacgcuGACCGuacgauaaagcuuagacguuccuagaaucaggacgcgggu
gacUGCUACcuuacgcggggauugcguagaGGGACCAgacgaugcgcagcggacccgauugunuaagcc
acCGCUACcacagggcuguGGGACCGgcgguagagauguagugacuuccccauacuggauaacuagcgc
uuCGCCACucacagcgcugagcgugGGGACCGacgauagggggguuuugggggugagauaugaagcuugggc
uugggcgacgggaucuacgaagucaggucgugguacaacgAAACccucuaguugaagGCUAacggaaug
cuggacgacggggaCGCCACuggacuagguaagccAGGACCGuacgucgggagccgucagaau
GGACCGucacgggaugaauaguauggcuguggaaCGCCACccagcaugggcgggauccugaucgau
gaagcaguagAgggAcuuccgcuccgaauaggggcgcgaaggauggaguaaggauucaguacg
```

Arginine:

```
cccgacagaucggcAaCGCCauguu ugAgACacc
gacgagaAGGAGCGcugguucuacuagCAGGUAGGuCacucguc
uggugcgugcaggaCGUCGAUCGAAUCCGC
agcGGUCGAaauccgucaugugcacugcua
augauAAAccgAugcugggcgAuucuccugaaguaggggaagAguugucauguauggg
```

Fig. 1. Amino acid aptamer sequences used for the calculations, primers omitted. Sequences are shown grouped by specificity, binding sites are indicated by uppercase letters.

Table 1. Probabilities for experimental associations between codons and binding sites, anticodons and binding sites, and both codons and anticodons and binding sites for each amino acid separately and all amino acids together

Amino acid	Codon	Anticodon	Both C & AC	Aptamers/total nt
Phenylalanine	0.72	3.4×10^{-5}	2.9×10^{-4}	2/160
Isoleucine	1.2×10^{-3}	1.0×10^{-6}	2.7×10^{-8}	6/320
Histidine	0.999	6.9×10^{-4}	5.7×10^{-3}	12/809
Leucine	0.27	4.5×10^{-4}	1.2×10^{-3}	1/78
Glutamine	0.042	0.99	0.17	2/156
Arginine	3.4×10^{-8}	0.045	3.3×10^{-8}	5/197
Tryptophan	0.99	1.8×10^{-4}	1.7×10^{-3}	12/800
Tyrosine	4.8×10^{-3}	1.6×10^{-6}	1.6×10^{-7}	3/271
Overall	6.6×10^{-3}	1.1×10^{-10}	5.4×10^{-11}	43/2791

Note. See text for discussion.

Evaluation of Code Optimality for Error Minimization

We used the code error value as previously defined (Haig and Hurst 1991). The error value is the sum over all possible single-base changes of the difference in amino acid properties before and after the change, We used the Polar Requirement measure (Woese 1973) to measure differences between amino acids, and no transition bias, codon usage bias, or positional weighting. Small tests altering the transition bias indicated that this factor was unlikely to affect the results (data not shown).

Optimization for Better Codes

To optimize codes with a certain fraction of the codons fixed (e.g., by stereochemistry), we used the Great Deluge Algorithm (Dueck 1993) Specifically, we fixed a fraction of the codons to their values in the canonical code and swapped sets of three amino acid assignments at a time to avoid becoming trapped in local optima.

Implementation Details

All software was written in the Python programming language on a Mandrake Linux system and is available from the authors on request. It relies on the Cogent (formerly PyEvolve) package and consists of three top-level classes; a genetic code randomizer and code error and coding triplet association tests.

The overall analysis used 100,000 random codes generated by each of the five randomization algorithms (see above), each run separately to permute and maintain the locations of the stop codons. This analysis took approximately 48 h to complete on a 2.4-GHz desktop computer with 1 GB of RAM.

Results

We begin by confirming that the phenomena we wish to compare actually exist in our present sample of codes. Table 1 shows the triplet/binding site associations for each amino acid specificity individually and combined over all amino acids. The associations are, in general, highly significant. Interestingly, an association between anticodons and binding sites is more strongly supported than an association between codons and binding sites with the new data set. The amino acid that contributes most significantly to the codon/binding site association is arginine, which was the first specificity for which aptamer sequences were available. However, the codons are also present surprisingly frequently in the minimal isoleucine-binding site, which has now been reselected hundreds of times independently in different experiments (Lozupone et al. 2003; Legiewicz and Yarus 2005). Figure 2 shows the canonical genetic code compared to random genetic codes under the most restrictive (left column) and least restrictive (right column) randomization models in terms of code error, codon association, anticodon association, and overall association. The canonical genetic code is far out on the tail of the distribution of random codes on each measure, even for the most restrictive randomization model, in all respects except codon association (for which about 10% random codes show larger associations than does the canonical genetic code across the different models). Results are intermediate for the other randomization models. Table 2 summarizes the fraction of random codes that appeared better than the canonical code on each association measure. Thus, the anticodon and overall associations are highly robust to the choice of model, while the codon associations are significant for all but one randomization (and in two individual groups of aptamers). The specific codon/site associations for arginine (Knight and Landweber 2000) and isoleucine remain individually highly significant, and the reselection of the codons in the minimal isoleucine site is a compelling second line of evidence (Yarus et al. 2005).

Next, we tested for correlations between error minimization and triplet/site association in each sample of random genetic codes. Figure 3 shows this relationship for each of the randomization models. The canonical genetic code, indicated by an asterisk, is clearly an extreme outlier on both measures and does not fall within any group of codes that are similarly near-optimal in both respects (see insets in Fig. 3 for the best 0.1% of codes on each measure, which would reveal a local correlation, if it existed for only the best codes, as a diagonal line of points surrounding the canonical genetic code). Thus, even the

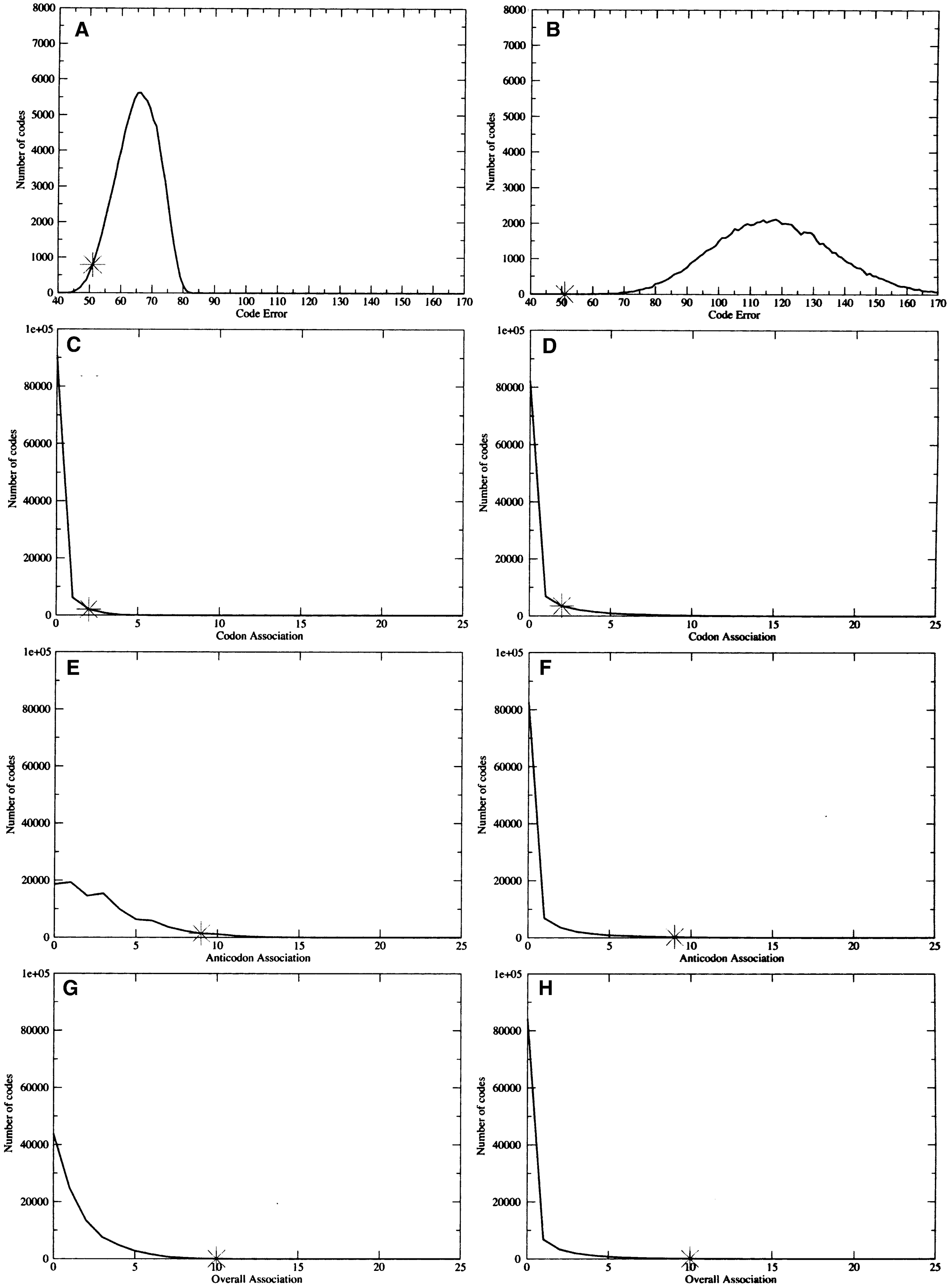

Fig. 2. The canonical genetic code (marked with an asterisk) compared to the distribution of random codes produced by the most restrictive model (the Quartet Shuffler, maintaining stop codon locations; left column) and the least restrictive model (AAAGALOC Shuffler; right column). The y axis shows the number of random codes in a particular range of x-axis values, which are code error (**a**, **b**), and the log of the p-values for codon association (**c**, **d**), anticodon association (**d**, **e**), and overall association (**g**, **h**). Data shown are for samples of 100,000 genetic codes in each model. The same sample of genetic codes is shown in the graphs for each column.

Table 2. Fraction of random codes showing greater triplet/site associations the canonical code under each model

Shuffler	Stops	Anticodon	Codon	Overall	Shared (mean/SD)
AAAGALOC	Random	0.00656	0.09994	0.00294	1.15/1.04
AAAGALOC	Preserved	0.00682	0.10492	0.00306	1.20/1.06
Codon	Random	0.0093	0.08695	0.00351	1.54/1.16
Codon	Preserved	0.00997	0.09208	0.00392	1.60/1.19
Amino Acid	Random	0.01138	0.1462	0.00283	1.15/2.13
Amino Acid	Preserved	0.0122	0.15463	0.00289	1.20/2.17
N-Block	Random	0.00008	0.12914	0.00005	2.50/2.43
N-Block	Preserved	0.00008	0.16385	0.00015	2.74/2.49
Quartet	Random	0.01182	0.07176	0.00065	2.00/2.36
Quartet	Preserved	0.02582	0.02454	0.00055	5.07/2.41

Note. Columns indicate the random model (Shuffler), whether the locations of stop codons was randomized, the fraction of codes with better anticodon associations over all amino acids, the fraction of codes with better codon associations over all amino acids, the fraction of codes with better combined codon and anticodon associations over all amino acids, and the mean and standard deviation of the number of codons (for amino acids with characterized aptamers) shared between a random code generated under each model and the canonical genetic code.

best codes in each respect show no tendency to have both properties simultaneously.

Although the graphs in Fig. 3 show no obvious association by eye, small but significant associations exist in the graphs for the Quartet, N-Block, and Amino Acid Shufflers when the locations of stop codons were preserved (r, from 0.008 to -0.02; p, from 0.004 to 10^{-10}). These small correlations, which would explain at most 0.04% of the variance, can be explained by preservation of substantial portions of the canonical genetic code in some of the randomized variants (Table 2). In particular, there was always a moderate positive correlation between the number of codon differences from the canonical genetic code (counting only codons for amino acids for which we have aptamers) and the p-value of the association between triplets and binding sites: in other words, codes that differed more from the canonical code were more likely to have larger, hence less significant, p-values when tested for association (r, from 0.10 to 0.38; $p < 10^{-200}$), and, similarly, codes that differed more from the canonical code were more likely to have higher error scores (r, from -0.0009 to 0.03; p, from 0.83 to 10^{-26}). Consequently, the small associations that do exist can be explained in terms of shared pieces of the canonical genetic code. Because the distribution of the data differs substantially from the bivariate normal distribution assumed by the Pearson correlation coefficient, the p-values should be treated with caution (although model distributions with the same visual appearance but with independent sampling of x and y, such as bivariate negative exponential distributions, give a correlation coefficient close to 0). This effect is illustrated in Fig. 4, in which highly significant associations between the number of shared codons and both the code error ($r = 0.011$, $p = 0.00035$) and the overall triplet/site association ($r = 0.10$, $p = 2.1 \times 10^{-220}$) interact to give an apparent association between code error and overall association. Similar effects give apparent associations between various combinations of the codon/site association, the anticodon/site association, the error value, and the overall association.

To test whether a substantial fraction of stereochemically determined codons would still allow optimization, we generated random codes through amino acid permutation in which a certain fraction of the amino acid assignments were fixed at their values in the canonical genetic code. We then used the Great Deluge Algorithm (Dueck 1993) to optimize this starting set as far as possible and compared the error values of the random codes to the error values of the optimized codes (Fig. 5). As expected, the error values of both the randomized and the optimized codes approached the error value of the canonical code as more codons were shared with the canonical code.

Discussion and Conclusion

Our results demonstrate that an error-minimizing genetic code does not necessarily display associations between coding triplets and binding sites, and that a code having such associations does not necessarily minimize the effects of errors. Taken together with the fact that the canonical genetic code performs surprisingly well on both measures, this suggests that error minimization and triplet/site associations both played important but independent roles in the evolution of the genetic code. The fact that these observations were supported over a wide range of randomization strategies, and hence hypotheses about what might have been conserved over the course of genetic code evolution, suggests that these conclusions (along with the consistent observation of both the error-minimizing and the coding triplet association properties) are robust to different assumptions about how the genetic code evolved. The results also suggest that the features of amino acids that are important for protein function may differ

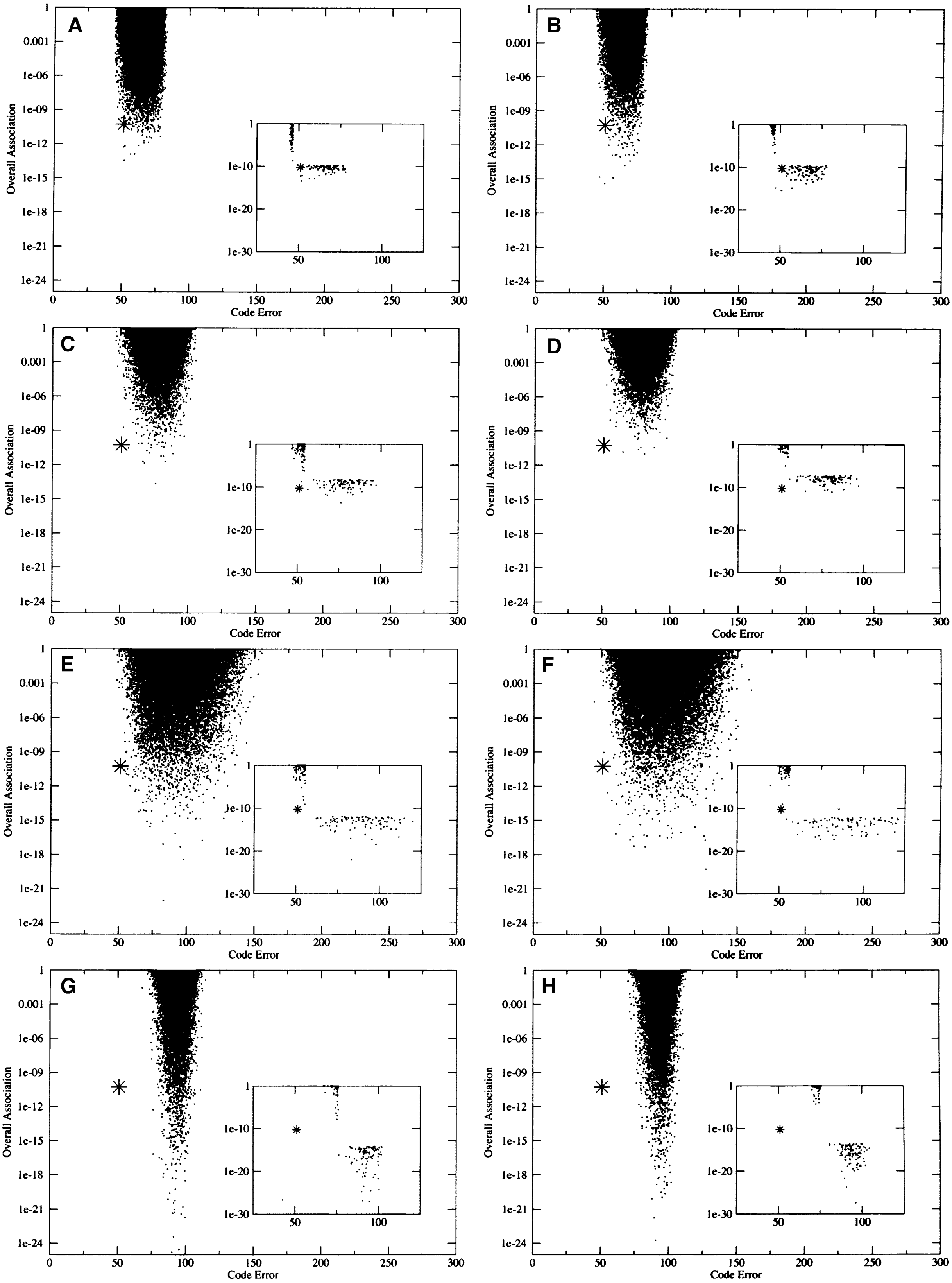

Fig. 3. Continued

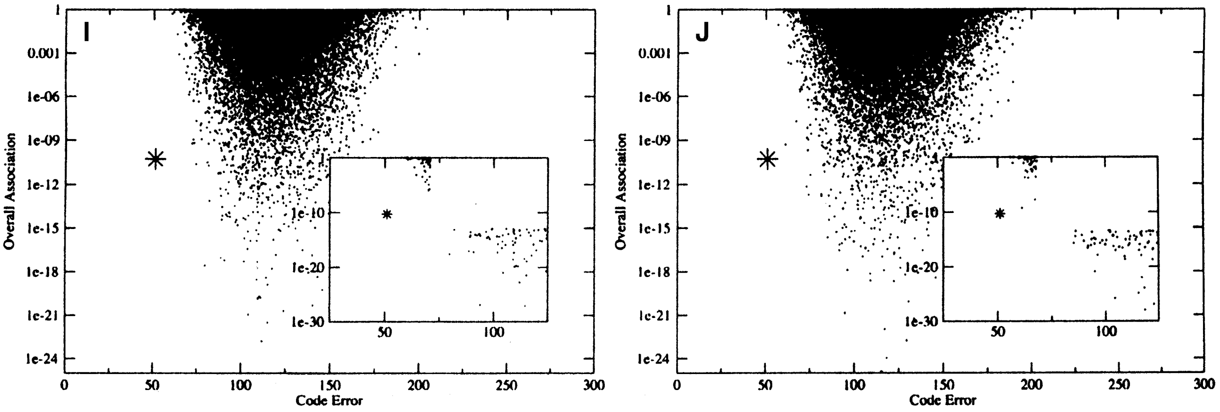

Fig. 3. Code error (x axis) plotted against the overall triplet/site association (y axis) for each randomization model. The inset within each graph shows just the best 0.1% of points on each measure, to test whether correlations exist for the best codes on either measure even if they do not exist for all codes. The left column shows models where the locations of stop codons are maintained; the right column shows models where the locations of stop codons are randomized. The plots are ranked from most restrictive to least restrictive model; Quartet (**A**, **B**), N-Block (**C**, **D**), Amino Acid (**E**, **F**), Codon (**G**, **H**), and AAAGALOC (**I**, **J**). Although small, statistically significant correlations actually do exist for some models despite their invisibility to the naked eye, they are explicable by random codes that share codons with the canonical genetic code. The canonical genetic code is indicated by an asterisk.

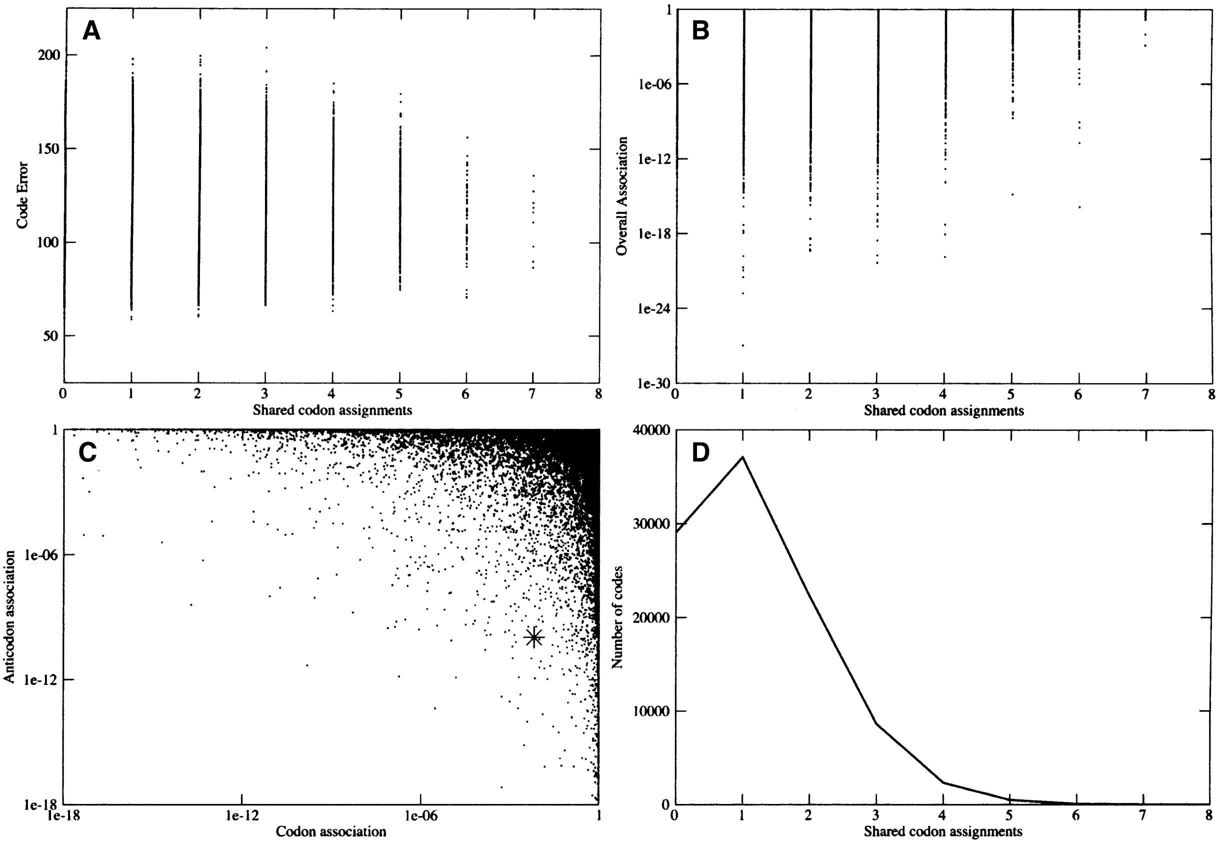

Fig. 4. Explanations for weak but significant correlations caused by sharing of codons with the canonical genetic code. **A** Number of shared codon assignments (x axis) against code error (y axis). **B** Number of shared codon assignments (x axis) against p-value of the association score (y axis). **C** Codon association (x axis) against anticodon association (y axis). **D** Number of shared codon assignments (x axis) against frequency of codes with that number of shared codon assignments (y axis). Each graph shows the same sample of 100,000 codes generated according to the least restrictive model (the AAAGALOC Shuffler, allowing stop codons to vary).

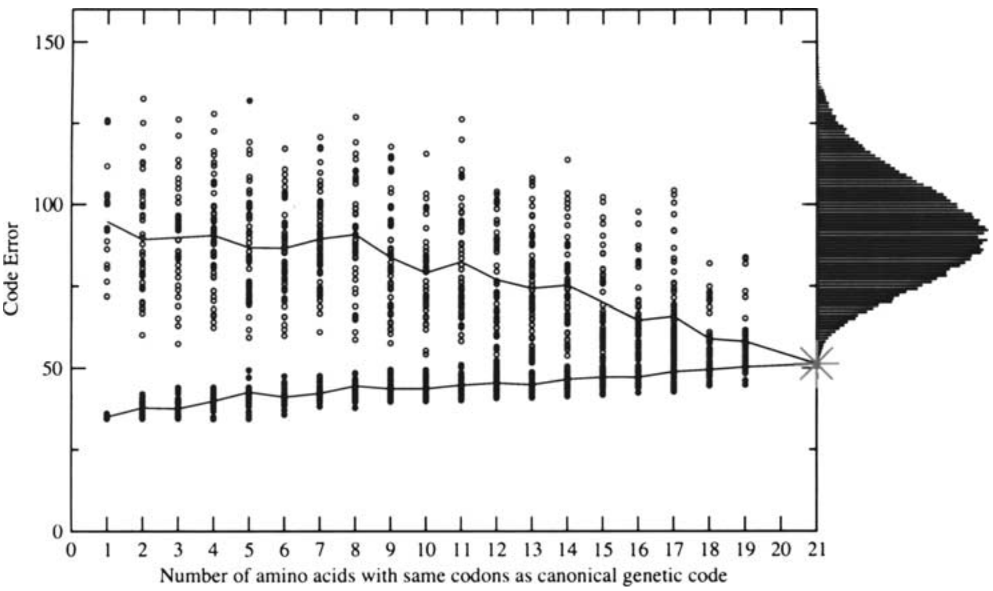

Fig. 5. Even a largely stereochemical code allows an adaptive result. This figure shows as open circles the error values of 50 random genetic codes (y axis) that share the codons for up to n amino acids with the canonical genetic code (x axis). The canonical genetic code is marked with an asterisk at the right, with 21 amino acids shared (the 20 canonical amino acids plus termination). For each random code, we applied a nonlinear optimizer, the Great Deluge Algorithm (Dueck 1993), to find the best possible code with that number of shared codons (filled circles). The histogram on the right shows the distribution of error values for 100,000 completely randomized codes generated by the same model (amino acid permutation among existing amino acid blocks). The means of the random and optimized codes are plotted as solid lines.

from those that are important for RNA-amino acid recognition, which may be unsurprising given the varied and different chemistry involved in the two types of interaction.

The present work confirms that significant associations exist between coding triplets and binding sites within aptamers for seven of the eight amino acids for which specific aptamers are now available (the exception, glutamine, being a plausibly late addition to the genetic code). However, nonspecific aptamers were earlier selected to valine (Majerfeld and Yarus 1994) and to phenylalanine and tryptophan (Zinnen and Yarus 1995). The more recent success in selecting specific aptamers to the latter two amino acids (Yarus et al. 2005) suggests that many initially intractable amino acids, such as lysine (M. Illangasekare, I. Majerfeld, and M. Yarus, pers. comm.), may eventually yield to SELEX. Thus, our results should be interpreted as providing a snapshot of current knowledge about triplet/binding site associations that is subject to modification as more aptamer sequences are determined. However, we demonstrate that selection for error minimization could have played a definite role in shaping the code even if most of the codon assignments were fixed by triplet/amino acid associations.

One scenario for the evolution of the genetic code is that a primordial set of codons fixed by stereochemistry was later added to through revolutionary processes and/or optimized by natural selection for error minimization (Knight and Landweber 1998). Accordingly, we tested whether a code in which a certain fraction of the modern codons was ''fixed'' could be optimized at least to the level of the canonical code. Figure 5 shows the plausibility of this scenario: even with a substantial fraction of the canonical code locked to the present codon assignments, codes that are even better than the canonical code at minimizing errors can be found through numerical optimization procedures such as the Great Deluge Algorithm. Consequently, the fact that the genetic code appears unusual in that it both minimizes errors and shows associations between coding triplets and binding sites is compatible with a pluralistic view of code evolution.

Acknowledgments. This work was supported in part by NIH Grant GM 48080 and NASA Center for Astrobiology Grant NCC2-1052 to M. Yarus. We would like to thank members of the Yarus and Knight labs, Eors Szathmary, and an anonymous reviewer for critical discussion of the manuscript.

References

Alff-Steinberger C (1969) The genetic code and error transmission. Proc Natl Acad Sci USA 64:584–591

Amirnovin R (1997) An analysis of the metabolic theory of the origin of the genetic code. J Mol Evol 44:473–476

Ardell DH (1998) On error minimization in a sequential origin of the standard genetic code. J Mol Evol 47:1–13

Connell GJ, Yarus M (1994) RNAs with dual specificity and dual RNAs with similar specificity. Science 264(5162):1137–1141

Di Giulio M (1991) On the relationships between the genetic code coevolution hypothesis and the physicochemical hypothesis. Z Naturforsch 46c:305–312

Di Giulio M (1998) The historical factor: the biosynthetic relationships between amino acids and their physiochemical properties in the origin of the genetic code. J Mol Evol 46:615–621

Di Giulio M (1999) The coevolution theory of the origin of the genetic code. J Mol Evol 48:253–255

Di Giulio M (2001) A blind empiricism against the coevolution theory of the origin of the genetic code. J Mol Evol 53:724–732

Di Giulio M, Medugno M (2000) The robust statistical bases of the coevolution theory of genetic code origin. J Mol Evol 20:258–263

Dueck G (1993) New optimization heuristics: the great deluge algorithm and the record-to-record travel. J Comput Phys 104:86–92

Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. Nature 346:818–822

Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000) Early fixation of an optimal genetic code. Mol Biol Evol 17:511–518

Geiger A, Burgstaller P, von der Eltz H, Roeder A, Famulok M (1996) RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. Nucleic Acids Res 24:1029–1036

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 33:412–417

Illangasekare M, Yarus M (2002) Phenylalanine-binding RNAs and genetic code evolution. J Mol Evol 54(3):298–311

Knight RD, Landweber LF (1998) Rhyme or reason: RNA-arginine interactions and the genetic code. Chem Biol 5:R215–R220

Knight RD, Landweber LF (2000) Guilt by association: the arginine case revisited. RNA 6:499–510

Knight RD, Landweber LF, Yarus M (2003) Tests of a stereochemical genetic code. Kluwer Academic/Plenum, New York, pp 115–128

Knight RD, Freeland SJ, Landweber LF (1999) Selection, history and chemistry: the three faces of the genetic code. Trends Biochem Sci 24:241–247

Lagerkvist U (1981) Unorthodox codon reading and the evolution of the genetic code. Cell 23:305–306

Legiewicz M, Yarus M (2005) A more complex isoleucine aptamer with a cognate triplet. J Biol Chem 280:19815–19822

Levy M, Ellington AD (2001) RNA world: catalysis abets binding, but not vice versa. Curr Biol 11:R665–R667

Lozupone C, Changayil S, Majerfeld I, Yarus M (2003) Selection of the simplest RNA that binds isoleucine. RNA 9(11):1315–1322

Majerfeld I, Yarus M (1994) An RNA pocket for an aliphatic hydrophobe. Nat Struct Biol 1:287–292

Majerfeld I, Yarus M (1998) Isoleucine: RNA sites with associated coding sequences. RNA 4(4):471–478

Manniron C, Scerch C, Fruscoloni P, Tocchini-Valentini GP (2000) Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif. RNA 6:520–527

Miseta A (1989) The role of protein associated amino acid precursor molecules in the organization of genetic codons. Physiol Chem Phys Med NMR 21:237–242

Robertson DL, Joyce GF (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. Nature 344:467–468

Ronneberg TA, Landweber LF, Freeland SJ (2000) Testing a biosynthetic theory of the genetic code: fact or artifact? Proc Natl Acad Sci USA 97:13690–13695

Sonneborn TM (1965) Degeneracy of the genetic code: Extent, nature and genetic implications. Academic Press, New York, pp 277–297

Szathmary E (1993) Coding coenzyme handles: A hypothesis for the origin of the genetic code. Proc Natl Acad Sci USA 90:9916–9920

Szathmary E (1999) The origin of the genetic code: amino acids as cofactors in an RNA world. Trends Genet 15(6):223–229

Tao J, Frankel AD (1996) Arginine-binding RNAs resembling TAR identified by in vitro selection. Biochemistry 35:2229–2238

Taylor FJR, Coates D (1989) The code within the codons. Bio Systems 22:177–187

Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505–510

Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552

Woese CR (1973) Evolution of the genetic code. Naturwissenschaften 60:447–459

Wong J-TF (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912

Yang Z (1996) Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol 42(2):294–307

Yarus M (1998) Amino acids as RNA ligands: A direct-RNA-template theory for the code's origin. J Mol Evol 47:109–117

Yarus M, Christian EL (1989) Genetic code origins. Nature 342:349–350

Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: The escaped triplet theory. Annu Rev Biochem 74:179–198

Zhu CT, Zeng XB, Huang WD (2003) Codon usage decreases the error minimization within the genetic code. J Mol Evol 57:533–537

Zinnen S, Yarus M (1995) An RNA pocket for the planar aromatic side chains of phenylalanine and tryptophane. Nucleic Acids Symp Ser 33:148–151