The genetic code is nearly optimal for allowing additional information within protein-coding sequences

Shalev Itzkovitz and Uri Alon

Genome Res. 2007 17: 405-412; originally published online Feb 9, 2007;
Access the most recent version at doi:10.1101/gr.5987307

Letter

The genetic code is nearly optimal for allowing additional information within protein-coding sequences

Shalev Itzkovitz[1,2] and Uri Alon[1,2,3]

[1]*Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel;* [2]*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel*

DNA sequences that code for proteins need to convey, in addition to the protein-coding information, several different signals at the same time. These "parallel codes" include binding sequences for regulatory and structural proteins, signals for splicing, and RNA secondary structure. Here, we show that the universal genetic code can efficiently carry arbitrary parallel codes much better than the vast majority of other possible genetic codes. This property is related to the identity of the stop codons. We find that the ability to support parallel codes is strongly tied to another useful property of the genetic code—minimization of the effects of frame-shift translation errors. Whereas many of the known regulatory codes reside in nontranslated regions of the genome, the present findings suggest that protein-coding regions can readily carry abundant additional information.

[Supplemental material is available online at www.genome.org.]

The genetic code is the mapping of 64 three-letter codons to 20 amino-acids and a stop signal (Woese 1965; Crick 1968; Knight et al. 2001). The genetic code has been shown to be nonrandom in at least two ways: first, the assignment of amino acids to codons appears to be optimal for minimizing the effect of translational misread errors. This optimality is achieved by mapping close codons (codons that differ by one letter) to either the same amino acids or to chemically related ones (Woese 1965). This feature has been attributed to an adaptive selection of a code, so that errors that misread a codon by one letter would result in minimal effects on the translated protein (Freeland and Hurst 1998; Freeland et al. 2000; Gilis et al. 2001; Wagner 2005b). Second, amino acids with simple chemical structure tend to have more codons assigned to them (Hasegawa and Miyata 1980; Dufton 1997; Di Giulio 2005).

There exist a large number of alternative genetic codes that are equivalent to the real code in these two prominent features (Fig. 1). Here we ask whether the real code stands out among these alternative codes as being optimal for other properties.

We consider the ability of the genetic code to support, in addition to the protein-coding sequence, additional information that can carry biologically meaningful signals. These signals can include binding sequences of regulatory proteins that bind within coding regions (Robison et al. 1998; Stormo 2000; Lieb et al. 2001; Kellis et al. 2003). Such binding sites are typically sequences of length 6–20 bp. In addition to regulatory proteins, there are binding sites of structural proteins such as DNA- and mRNA-binding proteins (Draper 1999). Histones, for example, bind with a code that has a periodicity of about 10 bp over a site of about 150 bp (Satchwell et al. 1986; Trifonov 1989; Segal et al. 2006). Other codes include splicing signals (Cartegni et al. 2002) that include specific 6–8 bp sequences within coding regions and mRNA secondary structure signals (Zuker and Stiegler 1981; Shpaer 1985; Konecny et al. 2000; Katz and Burge 2003). The latter often correspond to sequences of several dozen base pairs or longer. Since we do not know all of these additional codes, and different organisms can use a vast array of different codes, we tested the ability of the genetic code to support arbitrary sequences of any length in parallel to the protein-coding sequence.

We find that the universal genetic code can allow arbitrary sequences of nucleotides within coding regions much better than the vast majority of other possible genetic codes. We further find that the ability to support parallel codes is strongly correlated with an additional property—minimization of the effects of frame-shift translation errors. Selection for either or both of these traits may have helped to shape the universal genetic code.

[3]Corresponding author.
E-mail uri.alon@weizmann.ac.il; fax 972-8-934125.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5987307. Freely available online through the *Genome Research* Open Access option.

Results

Ability to include additional sequences

We first considered the ability of the genetic code to support, in addition to the protein-coding sequence, additional sequences that can carry biological signals. For this purpose, we studied the properties of all alternative genetic codes that share the known optimality features of the real code (Fig. 1). Each alternative code has the same number of codons per each amino acid and the same impact of misread errors as in the real code.

We tested the ability of the genetic codes to include arbitrary sequences, denoted n-mers, within protein-coding regions. As an example, consider the 5-mer "UGACA." This sequence may be a protein-binding site, which should appear within a protein-coding region. This 5-mer sequence can appear within a coding sequence in one of the three reading frames: UGA|CAN,

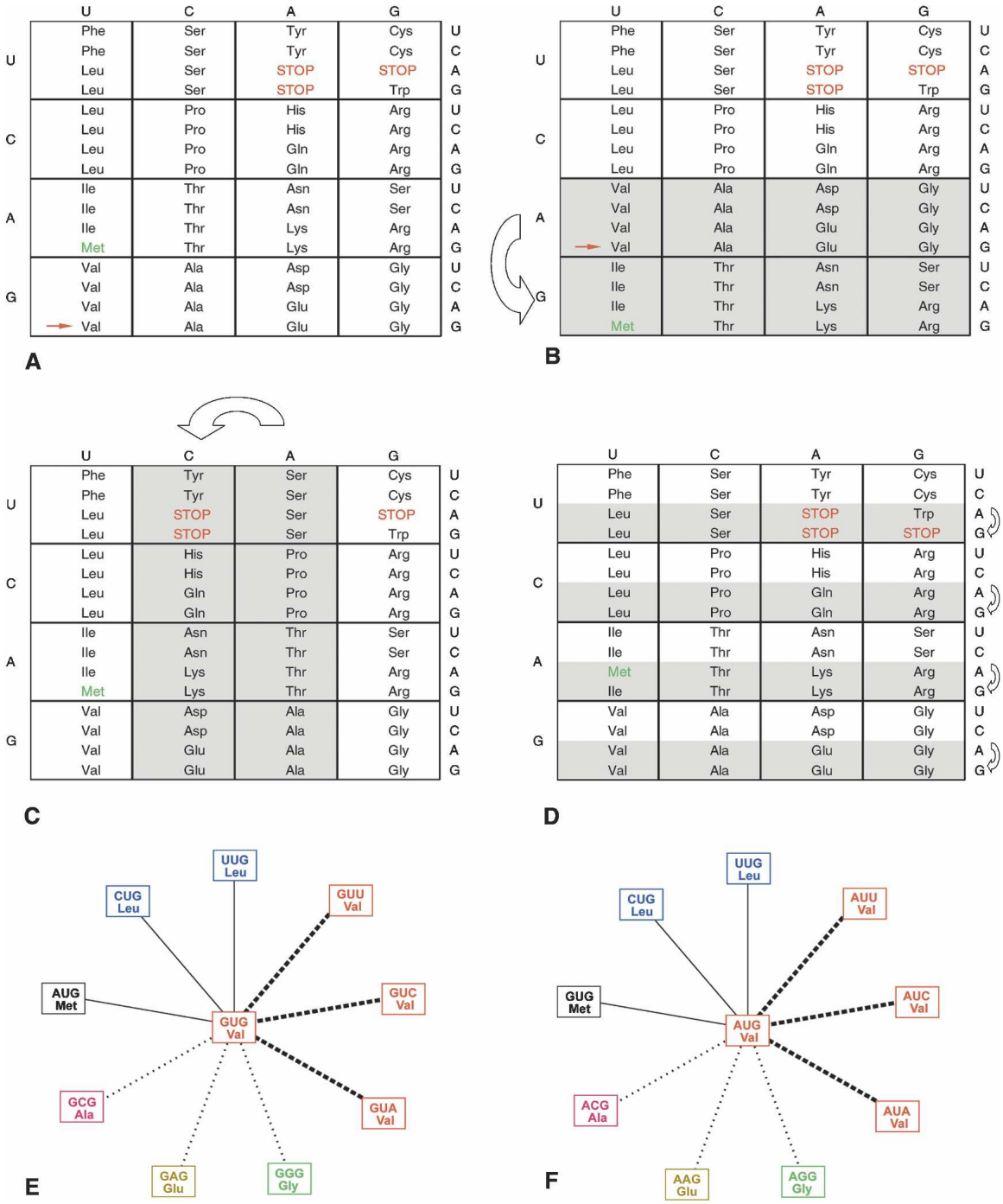

Figure 1. Alternative genetic codes. (*A*) The real code. (*B*) An alternative code obtained by an A↔G permutation in the first position. (*C*) An alternative code obtained by an A↔C permutation in the second position, and (*D*) A↔G permutation in the third position. Stop codons are marked in red, start (Met) codons in green. Codons that are changed relative to the real code are in gray. There are $4! \times 4! \times 2 = 1152$ alternative codes obtained by independent permutations of the nucleotides in each of the three codon positions. (*E,F*) Structural equivalence of real and alternative genetic codes. For example, (*E*) the nine neighboring codons of the Valine codon marked with a red arrow in the real code (shown in *A*) are the same as (*F*) the nine neighboring codons of the Valine codon marked with a red arrow in the alternative code shown in *B*. Solid lines connect codons differing in the first letter, dotted lines connect codons differing in the second letter, and dashed lines connect codons differing in the third letter. Different amino acids are displayed in different colors. This equivalence applies to all codons.

NNU|GAC|ANN, or NUG|ACA, where N denotes any nucleotide and the vertical lines separate consecutive codons. To assess the probability that this 5-mer appears in a coding region, one needs to sum over the three possible reading frames (Fig. 2A). In one of the frames, this sequence generates a stop codon, UGA. The 5-mer cannot appear in a coding region in this frame, because

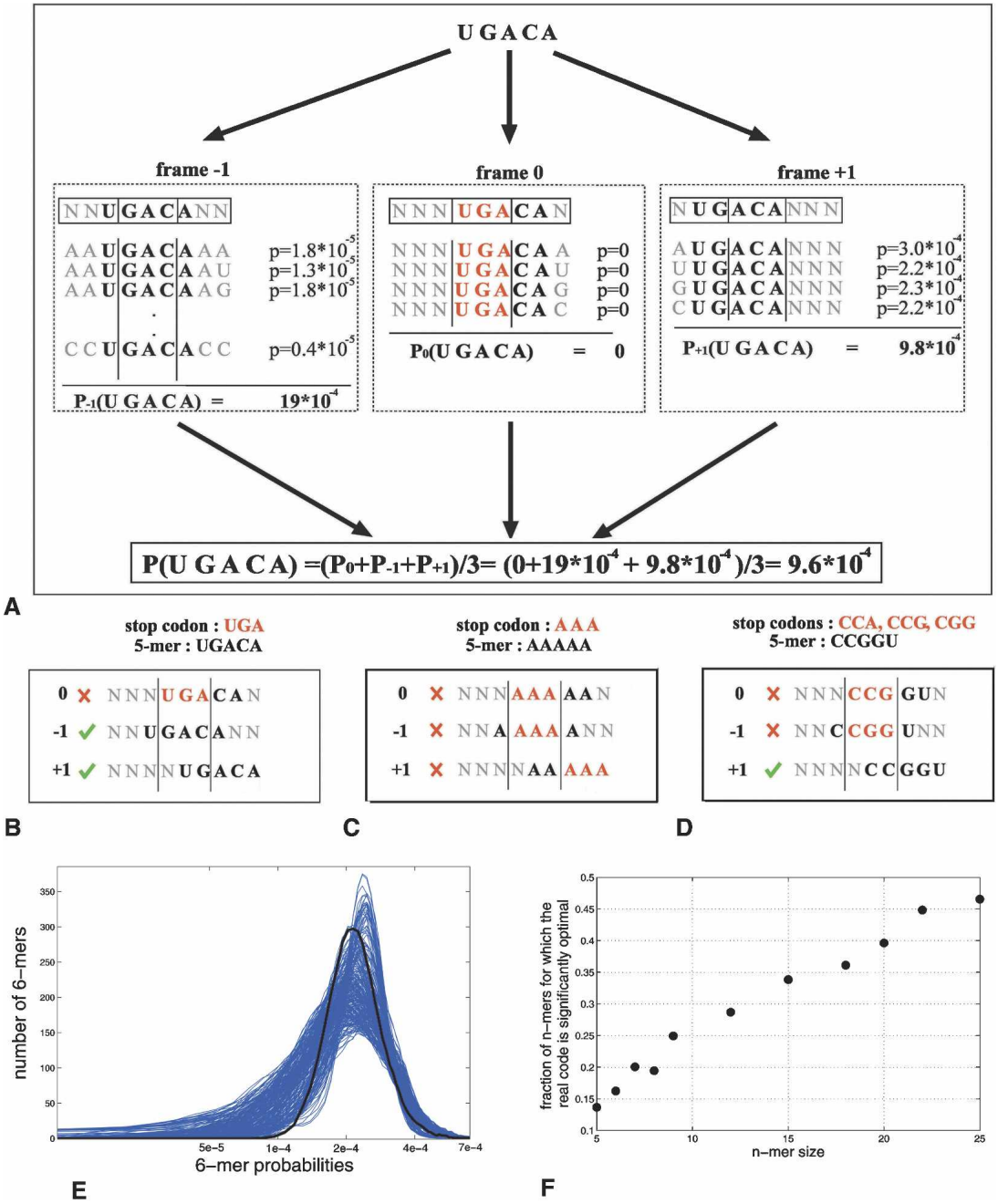

Figure 2. (*A*) Calculation of the probability that an n-mer sequence appears within a protein-coding region in the real genetic code. The 5-mer sequence S = UGACA can appear in one of the three reading frames. For each reading frame, the probabilities of all three codon combinations that contain S are summed up. Codon combinations with an in-frame stop (such as UGA) do not contribute to the n-mer probability since they cannot appear in a coding region. Vertical lines separate consecutive codons, stop codons are in red, P_0, P_{-1}, P_{+1} denote the probabilities of encountering S in the 0/−1/+1 frame. (*B,C,D*) Three examples of "difficult" n-mers in the real code and in alternative codes. (*B*) The 5-mer UGACA, which includes the stop codon UGA, can appear in a protein-coding sequence with the real genetic code in only two of the three possible reading frames (+1 and −1 frames). (*C*) In the alternative code shown in Figure 3D, whose stop codon AAA overlaps with itself, the 5-mer AAAAA cannot appear in a protein-coding sequence in any of the three reading frames. (*D*) In an alternative code with the overlapping stop codons CCG and CGG, the 5-mer CCGGU can only appear in one reading frame. The 5-mers are in bold text, stop codons are in red, N denotes any DNA letter, green v denotes a frame in which the n-mer can appear, red x denotes a frame in which the n-mer cannot appear. (*E*) Distribution of the probabilities of all 6-mers in the real code (bold black line) and in the alternative codes (light blue lines). The *x*-axis is the probability of obtaining 6-mers within protein-coding sequences; the *y*-axis is the number of 6-mers with this probability. In the real code there are significantly less "difficult" 6-mers (with low probabilities), relative to the alternative codes. (*F*) The fraction of n-mers that have a higher probability in the real code than in alternative codes increases with n-mer size. The *y*-axis shows the fraction of n-mers for which the average probability of appearing in the real genetic code is significantly higher than in the alternative codes.

coding regions have no in-frame stop codons. The sequence can, however, appear in one of the two other frames. Overall, the probability that this 5-mer appears in coding regions will tend to be lower than that of 5-mers that do not include stop codons.

Each genetic code has n-mer sequences, such as the above-mentioned sequence UGACA in the real genetic code, which are difficult to include in coding regions: these "difficult" sequences contain stop codons, and thus cannot appear in at least one of the three frames, since protein-coding regions do not contain stop codons. We find that the real genetic code is able to include even the most difficult n-mers because it has a special property: its stop codons, when frame shifted, tend to form abundant codons. Hence, n-mers that cannot be included in one frame-shift can be included with high probability in other frame shifts.

To understand the relation between the stop codons and the ability of the genetic code to include arbitrary n-mers, consider the 5-mer S = AAAAA (Fig. 2C). This 5-mer can appear within a coding sequence in one of the three reading frames: AAA|AAN, NNA|AAA|ANN, or NAA|AAA. Alternative genetic codes that assign one of their stop codons as AAA (Fig. 3D), can never include S in a protein-coding sequence. The problem is that the stop codon AAA overlaps with itself when frame shifted; hence, strings such as S include a stop codon in each of the three frames, precluding their presence in a coding region.

Another example is the 5-mer S = CCGGU. In an alternative code with stop codons CCA, CCG, and CGG, this n-mer can only appear in one of the three reading frames (Fig. 2D). This is because two of the stop codons, CCG and CGG, overlap each other. In contrast, the real genetic code has the stop codons UAA, UAG, and UGA that do not overlap with themselves or with each other, no matter how they are frame shifted. Furthermore, frame-shifted versions of the real stop codons overlap with the codons of the most abundant amino acids. For example, the UGA stop codon in a −1 frame-shift message results in the di-codon NNU|GAN, where N is any nucleotide (Fig. 2B). The GAN codons encode Asp and Glu, which are among the three amino acids

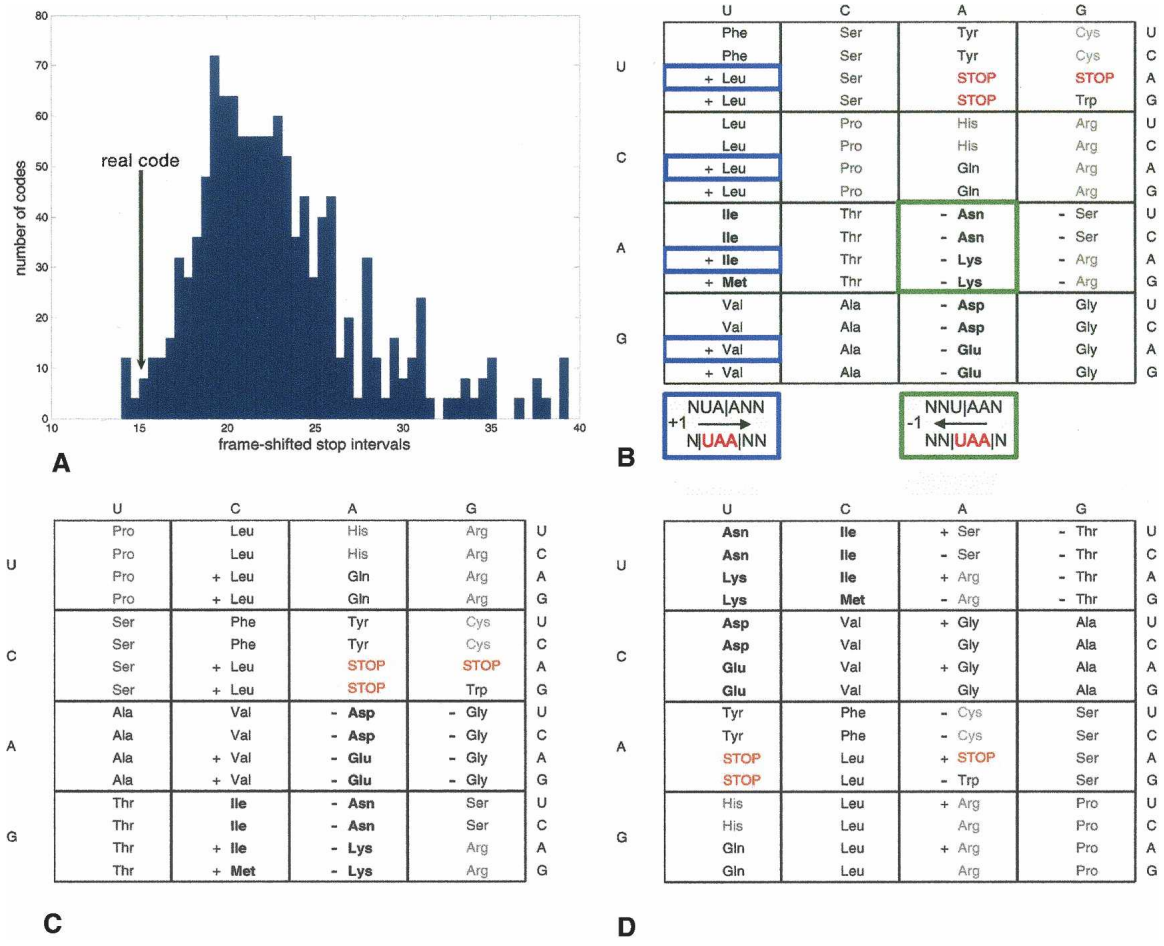

Figure 3. Optimality of the genetic code for minimizing the impact of frame-shift translation errors. (*A*) Distribution of average number of translated codons until a stop codon is encountered after a frame-shift event for the alternative genetic codes. This number corresponds to the mean length of the nonsense polypeptide translated after a frame-shift event, and is the inverse of the frame-shifted stop probability, averaged over the +1 and −1 frame-shifts. (*B*) In the real code, frame-shifted stop codons overlap with abundant codons. Codons with two-letter overlap with a stop codon are marked by + for a +1 frame-shift and – for a −1 frame-shift. Abundant codons are shown in heavier font. For example, the stop codon UAA, when frame shifted, results in codons such as AAN (green box), or NUA (blue boxes), which are relatively abundant. (*C*) The "best code," which achieves the highest frame-shifted stop probability both in a +1 frame-shift and in a −1 frame shift. Stop codons are CAA, CAG, and CGA. In the "best code," a stop codon has an overlap of two positions with codons of Glycine instead of codons of Serine and Arginine in the real code. (*D*) The "worst code" with the lowest frame-shifted stop probability. Stop codons are AUA, AUG, and AAA. Note that the stop codons overlap either with themselves (AAA) or with codons for nonabundant amino-acids (those with light font), in contrast to *B* and *C*.

with the most abundant codons (Table 1). Therefore, n-mers with the letters UGA can be included with high probability in protein sequences without generating an in-frame stop. The same idea applies to the other two stop codons in the real code; this property occurs in only very few of the alternative genetic codes. In short, optimality for including arbitrary n-mer sequences within coding regions is due to stop codons that do not overlap each other, but which do overlap codons for abundant amino acids.

We calculated the probability of including all n-mer sequences for each alternative genetic code by summing up, for every n-mer sequence, the probabilities of all codon combinations that contain it (Fig. 2A; for details see Methods). The codon probabilities were determined according to the known amino acid frequencies in proteins (Table 1). The results presented in the main text are for uniform codon usage, but they apply to a wide range of different codon usages (Supplemental material).

We find that the real code shows significantly higher probabilities to include arbitrary sequences. The average of the logarithm of all n-mer probabilities is significantly higher in the real code than in the vast majority of alternative codes (Table 2), with a P-value < 0.05 for n-mer sequences with n greater than seven. In addition, the real code shows significantly higher probabilities to include the most difficult sequences (n-mers with the lowest probability of appearing in a coding region) than the vast majority of alternative codes (Fig. 2E; Table 2; Supplemental Fig. 4). For example, the average probability of including the 20% most difficult sequences is exceeded by only 3% of the alternative codes for 8-mers and 1% of the alternative codes for 9-mers. This property can be seen when examining the distribution of the n-mer probabilities of appearing within protein-coding sequences. In the real code there are significantly fewer n-mers with low probabilities (Fig. 2E).

The optimality of the real genetic code relative to alternative codes seems to increase with the length of the n-mers (Fig. 2F). This is because as the length of the n-mers increases, the fraction of n-mers that include stop codons increases dramatically. Above $n = 16$, more than half of all n-mers include at least one stop codon. The real genetic code is able to include all n-mers with $n < 11$ in at least one, and often many combinations of amino acid codons. For n-mers of any length, the real code appears to exceed almost all of the alternative codes in its ability to include a large fraction of possible n-mers within coding regions (Fig. 2F; Table 2).

Table 1. Amino acid abundance (average amino acid frequency over 134 organisms, sorted in decreasing order by codon abundance)

amino acid	abundance	# codons	codon abundance
Glu	6.5	2	3.2
Lys	6.0	2	3.0
Asp	5.3	2	2.6
Met	2.3	1	2.3
Ile	6.8	3	2.3
Asn	4.4	2	2.2
Phe	4.3	2	2.1
Ala	8.2	4	2.0
Gln	3.6	2	1.8
Gly	6.9	4	1.7
Val	6.9	4	1.7
Leu	10.1	6	1.7
Tyr	3.3	2	1.6
Thr	5.3	4	1.3
Trp	1.1	1	1.1
Ser	6.5	6	1.1
Pro	4.3	4	1.1
His	2.1	2	1.0
Arg	5.2	6	0.9
Cys	1.1	2	0.6

Codon abundance is the amino acid frequency divided by number of codons for that amino acid.

Table 2. Significance of the genetic code in representing arbitrary sequences

n-mer size	P-value average log-probabilities	P-value 20%
5	0.110	0.054
6	0.097	0.045
7	0.083	0.028
8	0.049	0.031
9	0.043	0.010
12	0.028	0.004
15	0.016	0.004
18	0.012	0.006
20	0.026	0.006
22	0.021	0.004
25	0.029	0.009

Shown are the fractions of alternative codes for which the average of the logarithm of the probabilities of all n-mers is equal or higher to that of the real code. Also shown are the fraction of alternative genetic codes for which the average probability of the 20% most-difficult n-mer sequences is equal or higher than in the real genetic code. Similar results are obtained for larger fractions of the most difficult n-mer sequences. Results for $n > 8$ are based on 10^5 randomly sampled n-mers.

Robustness to translational frame-shift errors

How did such near optimality for parallel codes evolve? One possibility is that the ability to include parallel codes within protein-coding sequences conferred a selection advantage during the early evolution of the genetic code. Alternatively, the genetic code might have been fixed in evolution before most parallel codes existed. We therefore sought a different selection pressure on the code, which could have existed in the early stages of the evolution of the genetic code. One such inherent feature of protein translation is frame-shift translation errors (Parker 1989; Farabaugh and Bjork 1999; Seligmann and Pollock 2004). In these errors, the ribosome shifts the reading frame, either forward or backward. This results in a nonsense translated peptide, and usually loss of protein function. These errors occur in ribosomes nearly as frequently as misread errors (3×10^{-5} per codon, compared with misread errors of 10^{-4} per codon [Parker 1989]). These errors have a relatively large effect on fitness because they result in a nonsense polypeptide. Frame-shift errors may thus pose a selectable constraint on the genetic code: Codes that are able to abort translation more rapidly following frame-shift errors have an advantage (Seligmann and Pollock 2004).

To abort translation after a frame shift, the ribosome must encounter a stop codon in the shifted frame. It has been suggested that codon usage in some organisms may be biased toward codons that can form stop codons upon translational frame shift (Seligmann and Pollock 2004). Here, we consider whether robustness to translational frame-shift errors may be linked to the structure of the genetic code. We tested all alternative codes for the mean probability of encountering a stop in a frame-shifted protein-coding message. We find that the real genetic code encounters a stop more rapidly on average than 99.3% of the alternative codes (Fig. 3). The real code aborts translation eight codons ear-

lier than the average alternative code (15 codons vs. 23 codons). Conservative estimates suggest that such a difference, equivalent to a relative fitness advantage of about 10^{-4}, is readily selectable (see Methods).

Interestingly, the ability to abort translation after frame shift is closely related to the ability to include arbitrary parallel codes (Fig. 4). Robustness to frame-shift errors occurs because the frame-shifted codons for abundant amino acids overlap with the stop codons, hence increasing the probability that stop is encountered upon frame shift. As mentioned above, it is precisely this property that allows the real genetic code to include arbitrary sequences within protein-coding regions, including those with stop sequences, with a significantly higher probability than alternative codes.

The present optimality features are shared also by almost all of the nonuniversal codes such as those found in mitochondria (Osawa et al. 1992; Knight et al. 2001) (see Supplemental material). For example, the fraction of alternative genetic codes with higher probabilities for encountering frame-shifted stop codons is lower then 0.05 for all nonuniversal codes except for the flatworm mitochondrial code (see Supplemental Table 3). It is also found for a range of different codon usages (Muto and Osawa 1987), specifically those that represent GC content of <70% (see Supplemental material). This range of GC contents is also the range that supports the optimality of previously known features such as robustness to misread errors (Archetti 2004).

Discussion

In summary, we found that the genetic code is nearly optimal for encoding additional information in parallel to its main function of encoding for the amino acid sequence of proteins. This optimality is related to the identity of the stop codons in the universal code: when frame shifted, the stop codons overlap with codons of abundant amino acids. We showed that this optimality is strongly tied to a second useful property—minimization of the effect of translational frame-shift errors.

Robustness to frame-shift errors may be a reasonable inherent constraint on the early genetic code. One may therefore propose that the ability to carry parallel codes may have emerged as a side effect that was later exploited to allow genes and mRNA molecules to support a wide range of signals to regulate and modify biological processes in cells (Kirschner et al. 2005). Alternatively, the ability to include arbitrary parallel sequences within coding regions may have contributed to the selection of the early genetic code. For example, early RNA molecules that had the ability to both specify peptides and to include sequences that conferred useful RNA structure may have had an advantage over RNAs that were less effective at simultaneously fulfilling both objectives.

Whereas many of the currently known regulatory codes reside in nontranslated regions of the genome (Robison et al. 1998; Lieb et al. 2001), the present findings support the view that protein-coding regions can carry abundant parallel codes. It would be interesting to use information-theoretical approaches (Gusev et al. 1999; Wan and Wootton 2000; Troyanskaya et al. 2002) to search for such codes in genomes.

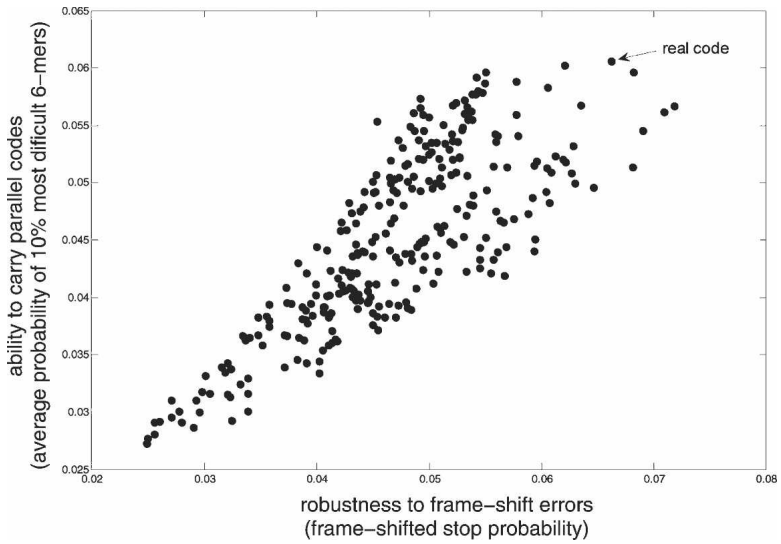

Figure 4. The parallel coding property is strongly tied to the translational frame-shift robustness property. Each point represents one of the alternative codes. The *x*-axis shows the probability of encountering a stop codon upon a frame-shifted event (average over +1 and −1 frame shift). The *y*-axis is the average probability of appearance of the 10% most difficult 6-mers. The arrow indicates the real code. The correlation between the two properties is 0.8. The real code is on the Pareto front, meaning that no alternative code is better than the real code in both properties. Similar results are obtained for n-mers of other sizes. Note that due to symmetries in the alternative codes with respect to the features studied (Supplemental material), multiple alternative codes often have the same values.

Methods

Alternative genetic codes

The alternative genetic codes were obtained by independently permuting the nucleotides in the three codon positions while preserving the amino acid assignment (Fig. 1). These permutations preserve both the number of codons per amino acid and the effect of misread errors on the translated protein, as defined in Freeland and Hurst (1998) and Gilis et al. (2001) (Fig. 1E,F). There are 4! = 24 possible permutations of the four nucleotides. There are, therefore, $24^3 = 13{,}824$ alternative codes. We additionally impose the wobble constraint for base pairing in the third codon position, which states that any two codons differing only in U-C in the third letter cannot be distinguished by the translation apparatus (Crick 1968; Osawa et al. 1992). This results in two allowed permutations in the third letter: the identity permutation and the A↔G permutation. The ensemble of alternative codes therefore contains $24 \times 24 \times 2 = 1152$ codes. In the Supplemental material, we show that relaxing the wobble constraint does not change any of the present conclusions (Supplemental Fig. 1).

Inclusion of arbitrary sequences within protein-coding sequences

We calculated the probability of encountering every n-mer in a coding sequence for each alternative code for $n = 4$–25. This was done by scanning all codon combinations in all three possible frame shifts, which can include the n-mer sequence, and summing the probabilities

of the codon combinations (Fig. 2A). Codon probabilities were calculated from average amino acid probabilities encountered in proteins of sequenced genomes (Pe'er et al. 2004) and uniform codon usage (Table 1). The probability of k consecutive codons $c_1,c_2, \ldots, c_k$ is: $P(c_1,c_2, \ldots, c_k)=\Pi p(a(c_i))/N(a(c_i))$, where $p(a(c_i))$ is the average frequency within protein-coding sequences of the amino acid assigned in the real code to the codon c_i, taken as an average over the amino acid probabilities in the proteome of 134 organisms (Pe'er et al. 2004), and $N(a(c_i))$ is the number of codons assigned to that amino acid. The same results were found when using estimated amino acid frequencies for early genetic code development (Brooks et al. 2004), as well as when amino acid frequencies were varied around their mean frequency with a standard deviation of up to 70% of the mean. Adding correlations between consecutive codons does not change the present results (Supplemental material).

For each code we calculated the average logarithm of the probabilities of all n-mer sequences. To avoid singularities, a small number ε was added to all probability values before taking the logarithm (the results do not depend on ε). The *P*-value is the fraction of alternative codes for which the average logarithm of all n-mer probabilities is higher than in the real code (Table 2). Note that the average logarithm measure is appropriate to situations in which many n-mers need to be independently encoded, so that the product of their probabilities is the biologically significant parameter (e.g., distinct sequences within an RNA that affect stability typically have an approximately multiplicative effect on the total stability of the RNA [Zuker and Stiegler 1981]).

In addition to an average of the logarithm of all n-mer probabilities, for each alternative genetic code we calculated the arithmetic average probability of obtaining the fraction x of n-mers, sorted from the most difficult to the easiest (lowest to highest probability). For every x, we assigned a *P*-value to the real code, which is the fraction of alternative codes for which the average probability of the x most difficult n-mers is equal or higher than in the real code (Supplemental Fig. 4). Table 2 shows the *P*-value for the average probability of obtaining the x most difficult n-mers for different n-mer sizes, with x = 20%. The values of x for which small *P*-values are found increases with the size of the n-mers under consideration (see below).

The FDR method was used to determine the range of difficult n-mers for which the average probability in the real code is significantly higher than in alternative codes, with a threshold that corresponds to a false discovery rate of 15% (Supplemental Fig. 4). For $n > 8$, the calculations were based on 10^5 randomly sampled n-mers.

We find that in the real code, all sequences with $n \leq 10$ can appear within protein-coding sequences, a feature shared by 37% of the alternative codes. For $n > 10$, some sequences cannot appear, since they contain nonoverlapping stop codons in each of the three reading frames (such as the 11-mer UAANUAANUAA).

The probability of encountering a frame-shifted stop

For each alternative code, we calculated the probability of encountering any one of the three stop codons following a frame-shift event. For this we examined all of the possible 61×61 di-codon combinations. A frame-shifted stop upon a +1 translational frame-shift codon can be encountered at positions 2–4 of a di-codon. A frame-shifted stop upon a -1 translational frame-shift codon can be encountered at positions 3–5 of a di-codon. The overall $+1/-1$ frame-shift stop probabilities were obtained for each code by summing the probabilities of all di-codons containing a stop signal at the appropriate position. Codon probabilities were calculated from average amino acid probabilities encountered in proteins of sequenced genomes (Pe'er et al. 2004) and uniform codon usage (Table 1). The present results also apply for a wide range of codon usages (Supplemental material).

Selection pressure of frame shift errors

A translational frame-shift event is estimated to occur at a probability of about 1/30,000 codons (Parker 1989; Farabaugh and Bjork 1999). The average alternative genetic code encounters a stop signal 23 codons on average after a frame-shift event (Fig. 3A), whereas in the real code a stop is encountered 15 codons on average after such an event. It is believed that the number of peptide bonds produced per unit time is one of the main selection pressures in growing microorganisms (Dekel and Alon 2005; Wagner 2005a; Alon 2006). The real code "saves" about 23–15 = 8 extra peptide bonds for each 30,000 translated peptide bonds, conveying an advantage of $8/30{,}000 \sim 2 \times 10^{-4}$, and hence, saving 0.02% of the peptide bonds made by the organism. This relative fitness advantage is much higher than minimal selectable fitness differences in microorganisms (Wagner 2005a), which is on the order of 10^{-7} to 10^{-8}. The reduction in length of the frame-shifted peptide can also have additional beneficial effects, such as reducing potential toxicity of the nonsense peptide and reducing the chances of misfolded protein aggregates. It is possible that frame-shift errors could have been even more common in the early translation apparatus in which the genetic code evolved (Woese 1998).

Acknowledgments

We thank James Shapiro for suggesting this problem, Orna Man for the amino acid probabilities data, and Tsvi Tlusti, Eran Segal, Liran Shlush, and all members of our lab for useful comments. We thank Minerva, HFSP and the Kahn family Foundation for support. S.I. acknowledges support from the Horowitz Complexity Science Foundation.

References

Alon, U. 2006. *An introduction to systems biology*. CRC Press, London, UK.
Archetti, M. 2004. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *J. Mol. Evol.* **59:** 258–266.
Brooks, D.J., Fresco, J.R., and Singh, M. 2004. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics* **20:** 2251–2257.
Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **3:** 285–298.
Crick, F.H. 1968. The origin of the genetic code. *J. Mol. Biol.* **38:** 367–379.
Dekel, E. and Alon, U. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436:** 588–592.
Di Giulio, M. 2005. The origin of the genetic code: Theories and their relationships, a review. *Biosystems* **80:** 175–184.
Draper, D.E. 1999. Themes in RNA-protein recognition. *J. Mol. Biol.* **293:** 255–270.
Dufton, M.J. 1997. Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? *J. Theor. Biol.* **187:** 165–173.
Farabaugh, P.J. and Bjork, G.R. 1999. How translational accuracy influences reading frame maintenance. *EMBO J.* **18:** 1427–1434.
Freeland, S.J. and Hurst, L.D. 1998. The genetic code is one in a million. *J. Mol. Evol.* **47:** 238–248.
Freeland, S.J., Knight, R.D., Landweber, L.F., and Hurst, L.D. 2000. Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17:** 511–518.
Gilis, D., Massar, S., Cerf, N.J., and Rooman, M. 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol.* **2:** research0049.
Gusev, V.D., Nemytikova, L.A., and Chuzhanova, N.A. 1999. On the

complexity measures of genetic sequences. *Bioinformatics* **15:** 994–999.
Hasegawa, M. and Miyata, T. 1980. On the antisymmetry of the amino acid code table. *Orig. Life* **10:** 265–270.
Katz, L. and Burge, C.B. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **13:** 2042–2051.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kirschner, M., Gerhart, J.C., and Norton, J. 2005. *The plausibility of life: Resolving Darwin's dilemma*. Yale University Press, New Haven, CT.
Knight, R.D., Freeland, S.J., and Landweber, L.F. 2001. Rewiring the keyboard: Evolvability of the genetic code. *Nat. Rev. Genet.* **2:** 49–58.
Konecny, J., Schoniger, M., Hofacker, I., Weitze, M.D., and Hofacker, G.L. 2000. Concurrent neutral evolution of mRNA secondary structures and encoded proteins. *J. Mol. Evol.* **50:** 238–242.
Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28:** 327–334.
Muto, A. and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci.* **84:** 166–169.
Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A. 1992. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56:** 229–264.
Parker, J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.* **53:** 273–298.
Pe'er, I., Felder, C.E., Man, O., Silman, I., Sussman, J.L., and Beckmann, J.S. 2004. Proteomic signatures: Amino acid and oligopeptide compositions differentiate among phyla. *Proteins* **54:** 20–40.
Robison, K., McGuire, A.M., and Church, G.M. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284:** 241–254.
Satchwell, S.C., Drew, H.R., and Travers, A.A. 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191:** 659–675.
Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772–778.
Seligmann, H. and Pollock, D.D. 2004. The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* **23:** 701–705.
Shpaer, E.G. 1985. The secondary structure of mRNAs from *Escherichia coli*: Its possible role in increasing the accuracy of translation. *Nucleic Acids Res.* **13:** 275–288.
Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.
Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. *Bull. Math. Biol.* **51:** 417–432.
Troyanskaya, O.G., Arbell, O., Koren, Y., Landau, G.M., and Bolshoy, A. 2002. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics* **18:** 679–688.
Wagner, A. 2005a. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22:** 1365–1374.
Wagner, A. 2005b. *Robustness and evolvability in living systems*. Princeton University Press, Princeton, N.J.
Wan, H. and Wootton, J.C. 2000. A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. *Comput. Chem.* **24:** 71–94.
Woese, C. 1998. The universal ancestor. *Proc. Natl. Acad. Sci.* **95:** 6854–6859.
Woese, C.R. 1965. Order in the genetic code. *Proc. Natl. Acad. Sci.* **54:** 71–75.
Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.

Received September 22, 2006; accepted in revised form November 29, 2006.