July 2006

Is there a twenty third amino acid in the genetic code?

Alexey V. Lobanov
University of Nebraska - Lincoln

Gregory V. Kryukov
University of Nebraska - Lincoln

Dolph L. Hatfield
National Institutes of Health, Bethesda, MD

Vadim Gladyshev
University of Nebraska - Lincoln, vgladyshev1@unl.edu

Genome Analysis

Is there a twenty third amino acid in the genetic code?

Alexey V. Lobanov[1], Gregory V. Kryukov[1], Dolph L. Hatfield[2] and Vadim N. Gladyshev[1]

[1]Department of Biochemistry, University of Nebraska, Lincoln, NE 68588, USA
[2]National Cancer Institute, National Institutes of Health, Bethesda, MD 20892 USA

The universal genetic code includes 20 common amino acids. In addition, selenocysteine (Sec) and pyrrolysine (Pyl), known as the twenty first and twenty second amino acids, are encoded by UGA and UAG, respectively, which are the codons that usually function as stop signals. The discovery of Sec and Pyl suggested that the genetic code could be further expanded by reprogramming stop codons. To search for the putative twenty third amino acid, we employed various tRNA identification programs that scanned 16 archaeal and 130 bacterial genomes for tRNAs with anticodons corresponding to the three stop signals. Our data suggest that the occurrence of additional amino acids that are widely distributed and genetically encoded is unlikely.

Introduction

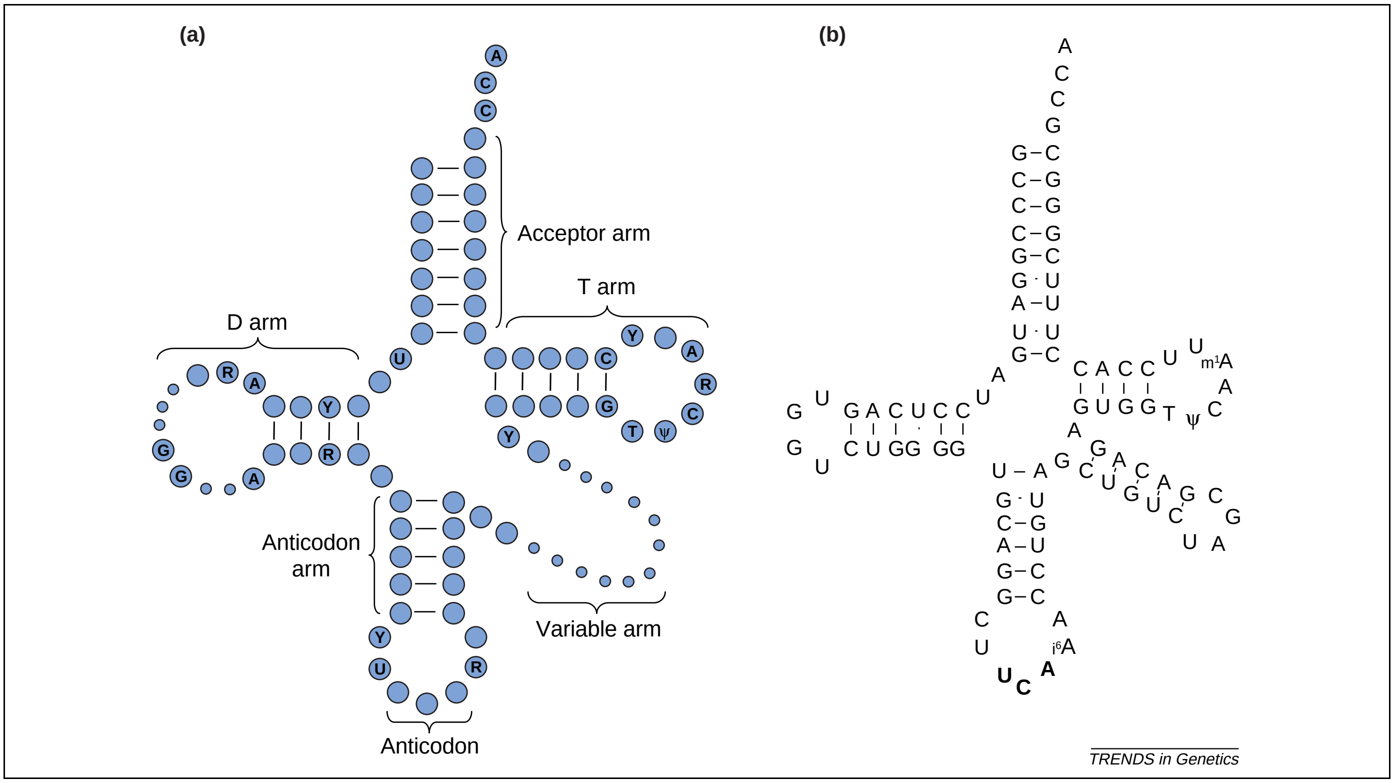
Even the most diverged organisms use the same set of 20 canonical amino acids for *de novo* synthesis of virtually all proteins. Although numerous amino acids resulting from post-translational modifications occur in mature proteins, only two additional amino acids joined the ‘exclusive club’ of genetically encoded amino acids that are incorporated into nascent polypeptide chains specifically and co-translationally. Selenocysteine (Sec), regarded as the twenty first amino acid (reviewed in Ref. [1]), was the first addition to the genetic code since this code was deciphered in the 1960s [2]. Sec is used in all three domains of life (bacteria, archaea and eukaryotes), suggesting that its origin predates their separation. It is inserted into nascent polypeptides in response to TGA codons and this process depends on unique *cis*- and *trans*-acting factors that help recode TGA from stop to Sec insertion. In particular, Sec insertion is dependent on an unusual Sec tRNA containing a TCA anticodon (Figure 1).

Although the co-translational nature of Sec insertion made this amino acid a true addition to the genetic code, for many years Sec was regarded as the only exception. However, four years ago pyrrolysine (Pyl), the twenty second amino acid, was discovered [3,4]. Pyl is encoded by TAG, another codon that usually functions as a stop signal. Pyl insertion requires Pyl-specific tRNA containing a CTA anticodon. However, distribution of Pyl is much more limited compared with that of Sec. Currently, only one bacterium and five archaea, all of which are methanogens, are known to use this residue. It is also clear that both Sec and Pyl traits can be transferred to other organisms. Sec can emerge by rare horizontal gene transfer events [5], whereas Pyl can be programmed by providing exogenous Pyl and expressing Pyl tRNA and the corresponding amino acid synthetase in *Escherichia coli* [6].

The discoveries of Sec and Pyl revealed that the extension of the genetic code to include these twenty first and twenty second amino acids required only a few new genes. Moreover, various unnatural amino acids can be incorporated into protein [7] with relative ease, thus posing an important question – is there a natural twenty third, currently undiscovered, amino acid in the genetic code? The most characteristic feature that distinguishes co-translationally incorporated amino acids is the

Corresponding author: Gladyshev, V.N. (vgladyshev1@unl.edu).

Available online 19 May 2006



TRENDS in Genetics

Figure 1. The cloverleaf tRNA structure. **(a)** Conventional cloverleaf structure. The invariant and semi-invariant nucleotides (A, C, G, T, U, Ψ, R and Y) are indicated. Nucleotides shown as smaller circles are neither conserved nor required for function. **(b)** Sec tRNA from *Bos taurus* (GenBank accession number X74110). Abbreviations: Ψ, pseudouridine; R, purine; Y, pyrimidine.

presence of tRNAs specific for these amino acids. Identification of these non-canonical tRNAs could be a direct way of finding amino acids generated as an extension of the genetic code (Box 1).

A search for the twenty third amino acid

To search for the putative twenty third amino acid, we employed currently available programs (Box 1) and developed a program (details are in supplementary information) to identify most (if not all) tRNA-like structures that might be missed if standard approaches are used (supplementary Table 1 online). This tool is based on RNABOB 2.1 (http://www.genetics.wustl.edu/eddy/software/)*, and is a highly sensitive program but has low selectivity; however, the data were still manageable. We analyzed 130 bacterial and 16 archaeal genomes (supplementary Table 2 online) using these programs. Although the number of false positives in the initial step of the search was significantly greater than that of the currently available tRNA identification programs, our program was more efficient in detecting Sec and Pyl tRNA than either default version of tRNAscan-SE or ARAGORN [8,9]. All tRNAs could also be found in the 146 genomes with the COVE model and appropriate matrix and threshold values (see supplementary Tables 1 and 2 online).

ARAGORN and tRNAscan-SE (in the default modes) are well suited for genome-wide tRNA searches because they are fast, selective and possess reasonably high sensitivity. In our searches, ARAGORN performed slightly better than tRNAscan-SE (supplementary Table 3 online). However, the users cannot adjust the parameters, thus decreasing the flexibility of the searches for unusual tRNAs. Additional modes are available in tRNAscan with the 'COVE only' mode being the most sensitive. Currently, all known tRNAs can be identified using this mode. The run time of the 'COVE only' model is comparable with that of our search tool, although both approaches require significant computational resources. A potential drawback of the 'COVE only' mode is that it relies on tRNA-specific matrices (e.g. see supplementary Table 1 online), which might not identify tRNAs that are significantly different from the known tRNAs.

Although our program did not use matrices that were specific for any particular tRNA, it did identify all Sec and Pyl tRNAs. Some of these tRNAs were misannotated in sequence databases; however, each detected candidate corresponded well to the consensus Sec or Pyl tRNA structures and its occurrence matched that of Sec- or Pyl-containing proteins and the corresponding systems for biosynthesis of these amino acids. All other putative candidates could be reliably filtered. Thus, these searches did not identify any additional tRNA that could insert non-canonical amino acids.

Is there a twenty third amino acid?

There would seem to be three reasons why novel tRNAs corresponding to a hypothetical twenty third amino acid were not identified in our analysis. First, such an amino acid might not exist in the 146 analyzed genomes.

* RNABOB: a program to search for RNA secondary structure motifs in sequence databases was written by Sean Eddy.

Box 1. How to identify a new amino acid?

A search for non-canonical tRNAs seems to be the most direct way of identifying new genetically-encoded amino acids. Currently, the *de facto* standard program for tRNA gene identification in genomic sequences is tRNAscan-SE [8]. An alternative computer program for detection of tRNA and tmRNA genes, ARAGORN, is also available [9]. Both tools perform well in recognition of standard tRNAs [10], however, atypical structures can be missed. For example, *Methanosarcina barkeri* Pyl tRNA [3] was not detected by either ARAGORN or tRNAscan-SE using their default settings (this tRNA was first discovered using a relatively slow 'maximum sensitivity' mode of tRNAscan-SE). As a result, this tRNA is absent in the genomic tRNA database (http://lowelab.ucsc.edu/GtRNAdb), which contains tRNAs identified with tRNAscan-SE. Apparently, many other non-canonical tRNAs are missing in the current genome annotations.

A tRNA for a new amino acid, if such an amino acid exists, might have an unusual primary sequence and secondary structure, as can be seen in Sec and Pyl tRNAs (e.g. a long variable arm and several unprecedented features such as a 9-bp acceptor arm, a 4-bp T-arm stem and a 6-bp D-arm stem in Sec tRNA). To identify such tRNAs, a program is required that is extremely sensitive, even if this is at the cost of having low selectivity and many false positives relative to other programs. The more relaxed settings can be compensated in part by the use of existing genomic annotations (e.g. filtering out known genes). Finally, the search can be restricted to identifying tRNAs corresponding to stop codons. Indeed, if one of the 61 codons for the 20 standard amino acids is used to encode the twenty third amino acid, while preserving its canonical function, the dual meaning of the codon will probably slow down the translation process, whereas this is less of an issue when the insertion of an unusual amino acid competes with the stop signals or a stop signal is completely reprogrammed to encode a new amino acid. Both Sec and Pyl are coded by stop signals, consistent with this logic. Suppressor tRNAs, although having a similar anticodon, have high nucleotide sequence similarity to canonical tRNAs and therefore can be distinguished from novel tRNAs.

If any novel tRNA gene is identified in a particular genome, the next step would be to search for ORFs containing the corresponding in-frame stop codons in the same genome. Sec and Pyl again can be used as true positives because these amino acids are the only known additions to the genetic code and are inserted at UGA and UAG codons, respectively, by the corresponding tRNAs. Both Sec and Pyl are considered as late additions to the genetic code that were acquired by recoding one of the stop codons.

Second, the twenty third amino acid might have a narrow phylogenetic distribution. Although our program detected Pyl tRNA, which was present in only five of the 146 genomes analyzed, a tRNA with the distribution comparable to that of the Pyl tRNA might have been missed during our analysis because the corresponding genomes were not among those included in the search. The twenty third amino acid might be used by organisms inhabiting isolated and difficult to reach environments untapped by the previous sequencing programs, such as an ocean floor or the subglacial lake Vostok (in Antarctica). Third, some of our assumptions could be incorrect. Our approach does not recognize tRNAs with introns and tRNAs in which the anticodon is edited to yield a target anticodon. Moreover, we assumed that the novel tRNA should share the overall cloverleaf structure with canonical tRNAs. However, a novel tRNA could be so dissimilar to the canonical tRNAs that even our high sensitivity program could fail to detect these sequences. There is also a possibility that one of the 61 non-stop codons is exclusively used to code for a novel amino acid in some of the organisms. In this case, novel tRNAs would also be missed in our analysis.

Thus, our data cannot be used to exclude the possibility that additional non-canonical tRNAs exist. However, considering the performance of our search tool in regard to Sec and Pyl tRNA detection, we can conclude that if the distribution of the twenty third amino acid is at least comparable to that of Sec, it would have been easily discovered in our searches. Fully sequenced genomes still represent only a fraction of the enormous prokaryotic diversity. If there is a twenty third amino acid, it is possible that it is used by organisms whose genomes have yet to be analyzed.

Although we did not identify non-canonical tRNAs, our search strategy and the new tool can be applied to additional genomes (including eukaryotic genomes and environmental genome projects which have not been analyzed in this study) to identify tRNAs that can insert novel genetically-encoded amino acids. This procedure should also be useful in identifying Sec and Pyl tRNAs in genomic sequences.

Concluding remarks

The number of completely sequenced genomes has increased dramatically in recent years. Since its development in 1997, tRNAscan-SE has become a tool of choice for tRNA prediction. This program is sensitive and highly selective, and all currently known tRNAs could be found using its 'maximum sensitivity' mode with the correct matrix. However, it might miss unusual tRNAs for which no matrix has been developed. To identify such tRNAs, we developed an alternative approach that, while characterized by decreased selectivity, could efficiently recognize non-canonical tRNAs. We scanned multiple available prokaryotic genomes with the purpose of detecting tRNA genes that might encode a putative twenty third amino acid. Although the program detected all known tRNAs that recognize stop codons, no additional atypical tRNAs were found. These data suggest that if a twenty third amino acid exists, it is likely to have limited distribution. Our program, however, should prove useful in examining newly sequenced genomes for the presence of non-standard amino acids within the genetic code and for annotating Sec and Pyl tRNAs.

Acknowledgements

This work was supported by GM061603 (to V.N.G.).

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2006.05.002

References

1 Hatfield, D.L. and Gladyshev, V.N. (2002) How selenium has altered our understanding of the genetic code. *Mol. Cell. Biol.* 22, 3565–3576

2 Nirenberg, M. (1965) Protein synthesis and the RNA code. *Harvey Lect.* 59, 155–185

3 Srinivasan, G. *et al*. (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* 296, 1459–1462

4 Hao, B. *et al*. (2002) A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* 296, 1462–1466

5 Romero, H. *et al.* (2005) Evolution of selenium utilization traits. *Genome Biol.* doi:10.1186/gb-2005-6-8-r66 (http://genomebiology.com/2005/6/8/R66)

6 Blight, S.K. *et al*. (2004) Direct charging of tRNA(CUA) with pyrrolysine *in vitro* and *in vivo*. *Nature* 431, 333–335

7 Cropp, T.A. and Schultz, P.G. (2004) An expanding genetic code. *Trends Genet.* 20, 625–630

8 Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964

9 Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16

10 Sprinzl, M. and Vassilenko, K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* 33, D139–D140

0168-9525/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2006.05.002