Pyrrolysine and Selenocysteine Use Dissimilar Decoding Strategies

Yan Zhang
University of Nebraska - Lincoln, yzhang3@unl.edu

Pavel V. Baranov
University of Utah, Salt Lake City, Utah

John F. Atkins
University of Utah, Salt Lake City, Utah

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

Follow this and additional works at: http://digitalcommons.unl.edu/biochemgladyshev

Part of the Biochemistry, Biophysics, and Structural Biology Commons

http://digitalcommons.unl.edu/biochemgladyshev/71

This Article is brought to you for free and open access by the Biochemistry, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Vadim Gladyshev Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Pyrrolysine and Selenocysteine Use Dissimilar Decoding Strategies

Yan Zhang*, Pavel V. Baranov§, John F. Atkins§,‡, and Vadim N. Gladyshev*,†

*Department of Biochemistry, University of Nebraska–Lincoln, Lincoln, Nebraska 68588-0664
†Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112
‡Biosciences Institute, University College Cork, Cork, Ireland
Corresponding author: Tel.: 402-472-4948; Fax: 402-472-7842; Email: vgladyshev1@unl.edu

Abstract: Selenocysteine (Sec) and pyrrolysine (Pyl) are known as the 21st and 22nd amino acids in protein. Both are encoded by codons that normally function as stop signals. Sec specification by UGA codons requires the presence of a cis-acting selenocysteine insertion sequence (SECIS) element. Similarly, it is thought that Pyl is inserted by UAG codons with the help of a putative pyrrolysine insertion sequence (PYLIS) element. Herein, we analyzed the occurrence of Pyl-utilizing organisms, Pyl-associated genes, and Pyl-containing proteins. The Pyl trait is restricted to several microbes, and only one organism has both Pyl and Sec. We found that methanogenic archaea that utilize Pyl have few genes that contain in-frame UAG codons, and many of these are followed by nearby UAA or UGA codons. In addition, unambiguous UAG stop signals could not be identified. This bias was not observed in Sec-utilizing organisms and non-Pyl-utilizing archaea, as well as with other stop codons. These observations as well as analyses of the coding potential of UAG codons, overlapping genes, and release factor sequences suggest that UAG is not a typical stop signal in Pyl-utilizing archaea. On the other hand, searches for conserved Pyl-containing proteins revealed only four protein families, including methylamine methyltransferases and transposases. Only methylamine methyltransferases matched the Pyl trait and had conserved Pyl, suggesting that this amino acid is used primarily by these enzymes. These findings are best explained by a model wherein UAG codons may have ambiguous meaning and Pyl insertion can effectively compete with translation termination for UAG codons obviating the need for a specific PYLIS structure. Thus, Sec and Pyl follow dissimilar decoding and evolutionary strategies.

Abbreviations: Pyl, pyrrolysine; Sec, selenocysteine; MtmB, monomethylamine methyltransferase; MtbB, dimethylamine methyltransferase; MtB, trimethylamine methyltransferase; pylT, tRNA\textsuperscript{Pyl} gene; pylS, pyrrolysyl-tRNA synthetase; SECIS, selenocysteine insertion sequence; PYLIS, pyrrolysine insertion sequence; ORF, open reading frame; UTR, untranslated region; RF1, class I release factor; RF2, release factor 2; selA, Sec synthase gene; SelB, Sec-specific elongation factor; EFSec, eukaryotic Sec-specific elongation factor; sefC, tRNA\textsuperscript{Sec} gene; SelD, selenophosphate synthetase; SBP2, SECIS-binding protein 2; nt, nucleotide(s).

INTRODUCTION

Pyrrolysine (Pyl) has recently been identified in the active site of monomethylamine methyltransferase (MtmB) from Methanosarcina barkeri, and sequences encoding Pyl-containing homologs of this protein were found in several other methanogenic archaea, including Methanosarcina acetivorans, Methanosarcina mazei, and Methanosarcina thermophila (1–3). Methylamine methyltransferase genes from these organisms contain in-frame UAG codons, which do not halt translation, but encode Pyl. Following this discovery, additional Pyl-containing methyltransferases have been identified in Methanosarcina, and to date three classes of Pyl-containing methylamine methyltransferase genes are known: mtmB, dimethylamine methyltransferase (mttB), and trimethylamine methyltransferase (mtiB) (1, 2). Some Methanosarcina contain several paralogs of each methyltransferase family. Using this information, various genome sequences were scanned for genes encoding homologous Pyl-containing proteins. This search identified an mtiB homolog in a Gram-positive bacterium Desulfotobacterium hafniense (2). More recently, an Antarctic archaeon, Methanococcoides burtoni, has also been reported to utilize Pyl (4). In contrast, no Pyl-containing methyltransferases have been reported in eukaryotes. It is also not known whether the utilization of Pyl is restricted to methyltransferases or other Pyl-containing proteins exist.

Although the mechanism of Pyl biosynthesis and incorporation into protein is not fully understood, the presence of a Methanosarcina tRNA\textsuperscript{Pyl} gene (pylT) with the CUA anticodon and of class II aminoaic-tRNA synthetase gene (pylS) argued for cotranslational incorporation of Pyl (2). A recent study suggested that pylT and pylS are the only foreign genes necessary for translating UAG as Pyl in Escherichia coli, when the cells are supplemented with exogenous Pyl (5). In addition, it was reported that PylS could activate and ligate Pyl directly onto tRNA\textsuperscript{Pyl} (5, 6) and that tRNA\textsuperscript{Pyl} is directly recognized by the standard elongation factor EF-Tu (7). Analysis of the genomic context of pylT and pylS identified pylB, pylC, and pylD, which were suggested to participate in Pyl biosynthesis or insertion into protein (2). pylT, -S, -B, -C, and -D genes constitute a Pyl gene cluster (or Pyl operon), and pylT and pylS genes are considered as the Pyl utilization signature.

Because Pyl is inserted in response to a codon that in most organisms functions as a terminator, there are three distinct possibilities for how Pyl insertion can be achieved: (i) redefinition of a subset of UAG stop codons by a cis-acting mRNA signal to encode Pyl; (ii) realignment of all UAG codons to encode Pyl; and (iii) ambiguous meaning of UAG codons, e.g. a competition between readthrough and termination such that a fraction of ribosomes translating the UAG codon incorporate Pyl, whereas the rest support termination (8). However, the attention of researchers has previously focused on the first possibility, because of the analogy between Pyl and sele-
nocysteine (Sec) (9). Both Pyl and Sec are encoded by “termination” codons and are the only known additions to the pool of 20 universal, directly encoded, amino acids. Therefore, Sec and Pyl are known as the 21st and 22nd amino acids.

The mechanism of Sec insertion is known in much detail (10–13). Incorporation of Sec requires the presence of selenocysteine insertion sequence (SEICIS) element, a hairpin structure residing in 3'-untranslated regions (3'-UTRs) of selenoprotein mRNAs in eukaryota and archaea, or immediately downstream of Sec UGA codons in eubacteria (13–15). SEICIS is essential for Sec insertion, whereas in its absence UGA serves as terminator (16). Several attempts have been made to search for analogous stem-loop structures in mRNAs encoding Pyl-containing proteins. A putative secondary structure was predicted 5–6 nucleotides downstream of the Pyl-encoding UAG codon. This predicted structure has not been tested experimentally for functional relevance.

Identification of genes encoding Sec- and Pyl-containing proteins in genomic sequences is challenging, because standard annotation tools interpret UGA and UAG as stop signals. For example, most methylamine methyltransferases in Methanosarcina are incorrectly annotated. At present, no tools are available for prediction of Pyl-containing proteins, and previous in silico approaches were limited to manual analyses and BLAST searches (2). In the case of Sec, tools have been developed and successfully used to identify selenoprotein genes by searching for SECIS elements (18–20) and Sec/Cys pairs in homologous sequences (21, 22).

In this study, we used bioinformatics approaches to analyze Pyl-utilizing organisms and Pyl-containing proteins, and to examine possible mechanisms of Pyl insertion. Our data suggest that indiscriminately Pyl insertion at UAG may be tolerated in Pyl-utilizing archaea and that Pyl decoding processes are different from those of Sec.

**EXPERIMENTAL PROCEDURES**


Identification of Pyl Gene Cluster Homologs and Known Pyl-containing Proteins—pylT and pylS sequences from M.arkeri (accession number AY064401 [GenBank]) were used as queries to search genomic databases for Pyl-utilizing organisms with blastn and blastp. A non-Pyl-utilizing archaeon, Methanococcus jannaschii, was also analyzed using the same approach. It served as the control in searches involving Pyl-utilizing archaea.

Analysis of Overlaps between Elongated UAG-containing Genes and Downstream Genes—Overlaps between genes are common in prokaryotic genomes (26). To examine how extensions of UAG-containing genes relate to the extent of the overlap, we analyzed overlapping genes before and after sequence elongation downstream of predicted stop codons in M. acetivorans and M. mazei. A simple Perl script was developed for this analysis (available upon request). We first identified overlapping genes in the original genome annotations, determined the number of overlaps in each genome, and measured overlap lengths. The longest overlap in a genome was defined as an overlap threshold. We then extended genes terminated at UAG until the next non-UAG stop signal using the approach described above and repeated the overlap analysis procedure. We reasoned that if no significant increase in the number of genes whose overlap was longer than the threshold would be observed, the situation would be consistent with the use of UAG as either a terminator or a Pyl codon. However, if the sequence extension procedure generated many genes with large overlaps with the downstream genes, the situation would be consistent with the use of UAG codon as terminator. In addition, the gene overlaps involving genes terminated at UAA and UGA codons were analyzed using the same strategy (e.g. before and after extension). These served as controls.

Identification of Genes Associated with Pyl Utilization—All predicted ORFs in M. acetivorans and M. mazei genomes were searched for exclusive occurrence in genomes that utilize Pyl. The tblastn program was used to search these sequences against 260 completely sequenced prokaryotic genomes, non-redundant nucleotide data base and unfinished microbial contigs with an e value below 0.05. A simple script was developed to parse the tblastn output and examine presence/absence of homologs in analyzed genomes. A pairwise alignment tool, blast2seq, was then used with an e value cutoff set to 0.001 to cluster protein sequences into different families. The occurrence of these proteins in D. hafniense was then analyzed.

**RESULTS AND DISCUSSION**

Distribution of Pyl-utilizing Organisms and Pyl-containing Proteins

Available completely and incompletely sequenced prokaryotic genomes were screened for tRNA^Pyl (pylT) and pyrrolysyl-tRNA synthetase (pylS) sequences, and their patterns of occurrence were compared with those of other Pyl genes (pylB, -C, and -D). We found that the products of pylB (biotin synthase homolog) and pylC (carbamoyl-phosphate synthetase homolog) have close homologs in a wide variety of organisms. In contrast, pylS, pylT, and pylD (nucleoside-diphosphate sugar epimerase homolog) are specific for methanogenic archaea and D. hafniense (Figure 1). In D. hafniense, pylSn and pylSc encode the N- and C-terminal parts of PylS (2), and small overlaps occur between M. burtonii, and D. hafniense. This procedure was implemented using simple Perl scripts (available upon request). First, genes terminating with UAG were extracted from the original annotation files and extended until the next non-UAG stop signal (UAA/UAG). ORFs translated from the elongated genes were analyzed by tblastn against non-redundant and microbial genome databases. We also screened for conservation of UAG codons in nucleotide sequences and of UAG-flanking regions in protein sequences. This procedure assigned each UAG codon to one of three categories as follows: (i) A UAG was interpreted as a terminator if an elongated sequence was sufficiently long (>30 nucleotides), and all of its homologs had a true stop signal (either a non-UAG terminator in Pyl-utilizing organisms or any termination signal in other organisms) that corresponded to the UAG codon. (ii) A UAG was interpreted as a candidate Pyl codon if an elongation was >30 nt, and >50% homologs extended beyond the UAG and terminated near the termination site of the elongated gene. All identified sequences were then analyzed for conservation of UAG in Pyl-utilizing organisms with blastn and blastp. (iii) A UAG was not assigned a function if the two situations discussed above could not be satisfied (for example, if we observed short elongations beyond UAG codons, lack of sequence similarity between homologs in regions flanking UAG, or a small number of homologs extending beyond the UAG).

A non-Pyl-utilizing archaeon, Methanococcus jannaschii, was also analyzed using the same approach. A simple Perl script was developed for this analysis (available upon request). We first identified overlapping genes in the original genome annotations, determined the number of overlaps in each genome, and measured overlap lengths. The longest overlap in a genome was defined as an overlap threshold. We then extended genes terminated at UAG until the next non-UAG stop signal using the approach described above and repeated the overlap analysis procedure. We reasoned that if no significant increase in the number of genes whose overlap was longer than the threshold would be observed, the situation would be consistent with the use of UAG as either a terminator or a Pyl codon. However, if the sequence extension procedure generated many genes with large overlaps with the downstream genes, the situation would be consistent with the use of UAG codon as terminator. In addition, the gene overlaps involving genes terminated at UAA and UGA codons were analyzed using the same strategy (e.g. before and after extension). These served as controls.

Identification of Genes Associated with Pyl Utilization—All predicted ORFs in M. acetivorans and M. mazei genomes were searched for exclusive occurrence in genomes that utilize Pyl. The tblastn program was used to search these sequences against 260 completely sequenced prokaryotic genomes, non-redundant nucleotide data base and unfinished microbial contigs with an e value below 0.05. A simple script was developed to parse the tblastn output and examine presence/absence of homologs in analyzed genomes. A pairwise alignment tool, blast2seq, was then used with an e value cutoff set to 0.001 to cluster protein sequences into different families. The occurrence of these proteins in D. hafniense was then analyzed.
pylT and pylSc (4 nt) and between the pylB and pylC genes (52 nt). pylS, pylT, and pylD always cluster with pylB and pylC, and the overall Pyl gene cluster has identical mutual organization of these sequences, except that D. hafniense pylSn is located at the end of the cluster (Figure 1). Thus, the five Pyl genes define the Pyl gene cluster, but only pylT and pylS (and perhaps pylD) sequences can be used as the signature for the Pyl trait.

A search of completely and incompletely sequenced prokaryotic genomes for Pyl genes revealed only six organisms that could utilize Pyl, including four members of Methanosarcina genera, M. burtonii, and D. hafniense. Methanosarcina species and M. burtonii belong to Methanosarcinales, suggesting that Pyl is encoded by a UAG codon in a restricted group of phylogenetically related organisms that occupy a specific environmental niche. High conservation of the Pyl gene cluster and the small number of organisms that utilize Pyl suggest its relatively recent origin.

In the 6 Pyl-utilizing organisms, a total of 29 Pyl-containing methylamine methyltransferase genes was identified (Table I). They are distributed in three enzyme families that do not share significant sequence similarity. Figure 2 shows the occurrence of these genes in genomes and contigs. Only mttB genes cluster with the Pyl operon genes (in three Methanosarcina organisms). In M. mazei, two distant duplicate mttB genes are present. In M. barkeri, two duplicate mttBl genes cluster together and are on the opposite strands with the Pyl cluster.

Further analyses of the three methylamine methyltransferase protein families revealed conservation of Pyl in MtmB and MtIbB (i.e. no MtmB and MtIbB homologs were detected, in which Pyl is replaced with other residues or in which the Pyl-encoding UAG codon is replaced with a non-UAG stop signal). On the other hand, multiple MtIbB homologs were detected, in which Pyl is not conserved and replaced with various amino acids (Figure 3). This situation is in contrast to Sec, which is highly conserved. In addition, most selenoproteins have homologs, in which Sec is replaced with cysteine (Cys). In fact, the Sec/Cys pair in homologous sequences is a feature that is used for identification of selenoproteins in genomic databases (21, 22).

Phylogenetic analyses of the three methylamine methyltransferase families as well as of the pylT and pylS genes typically placed D. hafniense genes as outliers (Figure 4). In Pyl-utilizing archaea, all mttB, mttB, and mttB genes encode Pyl-containing proteins. Conversely, D. hafniense possessed mttB genes encoding proteins with and without Pyl. MttB homologs that did not have Pyl were broadly distributed in other bacteria.

Although the use of Pyl appears to be prevalent in methanogenic archaea, from our data it could not be established with certainty whether the Pyl trait evolved in these organisms or in bacteria. Both Pyl operon sequences and mttB have optimal codon usage in organisms in which they are present (data not shown), arguing against a recent (traceable) lateral transfer of the Pyl trait between methanogenic archaea and D. hafniense.

**Analyses of Candidate PYLIS Elements**

By analogy to SECIS, a stem-loop structure in selenoprotein mRNAs that reprograms a small fraction of UGA codons in a genome to serve in Sec insertion (11, 13), the occurrence of putative PYLIS structures was proposed (9, 17). However, we found that the RNA structures downstream of UAG codons in several mttB and mttB mRNAs that were recently suggested as putative PYLIS elements (17) are dissimilar and do not occur in mttB. Manual analyses of sequences downstream of UAG codons, as well as sequences in UTRs revealed no obvious common structure shared by members of all three methylamine methyltransferase families. A possibility remains that structures downstream of UAG codons inhibit termination.

**Functions of UAG Codons in Pyl-utilizing Organisms**

Because UGA has a dual function in Sec-utilizing organisms (Sec insertion and translation termination), there is a possibility that UAG, in a similar fashion, also serves two functions (Pyl insertion and translation termination) in Pyl-utilizing organisms. If PYLIS is absent, what might be a mechanism for discriminating between the two functions of UAG codons? An alternative possibility would be a reassignment of all UAG codons from stop to Pyl. To address these possibilities, we analyzed the distribution of the three stop codons in Pyl-utilizing organisms (Table II). In D. hafniense, the percentage of the genes predicted to terminate at UAG is 22.5%, which is similar to the proportion of genes predicted to terminate at UGA (28.4%). The extensive utilization of UAG in this bacterium suggests that this codon likely has a dual role in protein synthesis.

In contrast, the proportion of genes that are predicted to terminate at UAG in Pyl-containing archaea is <5.0%. For example, only 126 (including all incorrectly annotated methylamine methyltransferases) of 3,371 genes are predicted to use UAG terminator in M. mazei. This value is much lower than the proportion of UAA or UGA terminators, suggesting that UAG might be functionally distinct from UAA and UGA in these archaea.

**FIGURE 1. Pyl gene cluster (operon).** The Pyl gene cluster includes five neighboring genes: pylT, pylS, pylB, pylC, and pylD. In archaeal genomes, these genes show identical organization. In D. hafniense, two genes (pylSn and pylSc) code for the N- and C-terminal parts of PylS, and there are small overlaps between pylT and pylSc and between the pylB and pylC genes as shown.
We further used a homology-based search strategy to examine the coding potential of UAG codons as an additional test to characterize the function of these codons in archaea. Our approach was to extend the reading frames of all genes predicted to contain UAG codons until the next non-UAG stop signal. For each elongated UAG-containing gene, the tblastn program was used to identify candidate homologs in other organisms and to examine the conservation of UAG-flanking regions within translated sequences. We reasoned that if a sequence is sufficiently extended (>30 nt) beyond the UAG, and all of its homologs in other organisms have true stop signals that corresponded to the UAG codon (that is, only the sequence upstream of UAG is conserved, whereas sequence similarity is absent downstream of UAG), the UAG should be a terminator (Figure 5A). Using this strategy, we could reliably identify UGA stop signals in Sec-utilizing organisms and distinguish them from UGA codons for Sec (data not shown).

Conversely, if the extension is long (>30 nt), and most candidate homologs extend beyond the UAG to end near (or after) the site corresponding to the non-UAG stop codon in the elongated sequence, the UAG codon in the sequence of interest is considered a Pyl codon candidate. In addition to testing UAG function, this strategy could also be used for identification, in Pyl-encoding organisms, of candidate Pyl-containing proteins. To avoid the possibility of dealing with a sequencing error or a pseudogene, we required the presence of sequences encoding a candidate Pyl-containing protein in two or more genomes of Pyl-utilizing organisms (Figure 5B). In this case, methyamine methyltransferases served as true positives, because they can be extended beyond their Pyl UAG codons, share homology with other proteins in sequences downstream of their UAG codons, and occur in at least 4 Pyl-utilizing organisms as Pyl-containing forms (Figure 3).

In other situations (see Figure 5C for specific examples), we could not distinguish between Pyl-encoding functions and stop signals. However, if the UAG was followed with a nearby stop codon, either Pyl insertion or translation termination could presumably be tolerated.

We further extended the reading frames of all genes predicted to contain UAG codons until the next non-UAG stop signal. For each elongated UAG-containing gene, the tblastn program was used to identify candidate homologs in other organisms and to examine the conservation of UAG-flanking regions. We reasoned that if a sequence is sufficiently extended (>30 nt) beyond the UAG, and all of its homologs in other organisms have true stop signals that corresponded to the UAG codon (that is, only the sequence upstream of UAG is conserved, whereas sequence similarity is absent downstream of UAG), the UAG should be a terminator (Figure 5A). Using this strategy, we could reliably identify UGA stop signals in Sec-utilizing organisms and distinguish them from UGA codons for Sec (data not shown).

Conversely, if the extension is long (>30 nt), and most candidate homologs extend beyond the UAG to end near (or after) the site corresponding to the non-UAG stop codon in the elongated sequence, the UAG codon in the sequence of interest is considered a Pyl codon candidate. In addition to testing UAG function, this strategy could also be used for identification, in Pyl-encoding organisms, of candidate Pyl-containing proteins. To avoid the possibility of dealing with a sequencing error or a pseudogene, we required the presence of sequences encoding a candidate Pyl-containing protein in two or more genomes of Pyl-utilizing organisms (Figure 5B). In this case, methyamine methyltransferases served as true positives, because they can be extended beyond their Pyl UAG codons, share homology with other proteins in sequences downstream of their UAG codons, and occur in at least 4 Pyl-utilizing organisms as Pyl-containing forms (Figure 3).

In other situations (see Figure 5C for specific examples), we could not distinguish between Pyl-encoding functions and stop signals. However, if the UAG was followed with a nearby stop codon, either Pyl insertion or translation termination could presumably be tolerated.

Surprisingly, among all the genes with predicted in-frame UAG codons in the Pyl-utilizing archaea, we could not detect a single unambiguous candidate containing UAG as its terminator (Table III). Instead, we found that in most genes either UAG codons are followed with additional nearby non-UAG stop signals (40.3% of UAGs), or UAG-containing genes can be extended to generate conserved sequence alignments downstream of UAGs (24.3%). The former situations cannot
distinguish between termination and Pyl insertion, whereas the latter correspond to candidate Pyl-containing proteins. Interestingly, these searches revealed only four protein families, including three known methylamine methyltransferase families and a family of transposases, as candidate Pyl-containing proteins encoded in at least two genomes. UAG-containing forms of transposases were identified in *M. acetivorans* (4 sequences) and *M. mazei* (16 sequences), and all had a single UAG codon at the same position. These data suggest that transposases are a family of novel Pyl-containing proteins (Figure 5B).

However, only methylamine methyltransferases were found in three or more genomes, and these enzymes provided the best match to the Pyl trait. This observation suggests that Pyl utilization was conserved during evolution for its use by methylamine methyltransferases. At least in *MtmB*, Pyl is located at the enzyme active site, where it was suggested to be directly involved in catalysis serving as a strong electrophile (3). It cannot be excluded that additional, species-specific proteins that use catalytic Pyl residue occur in organisms capable of Pyl insertion. However, our searches suggest that the Pyl trait is associated exclusively with methylamine methyltransferases and not with any other family of Pyl-containing proteins.

The bias against unambiguous assignment of a fraction of the occurrences of a codeword as terminator that we observed in Pyl-utilizing archaea, was not seen in either Sec-utilizing organisms (data not shown) or the Pyl-decoding bacterium *D. hafniense*. In *D. hafniense*, we detected a number of “true” UAG stop signals (221 hits, 19.6% of all UAG codons). In addition, 33 *D. hafniense* sequences, including one incorrectly annotated *mttB* gene, have a predicted in-frame UAG codon. However, except for *mttB*, these are present in single copies and so are likely false positives. Overall, these findings are consistent with the hypothesis that UAG has a dual function in *D. hafniense*, but no evidence for such a dual function was observed in archaea.

Using the same approach, we analyzed a non-Pyl-utilizing archaeon, *Methanococcus jannaschii*, in which UAG is known to have unambiguous stop codon meaning. Like *Methanosarcina*, *M. jannaschii* has a small number of UAG codons (UAG: UGA:UAA = 164:222:1343). However, we detected 52 genes (31.7% of all UAGs), in which these UAGs could be classified as true stop signals (Figure 5A). Only 5 genes showed sequence homology downstream of their UAG, but these were represented by single UAG-containing sequences. Thus, there is a clear difference in the use of UAG codons between Pyl-utilizing and non-Pyl-utilizing archaea (Table III).
Partially overlapping genes are common in prokaryotic genomes (26, 27), however, most overlaps are short. We reasoned that if UAG codons function as terminators, extensions of these sequences downstream of UAGs would result in some overlaps that are abnormally long. In contrast, if UAG codons function in Pyl insertion, the extended sequences would not result in overlaps or would produce mostly short overlaps with the downstream genes.

To examine these possibilities, we analyzed overlaps involving previously annotated genes and genes whose ORFs were extended beyond the annotated stop signals for all three stop codons in *M. acetivorans* and *M. mazei* (Table IV). The annotated genes showed a maximum of a 98 nucleotide overlap in *M. acetivorans* and a 241 nucleotide overlap in *M. mazei*, and these values were set as thresholds to evaluate the overlaps obtained after the genes were extended to the next stop signal.

We found that, among all UAG-extended genes, 27 *M. acetivorans* and 43 *M. mazei* genes overlap with downstream genes that are in different reading frames (genes that overlapped in the same reading frame were excluded as these produced gene fusions rather than overlaps and could correspond to Pyl-containing proteins). However, in *M. mazei*, the lengths of all overlaps are below the threshold (241 nt). In *M. acetivorans*, only 3 genes have overlaps longer than the threshold (98 nucleotides), but these (including the longest overlap of 172 nucleotides) are still shorter than the *M. mazei* threshold. Thus, extension of UAG-containing genes until the next stop signal does not modify the overall composition of the overlapping genes.

In contrast, extension of UAA- and UGA-containing genes generated many genes that overlap above the thresholds (e.g. 117 genes for UAA and 195 genes for UGA in *M. acetivorans*), with some overlaps being above 1000 nt (the longest overlap is 1934 nt for UGA in *M. acetivorans*, Table IV). Thus, both the number of abnormal gene extensions and the longest observed overlaps involving UGA- and UAA-containing sequences are much higher than those of the UAG-containing sequences, suggesting that the usage of UAA and UGA is functionally distinct from that of UAG in Pyl-utilizing archaea and that read-through of UAG codons should be tolerated better than read-through of UAA or UGA stop codons.

### Analysis of Class I Release Factors in Archaea

Most eukaryotes and archaea possess a single class I release factor (RF1) responsible for the recognition of all stop codons in mRNA (28, 29). Structural studies of human RF1 suggest that its N-terminal domain interacts with mRNA during stop codon recognition (30). Numerous studies demonstrated that the sequences near the highly conserved NIKS motif in this domain are critical for stop codon recognition (31–33). If UAG codons were reassigned to code for Pyl in Pyl-containing archaea, their release factors would have specific changes responsible for the alteration of their specificity. We have searched for release factors in all sequenced Euryarchaeota. Unexpectedly, we found that two Pyl-utilizing archaea, *M. acetivorans* and *M. barkeri*, encode two non-identical RF1s. Figure 6 shows a phylogenetic tree of RF1s from Euryarchaeota. All RF1s from Pyl-containing archaea have specific amino acid changes in the area surrounding the NIKS motif (Figure 7). Interestingly, in organisms with two release factors, additional

---

**Table II.** Distribution of annotated UAA, UAG, and UGA codons in the genomes of Pyl-utilizing organisms

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (nt)</th>
<th>Annotated genes</th>
<th>Stop codon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>UAG</td>
</tr>
<tr>
<td>Archaea</td>
<td></td>
<td></td>
<td>224 (4.9%)</td>
</tr>
<tr>
<td><em>M. acetivorans</em></td>
<td>6,514,492</td>
<td>4,540</td>
<td>126 (3.7%)</td>
</tr>
<tr>
<td><em>M. mazei</em></td>
<td>2,690,345</td>
<td>3,371</td>
<td>131 (4.7%)</td>
</tr>
<tr>
<td><em>M. barkeri</em></td>
<td>&lt;5.6 (unfinished)</td>
<td>2,792</td>
<td>1127 (22.5%)</td>
</tr>
</tbody>
</table>

---

**Figure 4.** Phylogenetic analysis of methyltransferases, *pytT* and *pytS*. A, phylogenetic analysis of methyltransferase genes. mtB homologs that do not encode Pyl are shown in red. B, phylogenetic analysis of *pytT* and *pytS* genes.
RF1s contain amino acid substitutions in the NIKS motif itself. These data suggest that RF1s from Pyl-containing archaea may have indeed evolved to change their stop codon recognition specificity.

Several organisms are known, in which stop codons are reassigned to sense codons (34, 35). Although in bacteria (in which there are two semi-specific class I release factors) stop codon reassignment involves loss of one release factor, in eukaryota stop codon reassignment involves changes in the specificity of release factors. For example, in the ciliate *Euplotes*, UGA was reassigned to encode Cys (36), and it has been shown that its RF1 does not recognize UGA stop codons (37).
Interestingly, there are also two non-identical RF1s in *Euplotes* (38, 39). An attractive hypothesis is that codon reassignment involves extensive mutational alteration of release factors, including duplication of their corresponding genes. If so, the fact that there are two release factors in some Pyl-encoding archaea can be also used to support the proposition that UAG function might be partially or fully reassigned from stop to Pyl.

### Putative Mechanisms for Discrimination between Pyl Decoding and Stop Signal Functions of UAG Codons

The several lines of evidence described above imply that, at least in Pyl-decoding archaea, UAG codons do not function as standard terminators. Instead, there appears to be either a complete reassignment of UAG codons from stop to Pyl, or a competition between Pyl insertion and translation termination that favors Pyl insertion.

Reassignment of codon meaning has been documented for genetic codes of several prokaryotic and eukaryotic organisms, and it is highly common in genetic codes of organelles (34, 40, 41). Most frequently stop rather than sense codons are reassigned. This is probably because it is less deleterious for a cell, because stop codon usage is lower than the usage of sense codons. In addition, terminators generally do not correspond to crucial parts of the protein as is the case with many sense codons. In case of UAG, a complete reassignment of this terminator to Pyl codon does not seem to result in adverse functional consequences as there only a few UAG codons in Pyl-utilizing archaea and many of these would simply be slightly extended at the C terminus if UAG codes for Pyl.

The second possibility is that the meaning of UAG is ambiguous and specifies both termination and Pyl, making UAG the polysemous codon in methanogenic archaea. There is one known example of such ambiguous meaning of a codon. In several members of the *Candida* species a standard leucine (Leu) CUG codon is translated as serine (Ser). tRNA<sub>Ser</sub>, which has 1-methylguanosine at position 37, can be charged with either both in vitro and in vivo and incorporate both Ser and Leu at CUG codons in these organisms (42, 43). A recent genetic exercise to manipulate the editing activity of *E. coli* isoleucyl-tRNA synthetase has generated a strain of *E. coli* where Cys is misincorporated at isoleucine (Ile) codons (44). This results in generation of “statistical proteins” where individual members of a protein pool have either Cys...
or Ile in the same positions. It has been shown that generation of heterogeneous proteins could have a growth yield advantage and be beneficial at some steps of evolution (42–44).

An attractive possibility is that the competition between termination and Pyl insertion for UAG codons is influenced by physiological stimuli, e.g., this is a regulated competition or a partial reassignment of UAG codons. Considering that only one new family of candidate Pyl-containing proteins (transposases) could be identified, and it occurred only in two archaea, methyamine methyltransferases emerged as the key enzymes that utilize Pyl. MtmB and MtB families of methyamine methyltransferases strictly conserve Pyl, whereas the occurrence of Pyl-containing mttB methylamine methyltransferases strictly matches that of the Pyl trait. Thus, it seems likely that Pyl insertion is maintained, because it is required for use by these enzymes. It is possible that the efficiency with which UAG is decoded as Pyl depends on specific environmental conditions, such as growth on methyamines (Figure 8). In this case, methyamines may activate Pyl biosynthesis and expression of methyamine methyltransferase genes. Under these conditions, Pyl insertion may out-compete translation termination at many or all UAG codons. However, because methyamine methyltransferases are extremely abundant enzymes (1–3), Pyl will primarily serve the methyltransferase UAG codons.

It is also possible that the rate with which Pyl is inserted at UAG codons is gene-specific, e.g., it depends on strength or leakiness of the UAG codon as a terminator. It is known that stop signals are extended elements with upstream and downstream sequences contributing to efficiency of decoding, terminating, and read-through functions (45–47). These stop-codon-flanking sequences were reported to cross-link with release factors and influence termination efficiency (48, 49). Thus, the context of UAG codons may be an important feature that influences the outcome of UAG decoding. However, the small number of true Pyl-containing proteins (e.g., methyltransferases) was insufficient for us to make definitive conclusions regarding the presence of sequence signals than flank UAG and discriminate between the two coding functions of UAG codons.

If metabolites influence competition between Pyl insertion and termination, in the absence of methyamines, Pyl might not be synthesized, and termination at UAG codons may prevail, because there would be no Pyl-tRNA\textsuperscript{Pyl} to compete with termination. If so, many UAG codons could serve as stop signals under certain growth conditions (e.g., no methyamines), but will insert Pyl when methyamines are present. This hypothesis is also consistent with the identical pattern of occurrence of methyamine methyltransferase genes (mtmB and mtbB) and the Pyl cluster.

In any case, the data suggest that there appears to be no need for a highly specific PYLIS element in archaeal Pyl-containing mRNAs. For example, if UAG codons are primarily decoded as Pyl, there might be only a small fraction of truncated forms of methyltransferases due to competition with termination. In addition, insertion of Pyl at other UAG codons should be well tolerated, because this would result in only slight increases in protein masses. At the same time, the small number of UAG-containing genes in Pyl-decoding archaea suggests that Pyl is not a common amino acid in protein in these organisms. Presumably, either utilization of Pyl could affect functions of some proteins, or the flux through the Pyl biosynthetic pathway could not satisfy the demand for this amino acid, if it is to be commonly used. Further research is needed to address these hypotheses.

Searches for Genes Associated with the Pyl Trait

The availability of multiple genomic sequences of Pyl-utilizing organisms provides an opportunity to identify genes associated with the Pyl trait by analyzing patterns of gene occurrence. Our strategy

FIGURE 7. Alignment of RF1 sequences in the region surrounding the NIKS motif. The specific amino acid changes of RF1s from Pyl-containing archaea are highlighted in red.

FIGURE 8. A model of Pyl insertion and termination of translation at UAG codons in Pyl-utilizing archaea. Decoding UAG codons as Pyl may be predominant under conditions that require Pyl (e.g., growth on methyamines), whereas in other situations, UAG codons may serve as terminators. Such competition between Pyl insertion and translation termination may occur in many or all UAG-containing genes, or it may be heavily shifted toward exclusive use of UAG as Pyl codon.
was to test each of the annotated genes present in *M. acetivorans* and *M. mazei* (a total of 7911 genes) for their possible exclusive presence in Pyl-utilizing organisms (Table V). We first analyzed the presence of these genes in Pyl-utilizing archaea and absence in other archaea. This search identified 414 genes, which belonged to 168 protein families and included *mtmB, mtbB, pylS*, and *pylD*. These sequences were then analyzed for their occurrence in *D. hafniense*. Interestingly, only 15 proteins, which belonged to 9 families, were identified, including *PylS, PylD*, and a family represented by NP_632172 [GenBank] in *M. mazei*, which belonged to COG2043 (uncharacterized protein conserved in archaea). Other protein families were represented by hypothetical proteins with no known function or conserved motifs, including NP_633144 [GenBank] in *M. mazei*. The identified sequences may contain proteins associated with the Pyl trait. The possible functions may include roles in Pyl biosynthesis, insertion, or degradation, as well as functions associated with regulation and utilization of Pyl and Pyl-containing proteins. Since 4 out of the 15 proteins were true positives (PylS and PylD in each genome), our approach is clearly capable of identifying Pyl-associated genes. The availability of additional genomes with the Pyl trait will help in further characterization of the identified genes.

### Parallels and Differences in Sec and Pyl Insertion Systems

A hypothesis of a parallel between Sec and Pyl insertion systems was suggested (3, 9, 50) based on similarities at various steps in the two translation pathways. However, our analyses described above argue against this possibility. We further analyzed and compared the known features of Pyl and Sec biosynthesis and insertion (Table VI).

### Distribution of Sec and Pyl Traits—Sec is used by approximately a quarter of prokaryotic organisms, which have been characterized by genome sequencing, and by many eukaryotic organisms, with the exception of yeasts and higher plants (1, 13, 21). At least 26 eukaryotic and 25 prokaryotic selenoprotein families are known; most are present in proteins with distinct structures and functions (20, 21). This relatively widespread use of Sec contrasts with the rare occurrence of Pyl. We compared the Sec and Pyl decoding traits and found that only one organism, *D. hafniense*, utilizes both rare amino acids. Searches against all known selenoproteins identified only one selenoprotein, formate dehydrogenase α subunit, in *D. hafniense* (data not shown). Formate dehydrogenases contain one Sec, which is present in the active site, coordinates molybdenum, and is directly involved in oxidation of formate to carbon dioxide (51, 52). Thus, it appears that *D. hafniense* has only one Sec residue and one Pyl residue in its set of proteins. Whereas the use of Pyl and Sec is limited in *D. hafniense*, this organism appears to use all 22 natural amino acids currently known. The use of both Sec and Pyl by this organism is clearly supported by the presence of corresponding biosynthetic and insertion machineries.

### Comparison of Sec and Pyl Biosynthetic Pathways—One distinctive feature of Sec biosynthesis is that its synthesis occurs on tRNAsec (the selC gene product) (53, 54). The tRNAsec has features that distinguish it from canonical tRNAs, including its length (typically 90 nucleotides; more than any other tRNAs), few post-transcriptional modifications, an unusually long variable arm, and the presence of 13 nucleotides in the acceptor and TΨC stems (13, 55, 56). Serine is initially acylated to tRNAsec by seryl-tRNA synthetase and then the selA gene product, Sec synthase, converts Ser-tRNAsec to Sec-tRNAsec in a two-step reaction (54). The selenophosphate donor for this reaction is synthesized by the selD gene product, selenophosphate synthetase (54,}
Although the biosynthesis of Pyl has not been completely characterized, pylB, -C, and -D genes are candidates to play a role in Pyl synthesis (5, 58–60). Recent data suggest that, like the 20 common amino acids, Pyl is synthesized prior to its attachment to tRNA<sub>Pyl</sub> (5, 56). The pyl<sub>S</sub> gene product is dedicated to acylating Pyl to tRNA<sub>Pyl</sub>. Direct charging of Pyl onto its tRNA contrasts with the tRNA-based pathway in Sec biosynthesis. In addition, like some canonical tRNAs, tRNA<sub>Pyl</sub> lacks the variable stem.

Comparison of Sec and Pyl Insertion Pathways—The mechanism of Sec insertion in response to UGA has been most thoroughly elucidated in <i>E. coli</i> (53, 61, 62). Sec-tRNA<sub>Sec</sub> forms a complex with the Sec-specific elongation factor SelB and GTP, and subsequently binds the SECIS element within ribosome-bound mRNAs (10). The resulting quaternary complex directs the insertion of Sec at in-frame UGA codons (54, 63, 64). The mechanism of Sec insertion in archaea and eukaryotes differs from that in bacteria in two aspects: (i) in bacteria, the SECIS element is located immediately downstream of the inframe UGA codon, but it is present in 3'-UTRs in archaea and eukaryotes; (ii) SelB in bacteria binds both Sec-tRNA<sub>Sec</sub> and SECIS, whereas in eukaryotes (and possibly archaea), SelB (EFSec) binds Sec-tRNA<sub>Sec</sub> and is associated with SECIS via SECIS-binding protein 2 (SBP2) (13, 65, 66). Release factor RF2 can recognize UGA efficiently if any component of the Sec incorporation machinery is missing (45).

Understanding the mechanism of Pyl insertion has lagged behind that of Sec. As discussed above, we neither could identify true UAG stop codons nor find conserved stem-loop structures (PYLIS) immediately downstream of the in-frame UAG codons or in 3'-UTRs of methionyltransferase genes in methanogenic archaea. Our data, viewed as a whole, argue for differences between Sec and Pyl insertion pathways.

CONCLUSIONS

The nature of the processes that extend the universal genetic code is still being debated. In recent years, many non-standard amino acids in proteins have been identified (67–69), but almost all are formed by post-translational modifications. Sec and Pyl are the only two exceptions identified to date. Both are encoded by canonical stop codons using specific tRNAs. These properties may have evolutionary significance in regard to the extension of the genetic code (8, 50).

In this study, various genomic sequences were scanned to identify and scrutinize Pyl-utilizing organisms and Pyl-containing proteins. Analysis of the distribution of stop codons revealed that UAG is a rare codon in Pyl-encoding archaea, but common in the bacterium <i>D. hafniense</i>. It appears that in <i>D. hafniense</i> UAG codons must serve two functions, stop and Pyl insertion. However, having only a single Pyl-containing protein in this organism precludes computational searches for possible cis-elements that modify the function of the UAG codon. In contrast, in Pyl-utilizing archaea, the lack of unambiguous candidates for UAG stop codons, the relatively large proportion of genes that could only slightly be extended beyond UAG and the acceptable overlaps for extended UAG-containing genes suggest that UAG codons might be decoded as Pyl in many or all UAG-containing genes. These features also suggest that insertion of Pyl could be well tolerated in these organisms, even though the number of proteins that require Pyl for their function is small and might be limited to methyamine methytransferases. Thus, although both Sec and Pyl insertion may compete with termination, there is a difference between decoding strategies employed by Pyl and Sec, with Pyl resembling the standard amino acids and Sec having its own biosynthetic and decoding mechanisms.

Our findings are also consistent with analyses of release factor sequences in archaea as well as with the unsuccessful search for PYLIS elements in methyamine methyltransferase genes. If some discrimination between Pyl decoding and stop signal functions of UAG codons is needed, it could potentially be provided by the context of the UAG codons or structural elements that inhibit termination thus allowing Pyl insertion. Alternatively, UAG codons could serve as Pyl codons under conditions that stimulate Pyl synthesis (e.g., during growth of cells on methylamines) by out-competing the weak terminator function of UAG. However, in situations where methyltransferase genes are not required, Pyl biosynthesis may be inhibited resulting in depletion of the pool of Pyl-tRNA<sub>Pyl</sub>, which would favor termination at UAG codons. For most genes, Pyl insertion or termination may be nearly equivalent, as gene products would only slightly differ in C-terminal regions. Combined with the fact that only a small number of genes utilize UAG codons in archaea, this change in UAG coding function could be well tolerated by these organisms.

Acknowledgments: We thank the Research Computing Facility of the University of Nebraska–Lincoln for the use of the Prairiefire Beowulf cluster supercomputer.

REFERENCES


Zhang et al. in Journal of Biological Chemistry 280 (2005)