Pyrrolysyl-tRNA Synthetase: an ordinary enzyme but an outstanding genetic code expansion tool

Wei Wan, Jeffery M. Tharp, and Wenshe R. Liu*
Department of Chemistry, Texas A&M University, College Station, TX 77845

Abstract

The genetic incorporation of the 22nd proteinogenic amino acid, pyrolysine (Pyl) at amber codon is achieved by the action of pyrrolysyl-tRNA synthetase (PylRS) together with its cognate tRNA\textsubscript{Pyl}. Unlike most aminoacyl-tRNA synthetases, PylRS displays high substrate side chain promiscuity, low selectivity toward its substrate \(\alpha\)-amine, and low selectivity toward the anticodon of tRNA\textsubscript{Pyl}. These unique but ordinary features of PylRS as an aminoacyl-tRNA synthetase allow the Pyl incorporation machinery to be easily engineered for the genetic incorporation of more than 100 non-canonical amino acids (NCAAs) or \(\alpha\)-hydroxy acids into proteins at amber codon and the reassignment of other codons such as ochre UAA, opal UGA, and four-base AGGA codons to code NCAAs.

1. Introduction

With few exceptions, all biological systems code strictly 20 canonical amino acids. Given the limited functionalities of 20 amino acids, biochemists have longed for an approached for the production of proteins with unique biochemical and biophysical handles. Early efforts focused on the chemical aminoacylation of tRNAs that could be applied to undergo \textit{in vitro} translation or injected into cells for the synthesis of proteins with non-canonical amino acids (NCAAs).[1] Pioneered by Schultz and others, engineered aminoacyl-tRNA synthetase (AARS)-tRNA pairs were later introduced to genetically encode NCAAs mainly at amber UAG codon in living cells, making the bulk expression of proteins with NCAAs possible. [2–4] Although artificial systems have been delicately crafted to selectively incorporate NCAAs at amber codon in living cells, nature itself has a very competitive system for reassigning amber codon to code the 22nd proteinogenic amino acid, pyrolysine (Pyl). In this review, we will discuss the Pyl incorporation system from its discovery to its applications in the genetic incorporation of close to 100 NCAAs or \(\alpha\)-hydroxy acids. This review intends to discuss the Pyl system alone. For more general summary of the genetic code expansion technique, please find reviews published elsewhere.
2. Pyl

2.1. Discovery

Although methanogens are phylogenetically diverse throughout the archaea domain, the methanogenic reaction in these organisms is almost universal in reducing a single substrate, carbon dioxide, to methane.[5] However, there is one order of methanogenic archaea, the Methanosarcinales, which has evolved the ability to reduce other compounds, methylamines, to methane. When sequencing clustered genes in *Methanosarcina barkeri* responsible for methanogenesis from monomethylamine (MMA), Krzycki et al. surprisingly discovered that the open reading frame of *mtmB* that codes monomethylamine methyltransferase (MMAMT) is interrupted by a single in-frame canonical amber UAG stop codon that does not appear to prevent translation of the full-length product.[6] MMAMT is a 52 kD polypeptide and is one of the most predominant proteins in MMA-grown cell extracts of *M. barkeri*. The truncated protein with a premature termination at the in-frame amber codon was not as abundant. Since basal levels of nonsense suppression occur at low frequencies to maintain translation integrity inside cells, this observation clearly specified that there is an endogenous but efficient amber suppression mechanism for recoding this amber stop codon in *M. barkeri*. The later sequencing analysis of other methylamine (trimethylamine and dimethylamine) methyltransferase genes in *M. barkeri* MS all showed an in-frame amber stop codon that is also well conserved in methylamine methyltransferase genes in other Methanosarcina strains.[7] This in-frame amber codon is apparently reassigned to code an amino acid that is indicatively critical for functional conservation among methylamine methyltransferases.

The crystal structure of MMAMT was later determined to a high resolution of 1.55 Å.[8] The in-frame UAG-encoded amino acid site displayed a unique electron density that was distinctive from any of the 20 canonical amino acids and selenocysteine (Sec). This electron density was refined to show a lysine (Lys) core with a modification at its side-chain amine. The identity of this modification was determined to be a (4R,5R)-4-methyl-pyrroline-5-carboxylate, with the carboxylate in the amide linkage to the Lys side chain. This novel amino acid was coined as pyrrolysine (three-letter abbreviation: Pyl; one-letter abbreviation: O). Although the installation of Pyl might result from two possible routes, one being a posttranslational modification of a pre-incorporated Lys and the other being co-translational incorporation of Pyl itself, the existence of a unique AARS and a unique tRNA in the *pyl* gene cluster adjacent to a methylamine methyltransferase gene cluster is most consistent with the co-translational installation route.[9] From these observations, Pyl clearly represents the 22nd naturally occurring genetically encoded amino acid to be discovered.

2.2. Incorporation and biosynthesis

Using tRANScanSE[10] to search the *M. barkeri* Fusaro genomic database, Krzycki *et al.* revealed a unique tRNA gene, *pylT*, near a methylamine methyltransferase gene cluster. *PylT* codes a special tRNA with a CUA anticodon.[9] Along with its special CUA anticodon for the recognition of amber codon, the *pylT* transcript, tRNA\textsuperscript{Pyl}, has a distinct anticodon stem of six base pairs instead of five base pairs as observed in most tRNAs, a single base between D and anticodon stems, a single base between D and acceptor stems, and a three-
base small variable arm. PylT resides in the pyl gene cluster that also contains four other genes, pylS, pylB, pylC, and pylD. Sequence alignment indicated that pylS codes a class II AARS but with a low sequence identity to all AARSs in the database. Since Pyl is a Lys derivative, the recombinant pylS gene product was later tested for lysyl-tRNA synthetase (LysRS) activity. In the presence of the tRNA pool from M. barkeri, the recombinant pylS gene product succeeded in ligating [14C]Lys to tRNA. This observation led to the hypothesis that tRNA^Pyl is charged with Lys first and then undergoes modification at the Lys side chain to form Pyl-tRNA^Pyl for the co-translational incorporation of Pyl. However, following experiments with pure tRNA^Pyl clearly showed that the recombinant pylS gene product definitely charged tRNA^Pyl with Pyl instead of Lys.[11–13] Therefore, the pylS gene product is not LysRS and should be named as pyrrolysyl-tRNA synthetase (PylRS);

Although it was original believed that the incorporation of Pyl requires a mRNA secondary structure similarly as observed for the Sec incorporation,[14, 15] this mRNA secondary structure was definitely not necessary for the incorporation of Pyl at an amber codon.[16–19] Therefore, the incorporation machinery of Pyl requires simply an AARS, a tRNA, a codon, and an amino acid, similar to the incorporation systems of the 20 canonical amino acids in most organisms (Figure 1).

Kryzcki et al. also demonstrated that lateral gene transfer of the whole pyl gene cluster from M. barkeri to a naïve organism, Escherichia coli, drove the recombinant expression of functional MMAMT. The expressed MMAMT had an intact Pyl at the in-frame amber stop codon site, indicating the pyl gene cluster codes all necessary enzymes for the biosynthesis of Pyl from E. coli metabolites. The removal of each of the three genes pylB, pylC and pylD dismissed the cluster’s ability to reassign the amber codon to code Pyl. Therefore, pylB, pylC and pylD form an essential gene cassette for the biosynthesis of Pyl.[20] Following studies revealed that pylB codes a radical S-adenosyl-L-methionine enzyme.[21, 22] This enzyme catalyzes the synthesis of (3R)-3-methyl-D-ornithine from Lys whose carboxylate is then ligated to a Lys side chain amine by the action of the pylC gene product, a member of the carbamoyl phosphate synthetase family with the expense of adenosyl 5′-triphosphate (ATP). The synthesized (3R)-3-methyl-D-ornithyl-Nε-L-Lys then undergoes the oxidation at the δ-amine of its (3R)-3-methyl-D-ornithine component by the action of the pylD gene product, a NAD-dependent dehydrogenase, a simultaneous hydrolysis to from a δ-aldehyde, and then a simultaneous intramolecular condensation to form Pyl, finishing the whole biosynthetic pathway (Figure 2).[23] When D-ornithine was provided in the growth medium, the clustered genes of pylT, pylS, pylC, and pylD were able to mediate amber suppression. In this situation, the UAG-coded amino acid was desmethyl-Pyl, indicating pylC and pylD gene products are able to synthesize desmethyl-Pyl from D-ornithine and lysine and that PylRS tolerates alternative substrates.[22, 24, 25]

3. PylRS

Sequence alignment indicated that full-length PylRS contains a C-terminal class II AARS catalytic core and an N-terminal domain that apparently does not share sequence homology with any structurally known protein domains. PylRS proteins from different Methanosarcina strains have variable lengths of N-terminal domains but their C-terminal catalytic cores with about 270 amino acids are highly conserved. Efforts to crystalize full-length PylRS were not
successful due to the low solubility of the protein. The N-terminal domain is highly insoluble and aggregates full-length PylRS.\[26, 27\] Using the truncated C-terminal catalytic core of PylRS from *Methanosarcina mazei*, apo-PylRS and PylRS complexes with different ligands were successfully determined, attributed mostly to contributions from Yokoyama, Söll, Steitz, and their coworkers.\[28–32\] The three dimensional organization of the PylRS catalytic core resembles that of other synthetases from the Class II AARS family. It has a typical β-sheet core surrounded by several helices, forming a Rossmann fold for the binding of ATP. Like most other class II aminoacyl-tRNA synthetases, PylRS also forms an obligate dimer. Each subunit has an active site. Although sharing relatively low sequence identity (<30%), the PylRS catalytic core and bacterial phenylalanyl-tRNA synthetase (PheRS) are structurally similar (Figure 3).\[31, 33–35\] The two structures are highly superimposable and have similarly organized hydrophobic pockets for binding amino acid substrates, possibly indicating divergent evolution of the two enzymes from the same origin.

### 3.1. High substrate side chain promiscuity

The structure of the *M. mazei* PylRS catalytic core revealed a deep hydrophobic pocket for the binding of Pyl (Figure 4). A302, L305, Y306, L309, N346, C348, and W417 form a bulky cavity for the binding of the side chain (4R,5R)-4-methyl-pyrroline-5-carboxylate of Pyl.\[30, 31\] The Pyl side chain also forms two hydrogen bonds at the PylRS active site, with one involving the side chain amide nitrogen of N346 and the Pyl side chain amide oxygen and the other involving the pyrroline nitrogen and the phenolic oxygen of Y384. Y384 is at a flexible loop region that is random in the absence of Pyl but serves as a cap for the binding of Pyl to the active site. Although PylRS binds tightly to Pyl, with a \( K_d \) value (53 μM) relatively comparable to other AARSs, PylRS displays remarkably high tolerance toward variations of the substrate side chain, especially when a variation is at the pyrroline region.\[36\] As discussed above, PylRS recognizes desmethyl-Pyl and is able to direct its incorporation at amber codon when in coordination with tRNA\text{Pyl}. Other NCAAs that are structurally similar to Pyl such as D-prolyl-L-Lys and \( N^\varepsilon \)-cyclopentylloxycarbonyl-L-Lys can also serve as good substrates of PylRS.\[36\] Surprisingly, NCAAs such as \( N^\varepsilon \)-(tert-butyloxycarbonyl)-L-Lys and \( N^\varepsilon \)-allyloxycarbonyl-L-Lys that are structurally dramatically different from Pyl also function well as PylRS substrates.\[29, 37\] This high substrate side chain promiscuity of an AARS is very unique. To maintain high translation fidelity, most AARSs have evolved to specifically interact with their amino acid substrates and kinetically favor the correct tRNA aminoacylation reactions. Many AARSs also contain an editing domain that can hydrolyze misacylated tRNAs.\[38–40\] Therefore, in general, even a slight variation of the amino acid substrate structure will cause a typical AARS to reject it from slipping into the translation process. However, PylRS recognizes the Pyl side chain chiefly with relatively non-specific hydrophobic interactions and in addition it does not have an editing domain, explaining its high substrate side chain promiscuity. This unique but ordinary feature of PylRS as an AARS might be due to relatively relaxed evolutionary pressure in its natural hosts. Unlike the 20 canonical amino acids that are structurally and spatially similar to one another, Pyl is exceptionally large. To the best of our knowledge, non-proteinogenic amino acid metabolites with a similar structure rarely exist in cells. Therefore, there is no selection pressure for PylRS to evolve strict and exclusive recognition
of Pyl. A relatively high Pyl binding affinity of PylRS alone is enough to maintain translation integrity.

3.2. Low selectivity toward α-amine

Another striking feature of PylRS is its low selectivity toward the Pyl α-amine. In the active site, the Pyl α-amine forms two hydrogen bonds, one to the phenolic oxygen of Y384 and the other to a water molecule that has three additional hydrogen bond interactions with the side chain amide oxygen of N346 and two backbone amide nitrogens of L301 and A302 (Figure 5A). Since Y384 is in a flexible loop region that does not fold back to cap the active site until the binding of Pyl, it is possible that the hydrogen bond between the Pyl α-amine and the Y384 phenolic oxygen is flexibly formed and not essential for the catalytic activity of PylRS.[30, 31] This has been supported by the fact that several PylRS mutants with phenylalanine (Phe) or tryptophan at Y384 display higher activities toward their substrates than corresponding Y384-containing enzymes.[37, 41–43] In the PylRS catalytic core, residues that are close to the Pyl α-amine include M300 and A302. However, side chains of these two residues are too distant from the Pyl α-amine to form strong van der Waals contacts. Therefore, its hydrogen bond to the water molecule is probably the only essential interaction for the loose binding of the Pyl α-amine to the active site. Based on the hydrogen bond pattern of the water molecule, we can also conclude that the Pyl α-amine binds to the active site uncharged. Two backbone nitrogens are juxtaposed at proper positions to serve as two hydrogen bond donors for the water molecule. This leaves the water molecule with no vacancy to accept an additional hydrogen bond donor. Therefore the Pyl α-amine has to be uncharged to interact with the water molecule as a hydrogen bond acceptor. Since PylRS doesn’t involve strict recognition of the Pyl α-amine and also binds its uncharged form, presumably changing the α-amine to a α-hydroxide, a neutral amine homolog that can potentially maintain hydrogen bond interactions with both Y384 and the water molecule may not significantly change the activity of PylRS. As a matter of fact, PylRS displays high tolerance toward α-hydroxy acids. It has been demonstrated that several α-hydroxy acids serve as substrates for PylRS and its engineered mutants that, together with tRNA\textsubscript{Pyl}, mediate direct incorporation of these α-hydroxy acids at amber codon in bacterial cells.[44, 45]

The loose recognition of α-amine at the active site of PylRS is in striking contrast to the strict α-amine recognition feature of most AARSs. For instance, PheRS, a structural homolog of PylRS, involves unique α-amine interactions to bind Phe at its active site (Figure 5B). The α-amine forms three hydrogen bonds with side chains of S180, Q218, and E220. Since the side chain of E220 needs to be negatively charged for a strong charge-quadruple interaction with the phenyl side chain of the substrate, there is no doubt that the substrate α-amine is positively charged to generate a very strong charged hydrogen bond (salt bridge) with the E220 side chain. In addition, the positively charged substrate α-amine serves as a hydrogen bond donor in all three hydrogen bond interactions. Both the preferential binding of its charged state and three specific hydrogen bond interactions contribute to the strict recognition of Phe at the PheRS active site. Simply changing the α-substitute to a hydroxide will not only dismiss the strong charged hydrogen bond interaction but also destroy the hydrogen bond pattern in the active site since a neutral hydroxide can
serve as both a hydrogen bond donor and a hydrogen bond acceptor but a charged amine can only be a hydrogen bond donor. This strict recognition of Phe can efficiently ward off the mistaken incorporation of phenyllactic acid into proteins during the translation process. Given that α-hydroxy acids that are structurally similar to canonical amino acids are regular metabolites in cells, most AARSs stay in cellular environments with high evolution pressure to maintain strict recognition of α-amine to prevent cell functions from going awry.[46–50] However, this is probably not the case for PylRS. An α-hydroxy acid with an extended side chain like the one in Pyl may not exist at all in Methanosarcinales, leaving PylRS with no pressure for evolving strict recognition of α-amine.

3.3. Low selectivity toward the tRNA anticodon

The third unique feature of PylRS is its low selectivity toward the tRNA anticodon. A recently determined crystal structure of the Desulfitobacterium hafniense PylRS complex with tRNA\textsubscript{Pyl} showed that PylRS has a unique tRNA-binding domain, a special C-terminal tail, a loop region, and a bulge domain that is involved in the specific recognition of the anticodon, D, and acceptor stems of tRNA\textsubscript{Pyl} (Figure 6).[32] However, D. hafniense PylRS does not directly interact with the three anticodon nucleotides. Although D. hafniense PylRS corresponds only to the C-terminal catalytic core of PylRS from Methanosarcinales, multiple observations have suggested that M. mazei PylRS does not involve specific interactions with the anticodon of tRNA\textsubscript{Pyl}. Mutating the anticodon of tRNA\textsubscript{Pyl} alone has not shown to abolish the aminoacylation activity of M. mazei PylRS.[18, 30] This feature is also in striking contrast to the direct synthetase-anticodon interactions that have been observed for the majority of AARSs. Anticodons in tRNAs are typically used as recognition elements of AARSs for strictly pairing amino acids with their corresponding codons to avoid misincorporation during translation.[51] Two notable exceptions are the E. coli leucyl- and seryl-tRNA synthetases.[52–56] Both leucine and serine have six corresponding codons, multiple isoacceptor tRNAs for pairing with six codons, but only a single corresponding AARS. For one synthetase to recognize multiple isoacceptor tRNAs with distinct anticodon sequences, a reasonable evolutionary route is to avoid direct interactions with the tRNA anticodon region. Therefore the evasion of direct tRNA anticodon interactions for leucyl- and seryl-tRNA synthetases might be an evolution compromise. Since Pyl has only one corresponding codon, one tRNA, and one synthetase, it is quite puzzling why PylRS has not evolved to directly interact with the anticodon of tRNA\textsubscript{Pyl} for strictly pairing Pyl with an amber codon. A single mutation at the anticodon region will lead to the misincorporation of Pyl at another codon, which may be detrimental to cells. This non-selective recognition of the tRNA\textsubscript{Pyl} anticodon by PylRS might be rationalized considering the pyl gene cluster is probably highly regulated and only turned on in the presence of methylamines. Its noncontinuous or even rare expression would leave PylRS with no constant evolution pressure for evolving a more optimized feature.

4. An outstanding genetic code expansion tool

Due to its unique structure, tRNA\textsubscript{Pyl} has been shown to be orthogonal in commonly used bacterial and eukaryotic cells and functional to reassign amber codon when coupled with PylRS in these cell strains. Since it was first reported as a genetic code expansion tool in E.
coli in 2004, the PyrRS-tRNA\textsuperscript{Pyl} pair and its mutant forms have been used to site-specifically incorporate NCAAs into proteins at amber codon mutation sites in \textit{E. coli}, \textit{Saccharomyces cerevisiae}, \textit{Caenorhabditis elegans}, \textit{Drosophila melanogaster}, and mammalian cells.[9, 29, 57–60] In \textit{E. coli}, the genetic incorporation of more than 100 NCAAs and several \(\alpha\)-hydroxy acids have been demonstrated. The PyrRS-tRNA\textsuperscript{Pyl} pair has also shown high codon flexibility as several alternative codons for coding NCAAs have been demonstrated.[11, 61, 62] This remarkable success is largely attributed to the three unconventional features of PyrRS.

4.1 The native enzyme

Shown in Figure 4A, the binding of the side chain pyrroline of Pyl to the PyrRS active site involves essentially van der Waals interactions. Replacing the side chain pyrroline with a similar size chemical component with a hydrophobic nature might retain the PyrRS activity to aminoacylate tRNA\textsuperscript{Pyl}. This potential has been explored extensively. NCAAs including Pyl and desmethyl-Pyl (1–2) which have been demonstrated to be PyrRS substrates are presented in Figure 7. Söll \textit{et al.} pioneered the search for alternative substrates of PyrRS due to the inconvenient synthesis of Pyl and their need of a substrate to characterize PyrRS.[36, 63] They showed that 3 and 4 could both serve as PyrRS substrates with about 10–fold weaker binding affinities than Pyl itself. 4 was also successfully used to demonstrate that a mRNA secondary structure is not necessary for the PyrRS-tRNA\textsuperscript{Pyl} pair to reassign amber codon to code a NCAA.[19] Two NCAAs (5–6) were discovered as PyrRS substrates when Yokoyama \textit{et al.} soaked PyrRS crystals with different ligands for structure determination. Both 5 and 6 show strong interactions with PyrRS and their electron densities can be clearly observed in the active site of PyrRS.[29, 37] 5 is a lysine derivate with a \(N^\varepsilon\)-Boc protecting group that can be easily removed in acidic conditions. Chin \textit{et al.} have taken advantage of this chemistry to synthesize native diubiquitins and histones with a site-selectively installed \(N^\varepsilon,N^\varepsilon\)-dimethyllysine.[64, 65] The generated diubiquitins have been used successfully to reveal Lys29-isopeptide selectivity of an OIT deubiquitinase, TRABID.[65] 6 has a terminal alkene that can potentially undergo thio-ene photoclick, olefin metathesis, Diels-Alder, and 1,3-dipolar cycloaddition reactions. Selective labeling of proteins incorporated with 6 and another terminal alkene-containing PyrRS substrate 7 using the thio-ene photoclick reaction has been demonstrated.[66] Kryzcki \textit{et al.} also sought alternative PyrRS substrates independently. They discovered that 8 maintained about 9% of the activity of Pyl as a PyrRS substrate.[67] Encouraged by the results of 8, three NCAAs (9–11), each with a terminal alkene, were designed and synthesized. They all worked as PyrRS substrates and their incorporation into proteins allowed for site-selective modifications with azide via the Cu(I)-catalyzed azide-alkyne 1,3-dipolar cycloaddition (CuAAC) reaction.[68–70] Concurrent efforts by Chin \textit{et al.} revealed another two click substrates (12–13) of PyrRS. Although their original goal was to select PyrRS mutants that can selectively recognize 12 and 13, the evolution resulted in the native enzyme.[71] 12 and 13 contain alkene and azide, respectively. Their incorporation into proteins allowed specific labeling with azide or alkene dyes using the CuAAC reaction. The tetrazine-alkene Diels-Alder cycloaddition reaction is a click reaction type that has been actively performed in the chemical biology research area due to its rapid reaction nature.[72–74] Chin \textit{et al.} discovered that a norbornene-containing NCAA (14) could be efficiently recognized by PyrRS and incorporated into proteins at
amber codon in both *E. coli* and mammalian cells mediated by the PylRS-tRNA<sub>Pyl</sub> pair. Proteins incorporated with 14 in mammalian cells were rapidly labeled with tetrazine dyes. [75] Two NCAAs (15–16) with a D- or L-cystyl side chain and one NCAA (17) that can be deprotected to recover a L-cystyl side chain can be weakly recognized by PylRS and genetically incorporated into proteins. Their unique abilities to undergo native chemical ligation with a peptide thioester and selective condensation with cyanobenzothiazole have been harnessed to synthesize ubiquitinated proteins and proteins selectively labeled with cyanobenzothiazole dyes.[76–80] Two NCAAs that have an additional methyl group on the side chain carbamate nitrogen also serve as PylRS substrates and have been incorporated into proteins. It has been demonstrated that both carbamate groups in the two NCAAs could be hydrolyzed in relatively mild conditions to recover Nε-methyllysine, allowing the genetic installation of a monomethylated Lys residue in proteins.[81, 82] Recently Carell et al. showed that PylRS weakly recognized three NCAAs (20–22) and mediated their incorporation into proteins at an amber codon in coordination with tRNA<sub>Pyl</sub>, allowing the synthesis of proteins with site-specific lysine propionylation, butylation, and crotonylation. [83] PylRS mutants with much improved binding affinities toward 20–22 were developed concurrently in the Schultz group and in our group. They will be discussed in the next section. Another NCAA (23) that can be recognized by PylRS is a photocrosslinking amino acid.[84] Applications of this NCAA and other photocrosslinking NCAAs will also be discussed in the next section.

### 4.2 Engineered enzymes for lysine derivatives

Although the native PylRS displays very high substrate side chain promiscuity, its activities on certain NCAAs shown in Figure 7 such as 15–17 and 20–22 are weak. Efforts have been made in several research groups to identify PylRS mutants that have improved activities and also recognize Lys derivatives that could be targeted by the native PylRS. The distal location of the deep pyrroline binding pocket to the catalytic center of the enzyme makes the engineering process relatively straightforward without too much concern that a mutation will dramatically affect the enzymatic catalyzed aminoacylation of tRNA<sub>Pyl</sub>. To date, more than 30 lysine derivatives have been genetically incorporated into proteins at amber codon in living cells using engineered PylRS-tRNA<sub>Pyl</sub> pairs. Their structures are presented in Figure 8 and the sequence information of engineered PylRS mutants are shown in Table 1. For NCAAs 6–7, 14–17, and 20–22, all engineered mutants display higher activities for their genetic incorporation than the native PylRS. Other NCAAs in Figure 8 show very weak or undetectable activities toward the native PylRS. Applications of these genetically encoded NCAAs span many research areas. 6–7 and 24 have a terminal alkene that has been demonstrated to undergo thio-ene photoclick protein labeling.[66] 15–17 can be applied to undergo native chemical ligation for the synthesis of ubiquitinated proteins and a biocompatible condensation for protein labeling.[78–80] 25 and 27 has an azide that selectively undergoes the Staudinger ligation, CuAAC click, and Cu-free click reactions.[37, 85] 26 was an intermediate NCAA for the development of a 27-specific PylRS mutant. Both 28 and 31 undergo selective Cu-free click labeling with azide. They have been applied to selectively label proteins incorporated with them in living cells.[86, 87] NCAAs 14 and 28–32 react rapidly with tetrazine, allowing rapid turn-on fluorescent labeling of proteins containing these NCAAs in mammalian cells.[88–90] 33–34 have a terminal alkyne. They
can be well applied for CuAAC reactions with azide and thiol-yne photoclick reactions. They also undergo biocompatible Sonogashira cross-coupling reactions with halogenated benzene derivatives in the presence of palladium catalysts. 35 undergoes a similar cross-coupling reaction.[91–93] Two groups concurrently worked out the genetic incorporation of 36. 36 has an acrylamide functionality that can selectively undergo Michael addition with a thiol, radical polymerization with an alkene, and 1,3-dipolar nitrilimine-alkene cycloaddition. We have demonstrated that a genetically incorporated 36 allowed selectively turn-on fluorescent labeling of proteins in living cells.[42, 94] Two photocrosslinking NCAAs 23 and 37 have been genetically encoded using native and engineered PylRS-tRNA\textsuperscript{Pyl} pairs and applied successfully to covalently crosslink associate proteins. Chen \textit{et al.} has pushed the technique further in using 37 to successfully profile \textit{in vivo} substrates of a major acid-protection chaperone, HdeA, in \textit{E. coli} periplasm.[95–97] Two photocaged lysines 38–39 have been genetically encoded. The genetic incorporation of 39 into mammalian proteins and its following photoactivation have allowed the spatiotemporal control of p53 cellular localization, kinase activity, and gene transcription in human cells.[41, 98–100] Besides 18 and 19, two other NCAAs, 40 and 41 have been genetically encoded using mutant PylRS-tRNA\textsuperscript{Pyl} pairs for the indirect synthesis of proteins with site-specifically installed \(N^\varepsilon\)-methyllysine. In comparison to 18 which needs to be deprotected under acidic conditions, 19 which needs ruthenium catalysts for deprotection, and 40 which can be potentially deprotected using palladium-catalyzed hydrogenation, 41, a photocaged \(N^\varepsilon\)-methyllysine, can be facilely deprotected under UV irradiation. Its deprotection condition is probably the most mild. Using 41 to synthesize proteins incorporated with \(N^\varepsilon\)-methyllysine has been demonstrated.[82, 101, 102] In order to undergo traceless synthesis of diubiquitins with native isopeptide bonds, Chin \textit{et al.} evolved PylRS mutants that can selectively recognize 43 and 46 for their genetic incorporation into ubiquitin. 43 and 46 can selectively remove their \(N^\varepsilon\) protection groups to recover \(\delta\)-mercapto-Lys that undergoes native chemical ligation with ubiquitin thioester to generate diubiquitins. The \(\delta\)-thiol auxiliary group can then be removed. Using this approach, several native diubiquitins were synthesized. In the process of evolving PylRS mutants for 43 and 46, PylRS mutants that recognize 42, 44, and 45 were also identified.[43] Four posttranslationally modified lysines 47 and 20–22 have been genetically encoded using PylRS mutants. PylRS mutants displayed much higher activities for the genetic incorporation of these four NCAAs than the native enzyme. 47 represents the most abundant posttranslational Lys acylation type in mammalian cells. Its genetic incorporation has been applied to study functional roles of Lys acetylation in nucleosome and p53.[103–111] Alternatively, posttranslational lysine modification mimics could be indirectly installed into proteins through the genetic incorporation of 49 followed by oxidation to form dehydroalanine and then Michael addition with \(N\)-substituted cysteamine derivatives.[112] 49 and another NCAA 48 were genetically encoded using a PylRS mutant originally evolved for 26. Another two lysine derivatives that have been incorporated into proteins using evolved PylRS mutants together with tRNA\textsuperscript{Pyl} are 50 and 51.[113, 114] 50 has a furan at its side chain that can be sensitized for the formation of protein-nucleotide complexes and 51 has a spin label that has been applied to measure intramolecular distance distributions of two protein residues.
4.3 Engineered enzymes for Phe derivatives

PyRS displays a high structural similarity to bacterial PheRS, though they share low sequence identities. This high structural similarity made us wonder whether we would be able to evolve PyRS to selectively recognize Phe. The evolution turned out to be very simple. After a PyRS library with randomization at six active site residues was selected for amber suppression ability in E. coli in LB medium, multiple clones responsive to Phe were identified.[115, 116] This result opened a new avenue of the PyRS research area which is to develop PyRS mutants for the genetic incorporation of Phe derivatives. Although a number of Phe derivatives have been genetically encoded in bacteria and eukaryotes using engineered tyrosyl-tRNA synthase-tRNA\textsuperscript{Pyl} pairs that were originally from Methanocaldococcus jannaschii and E. coli,[3, 117–119] the relatively relaxed substrate binding makes the engineering of PyRS more straightforward and the high orthogonally of the native Pyl system in different cell strains also allows the more convenient transfer of engineering Pyl systems between different cells. As a demonstration, we evolved PyRS mutants that can selectively incorporate 53 and 54 at amber codons in E. coli in coordination with tRNA\textsuperscript{Pyl}.[115] When the sequence information of PyRS clones for 52 was available, we noticed two residues N346 and C348 had convergent mutations. N346 was mutated to either alanine or serine and C348 was always mutated to a large amino acid. The structure of the PyRS complex with Pyl-AMP showed one hydrogen bond between the N346 side chain and the Pyl side chain amide. We suspected that removing the side chain of N346 dismisses this hydrogen bond and also leaves space for the binding of Phe. A large residue at C348 might provide strong van der Waals interactions with Phe for its binding to the active site. Therefore, removing the side chain of C348 could potentially leave space for adding substituents to the phenylalanine side chain. We constructed the N346A/C348A mutant of PyRS and tested this mutant on a series of \( p \)-alkoxy-Phe molecules 55–61. They all served as substrates and could be genetically incorporated.[120] Surprisingly this mutant also recognized \( m \)-substituted Phe derivatives 62–73.[121] Encouraged by these results, we synthesized Phe derivatives 74–87 to test the substrate scope of this mutant. All of these NCAAs can be readily taken and selectively incorporated at amber codons in E. coli cells expressing the N346A/C348 mutant and tRNA\textsuperscript{Pyl}.[122] Our most recent efforts also demonstrated that the N346A/C348 mutant recognizes \( o \)-substituted Phe derivatives 88–95 and the N346A/C348 mutant–tRNA\textsuperscript{Pyl} pair is able to mediate site-selective incorporation of these NCAAs into proteins in human cells.[123] NCAAs 52–95 contain reactive functional groups such as alkyne, azide, ketone, and halobenzene. As we demonstrated, their incorporation into proteins allowed the use of multiple click reaction types for selective protein modifications. Similar to the native PyRS, the N346A/C348A mutant represents a single enzyme that allows the installation of a large variety of functionalities into proteins. Efforts to identify PyRS mutants selective for Phe derivatives were also made in other groups. Chin \textit{et al.} showed that a PyRS mutant together with tRNA\textsuperscript{Pyl} was able to genetically incorporate 77 into a tyrosine residue in STAT1 for regulating its phosphorylation.[124] Wang \textit{et al.} showed that PyRS mutants selective for 56, 96, and 97 could be evolved.[125, 126]
4.4 α-Hydroxy acids

Yokoyama et al. first noticed the ability of PylRS to recognize substrates with α-substituents different from amine. In vitro aminoacylation analysis revealed that PylRS was able to charge tRNA^{Pyl} with 5 analogs that were α-hydroxy acid, non-α-amino-carboxylic acid, N\(^{\alpha}\)-methylamino acid, and D-amino acid. By expressing PylRS and tRNA^{Pyl} in E. coli, the UAG amber codon was successfully reassigned to code the α-hydroxy acid analog of 5 (98).[44] In comparison to α-amine which forms a backbone amide bond and serves as a hydrogen bond donor in proteins, α-hydroxy acid forms a relatively weaker ester bond and can solely function as a hydrogen bond acceptor. A readily hydrolysable ester bond in a protein may allow for easy removal of its C-terminal tag. Yokoyama et al. showed that, after its site-specific incorporation into glutathione S-transferase, the C-terminal side of 98 could be removed with the incubation with aqueous ammonia. The weak bond nature of the protein ester bond also potentially allows appending additional chemical functionalities to protein C-terminus. Chen, Liu, and their coworkers showed that α-hydroxy acids 98–101 could be selectively incorporated into proteins using an evolved PylRS-tRNA^{Pyl} pair. A synthesized protein, HdeA with an ester bond could undergo aminolysis with hydrazine to append a C-terminal hydrazide that could further undergo alternative native chemical ligation to a peptide with a N-terminal cysteine.[45] Another potential application of the α-hydroxy acid incorporation to be demonstrated is to probe the contribution of a single hydrogen bond to protein stability. Incorporation of α-hydroxy acid to an α-helix will remove one hydrogen bond and may dissolve the α-helical folding.

4.5 Alternative codons

Previous studies have indicated that PylRS does not specifically recognize the CUA anticodon of tRNA^{Pyl}. By simply mutating the tRNA^{Pyl} anticodon, we and others have demonstrated that the system can be well applied to reassign other codons such as ochre UAA, opal UGA, and four-base AGGA codons.[61, 62, 127] The PylRS-tRNA^{Pyl} pair also displays remarkable orthogonality toward two engineered tyrosyl-tRNA synthase-tRNA^{Pyl} pairs that were originally from M. jannaschii and E. coli and used to expand the genetic code of bacteria and eukaryotes. By coupling a PylRS-tRNA^{Pyl} pair with these two pairs, two distinctive NCAAs have been genetically encoded at two separate codons namely amber UAG and ochre UAA codons in E. coli and mammalian cells.[61, 128, 129] We recently showed that p-azido-Phe which served as a substrate of an evolved M. jannaschii tyrosyl-tRNA synthetase−tRNA^{Pyl} pair and 102 which served as a substrate of an evolved PylRS−tRNA^{Pyl}_{UA} pair could be site-specifically incorporated at amber and ochre mutation sites respectively into one protein that allow site-selectively dual labeling of a protein in a catalyst-free and one-pot fashion (Figure 11).[130, 131] This development provides a readily adoptable platform for generating proteins labeled with two dyes for their folding/dynamic analysis using Förster resonance energy transfer.

5. Conclusion remarks

In about a decade after the Pyl incorporation mechanism was discovered, the system has been transferred to bacteria, yeast, and mammalian cells for the genetic incorporation of
more than 100 NCAAs or α-hydroxy acids into proteins. These NCAAs or α-hydroxy acids contain many chemically unique reactive groups that allow for the installation of a large variety of biochemical and biophysical probes for protein function investigation and biotechnological development. The system has also allowed access to proteins with multiple posttranslational modifications mainly on Lys both in vitro and in vivo. This makes it possible to tackle important issues such as how these modifications regulate protein functions, enzyme activities, and nucleosome interactions with transcription factors, addressing a lot of questions related to epigenetics. With these magnificent tools developed, we envisage the field will soon diverge to focusing on a large variety of applications from basic study to biotechnological development.

Acknowledgments

Financial support of the PyIRS engineering work in the authors’ laboratory is kindly provided from the National Institute of Health (grant 1R01CA161158), the National Science Foundation (grant CHEM-1148684), and the Welch Foundation (grant A-1715).

Abbreviations used

- **Pyl**: Pyrrolysine
- **PylRS**: pyrrolysyl-tRNA synthetase
- **AARS**: aminoacyl-tRNA synthetase
- **NCAA**: non-canonical amino acid
- **MMA**: monomethylamine
- **MMAMT**: monomethylamine methyltransferase
- **LysRS**: lysyl-tRNA synthetase
- **Lys**: lysine
- **Sec**: selenocysteine
- **ATP**: adenosine 5′-phosphate
- **PheRS**: phenylalanyl-tRNA synthetase
- **Phe**: phenylalanine
- **CuAAC**: Cu(I)-catalyzed azide-alkyne 1,3-dipolar cycloaddition

References

5. Gao B, Gupta RS. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics. 2007; 8:86. [PubMed: 17394648]

Biochim Biophys Acta. Author manuscript; available in PMC 2015 June 01.


Figure 1.
The essential four components of the Pyl incorporation machinery: pyrrolysyl-tRNA synthetase, tRNA\textsubscript{Pyl}\textsuperscript{CUA}, Pyl, and amber codon.
Figure 2.
The biosynthetic pathway of Pyl.
Figure 3. 
(A) The structure of *Thermus thermophilus* PheRS with phenyl-AMP bound at the active site (PDB entry: 1B7Y). (B) The structure of *M. mazei* PylRS with pyrrolysyl-AMP bound at the active site (PDB entry: 2ZIM). (C) The overlay of the two structures.
Figure 4.
(A) The active site of Pyl. (B) The contoured deep substrate binding cavity of PylRS. Pyl-AMP is shown in sticks.
Figure 5.
(A) The recognition of the substrate α-amine at the active site of PylRS. Side chains of residues 300–302 are not shown for a clear view of the hydrogen bond network of α-amine.
(B) The recognition of the substrate α-amine at the active site of *T. thermophilus* PheRS.
Figure 6.
The *D. hafniense* PylRS complex with tRNA\textsuperscript{Pyl}. PylRS forms a homodimer that associates with two tRNA\textsubscript{CUA}\textsuperscript{Pyl} molecules. Three anticodon nucleotides in both tRNA\textsubscript{CUA}\textsuperscript{Pyl} molecules are shown in spheres and colored in magenta.
Figure 7.
NCAAs that serve as substrates of the native PylRS.
Figure 8.
Lysine derivatives that have been genetically incorporated into proteins using engineering PylRS mutants in coordination with tRNA$^{Pyl}$. 
Figure 9.
Phenylalanine derivatives that have been genetically incorporated into proteins using engineering PylRS mutants in coordination with tRNA$^{\text{Pyl}}$. 

*Biochim Biophys Acta.* Author manuscript; available in PMC 2015 June 01.
Figure 10.
α-Hydroxy acids that have been genetically incorporated into proteins using engineering PylRS mutants in coordination with tRNA\textsubscript{Pyl}.
Figure 11.
A general platform of incorporating two different NCAAs into one protein for a catalyst-free and one-pot dual labeling.
Table 1

Sequence information of PylRS mutants that recognize a variety of NCAAs as substrates.

| Sequence   | M276 | L301 | A302 | L305 | Y306 | L309 | N346 | C348 | M350 | Y384 | V401 | W417 | M. mazei<sup>1</sup> | M. barkeri<sup>1</sup>
|------------|------|------|------|------|------|------|------|------|------|------|------|------|---------------------|---------------------
<p>| Substrate  | M241 | L266 | A267 | L270 | Y271 | L274 | N311 | C313 | M315 | Y349 | V367 | W383 |                     |                     |
| 6, 7, 14, 25, 98–101 |      |      |      |      |      |      |      |      |      |      |      |      | M. barkeri          |                     |
| 15, 16, 17 |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 17         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 26, 27, 30 |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 26         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 14, 28, 29, 30, 31 |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 29         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 31, 32     |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 30, 32     |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 33, 34, 35, 37 |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 36, 20     |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 23         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 38         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 39         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 18         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 26, 38, 40, 41, 48, 49 |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. mazei            |
| 41         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 42, 43     |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 44         |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |
| 45, 46     |      |      |      |      |      |      |      |      |      |      |      |      |                     | M. barkeri          |</p>
<table>
<thead>
<tr>
<th>Sequence</th>
<th>M276</th>
<th>L301</th>
<th>A302</th>
<th>L305</th>
<th>Y306</th>
<th>L309</th>
<th>N346</th>
<th>C348</th>
<th>M350</th>
<th>Y384</th>
<th>V401</th>
<th>W417</th>
<th>M. mazei&lt;sup&gt;1&lt;/sup&gt;</th>
<th>M. barkeri&lt;sup&gt;1&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substrate</td>
<td>M241</td>
<td>L266</td>
<td>A267</td>
<td>L270</td>
<td>Y271</td>
<td>L274</td>
<td>N311</td>
<td>C313</td>
<td>M315</td>
<td>Y349</td>
<td>V367</td>
<td>W383</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20, 36, 47, 102</td>
<td>M</td>
<td>L</td>
<td>A</td>
<td>F</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>L</td>
<td>L</td>
<td>S</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>I</td>
<td>F</td>
<td>A</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>I</td>
<td>L</td>
<td>A</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>I</td>
<td>F</td>
<td>A</td>
<td>M. barkeri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>A</td>
<td>M</td>
<td>F</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21, 22</td>
<td>A</td>
<td>F</td>
<td>F</td>
<td>M. barkeri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>A</td>
<td>L</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>A</td>
<td>K</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53, 54</td>
<td>L</td>
<td>M</td>
<td>S</td>
<td>L</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>L</td>
<td>T</td>
<td>F</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>L</td>
<td>S</td>
<td>S</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55-95</td>
<td>A</td>
<td>A</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>T</td>
<td>V</td>
<td>W</td>
<td>F</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>F</td>
<td>M</td>
<td>G</td>
<td>M. barkeri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>I</td>
<td>Y</td>
<td>M. mazei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup> Native enzyme origins.