Expanding the Genetic Code for Biological Studies

Qian Wang,1 Angela R. Parrish,1 and Lei Wang1,*
1The Jack H. Skirball Center for Chemical Biology and Proteomics, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037
*Correspondence: lwang@salk.edu
DOI 10.1016/j.chembiol.2009.03.001

Using an orthogonal tRNA-synthetase pair, unnatural amino acids can be genetically encoded with high efficiency and fidelity, and over 40 unnatural amino acids have been site-specifically incorporated into proteins in Escherichia coli, yeast, or mammalian cells. Novel chemical or physical properties embodied in these amino acids enable new means for tailored manipulation of proteins. This review summarizes the methodology and recent progress in expanding this technology to eukaryotic cells. Applications of genetically encoded unnatural amino acids are highlighted with reports on labeling and modifying proteins, probing protein structure and function, identifying and regulating protein activity, and generating proteins with new properties. Genetic incorporation of unnatural amino acids provides a powerful method for investigating a wide variety of biological processes both in vitro and in vivo.

Introduction
The genetic code consists of 64 triplet codons specifying 20 canonical amino acids and 3 stop signals. This code is preserved in all three kingdoms of life with few reassignments (Knight et al., 2001). Although various amino acid derivatives have been identified in natural proteins, most of them are post-translational modifications of the canonical amino acids (Uly and Wold, 1977). Two additional amino acids, selenocysteine (Bock et al., 1991) and pyrrolysine (Srinivasan et al., 2002), are also found to be incorporated into proteins cotranslationally and are thus regarded as natural expansions of the genetic code (Ambrogelly et al., 2006). With these amino acids and with the help of postranslational modifications and cofactors, proteins are able to carry out various functions for sustaining life.

Given the important roles of proteins in biology, it is desirable to be able to manipulate proteins at will for understanding protein structure-function relationships, investigating protein-involved biological processes, and generating proteins and organisms with new properties. Therefore, tremendous efforts have been made regarding the incorporation of unnatural amino acids into proteins to introduce new functional groups apart from those found in the canonical amino acids, through chemical or biosynthetic means (Wang and Schultz, 2004). Chemical methods use chemistry to directly modify or prepare proteins. Early efforts focused on derivatizing reactive side chains of the protein, such as the thiol group of cysteine and the ε-amino group of lysine (Kaiser et al., 1985). The advance of stepwise solid-phase peptide synthesis makes possible the complete synthesis of peptides and small proteins (usually < 100 residues) containing any synthetically accessible unnatural amino acid (Kent, 1988). The protein size limitation is subsequently overcome by semi-synthetic protein ligation methods, in which two or more peptide fragments are chemically ligated to make the full-length protein (Dawson et al., 1994; Muir, 2003). By contrast, biosynthetic methods harness the endogenous cellular machinery to translate the target protein, in which the unnatural amino acid is incorporated cotranslationally. Bacterial strains auxotrophic for a particular amino acid have been used to globally replace the amino acid with a close structural analog (Cohen and Cowie, 1957; Minks et al., 2000; Link et al., 2003). For the site-specific incorporation of unnatural amino acids, a suppressor tRNA is chemically acylated with an unnatural amino acid. When added into cell extracts or Xenopus oocytes that support protein translation, the acylated tRNA incorporates the attached unnatural amino acid in response to a stop or extended codon (Noren et al., 1989; Bain et al., 1989; Nowak et al., 1995). All of these methods have proved to be extremely valuable, but each has limitations in the lack of site selectivity, the in vitro nature of the method, or the low incorporation efficiency (Wang and Schultz, 2004).

The ability to genetically encode unnatural amino acids in the same manner as the canonical amino acids would enable specific changes to be precisely made in proteins directly in vivo, thus providing novel tools for understanding biology in molecular terms in the native settings. Here, we review the methodology for genetically incorporating unnatural amino acids into proteins in live cells, and focus on recent advances in expanding the method to eukaryotic cells. Representative applications are discussed to illustrate the utility and scope of this technology in studying various biological questions.

Methodology
Similar to canonical amino acids, the genetic encoding of an unnatural amino acid requires a dedicated set of components including a tRNA, a codon, and an aminoacyl-tRNA synthetase (hereafter referred to as synthetase) (Figure 1) (Wang and Schultz, 2002). This tRNA/codon/synthetase set, termed the orthogonal set, must not crosstalk with endogenous tRNA/codon/synthetase sets and must be functionally compatible with other components of the translation apparatus. Specifically, the orthogonal tRNA should not be recognized by any endogenous synthetase, and it should decode the orthogonal codon, which is not assigned to any canonical amino acid. The orthogonal synthetase should not charge any endogenous tRNA but the orthogonal tRNA, and it should charge the orthogonal tRNA with the unnatural amino acid only. The unnatural amino acid cannot be a substrate for any endogenous synthetase and must be available in the cytoplasm of the host cell. Once expressed in cells, the orthogonal synthetase charges the
orthogonal tRNA with the desired unnatural amino acid, and the acylated tRNA incorporates the attached unnatural amino acid into proteins in response to the unique codon by utilizing the endogenous translational machinery. As all elements are genetically encodable, this method has the potential to be applied to almost all genetically tractable organisms.

**Method Development in Escherichia coli**

*Escherichia coli* was initially chosen as the host organism to develop this general method for genetically encoding unnatural amino acids in vivo, because its translational machinery had been extensively studied and genetic manipulation of *E. coli* is efficient and relatively straightforward. A unique codon is required to specify the unnatural amino acid, and the amber stop codon has been chosen for this purpose. There are three stop codons that do not code for amino acids but function as a signal for the termination of translation. By reassigning a “blank” stop codon to an unnatural amino acid any overlap and ambiguity with sense codons is avoided. The amber codon became the “blank” codon of choice in this approach because (i) it is the least used among the three stop codons in *E. coli*, and because (ii) amber suppressor tRNAs have been identified or engineered in *E. coli* to readthrough the amber codon with canonical amino acids, and they do not significantly affect the growth rates of *E. coli* cells (Normanly et al., 1990). Therefore, specifying an unnatural amino acid with the amber codon was expected to minimize perturbations to the genetic code and to the health of the host.

Endogenous tRNAs and synthetases recognize their cognate partners through delicate interactions between various regions (Ibba and Soll, 2000). As there is already a pool of tRNA/synthetase pairs inside cells, it is challenging to generate a new tRNA/synthetase pair orthogonal to existing ones and functional in translation. A successful strategy is to import a tRNA/synthetase pair from a species in a different kingdom of life, followed by fine tuning with selections. This strategy is based on early reports showing that in vitro cross-species aminoacylation is often low (Kwok and Wong, 1980). The first orthogonal tRNA/synthetase pair in *E. coli* generated from archaeal bacteria was derived from the tRNA$^{\text{tyr}}$/tyrosyl-tRNA synthetase (TyrRS) from *Methanococcus jannaschii* (Wang et al., 2000). The anticodon of the tRNA is mutated to CUA for recognizing the amber codon UAG. This *M. jannaschii* tRNA$^{\text{tyr}}_{\text{CUA}}$ and its cognate *Mj* TyrRS function efficiently in translation in *E. coli*, but there is a small degree of aminoacylation of this *Mj* tRNA$^{\text{tyr}}_{\text{CUA}}$ by endogenous *E. coli* synthetases. To make a completely orthogonal tRNA while maintaining its affinity for the cognate synthetase, a general strategy has been developed to evolve orthogonal tRNAs in *E. coli* (Figure 2A) (Wang and Schultz, 2001). A library of tRNA mutants is prepared by randomizing specific nucleotides, and the orthogonal tRNA is selected by using a combination of negative and positive selection. The negative selection removes nonorthogonal tRNAs from the library, and the subsequent positive selection removes nonfunctional tRNAs but allows only orthogonal tRNAs with high affinity for the cognate synthetase to pass. This approach was applied to the *Mj* tRNA$^{\text{tyr}}_{\text{CUA}}$, and the muttRNA$^{\text{tyr}}_{\text{CUA}}$ that is orthogonal in *E. coli* and functions efficiently with *Mj* TyrRS to translate the amber codon was identified.

![Figure 1. A General Method for Genetically Encoding Unnatural Amino Acids in Live Cells](image-url)
The next step is to change the substrate specificity of the orthogonal MjTyrRS from tyrosine to a desired unnatural amino acid, which is also achieved by using two-step selections (Figure 2B). A mutant synthetase library (>10⁹ members) is generated by randomizing five amino acid residues in the active site of the MjTyrRS. The selection of the residues to be randomized was enabled by the availability of the Bacillus stearothermophilus TyrRS crystal structure (Brick et al., 1989). The library is first positively selected in the presence of the unnatural amino acid, in which active synthetase mutants able to charge the orthogonal mutRNA_CUA with the unnatural amino acid or any common amino acid are identified. The active synthetases are then subjected to a negative selection, in which synthetase mutants charging a common amino acid are eliminated. In the end, only synthetase mutants specific for the unnatural amino acid will survive both selections. Using this approach, a mutant synthetase specific for the unnatural amino acid o-methyltyrosine (1, Figure 3) was identified from the MjTyrRS library. Together with the evolved orthogonal mutRNA_CUA described above, it enables the site-specific incorporation of o-methyltyrosine into proteins in response to the amber codon in E. coli. Western blot and mass spectrometric analyses of the expressed protein indicate that the translation fidelity for o-methyltyrosine is close to that of common amino acids (Wang et al., 2001).

This work represents the first expansion of the genetic code to include an unnatural amino acid by using engineering approaches. Structurally similar to tyrosine and phenylalanine, o-methyltyrosine provides an excellent case to demonstrate the high translational fidelity that can be achieved by using this approach. Meanwhile, it was important to test whether this method could be generally used to genetically encode various unnatural amino acids, especially those significantly deviating...
There are examples of eukaryotic organisms having natural changes in their genetic codes (Knight et al., 2001), suggesting that it is possible to artificially alter the genetic code in eukaryotes as well. The general strategy developed to genetically encode unnatural amino acids in E. coli should also be applicable to eukaryotic organisms. A number of tRNA/synthetase pairs derived from E. coli have been shown to be orthogonal in eukaryotic cells (Drabkin et al., 1996; Kowal et al., 2001), again due to the low crosskingdom aminoaacylation. In particular, the E. coli tRNA<sub>CUA</sub>/TyrRS pair is orthogonal in yeast (Edwards and Schimmel, 1990; Kwok and Wong, 1980). Moreover, the success in evolving the MjTyrRS to incorporate a wide range of unnatural amino acids in E. coli guided use of the E. coli tRNA<sub>CUA</sub>/TyrRS pair in yeast.

To evolve synthetases specific for unnatural amino acids in yeast, a general selection scheme analogous to that used in E. coli was set up, which also uses positive and negative selections with a mutant synthetase library. A synthetase library (10<sup>8</sup> in size) is similarly constructed by randomizing five active site residues in the E. coli TyrRS that correspond to the five residues randomized in the MjTyrRS system (Chin et al., 2003). Low incorporation efficiency results in reduced yield of the full-length protein containing the unnatural amino acid and in an increase of truncated protein products terminated at the introduced amber codon, thus creating difficulties with employing genetically encoded unnatural amino acids in yeast. To address this problem, new methods have been developed to efficiently express orthogonal prokaryotic tRNAs and to improve the target mRNA stability in yeast. Transcription and processing of tRNAs are different in prokaryotes and eukaryotes. A major distinction is that E. coli tRNAs are transcribed through promoters upstream of the tRNA gene, whereas eukaryotic tRNAs are transcribed through promoter elements within the tRNA known as the A- and B-box (Figure 4A) (Galli et al., 1981). The A- and B-box identity elements are conserved among eukaryotic tRNAs but are absent in many E. coli tRNAs. Creating the consensus A- and B-box sequences in E. coli tRNAs through mutation could cripple the tRNA (Sakamoto et al., 2002). The E. coli tRNA<sub>CUA</sub>/TyrRS has the B-box but no fully matched A-box. In the previously described system (Chin et al., 2003), only the tRNA structural gene was inserted in a plasmid, resulting in a basal level of tRNA expression likely driven by a cryptic promoter elsewhere (Chen et al., 2007).

Transition to Yeast

There are examples of eukaryotic organisms having natural changes in their genetic codes (Knight et al., 2001), suggesting that is possible to artificially alter the genetic code in eukaryotes as well. The general strategy developed to genetically encode unnatural amino acids in E. coli should also be applicable to eukaryotic organisms. A number of tRNA/synthetase pairs derived from E. coli have been shown to be orthogonal in eukaryotic cells (Drabkin et al., 1996; Kowal et al., 2001), again due to the low crosskingdom aminoaacylation. In particular, the E. coli tRNA<sub>CUA</sub>/TyrRS pair is orthogonal in yeast (Edwards and Schimmel, 1990; Kwok and Wong, 1980). Moreover, the success in evolving the MjTyrRS to incorporate a wide range of unnatural amino acids in E. coli guided use of the E. coli tRNA<sub>CUA</sub>/TyrRS pair in yeast.

To evolve synthetases specific for unnatural amino acids in yeast, a general selection scheme analogous to that used in E. coli was set up, which also uses positive and negative selections with a mutant synthetase library. A synthetase library (10<sup>8</sup> in size) is similarly constructed by randomizing five active site residues in the E. coli TyrRS that correspond to the five residues randomized in the MjTyrRS system (Chin et al., 2003). Low incorporation efficiency results in reduced yield of the full-length protein containing the unnatural amino acid and in an increase of truncated protein products terminated at the introduced amber codon, thus creating difficulties with employing genetically encoded unnatural amino acids in yeast. To address this problem, new methods have been developed to efficiently express orthogonal prokaryotic tRNAs and to improve the target mRNA stability in yeast. Transcription and processing of tRNAs are different in prokaryotes and eukaryotes. A major distinction is that E. coli tRNAs are transcribed through promoters upstream of the tRNA gene, whereas eukaryotic tRNAs are transcribed through promoter elements within the tRNA known as the A- and B-box (Figure 4A) (Galli et al., 1981). The A- and B-box identity elements are conserved among eukaryotic tRNAs but are absent in many E. coli tRNAs. Creating the consensus A- and B-box sequences in E. coli tRNAs through mutation could cripple the tRNA (Sakamoto et al., 2002). The E. coli tRNA<sub>CUA</sub>/TyrRS has the B-box but no fully matched A-box. In the previously described system (Chin et al., 2003), only the tRNA structural gene was inserted in a plasmid, resulting in a basal level of tRNA expression likely driven by a cryptic promoter elsewhere (Chen et al., 2007).
One method for increasing tRNA expression is to use the 5′- and 3′-flanking sequences of an endogenous yeast suppressor, tRNA_{SUP4}^CUA (Francis and RajBhandary, 1990; Drabkin et al., 1996; Kowal et al., 2001; Chen et al., 2007). The E. coli tRNA_{SUP4}^CUA without the -CCA trinucleotide is flanked by these sequences, and this cassette is repeated in tandem 3–6 times. A strong PGK1 RNA Polymerase II promoter is additionally placed upstream of the tandem arrangements (Chen et al., 2007). This combination gives an overall > 50 fold increase in tRNA levels compared to the previous tRNA alone scheme. Together with optimized expression of the synthetase, this approach improves the yield of unnatural amino acid-containing proteins to 6–8 mg/L yeast cells. A new method to efficiently express prokaryotic tRNAs in mammalian cells face the same challenges as in yeast. Specifically, a major challenge is how to efficiently express prokaryotic orthogonal tRNAs and to ensure they are translationally competent in mammalian cells, due to the aforementioned difference in tRNA transcription and processing. In addition, another major challenge is how to evolve mutant orthogonal synthetases to be specific for an unnatural amino acid for use in mammalian cells. The latter arises because the transfection efficiency of mammalian cells is much lower than the transformation efficiency in yeast.

Transition to Mammalian Cells

Genetic incorporation of unnatural amino acids into proteins in mammalian cells faces the same challenges as in yeast. Specifically, a major challenge is how to efficiently express prokaryotic orthogonal tRNAs and to ensure they are translationally competent in mammalian cells, due to the aforementioned difference in tRNA transcription and processing. In addition, another major challenge is how to evolve mutant orthogonal synthetases to be specific for an unnatural amino acid for use in mammalian cells. The latter arises because the transfection efficiency of mammalian cells is much lower than the transformation efficiency of E. coli or yeast, making it impractical to generate large synthetase mutant libraries inside mammalian cells. Moreover, survival-death selection in mammalian cells is not as effective as in E. coli and yeast. Therefore, it is difficult to evolve a specific
synthetase out of a large mutant synthetase library directly in mammalian cells.

The *E. coli* tRNA$_{CUA}^{\text{Tyr}}$/TyrRS pair is orthogonal in mammalian cells, but the *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ is not easily expressed in mammalian cells, presumably due to the lack of a matched A-box. Fortunately, the *B. stea- othermophilus* tRNA$_{CUA}^{\text{Tyr}}$ happens to contain consensus A- and B-box sequences, is not charged by mammalian synthetases, and works with the heterologous *E. coli* TyrRS in suppressing the TAG codon. The *B. stea- othermophilus* tRNA$_{CUA}^{\text{Tyr}}$ (lacking 3'-CCA) was linked to the 5'-flanking sequence of the human tRNA$_{CUA}^{\text{Thr}}$, and nine tandem repeats of this gene arrangement were used for tRNA expression (Sakamoto et al., 2002). A small collection of designed active-site variants of the *E. coli* TyrRS was screened in vitro, and a mutant synthetase using 3-iodotyrosine (3) more effectively than Tyr is identified. This mutant synthetase is used with nine tandem *B. stea- othermophilus* tRNA$_{CUA}^{\text{Tyr}}$ to incorporate 3-iodotyrosine into proteins in Chinese hamster ovary (CHO) cells and HEK293 cells with ~95% fidelity (Sakamoto et al., 2002). In another attempt, mutations were introduced into the A-box region of the *B. subtilis* tryptophan opal suppressor tRNA (tRNA$_{CUA}^{\text{Tyr}}$) to create a consensus A-box (Zhang et al., 2004a). The 3'-CCA-lacking mutant tRNA was expressed by using the 5'- and 3'-flanking sequences from the *Arabidopsis* tRNA$_{CUA}^{\text{Tyr}}$. Together with a rationally designed *B. subtilis* TrpRS mutant, this tRNA incorporates 5-hydroxytryptophan (4) into the foldon protein in HEK293 cells in response to the opal stop codon with ~97% fidelity.

Encouraging as these results are, it remained a challenge to incorporate various unnatural amino acids in mammalian cells with high efficiency and fidelity. The above-described approaches may not be generally applicable to other tRNA/ synthetase pairs or other unnatural amino acids. As for tRNA expression, most bacterial tRNAs do not have the consensus eukaryotic A- and B-box sequences, and heterologous pairing of the tRNA with synthetase often results in low activity. Since the A- and B-box lie at nucleotides involved in tertiary interactions that support the L-shaped tRNA structure, mutations to create such consensus sequences, although they may not dramatically decrease the activity of the *B. subtilis* tRNA$_{CUA}^{\text{Tyr}}$, greatly impair the suppression ability of the *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ (Sakamoto et al., 2002). In addition, using tandem repeats to boost the tRNA expression level may eventually lead to lower expression, especially in stable cell lines, as repeating sequences are frequently associated with gene silencing (Hsieh and Fire, 2000).

A new method has been developed that involves the use of an external type-3 Pol III promoter to efficiently express functional prokaryotic tRNAs in mammalian cells (Figure 4C) (Wang et al., 2007b). A type-3 Pol III promoter has its promoter elements exclusively upstream of the coding sequence and does not require any intragenic elements (such as the A- and B-box) for transcription (Willis, 1993). Therefore, tRNAs with or without the consensus A- and B-box sequences are both expected to be transcribed by this type of promoter in mammalian cells. In addition, the transcription initiation site of some type-3 Pol III promoters, such as the H1 promoter (Myslinski et al., 2001) and the U6 small-nuclear RNA promoter (Paule and White, 2000), is well defined, which could be used to generate the correct 5’ end of the tRNA without further posttranscriptional processing. Using a fluorescence-based translation assay, tRNA expression elements including the promoter, the 5’- and 3’-flanking sequences, and the 3’-CCA trinucleotide are linked to the *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ in various combinations and are systematically tested for their ability to generate functional tRNAs in mammalian cells. The gene arrangement that yields tRNAs with the highest translational activity is the H1 promoter followed by the *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ (without 3’-CCA) and then by the 3’-flanking sequence of tRNA$_{CUA}^{\text{Met}}$. *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ expressed by using this method is ~70-fold more active in translation than the one expressed with the 5’- and 3’-flanking sequences of human tRNA$_{CUA}^{\text{Met}}$. This method has also been used to express different *E. coli* tRNAs successfully, and it is effective in various mammalian cells such as HEK293 cells, HeLa cells, and primary neurons.

For generating unnatural amino acid-specific synthetase mutants, it is difficult to predict a priori which active site residues need to be mutated. More often, mutations at multiple sites are required to achieve high substrate specificity, suggesting that small collections of synthetase mutants may likely fail short. Mutants thus generated may still recognize common amino acids, as is the case with the synthetase used to incorporate 3-iodotyrosine.

To circumvent the challenge of evolving unnatural amino acid-specific synthetases directly in mammalian cells, a transfer strategy is used. Since the *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ (TyrRS pair is orthogonal in mammalian cells as well as in yeast, and the translational machinery of yeast is homologous to that of higher eukaryotes, it should be possible to evolve the specificity of this synthetase in yeast and transfer the optimized tRNA/synthetase pairs for use in mammalian cells. In one report, mutant synthetases that were evolved in yeast from the *E. coli* TyrRS specific for different unnatural amino acids (Deiters et al., 2003; Chin et al., 2003), together with the *B. stea- othermophilus* tRNA$_{CUA}^{\text{Tyr}}$, have been successfully used in CHO cells and HEK293 cells for incorporating the corresponding unnatural amino acids (Liu et al., 2007). A high translational fidelity for the unnatural amino acid is confirmed with mass spectrometric analysis, and proteins could be produced up to 1 µg per 2 × 10^6 cells. In another report, mutant synthetases evolved in yeast from the *E. coli* TyrRS are used together with the cognate *E. coli* tRNA$_{CUA}^{\text{Tyr}}$ expressed by the H1 promoter in mammalian cells (Wang et al., 2007b). Because the H1 promoter also enables the functional expression of the *E. coli* tRNA$_{CUA}^{\text{Leu}}$, mutant synthetases evolved in yeast from the *E. coli* LeuRS (Summerer et al., 2006) are also successfully transferred in mammalian cells to incorporate the fluorescent unnatural amino acid dansylalanine (Wang et al., 2007b). Incorporation efficiencies are in the range of 13%–41%, depending on the activity of the evolved synthetase. The incorporation fidelity is also high, as measured by a sensitive fluorescence assay based on the translation of GFP. Unnatural amino acids have been incorporated in HEK293 cells, HeLa cells, and even in primary neurons.

**Applications**

New chemical, physical, and biological properties can be introduced into proteins via the incorporation of unnatural amino acids, and thus unnatural amino acids have been exploited to study a wide range of biological problems involving proteins. In particular, the stability, specificity, and catalytic properties of multiple proteins have been extensively studied with unnatural amino acids incorporated with chemically acetylated tRNAs,
revealing fundamental principles of protein structure and function (Cornish et al., 1995). Microinjection of chemically acylated tRNAs into Xenopus oocytes enables studies of integral membrane proteins with unnatural amino acids. Various ion channels and neurotransmitter receptors have been probed with unnatural amino acids by using electrophysiological techniques, yielding novel insights into the function role of conserved amino acid residues and the gating mechanism of this important family of proteins (Beene et al., 2003; Lummis, et al., 2005). Recent years have also seen elegant studies of signal transduction, selectivity of ion channels, and chromatin biology by using unnatural amino acids that are introduced with native chemical ligation or expressed protein ligation (Hahn, et al., 2007; Valiyaveetil et al., 2006; McGinty, et al., 2008). There are already many excellent reviews on different aspects of this subject (Cornish et al., 1995; Hohsaka and Sisido, 2002; Beene et al., 2003; Link et al., 2003; Wang and Schultz, 2004; Schwarzer and Cole, 2005; Pellois and Muir, 2006), so here we will focus on the applications of unnatural amino acids that are genetically encoded by using the orthogonal tRNA/synthetase method. Representative examples are summarized to illustrate the usefulness of this technology in modifying and investigating proteins, regulating protein functions, and generating new protein properties. We note that some of the unnatural amino acids discussed below have also been incorporated into proteins by using other approaches, details of which can be found in the reviews cited above.

**Labeling and Modifying Proteins**

Reactive side chains of canonical amino acids such as cysteine and lysine have long been used for the selective chemical modification of proteins, but the intrinsic selectivity is low unless the amino acid is absent or can be mutated in the target protein. A nonproteinogenic chemical group embodied by an unnatural amino acid can serve as a unique chemical handle for bioorthogonal chemical reactions, through which a variety of reagents can be selectively appended to proteins in vitro. These reagents can be biophysical probes, tags, posttranslational modifications, and groups modifying protein stability or activity. Currently, site-specific labeling of a target protein via unnatural amino acids directly inside live cells is still challenging.

The keto group is not present in any of the canonical amino acids, and it reacts with hydrazides, alkoxyamines, and semicarbazides under aqueous, mild conditions to respectively produce hydrazone, oxime, and semicarbazone linkages that are stable under physiological conditions. The keto group has been genetically encoded in E. coli in the form of p-acetylphenylalanine (5) and is specifically labeled in vitro with fluorescein hydrazone and biotin hydrazide with greater than 90% yield (Wang et al., 2003). Similarly, fluorescent dyes are selectively appended to the m-acetylphenylalanine (6) introduced into the membrane protein LamB in E. coli (Zhang et al., 2003). p-Acetylphenylalanine and p-benzoylphenylalanine (21) have also been incorporated into rhodopsin, a transmembrane G protein-coupled receptor, in mammalian cells (Ye et al., 2008). The purified mutant rhodopsin is labeled in vitro with fluorescein hydrazide.

Natural glycoproteins are often present as a population of different glycoforms, which complicates glycan structure analysis and the study of glycosylation effects on protein structure and function. Generation of pure glycoprotein mimetics with defined glycan structures will be valuable for the systematic understanding of glycan function and the development of improved glycoprotein therapeutics. To prepare homogenous glycoprotein mimetics, sugars are synthesized in the aminoxy form and are attached to p-acetylphenylalanine incorporated at defined sites in proteins (Figure 5A) (Liu, et al., 2003). The attached sugar is subsequently elaborated by adding additional saccharides with glycosyltransferases to generate oligosaccharides with defined structures. Alternatively, an aminoxy-derivatized glycan can also be covalently coupled to the p-acetylphenylalanine-containing protein in one step.

PEGylation, the attachment of polyethylene glycol (PEG) to proteins, increases protein stability and solubility. In comparison with using common amino acids, PEGylation using a unique chemical handle of an unnatural amino acid increases the homogeneity of the final protein product and provides flexibility in choosing PEGylation sites for optimizing protein activity. PEG has been attached to p-acetylphenylalanine introduced in the human growth hormone to afford a protein that retained wild-type activity but had a considerably improved half-life in serum (P.G. Schultz, personal communication). This approach can be extended to other therapeutic proteins to generate purer and more stable protein drugs.

Azide and acetylene are two additional nonproteinogenic chemical groups that react with each other through a copper (I)-catalyzed [2+3] cycloaddition reaction, p-Azidophenylalanine (7) and p-propargyloxyphenylalanine (8) have been genetically incorporated into human superoxide dismutase-1 in E. coli and yeast (Deiters, et al., 2003, 2004). Purified mutant proteins are labeled in vitro with fluorescent dyes or PEG derivatized with the complementary functional group (Figure 5B). The requirement of Cu2+ as a catalyst may make it difficult to use this reaction inside live cells. The azide group can also react with phosphate derivatives through the Staudinger ligation. p-Azidophenylalanine has been incorporated into the Z-domain protein in E. coli or into peptides displayed on phage, and it can be labeled with fluorescein-derived phosphines (Figure 5B) (Tsao, et al., 2005).

Phenylselenide undergoes oxidative elimination in hydrogen peroxide followed by a Michael addition with thiols (Figure 5C). Phenylselenocysteine (9) is genetically incorporated into GFP in E. coli (Wang et al., 2007a). After treating with hydrogen peroxide, the resultant dehydroalanine is selectively labeled in vitro with thiol-containing mannopyranose to generate an analog of glycosylated GFP, or with n-hexadecylmercaptan to generate a palmitoylated GFP with a nonhydrolyzable linkage. Similarly, phenylselenocysteine has been used to append analogs of methyl- or acetyl-lysine onto the purified Xenopus histone H3, important histone modifications that contribute to chromatin structure and accessibility (Guo et al., 2008). The H3 protein with an acetyl-lysine 9 mimic is deacylated by a histone deacetylation complex and phosphorylated by Aurora B kinase, suggesting that such chemically labeled histones are likely functional in nucleosomes to facilitate the analysis of chromatin.

Besides using a chemical handle, changes can be directly incorporated into proteins in the form of unnatural amino acids, in particular, to mimic posttranslational modifications. These modifications are common in eukaryotic proteins playing important roles in protein stability, localization, and function, but they cannot be easily produced in bacterial cells. It is also difficult
to prepare homogenously modified proteins from eukaryotic cells. Unnatural amino acids with the properties of posttranslational modifications may allow for the generation of homogenous proteins with the desired modifications to uncover their functional roles and possibly to control protein function.

Glycosylation is a frequent modification of eukaryotic proteins. Homogeneous glycoprotein mimetics with defined glycan structures can be prepared by using the keto- or phenylselenide-containing unnatural amino acids, but it leaves a small unnatural linkage between the protein and the first sugar. This linkage does not prevent glycosyltransferases from attaching additional sugars (Liu et al., 2003), but whether it affects other glycoprotein functions awaits more research. To prepare a native glycosylated protein, mutant synthetases have been evolved that site-specifically incorporate β-N-acetylglucosamine-O-serine (10) or N-acetylgalactosamine-α-O-threonine (11) into proteins in *E. coli* (Zhang et al., 2004b; Xu et al., 2004). Since there is no endogenous glycosyl modification in *E. coli*, genetic incorporation of glycosylated amino acids will help improve the production of proteins that require glycosylation for proper folding and function.

Phosphorylation of serine, threonine, and tyrosine by kinases and reversion by phosphatases are key switches in various signal transduction cascades. A phosphorylation mimic resistant to phosphatase activity can be used to identify the role of a specific phosphorylated residue, and to create stably activated proteins for functional assays. p-Carboxymethylphenylalanine (pCMF, 12), a non-hydrolyzable analog of phosphotyrosine, was found to be capable of mimicking the phosphorylated state of Tyr (Xie et al., 2007b). Wild-type human STAT1, after being phosphorylated at Tyr701, forms a homodimer and strongly binds a DNA duplex. The mutant STAT1 with Tyr701 substituted with pCMF also binds the same DNA duplex tightly, suggesting that pCMF could replace phosphotyrosine in the generation of constitutively active phosphoproteins.

Acetylation is another important reversible protein modification. For instance, acetylation of lysine residues in histone proteins controls the secondary structure of chromatin as well as gene expression levels. One method for generating acetyl- and methyl-lysine analog-tagged histones is described above and uses phenylselenocysteine. N′-acetyllysine (13) is genetically incorporated into proteins in *E. coli* (Neumann et al., 2008). In the presence of a deacetylase inhibitor, manganese superoxide dismutase containing N′-acetyllysine at residue Lys44 is produced; however, acetylation of Lys44 does not appear to affect the enzyme activity. In addition, N′-acetyllysine and other lysine analogs are also incorporated into GRB2 in HEK293 cells (Mukai et al., 2008).

**Probing Protein Structure and Function**

The structure and function relationship of proteins can be studied with greater precision and accuracy when various biophysical probes are site-specifically incorporated into proteins. Genetically encoded biophysical probes further extend...
the potential of these studies into live cells, the native environment of proteins.

NMR spectra from large proteins or complexes are complicated to interpret. Specific NMR labels introduced at defined locations in a protein would greatly reduce complexity and simplify signal assignment. Tri-fluoromethylphenylalanine (tfm-Phe, 14) was incorporated at the binding interface of two obligate dimers, nitroreductase and histidinol dehydrogenase, both of which contain active sites at the dimer interface (Jackson et al., 2007). Substrate- or inhibitor-induced conformational changes are evident by monitoring the $^{19}$F NMR chemical shifts. In another report, the binding of a small-molecule ligand to the ferrous myoglobin, the Fe(III) ligand is greatly decreased upon denaturation (Summerer et al., 2006). At 3 M urea, coumarin placed in helices A and C shows substantial changes are evident by monitoring the $^{19}$F NMR chemical shifts.

These results demonstrate that the nitrile group is a sensitive probe for infrared (IR) spectroscopy (Schultz et al., 2006). The stretching vibration of its nitrile group has strong absorption and a frequency (vCN) at ~2200 cm$^{-1}$, which falls in the transparent window of protein IR spectra. $p$CNpe was incorporated into myoglobin at His64, a site close to the iron center of the heme group, to examine ligand-bound states of the heme group. When the Fe(III) ligand is changed from water to cyanide in ferric myoglobin, vCN shifts from 2248 cm$^{-1}$ to 2236 cm$^{-1}$. In the ferrous myoglobin, the linear Fe(II)CO complex shows a vCN absorption at 2239 cm$^{-1}$, whereas the bent Fe(II)NO and Fe(II)O$_2$ complexes show absorption at 2230 cm$^{-1}$. These results demonstrate that the nitrile group is a sensitive probe for ligand binding and local electronic environment.

Fluorescent unnatural amino acids may complement the widely used fluorescent proteins in imaging protein expression, dynamics, and function. Their smaller size can minimize potential perturbations to target proteins, and their fluorophores can be designed to report different microenvironmental changes. Dansylalanine (18) contains the dansyl fluorophore, whose fluorescence intensity increases in hydrophobic surroundings. When incorporated on the surface of a $\beta$ barrel in human superoxide dismutase, dansylalanine shows less change in fluorescence after protein denaturation; when placed at an internal site of the $\beta$ barrel, its emission wavelength is red shifted and intensity is greatly decreased upon denaturation (Summerer et al., 2006). Therefore, fluorescence change can be used to infer when a residue is exposed to solvent during unfolding. $\gamma$-(7-hydroxy-coumarin-4-yl) ethylglycine (19) bears the coumarin fluorophore, which is sensitive to pH and polarity. It was genetically incorporated into myoglobin at two different helices, A or C (Wang et al., 2006a). At 3 M urea, coumarin placed in helices A and C shows a similar increase in fluorescence. However, at 2 M urea, fluorescence increases only when coumarin is incorporated in helix A, suggesting that helix C is not destabilized until urea concentrations reach above 3 M. This result is consistent with previous NMR data, and it suggests that coumarin fluorescence can also report local protein unfolding. Another reporter, $p$-nitrophenylalanine (20) quenches the fluorescence of nearby tryptophans (Tsao et al., 2006). When it is incorporated in a leucine zipper protein, GCMH, the tryptophanyl fluorescence is quenched in a distance-dependent manner, making $p$-nitrophenylalanine a useful distance probe to monitor protein folding or conformational changes. Although still in its infancy, the use of genetically encoded fluorescent amino acids as real-time optical reporters and biosensors in live cells is promising.

Introducing heavy atoms into protein crystals is a critical step for phase determination in X-ray crystallography. A new method to introduce heavy atoms into proteins is to genetically incorporate $p$-iodophenylalanine (3) into proteins in E. coli or yeast. $p$-iodophenylalanine was incorporated into the T4 lysozyme, and the mutant protein was crystallized (Xie et al., 2004). Diffraction data were collected with a laboratory CuKα X-ray source, and the structure was solved by using single-wavelength anomalous dispersion phasing. A single iodinated amino acid among 164 residues results in a strong anomalous signal, about 3% of the total intensities, which compares favorably with the level achieved with selenomethionine by using synchrotron beams. $p$-iodophenylalanine causes little structural perturbation when substituted for Phe in the core of the T4 lysozyme. This approach ensures that the heavy atom iodine is quantitatively introduced at a specific site of the target protein. The strong anomalous signal, the possibility of incorporation at multiple sites and in different cell types, and the use of an in-house X-ray source should facilitate solving protein structures in a high-throughput manner.

The side chain size of unnatural amino acids can be conveniently altered at atomic precision to provide desired changes for probing the bulk effects of amino acids on protein structure and function. The fast inactivation mechanism of the voltage-gated K$^+$ channel Kv1.4 was examined by using unnatural amino acids in HEK293 cells (Wang et al., 2007b). The classic ball-and-chain model for channel inactivation proposes that the N-terminal inactivation peptide forms a ball-like domain to occlude the channel exit for ions. In contrast, a new model hypothesizes that the inactivation peptide threads through a side portal and extends into the inner pore to block ion flow. To experimentally test this, Tyr19 in the inactivation peptide was initially mutated to Phe or Trp, which shows no difference in channel inactivation. However, mutation to $\alpha$-methyltyrosine (1) results in a markedly slower inactivation, as does mutation to dansylalanine (18) (Figure 6). Modeling suggested that the diameter of the inactivation peptide, which is unchanged for Phe or Trp at this site but is larger for $\alpha$-methyltyrosine and dansylalanine, is important for channel inactivation. This is likely due to the narrow width of the side portal in the channel, supporting the new model for channel inactivation.

**Identifying and Regulating Protein Activity**

The identification of protein interactions and the control of protein activities in vivo would ideally require methods that are noninvasive and easy to administer. Application of light is an attractive approach. Photosensitive unnatural amino acids that can crosslink with nearby molecules, shed protecting groups, or change conformation have been developed and genetically incorporated into proteins. They enable the in vivo manipulation of proteins, which should have great potential in molecular and synthetic biology applications.
Crosslinking has been an important technique for uncovering protein interactions in cells. Site-specific photocrosslinking enabled by genetically encoded unnatural amino acids may help to pinpoint transient or weak interactions, to identify which region is involved, and to distinguish direct from indirect interactions. A photoremovable protecting group is installed onto the target protein activity by using light. A photoremovable protecting group is installed onto the target protein activity by using light. Photocaging of critical residues can be harnessed to regulate protein activity.

Photocaging prevents posttranslational modification of the selected residue until a desired moment, which is a more versatile method than mutating the residue to a null amino acid. After a 30 min exposure of bacterial cells to UV light, the enzyme regains activity to ~70% of wild-type levels. A photocaged serine, 4,5-dimethoxy-2-nitrobenzylserine (DMNB-Ser, 24), is incorporated at phosphoserine sites in the transcription factor Pho4 to control its phosphorylation in yeast (Lemke et al., 2007). DMNB-Ser blocks phosphorylation until photolyzed with 405 nm light. Serine is then regenerated and subsequently phosphorylated, triggering the nuclear export of Pho4 (Figure 7A). Photocaging prevents posttranslational modification of the selected residue until a desired moment, which is a more versatile method than mutating the residue to a null amino acid in examining signal transductions.

Photocaging of critical residues can be harnessed to regulate protein activity by using light. A photocaging technique involves the use of a light-sensitive group that can be activated by light to bind to the protein, thereby altering its activity. This technique can be used to regulate protein activity in a temporal and spatial manner.

**Generating New Protein Properties**

Natural evolution affords numerous proteins for life needs, whereas directed protein evolution and rational engineering further increase the diversity of protein functions. Gene encoding additional functional groups into proteins has the potential to generate totally novel and previously impossible protein properties, and initial progress is encouraging.

Metal ions bound in proteins participate in a large number of catalytic and electron transfer reactions in cells. Rationally
Engineering a metalloprotein is challenging, as it is difficult to predict and control the amino acid shells surrounding the metal ion. Multidentate, metal-chelating amino acids with coordinating atoms preoriented in the correct configuration may facilitate such an effort. Bipyridylalanine (BpyAla, 27) has been shown to reversibly bind copper ions when genetically incorporated into the T4 lysozyme (Xie et al., 2007a). The Cu(II)-binding property is exploited to oxidatively cleave the phosphosugar moiety of the nucleic acid backbone. BpyAla is incorporated into the catabolite activator protein (CAP) at Lys26, a site close to the protein-DNA interface (Lee and Schultz, 2008). In the presence of Cu(II) and 3-mercaptopropionic acid, the mutant CAP, originally a transcription factor, cleaves double-stranded DNA at its consensus sequence with high specificity. This method may be generally applied to other DNA-binding proteins to map their preferred DNA sequences.

The distance probe p-nitrophenylalanine (20) has recently been rediscovered to stimulate potent immune responses for novel immunogenic applications. An animal usually will not mount a substantial immune response against a self-protein. This immune self-tolerance has been shown to be broken by introducing p-nitrophenylalanine into the antigen (Grünewald et al., 2008). Incorporation of this unnatural amino acid into tumor necrosis factor α (TNF-α) leads to a strong immune response directed at the unnatural mutant TNF-α in mice, even in the absence of adjuvant. The generated antibodies crossreact with native mouse TNF-α. The immune response is to a common amino acid that the analog is charged by the cognate synthetase of the amino acid and misincorporated into the proteome (Hartman et al., 2007). The novel functional groups embodied by these genetically encoded new amino acids will enable an increasing number of new experiments for biological research. The utility of many unnatural amino acids has been demonstrated in proof-of-principle experiments, and they are expected to be applied to various biological systems for addressing challenging questions and uncovering unknowns in the coming years. Unnatural amino acids that are fluorescent, photoresponsive, bio-orthogonally reactive, or able to mimic various protein modifications would be particularly useful in addressing cell biology questions. Further improvement of the unnatural amino acid incorporation efficiency and detailed studies on the physiological effects of their incorporation would be important for effective application, especially in mammalian cells. The introduction of this technology into multicellular organisms would open the field to biological problems concerning how cells interact with one another in a variety of settings, allowing aspects of development, neural connectivity, and cellular signaling to be studied, to name only a few. The development of random unnatural amino acid mutagenesis may enable the evolution of protein properties that have never existed in nature. Lastly, the genetic incorporation of unnatural amino acids has implications on various aspects of the origin and evolution of the genetic code itself, which remains to be explored.

Figure 7. Regulation of Protein Activity with Photoresponsive Amino Acids
(A) Photocaged serine is used to control when the target serine is regenerated and phosphorylated. (B) The photoisomerizable p-azo phenyl-phenylalanine is used to control the binding of CAP to its promoter, which is determined by a gel mobility shift assay.
ACKNOWLEDGMENTS

L.W. acknowledges support from the Searle Scholar program, the Beckman Young Investigator Program, the March of Dimes Foundation (5-FY08110), the California Institute for Regenerative Medicine (RN1-00577-1), and the Office of the Director of the National Institutes of Health (DP2OD004744).

REFERENCES


