EXPANDING GENETIC CODE FOR PROTEIN LYSINE AND PHENYLALANINE MODIFICATIONS

A Dissertation

by

YANE-SHIH WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Chemistry

Expanding Genetic Code for Protein Lysine and Phenylalanine Modifications

EXPANDING GENETIC CODE FOR PROTEIN LYSINE AND PHENYLALANINE MODIFICATIONS

A Dissertation

by

YANE-SHIH WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Wenshe Liu
Committee Members,	Christian Hilty
	Frank Raushel
	Tadhg Begley
Head of Department,	David H. Russell

August 2012

Major Subject: Chemistry

ABSTRACT

Expanding Genetic Code for Protein Lysine and Phenylalanine Modifications. (August 2012)

Yane-Shih Wang, B.S.; M.S., National Cheng-Kung University, Taiwan

Chair of Advisory Committee: Dr. Wenshe Liu

The naturally occurring pyrrolysine (Pyl) incorporation machinery was discovered in methanogenic archaea and some bacteria. In these organisms, Pyl is cotranslationally inserted into proteins and coded by an in-frame UAG codon. Suppression of this UAG codon is mediated by a suppressor tRNA, $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$, that has a CUA anticodon and is acylated with Pyl by pyrrolysyl-tRNA synthetase (PylRS). The PylRS-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair can be directly applied to incorporate Pyl and other lysine derivatives into proteins at amber mutation sites in *E. coli* and mammalian cells. In the approach of amber codon suppression, evolved PylRSs were selected to synthesize the proteins genetically with lysine and phenylalanine derivatives which contain native or mimic of post-translational modifications (PTMs) or active chemical functional groups for protein labelling and protein folding studies.

A photocaged N^{ε}-methyl-L-lysine has been genetically incorporated into proteins at amber codons in *Escherichia coli* using an evovled PylRS-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair. Its genetic incorporation and following photolysis to recover N^{ε}-methyl-L-lysine at phsyiological pH provide a convenient method for the biosythesis of proteins with monomethylated lysines.

Using an evolved PylRS-$\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair, a *Se*-alkylselenocysteine was genetically incorporated in histone H3. The H3 with mimics of PTMs such as lysine methylation, lysine acetylation, and serine phosphorylation has been synthesized by selective oxidative elimination of *Se*-alkylselenocysteine and followed Michael addition reactions with different thiol-containing small molecules.

Using evolved PylRS - $\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pairs, L-phenylalanine, *p*-iodo-L-phenylalanine and *p*-bromo-L-phenylalanine have been genetically incorporated into proteins at amber mutation sites in *E. coli.* The drastic change of the substrate specificity of PylRS from an aliphatic amino acid to short aromatic amino acids indicates that the PylRS-$\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair can be evolved for genetic incorporation of a large variety of NAAs into proteins in *E. coli*. Inspired by the consistent mutations on N346 position, the mutants on N346 and C348 were constructed and evaluated with different L-phenylalanine derivatives. Using PylRS - N346A/C348A $\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair, more than 30 L-phenylalanine derivatives have been genetically incorporated into proteins at defined sites with amber mutation in *E. coli*. These breakthroughs and development greatly expand the inventory of genetically encoded NAAs and our abilities to do protein engineering in these cells.

DEDICATION

To my parent, my wife and my twin daughters

ACKNOWLEDGEMENTS

Completing my Ph.D. degree is one of my dreams in my life. I experience many difficulties and struggle to find the ways in my doctoral journey. It has been a great privilege to spend five years in the Department of Chemistry of Texas A&M University, and its members will always remain dear to me.

My first debt of gratitude must go to my advisor, Dr. Wenshe Liu. He patiently provided the vision, encouragement and advice necessary for me to proceed through the doctorial program and complete my dissertation. He has been a strong and supportive adviser to me throughout my graduate school career, but he has always given me great freedom to pursue independent work.

Special thanks to my committee, Dr. Christian Hilty, Dr. Tadhg Begley, and Dr. Frank Raushel for their support, guidance and helpful suggestions. Their guidance has served me well and I owe them my heartfelt appreciation.

Members of Dr. Wenshe Lab. also deserve my sincerest thanks, their friendship and assistance has meant more to me than I could ever express. I should also mention The Texas A&M chapter of Sigma Xi Society and Beta Beta Chapter at Texas A&M University of Phi Lambda Upsilon for allowing me to be part of a great professional communities.

My friends in US, Taiwan and other parts of the world were sources of laughter, joy, and support. Special thanks go to, members of Taiwanese Women Club, aka "Taiwanese F2 Club", at Texas A&M, members of Taiwanese Student Society of

Department of Chemistry at Texas A&M University. I am very happy that, in many cases, my friendships with you have extended well beyond our shared time in College Station.

I wish to thank my parents, Mr. Wan-Ded Wang and Mrs. Su-Chin Cheng and my brother and sister. Their love provided my inspiration and was my driving force. I owe them everything and wish I could show them just how much I love and appreciate them. My wife, Shih Chyi Low, whose love and encouragement allowed me to finish this journey. She already has my heart so I will just give her a heartfelt “thanks.” My lovely twin daughters’ birth, Mandy and Maegen, brought me to expreince happiness of the life during my graduate study. They are the best gift I ever have. I also want to thank to my in-laws for their unconditional support. Finally, I would like to dedicate this work to my grandma, Mrs. Yu-Yah Chen (1907-2012). I hope that this work makes you proud in the heaven.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION*

Protein Post-Translational Modifications

With few exceptions, the genetic code of all known organisms specifies the same 20 amino acid building blocks. However, it is clear that many proteins require additional chemical groups beyond what 20 amino acid building blocks can provide to carry out their native functions.[1, 2] Many of these chemical groups are installed on proteins through covalent posttranslational modifications of amino acid side chains. These modifications greatly expand the coding capacity especially of eukaryotic genomes and lead to proteomes two to three orders of magnitude more complex than the encoding genomes would predict. Prominent posttranslational modifications include serine, threonine, and tyrosine phosphorylation, tyrosine sulfation, lysine acetylation, lysine and arginine methylation, lysine ubiquitination, glycosylation, and hydroxylation, etc (**Figure I-1**). These modifications grant proteins with expanded opportunities for catalysis, regulating signal transduction, controlling gene expression, integration of information at many metabolic intersections, and alternation of cellular locations.

This dissertation follows the style of *Biochemistry*.

*Reprinted in part from "Synthesis of proteins with defined posttranslational modifications using the genetic noncanonical amino acid incorporation approach" Wenshe, R. Liu, Yane-Shih Wang, and Wei Wan. *Mol. Biosyst.* **2011**, 7, 38 - 47. Copyright 2011, with permission from The Royal Society of Chemistry.

Studies of posttranslational modifications will provide fundamental understanding of many cellular processes but have been greatly hampered by the difficulty to attain homogenously modified proteins. In cells, most posttranslational modifications are reversible and delicately regulated. It is usually impossible to drive a protein completely to a modified form.[3] Separation of themodified form from the unmodified counterpart is generally difficult because most posttranslational modifications provide small size changes to proteins and do not dramatically change the proteins' biophysical properties. For a lot of eukaryotic proteins, multiple modification sites and multiple modification types coexist, adding a great complexity not only to functional investigation of these proteins but also to separation of a specific modified form from a largely heterogeneous protein pool in their natural sources.[4-6] It is obvious that obtaining a homogeneously modified protein poses a big challenge for efforts to elucidate the precise regulatory role of a posttranslational modification. To resolve this major obstacle, several approaches have been developed to synthesize proteins with defined posttranslational modifications. A straightforward method is to directly incubate a target protein together with its corresponding enzyme or coexpress the two proteins together in living cells. Modified protein forms for both histones and p53 have been synthesized in this way.[7, 8] However, the separation of a modified form from its original protein is a big challenge for aforementioned reasons. For large-size modifications such as ubiquitination,[9, 10] the separation is readily achievable. But small-size modifications including acetylation and methylation often lead to modified protein forms inseparable from their unmodified

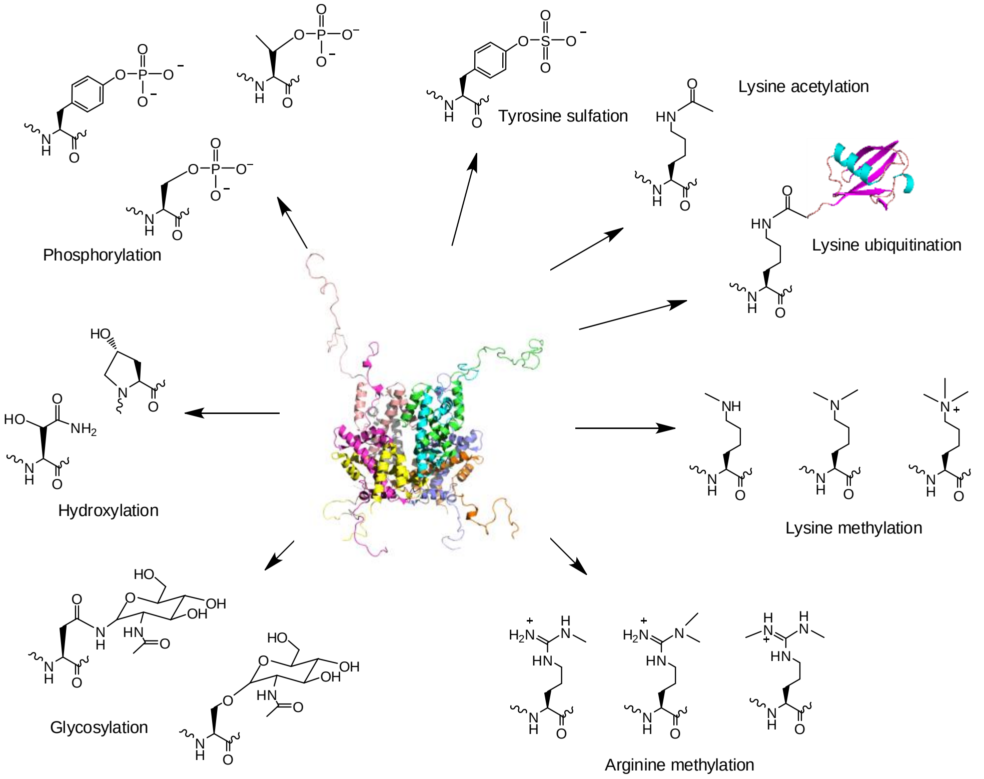


Figure I-1. Representative posttranslational modifications on proteins. The protein complex shown in the center is the histone octomer. The structure is based on the PDB entry: 1KX5.

precursors. Although antibodies have been used to isolate some modified protein forms, not every modified protein form has a specific antibody that is readily available.[11, 12] Substrate promiscuity of an enzyme (e.g. CBP/p300 catalyses acetylation of p53 at more than 8 sites) may lead to formation of heterogeneous protein forms that are hard to be isolated from each other even using antibodies.[3, 7, 13-15] One additional problem is that not every protein modification has an identified corresponding enzyme. The advances of proteomic techniques have led to the discovery of thousands of protein modifications. But identifications of enzymes responsible for these modifications have largely fallen behind. Alternative approaches for synthesis of proteins with defined modifications include selective chemical modifications of proteins, native chemical ligation, and the genetic NAA incorporation technique. Shokat and coworkers developed a semi-synthetic method in which a protein cysteine selectively reacts with an alkyl halide to form a N^{ε}-mono-, N^{ε},N^{ε}-di- or N^{ε},N^{ε},N^{ε}-trimethyllysine analog.[16, 17] Recently, Cole and coworkers extended this method to synthesize proteins with a lysine acetylation mimic.[18] Modified histones synthesized in this way function similarly as their native counterparts and have been applied to study regulatory roles of histone methylations in chromatin.[16] One limitation associated with the cysteine modification approach is that only a few posttranslational modification mimics can be obtained. In addition, non-targeted cysteine residues in a protein have to be mutated to other residues to avoid nonspecific modifications. Another semi-synthetic method, native chemical ligation together with its extension, expressed protein ligation has also been applied to synthesize proteins with posttranslational modifications and their mimics.[19] This method has been extensively used to synthesize modified histones for investigating epigenetic regulation of chromatin. Using this method, an elegant study carried out by Muir and coworkers demonstrated that H2B ubiquitination directly stimulates intranucleosomal methylation mediated by hDot1L.[20] Although powerful and straightforward, native chemical ligation requires synthesis of large quantities of peptide thioesters and is still challenging for most molecular biology and biochemistry labs. In addition, both the cysteine modification method and native chemical ligation are not applicable in living cells. Modified proteins have to be synthesized *in vitro* and introduced into cells by invasive approaches such as microinjection for their cellular function studies. In the past several years, Schultz, Yokoyama, Chin and Liu groups have focused on applying the genetic NAA incorporation approach to synthesize proteins with defined posttranslational modifications. One advantage of the NAA incorporation technique is that proteins with defined posttranslational modifications can be directly produced in living cells. This allows the direct characterization of phenotypic consequences of protein posttranslational modification events. Another advantage of the NAA incorporation technique is its simplicity. It combines the recombinant DNA technique and relatively simple organic synthesis. Most molecular biology and biochemistry groups can easily adopt it. In addition, genetic incorporation of a NAA is site-specific regardless of the incorporation site or protein size. It is an optimal choice for synthesis of large proteins with posttranslational modifications in internal sites, a process that is generally cumbersome for native chemical ligation.

The Genetic NAA Incorporation Approach

A general method for genetic incorporation of a NAA in living cells was first developed by Schultz and coworkers.[21-24] This method relies on the read-through of an in-frame amber (UAG) stop codon in mRNA by an amber suppressor tRNA ($tRNA_{CUA}$) that is specifically acylated with a NAA by an evolved aminoacyl-tRNA synthetase (aaRS). This system is equivalent to the naturally occurring pyrrolysine (Pyl) incorporation machinery discovered in some methanogenic archaea and a Gram positive bacterium *Desulfitobacterium hafniense*.[25, 26] In these organisms, Pyl is cotranslationally inserted into proteins by an in-frame amber codon (**Figure I-2**). Suppression of this amber codon is mediated by the Pyl amber suppressor tRNA ($tRNA^{Pyl}_{CUA}$), which has a CUA anticodon and is acylated with Pyl by pyrrolysyl-tRNA synthetase (PylRS). The PylRS-$tRNA^{Pyl}_{CUA}$ pair in these organisms is orthogonal to and therefore does not cross-interact with other aaRS-tRNA pairs in cells, ensuring the fidelity of the Pyl incorporation at amber codon. Interestingly, the encoded Pyl in the methylamine methyltransferase, MtmB, bind to ammonia at the carbon of the imine bond, and the pyrrolysine-methylammonium adduct serves to activate methylamines as substrates for nucleophilic attack by Cobalt (I) metalloprotein.[27] However, the roles of the Pyl encoding machinery in archaea and bacteria remain relatively unexplored. Similarly to the Pyl incorporation machinery, an orthogonal aaRS-$tRNA_{CUA}$ pair can be developed in which an aaRS is evolved to specifically charge its cognate $tRNA_{CUA}$ with a NAA and used to site-specifically incorporate this NAA into a protein at amber codon with high

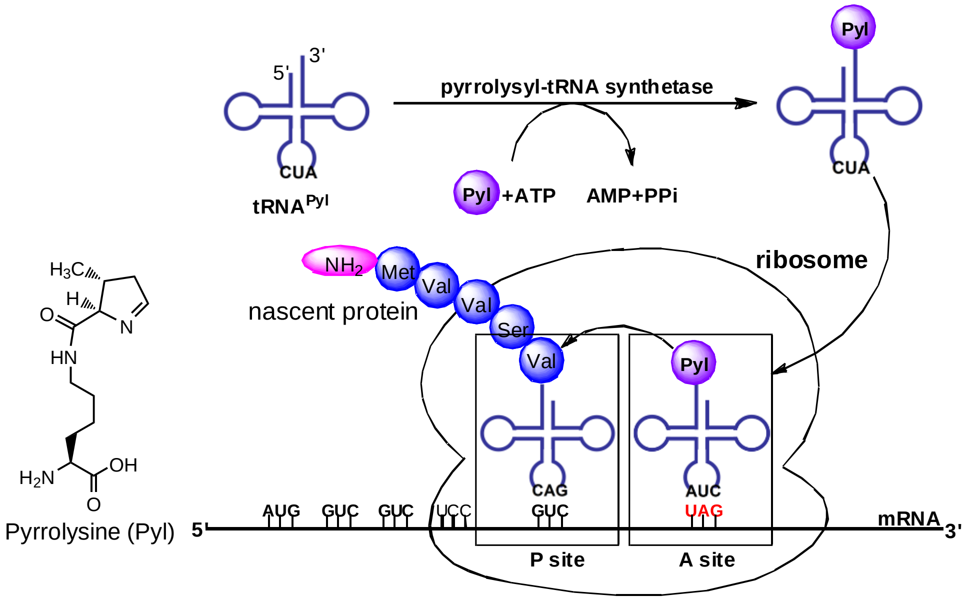


Figure I-2. The pyrrolysine incorporation machinery.

fidelity and efficiency in living cells. Since both the orthogonal aaRS-tRNA$_{\mathrm{CUA}}$ pair and the mRNA with a defined amber codon are genetically encoded, a protein incorporated with a NAA can be expressed by just simply growing cells in a medium supplemented with the NAA. Using an evolved *Mj*TyrRS-MjtRNA$_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pair that was derived from *Methanococcus jannaschii* tyrosyl-tRNA synthetase-tRNA$_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pair, Schultz and coworkers successfully achieved genetic incorporation of sulfotyrosine into proteins in *E. coli* and synthesized hirudin, a naturally occurring sulfated peptide secreted by leeches.[28, 29] This was the first successful demonstration that proteins with defined posttranslational modifications can be directly synthesized cotranslationally. Following this outstanding work, proteins with other modifications or their mimics have been synthesized using evolved *Mj*TyrRS- MjtRNA$_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pairs and PylRS- tRNA$_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pairs, as detailed in the following sections.

Protein Tyrosine Sulfation

The posttranslational tyrosine sulfation is an enzymatic process catalyzed by two tyrosylprotein sulfotransferases, TPST1 and TPST2 in human.[30] The sulfation occurs in the Golgi apparatus before newly synthesized proteins are sorted and shipped to their correct destinations. The natural occurrence of tyrosine sulfation was first discovered in 1954 in a peptide derived from bovine fibrinogen.[31] In the following four decades, sulfated peptide hormones and proteins were only sporadically identified. Owing to the emergence of new proteomic tools, many proteins with sulfated tyrosine residues have been found in the past decade, revealing that tyrosine sulfation is prevalent in human cells. It was also recently discovered that tyrosine sulfation plays a crucial role in human physiology and pathology. Numerous secreted and integral membrane proteins in human contain sulfated tyrosine residues. Among these proteins, CCR5 and CXCR4 are probably the most prominent because of their roles in the entry of human immunodeficiency virus type 1 (HIV-1) into host cells.[32, 33] Entry of HIV-1 requires its gp120 envelop glycoprotein to bind to two cell-surface receptors, CD40 and a coreceptor, either CCR5 or CXCR4. The interaction of either CCR5 or CXCR4 with gp120 requires sulfation of several tyrosine residues at its N-terminus. Coincidently, several human monoclonal HIV-1 antibodies also contain sulfated tyrosine residues and target the same coreceptor binding site on gp120.[34, 35] Other human proteins containing tyrosine sulfation include P-selectin glycoprotein ligand-1, most chemokine receptors, and some other G-protein coupled receptors.[36, 37] It is evident that tyrosine sulfation in these proteins is critical for their interactions with binding partners. Although tyrosine sulfation is important in both physiology and pathology of human, biochemical studies of this modification is very difficult to pursue because of low expression levels of most sulfated proteins. Synthesis of sulfated peptides using organic chemistry is also challenging because of the instability of the sulfate group.[38] A major technology breakthrough for biosynthesis of sulfated proteins was achieved in 2006. Using an evolved *Mj*TyrRS- $Mj\text{tRNA}_{\text{CUA}}^{\text{Tyr}}$ pair, Schultz and coworkers successfully incorporated sulfotyrosine into proteins at amber mutation sites in *E. coli* and achieved overexpression of the naturally occurring hirudin. Hirudin is secreted by leeches and contains 65 residues.[29] Its Tyr65 is naturally sulfated. Hirudin is the most potent naturally occurring direct thrombin inhibitor known and is used clinically as an anticoagulant. It exhibits a K_i towards thrombin of approximately 25 fM. Although the therapeutic form of hirudin have been produced recombinantly in *E. coli* and yeast, the recombinant hirudin is lack of sulfation at Tyr63 and exhibits a K_i of approximately 300 fM. The activity difference between the natural form and the recombinant form clearly arises from the additional sulfate group on the natural form. It was proposed that the enhanced binding of the natural hirudin to thrombin is due to the salt bridge interaction between the sulfate group and $K81^H$ of thrombin, but there was no direct structural evidence to support the claim. The natural hirudin is extracted from leech heads in very low yields, causing a huge obstacle to directly crystallize its complex with thrombin. Although structures of synthetic sulfated C-terminal peptide fragments of hirudin bound to thrombin were solved, they either showed distorted structures around the sulfate group or did not reveal direct interactions between the sulfate group and thrombin. Since the natural hirudin could be simply expressed in *E. coli*, a large quantity of this protein was later obtained and used directly to undergo X-ray crystallographic studies of its interactions with thrombin. The structure determined clearly shows a sulfate group covalently linked to Tyr63 and revealed a network of hydrogen bonding/salt bridge interactions between the sulfate group and thrombin (**Figure I-3**).[39] This work unprecedentedly demonstrated potential applications of the genetic NAA incorporation approach in basic biochemistry studies of protein posttranslational modifications. Given the prevalence of tyrosine sulfation in higher organisms, it is expected that the genetic sulfotyrosine incorporation will be used broadly to understand many basic biological questions related to tyrosine sulfation. In addition to its applications in basic biochemistry studies, the genetic sulfotyrosine incorporation may also be applied to evolve synthetic antibodies with enhanced binding potentials toward antigens. By coupling phage display and the genetic sulfotyrosine incorporation, Schultz and coworkers recently identified a single-chain variable fragment that shows a 10 fold higher binding potential toward gp120 than the monoclonal antibody 412d.[28]

Protein Tyrosine Phosphorylation

The addition of a phosphate group to a protein tyrosine phenolic hydroxyl group is catalyzed by a tyrosine kinase. This phosphorylation process plays a main role in a wide range of cellular processes in eukaryotes, regulating protein structure and function, modulating protein localization, and controlling protein complex formation and degradation. Accordingly, it affects every basic cellular process, including metabolism, growth, division, differentiation, motility, organelle trafficking, membrane transport, muscle contraction, immunity, learning and memory.[40] Protein tyrosine phosphorylation contributes to the development of many diseases such as cancer and diabetes when going awry. Because of its key role in many biological processes and human diseases, many efforts have been made to synthesize proteins with phosphorylated tyrosine residues to study individual phosphorylation events, mainly by the enzymatic method and native chemical ligation. The direct incorporation of phosphotyrosine into proteins using the genetic NAA incorporation approach has not been successful yet. The two negative charges of phosphotyrosine apparently prevent it from permeating through *E. coli* cell membrane. For this reason, several alternative methods based on the genetic NAA approach have been developed to synthesize proteins with tyrosine phosphorylation mimics. Sulfotyrosine and phosphotyrosine are highly structurally similar. Although sulfotyrosine contains only one negative charge instead of two on its side chain, its potential to form hydrogen bonding interactions with others is identical to phosphotyrosine. We expect genetic incorporation of sulfotyrosine will be soon adopted to study protein tyrosine phosphorylation events. Another phosphotyrosine mimic that has been genetically incorporated into proteins is *p*-carboxylmethyl-phenylalanine (*p*CMF) (**Scheme I-1**).[41, 42] *p*CMF was previously used to mimic phosphotyrosine in a pentapeptide-based inhibitor of the interaction between the Src SH3-SH2 domain and phosphorylated epidermal growth factor receptor. Both molecular modeling and X-ray crystallographic studies indicated that the side chain of *p*CMF has good spatial overlap with two of the phosphotyrosine oxygen atoms that interact with positively charged arginine residues in SH2 domains. The one negative charge nature of its side chain apparently does not prevent its transport into the *E. coli* cytoplasm. Genetic incorporation of *p*CMF in *E. coli* was also achieved using evolved *Mj*TyrRS-$Mj\text{tRNA}_{\text{CUA}}^{\text{Tyr}}$ pairs.[42] Using an evolved pair, STAT1 with a *p*CMF at Tyr701 was recombinantly expressed and purified to homogeneity. It is known that phosphorylation of STAT1 on Tyr701 leads to its homodimerization and the resulting dimer binds tightly to DNA containing M67 sites. The purified mutant STAT1 shows a similar feature. Since *p*CMF can be easily prepared and its incorporation is more efficient than sulfotyrosine, we expect its genetic incorporation will also have utility for a wide variety

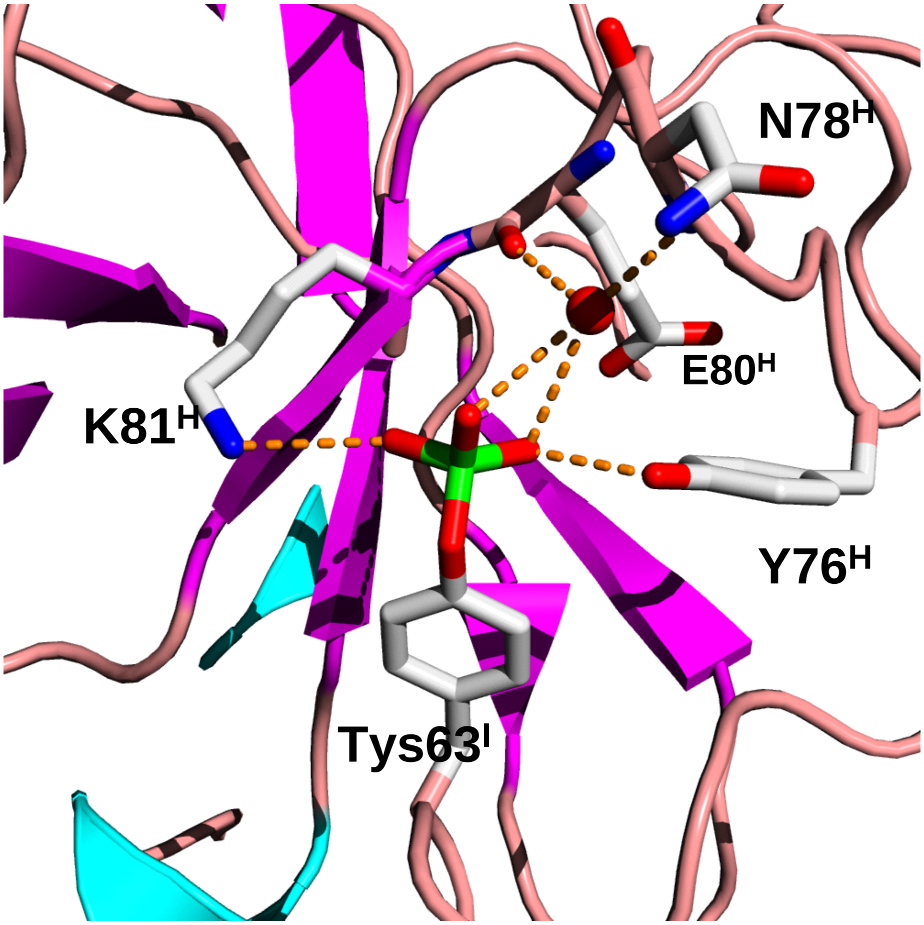
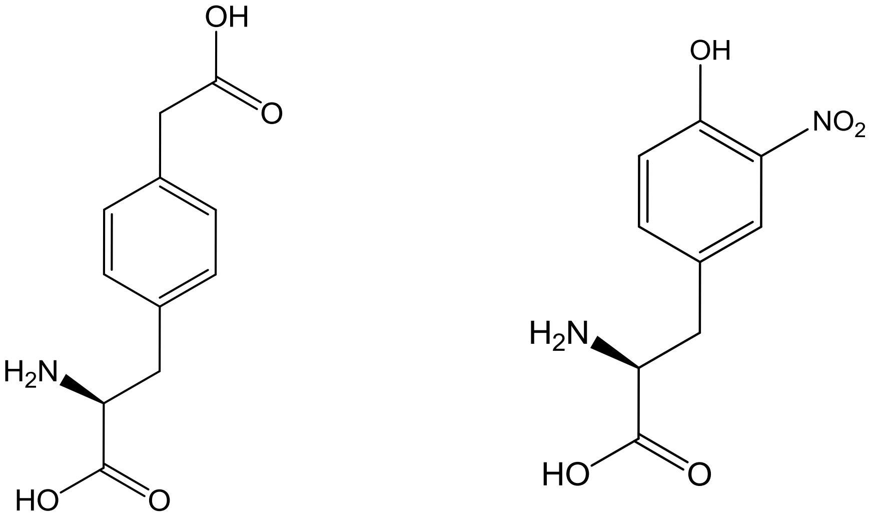


Figure I-3. Structure of sulfotyrosine (Tys) and amino acid residues that directly or indirectly interact with the sulfate group of Tys in the hirudin-thrombin complex. The structure is derived from the PDB entry: 2PW8. The superscripts in labels indicate which chains the residues are located in (I: hirudin; H: the heavy chain of thrombin).



p-carboxymethyl-phenylalanine (*p*CMF)

3-nitrotyrosine

Scheme I-1 Chemical structure of p-carboxymethyl-phenylalanine (pCMF) and 3-Nitrotyrosine of signaling experiments. Another interesting development to synthesize proteins with tyrosine phosphorylation mimics was to produce a phosphotyrosine mimic with two negative charges on its side chain by modifying a genetically incorporated *p*-azidophenylalanine (*p*AZF). The genetic incorporation of *p*AZF has been achieved in *E. coli*, yeast, and mammalian cells.[21, 43, 44] In *E. coli*, evolved *Mj*TyrRS- *Mj*tRNA$_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pairs allow efficient incorporation of *p*AZF at amber codon, leading to overexpression of proteins containing *p*AZF. As shown in **Figure I-4**, a selective modification of *p*AZF in a protein using a phosphite through Staudinger reaction followed by photolysis results in a *p*-(phosphonoamino)-phenylalanine at the *p*AZF site.[45] Since *p*-(phosphonoamino)-phenylalanine has two negative changes and maintains the hydrogen bonding/salt bridge interaction potential of phosphotyrosine, it closely resembles phosphotyrosine and is the best mimic of phosphotyrosine that can be genetically installed so far. Using this chemoselective phosphorylation generation strategy to study biologically relevant signaling processes has been proposed by the original developers of the technique and will be, to our conviction, adopted by other scientists.

Protein Tyrosine Nitration

Protein tyrosine nitration is a nonenzymatic covalent modification, introducing a nitro group to one of ortho carbons of the phenolic ring of tyrosine residues. (**Scheme I-1**) This modification primarily arises from reactive nitrosative species such as peroxynitrite that are generated by nitric oxide. The addition of a nitro group confers

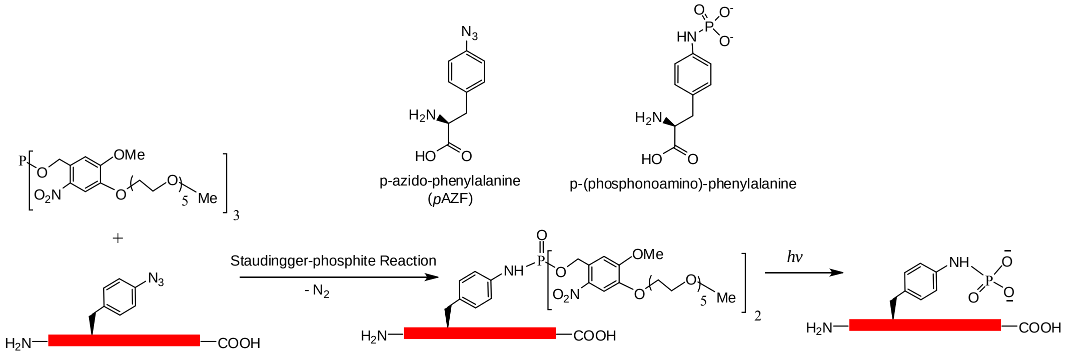

p-azido-phenylalanine
(pAZF)
p-(phosphonoamino)-phenylalanine
Staudingger-phosphite Reaction
- N2
hv
H2N
COOH
OMe
O2N
Me
5
3
2
N3
NH
HN
P
+

Figure I-4. Conversion of *p*AZF to *p*-(phosphonoamino)-phenylalanine.

particular physiological properties to tyrosine such as a much lower pK_a of the corresponding proteins. The impact of tyrosine nitration may affect protein function and structure, and change the rate of proteolytic degradation of nitrated proteins.[46, 47] Many recent studies have shown that tyrosine nitration is a pivotal physiological event implicated in numerous biological processes and a number of diseases such as cancer, neurodegenerative diseases, and cardiovascular injury.[48] The availability of a method for synthesis of proteins with defined tyrosine nitration is critical to elucidate detrimental effects of most tyrosine nitration events. Although protein tyrosine nitration can be achieved by chemically modifying an expressed protein with tetranitromethane in ambient conditions, this modification is not selective and adds nitro to most solvent accessible tyrosine residues.[49] Alternatively, a 3-nitrotyrosine can be selectively incorporated into a protein at amber codon using the genetic NAA incorporation approach. Chin and coworkers recently showed that proteins with defined tyrosine nitration can be expressed to large quantities in *E. coli* using evolved MjTyrRS-$Mj\mathrm{tRNA}^{\mathrm{Tyr}}_{\mathrm{CUA}}$ pairs.[50] Human manganese superoxide dismutase incorporated with 3-nitrotyrosine at Tyr34 synthesized using this method displays a diminished activity which is only 3% of that for the wild-type enzyme. Manganese superoxide dismutase is known to undergo near-quantitative nitration on Tyr34 in cardiovascular disease and aging and is also nitrated in acute and chronic inflammatory processes in both animal models and human diseases. The work by Chin and coworkers clearly showed a deleterious effect of a single nitration to an enzyme’s activity. Since many other proteins such as p53 and tau undergo tyrosine nitration in neurodegenerative disease and atherosclerosis, applications of the genetic 3-nitrotyrosine incorporation to understand basic functions of tyrosine nitration in these proteins are expected.[51]

Protein Lysine Acetylation

Protein lysine acetylation is a reversible enzymatic process catalyzed by histone acetyltransferases and histone deacetylases (HDACs). Originally identified in histones 40 years ago, lysine acetylation is now known to occur in more than 80 transcription factors such as p53 and p300/CBP, many other nuclear regulators, and various cytoplasmic proteins.[6] As a result, lysine acetylation is both crucial for the transcription regulation in the nucleus and important for regulating different cytoplasmic processes, including cytoskeleton dynamics, energy metabolism, endocytosis, autophagy, and even signaling from the plasma membrane. Recently the histone code that involves histone acetylation has garnered considerable attention for its role in regulating the chromatin structure and gene expression. To date, biochemical studies of histone acetylation have mainly been attributed to native chemical ligation. Recently genetic incorporation of N^{ε}-acetyllysine in *E. coli* for synthesis of proteins with lysine acetylation was also developed. The incorporation of N^{ε}-acetyllysine takes advantage of an evolved PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair originally from *Methanosarcina barkeri*.[52] Using this evolved pair, Chin and coworkers expressed homogenous histone proteins with acetylation at different sites and reconstituted nucleosomes bearing defined acetylated lysine residues. Studies of H3 Lys56 acetylation revealed that this modification does not directly affect the compaction of chromatin, has modest effects on remodeling by SWI/SNF and RSC, and increases

DNA breathing.[53] A most recent application of this technique was to elucidate the regulatory role of Lys125 acetylation in cyclophilin A (CypA), a ubiquitous cis-trans prolyl isomerase. CypA has a key role in immunity and is required for effective HIV-1 replication in host cells. Chin & James showed that Lys125 acetylation markedly inhibits CypA catalysis of *cis* to *trans* isomerization and stabilizes *cis* rather than *trans* forms of the HIV-1 capsid.[54] This modification also antagonizes the immunosuppressive effects of cyclosporine by inhibiting cyclosporine binding and calcineurin inhibition. Given that a lot of proteins including all four histone proteins, p53, p300/CBP, PGC-1α, and RIP140 contain multiple lysine acetylation sites, one intriguing interest is to study regulatory roles of concomitant lysine acetylation at multiple sites of a single protein.[6] To this end, our group has extended the genetic N^{ε}-acetyllysine incorporation technique to incorporate three N^{ε}-acetyllysine residues into a single green fluorescent protein in *E. coli*.[55] This was achieved by overexpressing a truncated large ribosomal protein L11 to enhance amber suppression in *E. coli*. In the future, we will apply this technique to study concomitant lysine acetylation at multiple sites in four histone proteins. One advantage of the PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair is its orthogonality in yeast and mammalian cells. This makes it possible to apply the evolved PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair specific for N^{ε}-acetyllysine directly in eukaryotic cells for cellular functional analysis of certain lysine acetylation events. However, lysine acetylation is a reversible process. After genetic incorporation of N^{ε}-acetyllysine into proteins in eukaryotic cells, its deacetylation by HDACs is anticipated. This will certainly complicate the *in vivo* functional analysis of lysine acetylation events. To resolve this obstacle, we designed an unhydrolysable N^{ε}-

acetyllysine analog, 2-amino-8-oxononanoic acid and took advantage of the evolved PylRS-tRNA$_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair specific for N^{ε}-acetyllysine to incorporate it into proteins.[56] Further tests need to be carried out to assess whether this analog can closely mimic N^{ε}-acetyllysine in mammalian cells.

Protein Lysine Methylation

Protein lysine methylation, a reversible enzymatic process catalyzed by histone methyltransferases and histone demethylases, was also originally discovered in histones.[57] It represents one of the most important posttranslational modifications and is crucial in modulating chromatin-based transcriptional control and shaping inheritable epigenetic programs in the eukaryotic cells.[58] Similarly to lysine acetylation, lysine methylation was also found in non-histone transcription factors such as p53 and p300/CBP in the nucleus.[59-62] A recent discovery of histone methyltransferase Ezh2 in the cytosol of various mouse and human T cells has led to expectation that lysine methylation is also prevalent in the cytosol.[63] There are three lysine methylation patterns, monomethylation, dimethylation and trimethylation, which may serve different regulatory roles.[64] As evidenced in the literature, distinctive histone methyltransferases are responsible for these three methylation patterns. So far, studies of histone lysine methylation have been attributed to native chemical ligation.[65] But several technological breakthroughs in 2009 could drive applications of the genetic NAA incorporation approach in this important research area to surge. Given the structural similarity between N^{ε}-methyllysine and N^{ε}-acetyllysine, one would think it is possible to evolve the PylRS-

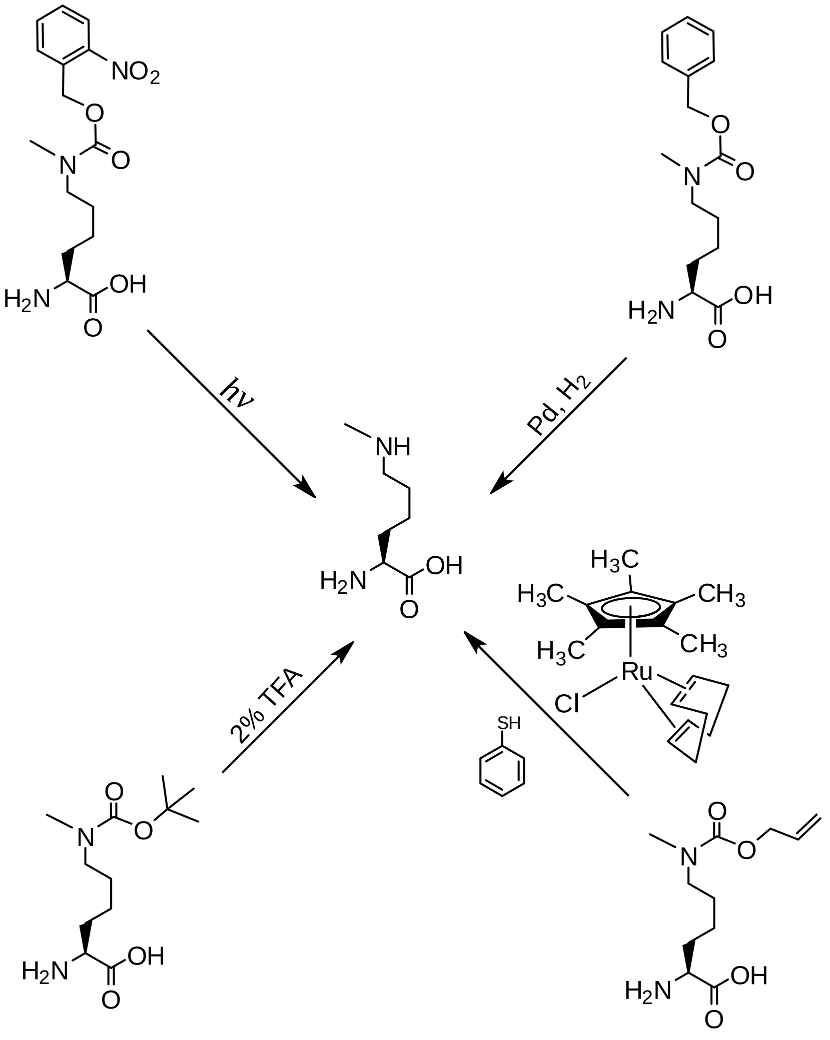


Figure I-5. N^{ε}-methyllysine precursors and their deprotection.

$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair for genetic incorporation of N^{ε}-methyllysine. However, creating a synthetase that recognizes N^{ε}-methyllysine but discriminates against lysine is difficult since the two amino acids differ only by a single methyl group. For this reason, we and other two groups developed indirect biosynthesis methods for production of proteins with lysine monomethylation as shown in **Figure I-5**. Directly using the wild-type PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair from *Mathanosarsina barkeri*, Chin and coworkers managed to incorporate N^{ε}-Boc-N^{ε}-methyllysine site-specifically at Lys9 of histone H3.[66] The deprotection of Boc from the incorporated N^{ε}-Boc-N^{ε}-methyllysine to recover N^{ε}-methyllysine was achieved using 2% TFA. Although the deprotection condition denatures proteins, the small proteins such as histones can be easily refolded afterward. H3K9me1 (monomethylation at Lys 9 of H3) synthesized in this way can be specifically recognized by heterochromatin protein 1, a chromodomain protein that binds H3K9me1 but not unmethylated H3. Liu and Schultz groups developed alternative methods that undergo deprotection to recover N^{ε}-methyllysine at mild conditions.[67, 68] Both groups independently evolved the PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair for genetic incorporation of N^{ε}-(*o*-nitrobenzyloxycarbonyl)-N^{ε}-methyllysine into proteins in *E. coli*. The *o*-nitrobenzyloxycarbonyl group is photolytic. Deprotection of this group from the incorporated NAA to recover N^{ε}-methyllysine can be simply achieved using UV light. This photo deprotection process (360 nm UV light, 30 min, room temperature) is very mild to proteins and has been performed on proteins incorporated with N^{ε}-(*o*-nitrobenzyloxycarbonyl)-N^{ε}-methyllysine in living mammalian cells. We have also achieved the incorporation of N^{ε}-Cbz-N^{ε}-methyllysine into proteins using an evolved PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair. Deprotection of Cbz to recover N^{ε}-methyllysine in mild conditions is in theory feasible by catalytic hydrogenation but have not been achieved yet in practice. Another protected N^{ε}-methyllysine that has been genetically incorporated into proteins is N^{ε}-allylcarbamoyl-N^{ε}-methyllysine. Its incorporation was achieved using a modified PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair. Deprotection of the allylcarbamoyl group can be carried out using chloro-pentamethylcyclopentadienyl-cyclooctadiene-ruthenium(II) ([Cp*Ru(cod)Cl]) under conditions compatible with living cells.[69] Besides the direct installation of N^{ε}-methyllysine in proteins, Schultz and coworkers also developed a method to site-specifically install a N^{ε}-methyllysine analog into proteins. This method takes advantage of the unique chemical reactivity of phenylselenocysteine that can be genetically incorporated into proteins in *E. coli* using evolved MjTyrRS- $Mj\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pairs.[70] After its incorporation into proteins, phenylselenocysteine can be efficiently converted to dehydroalanine under relatively mild conditions, followed by Michael addition reaction with a corresponding thiol-containing nucleophilic compound to form a N^{ε}-methyllysine analog (**Figure I-6**).[71] Using this method, histone H3 with a N^{ε}-methyllysine analog at Lys9 was homogenously synthesized. The same method has also been extended to synthesize proteins incorporated with $N^{\varepsilon}, N^{\varepsilon}$-dimethyllysine and $N^{\varepsilon}, N^{\varepsilon}, N^{\varepsilon}$-trimethyllysine analogs. Although the oxidation of phenylselenocysteine to dehydroalanine requires a high concentration of H_2O_2 that also modifies cysteine and

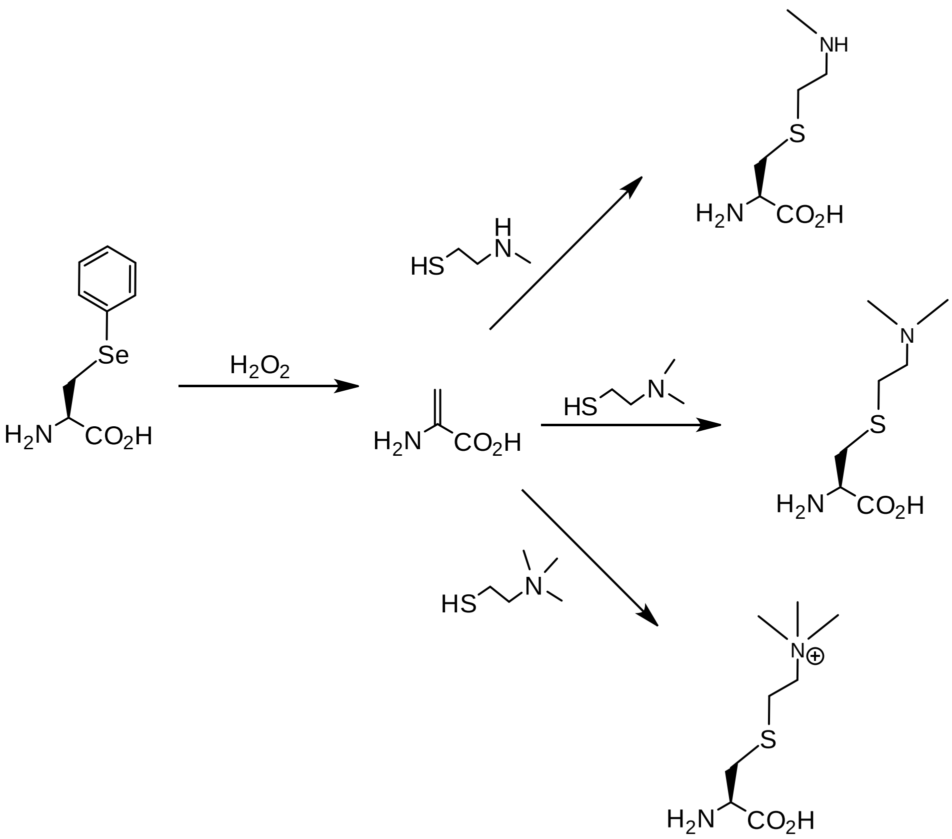


Figure I-6. Conversion of phenylselenocysteine to dehydroalanine and its following reactions to form three lysine methylation analogs.

methionine residues, this method is a very useful tool to study proteins in which mutating cysteine and methionine to other residues does not affect protein functions.

Protein Lysine Ubiquitination

Ubiquitination is another important posttranslational lysine modification. It is a process that covalently attaches the 76-residue protein ubiquitin to a target via an isopeptide bond formed between the C-terminal carboxylate of ubiquitin and the side chain amino group of a lysine.[72] This process consists of a series of steps that involve three enzymes E1, E2, and E3.[73] The most prominent function of ubiquitination is to label proteins for proteasomal degradation. Besides this function, ubiquitination also controls the stability, function, and intracellular localization of a wide variety of proteins.[74] Although functionally important, the study of protein ubiquitination is hampered by the difficulty of obtaining homogenous ubiquitinated proteins. Isolation from an *in vivo* source usually has a low yield. The *in vitro* reconstruction of an ubiquitinated protein using corresponding enzymes suffers from low productivity. Native chemical ligation has been proved an efficient way to synthesize homogenous ubiquitinated proteins,[75] but the synthesis is challenging and can not be easily handled in most biochemistry groups. Recently two groups assembled a relative straightforward method that combines cysteine chemistry and native chemical ligation to make ubiquitinated protein mimics.[76, 77] However, using a disulfide linkage to replace one methylene in the LysGly isopeptide generates a much less rigid linkage whose length is also 3 Å longer than the original LysGly isopeptide. To assemble a method for easy

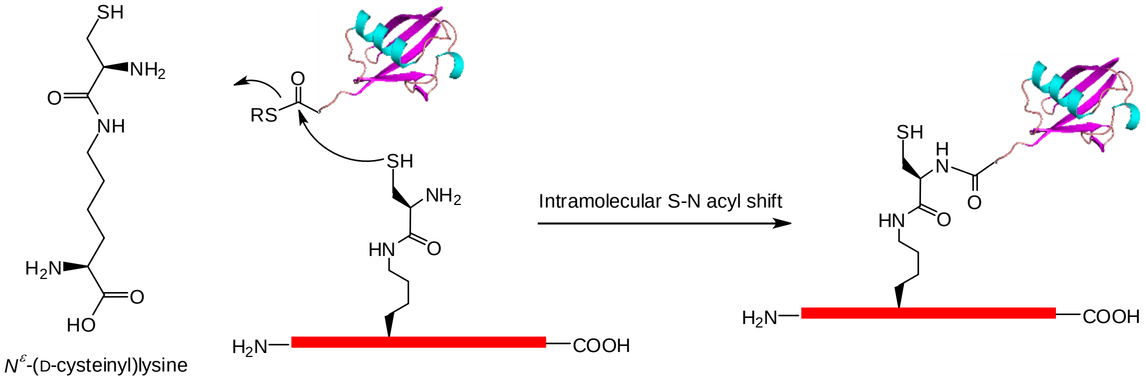


Figure I-7. Genetic incorporation of N^{ε}-(D-cysteinyl)lysine that undergoes expressed protein ligation to crosslink with a ubiquitin molecule.

preparation of proteins with defined ubiquitination, Chan and coworkers used the wild type PylRS- $\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair to genetically install a protein with N^{ε}-(D-cysteinyl)lysine that can be subsequently linked to ubiquitin by expressed protein ligation (**Figure I-7**).[78] N^{ε}-(D-cysteinyl)lysine closelymimics the LysGly isopeptide. Similar to its natural counterpart, an ubiquitinated calmodulin synthesized using this method displays an inhibitory effect on phosphorylase kinase. In order to synthesize ubiquitinated proteins with a natural LysGly isopeptide linkage, Chin and coworkers recently developed a method termed GOPAL (genetically encoded orthogonal protection and activated ligation) that combines genetic incorporation of N^{ε}-Boc-lysine and chemoselective protein chemistry (**Figure I-8**).[79] Using this method, diubiquitins with atypical linkages at Lys6 and Lys29 were synthesized. The structure of Lys6-linked ubiquitin reveals an asymmetric compact conformation distinct from other ubiquitin chain structures. The screening of human deubiquitinases against Lys29-linked diubiquitin revealed that TRABID cleaves the Lys29-linkage 40-fold more efficiently than the Lys63 linkage. Since enzymes catalyzing atypical ubiquitin linkages at Lys6, Lys11, Lys27, Lys29, and Lys33 are unknown, the method developed by Chin and coworkers is indispensable in promoting research interests in this area.

In summary, multiple methods based on the genetic NAA incorporation approach have been developed for the synthesis of proteins with defined posttranslational modifications and their mimics. The intrinsic simplicity of these methods will promote their adoption by many scientists and accelerate our understanding of protein posttranslational modifications.

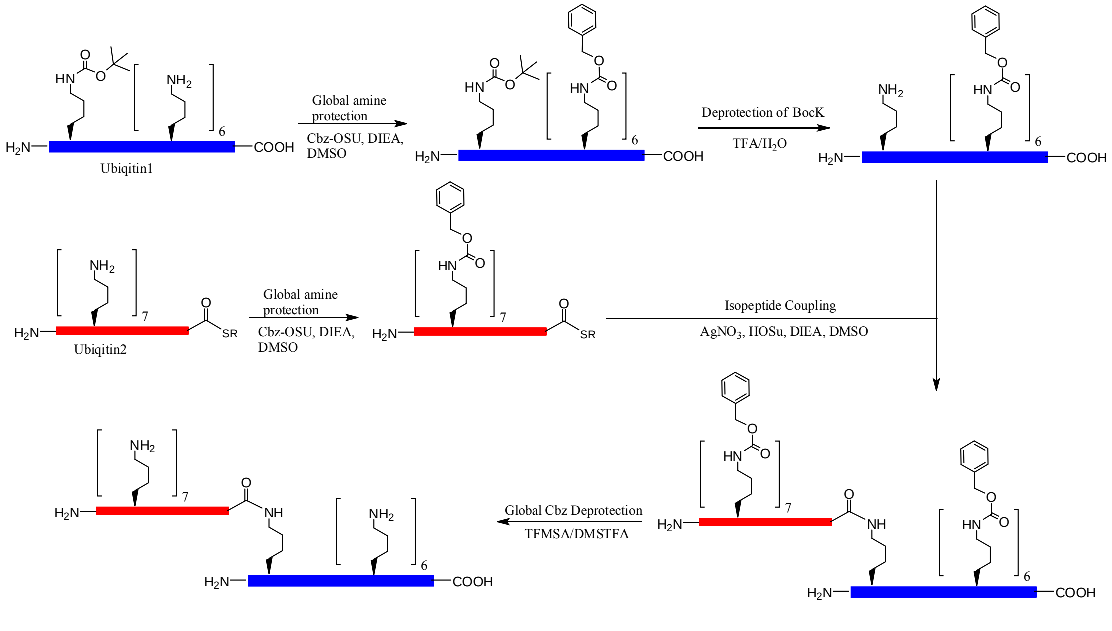


Figure I-8. Genetic incorporation of N^{ε}-Boc-lysine and its application to synthesize a diubiquitin protein by GOPAL.

Although considerable efforts have made several methods available for the synthesis of proteins with defined posttranslational modifications and their mimics using the genetic NAA incorporation approach, challenges still exist in this research area. Bothsulfotyrosine and phosphotyrosine mimics can only be incorporated into proteins in *E. coli*. Studies of cellular functions of these modifications require the use of invasive methods to introduce them to their eukaryotic cell hosts. A method that allows the incorporation of these NAAs directly into eukaryotic cells is necessary for functional studies of these modifications in cells for which microinjection is not readily available.

Protein Labeling

Site-selective covalent modification is an important biotechnological strategy to introduce new functionalities to proteins. Examples include the PEGylation of therapeutic proteins to prolong their *in vivo* half-lives, fluorescent labeling of proteins for protein folding/unfolding analysis and biosensor development, linking proteins with photocrosslinkers for protein-protein/DNA interaction analysis, and introducing NMR, EPR, and IR probes for protein functional investigations. Traditionally, protein modifications exploit the high reactivities of cysteine and lysine side chains in proteins. This approach has poor selectivity due to the existence of multiple copies of cysteine or lysine in a single protein. Modern techniques that attempt to achieve site-selective protein modification include the genetic noncanonical amino acid (NAA) incorporation, native chemical ligation and its extension expression protein ligation, enzymatic and chemical modifications of peptide tags, and specific modifications of protein N- and Ctermini. These techniques, in general, seek to introduce bioorthogonal functional groups into proteins whose reactions do not cross-interact with the 20 canonical amino acids (CAAs). Using orthogonal aminoacyl-tRNA synthetase (aaRS)-amber suppressing $\mathrm{tRNA}_{\mathrm{CUA}}$ pairs in *Escherichia coli*, *Saccharomyces cerevisiae*, and mammalian cells, several groups have successfully demonstrated that NAAs with bioorthogonal functional groups such as keto, alkyne, azide, and phenylhalide groups could be genetically incorporated into proteins at amber mutation sites. However, the genetic encoding of each NAA is usually required to use a uniquely evolved aaRS. AaRS-$\mathrm{tRNA}_{\mathrm{CUA}}$ pairs with different origins might have to be used in different cell lines due to unorthogonal nature of some aaRS-$\mathrm{tRNA}_{\mathrm{CUA}}$ pairs in certain cell lines.

The goals of the research are evolving the PylRS for lysine and phenylalanine derivatives incorporations genetically and applying the synthesized proteins for the protein functional studies according to the protein post-translational modifications, protein folding mechanism and protein labeling.

CHAPTER II

A GENETICALLY ENCODED PHOTOCAGED METHYLLYSINE*

Introduction

Protein lysine methylation is an enzymatic process that involves the transfer of a methyl group from *S*-adenosyl-L-methionine to a lysine side-chain amine in a protein. It represents one the most important posttranslational modification and is crucial in modulating chromatin-based transcriptional control and shaping inheritable epigenetic programs in the eukaryotic cells.[5, 80-82] There are three lysine methylation patterns, mono-, di- and trimethylation, which may serve different regulatory roles.[64] The study of protein lysine methylation is critical to understanding chromatin epigenetics and transcription factor regulation but has long been impeded by the challenge of synthesizing site-specifically and quantitatively methylated proteins.[6] Several methods including enzymatic protein methylation, native chemical ligation, and chemical modification of cysteines have been introduced to synthesize proteins with a defined lysine methylation. [3, 17, 65]However, they all suffer limitations. Enzymatic protein methylation hardly reaches completeness and the separation of the modified protein from the original one is difficult. Native chemical ligation requires a cysteine to mediate the

*Reprinted in part from “A genetically encoded photocaged N^{ε}-methyl-L-lysine”, Yane-Shih Wang, Bo Wu, Zhiyong Wang, Ying Huang, Wei Wan , William K. Russell, Pei-Jing Pai, Yin N. Moe, David H. Russell and Wenshe, R. Liu. *Mol. Biosyst.* **2010**, 6, 1557 - 1560. Copyright 2010, with permission from The Royal Society of Chemistry.

ligation and the installation of a methylated lysine in the middle of a protein is problematic. The cysteine modification method can only install a methylated lysine analog into a protein. We sought to resolve limitations associated with the methods discussed above by applying the genetic noncanonical amino acid (NAA) incorporation method to synthesize methylated proteins. Originally developed by Schultz *et al.*,[23, 24] the genetic NAA incorporation method relies on the read-through of an in-frame amber UAG codon in mRNA by an amber suppressor tRNA that is specifically acylated with a NAA by an evolved aminoacyl-tRNA synthetase (aaRS). This method is equivalent to the naturally occurring pyrrolysine (Pyl) incorporation machinery discovered in methanogenic archaea and some bacteria.[25, 26] In these organisms, Pyl is cotranslationally inserted into proteins and coded by an in-frame UAG codon. Suppression of this UAG codon is mediated by a suppressor tRNA, $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ that has a CUA anticodon and is acylated with Pyl by pyrrolysyl-tRNA synthetase (PylRS). Similarly, the PylRS-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair can be directly applied to incorporate Pyl and several other lysine derivatives into proteins at amber mutation sites in *E. coli*.[83-86] Together with $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$, evolved PylRSs have also been applied to incorporate N^{ε}-acetyl-L-lysine into proteins in both *E. coli* and mammalian cells.[52, 55, 85] Inspired by these advances, we thought it might be possible to incorporate either N^{ε}-methyl-L-lysine or its derivatives into proteins at amber mutation sites using an evolved PylRS-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair. This may allow the synthesis of proteins with monomethylated lysines for their functional investigations. Herein, we wish to show that a N^{ε}-methyl-L-lysine can be site-specifically installed into a protein by the genetic incorporation of a photocaged N^{ε}-methyl-L-lysine using an evolved PylRS-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair and the following deprotection under UV irradiation.

Experimental Section

General Experimental

All reactions involving moisture sensitive reagents were conducted in oven-dried glassware under an argon atmosphere. Anhydrous solvents were obtained through standard laboratory protocols. Analytical thin-layer chromatography (TLC) was performed on Whatman SiO_2 60 F-254 plates. Visualization was accomplished by UV irradiation at 254 nm or by staining with ninhydrin (0.3% w/v in glacial acetic acid/n-butyl alcohol 3:97). Flash column chromatography was performed with flash silica gel (particle size 32-63 μm) from Dynamic Adsorbents Inc (Atlanta, GA).

Specific rotations of chiral compounds were obtained at the designated concentration and temperature on a Rudolph Research Analytical Autopol II polarimeter using a 0.5 dm cell. Proton and carbon NMR spectra were obtained on Varian 300 and 500 MHz NMR spectrometers. Chemical shifts are reported as δ values in parts per million (ppm) as referenced to the residual solvents: chloroform (7.27 ppm for ^{1}H and 77.23 ppm for ^{13}C) or water (4.80 ppm for ^{1}H). A minimal amount of 1,4-dioxane was added as the reference standard (67.19 ppm for ^{13}C) for carbon NMR spectra in deuterium oxide, and a minimal amount of sodium hydroxide pellet was added to the NMR sample to aid in the solvation of amino acids which have low solubility in deuterium oxide under neutral or acidic conditions. ^{1}H NMR spectra are tabulated as follows: chemical shift, multiplicity (s = singlet, bs = broad singlet, d = doublet, t = triplet, q = quartet, m = multiplet), number of protons, and coupling constant(s). Mass spectra were obtained at the Laboratory for Biological Mass Spectrometry at the Department of Chemistry, Texas A&M University.

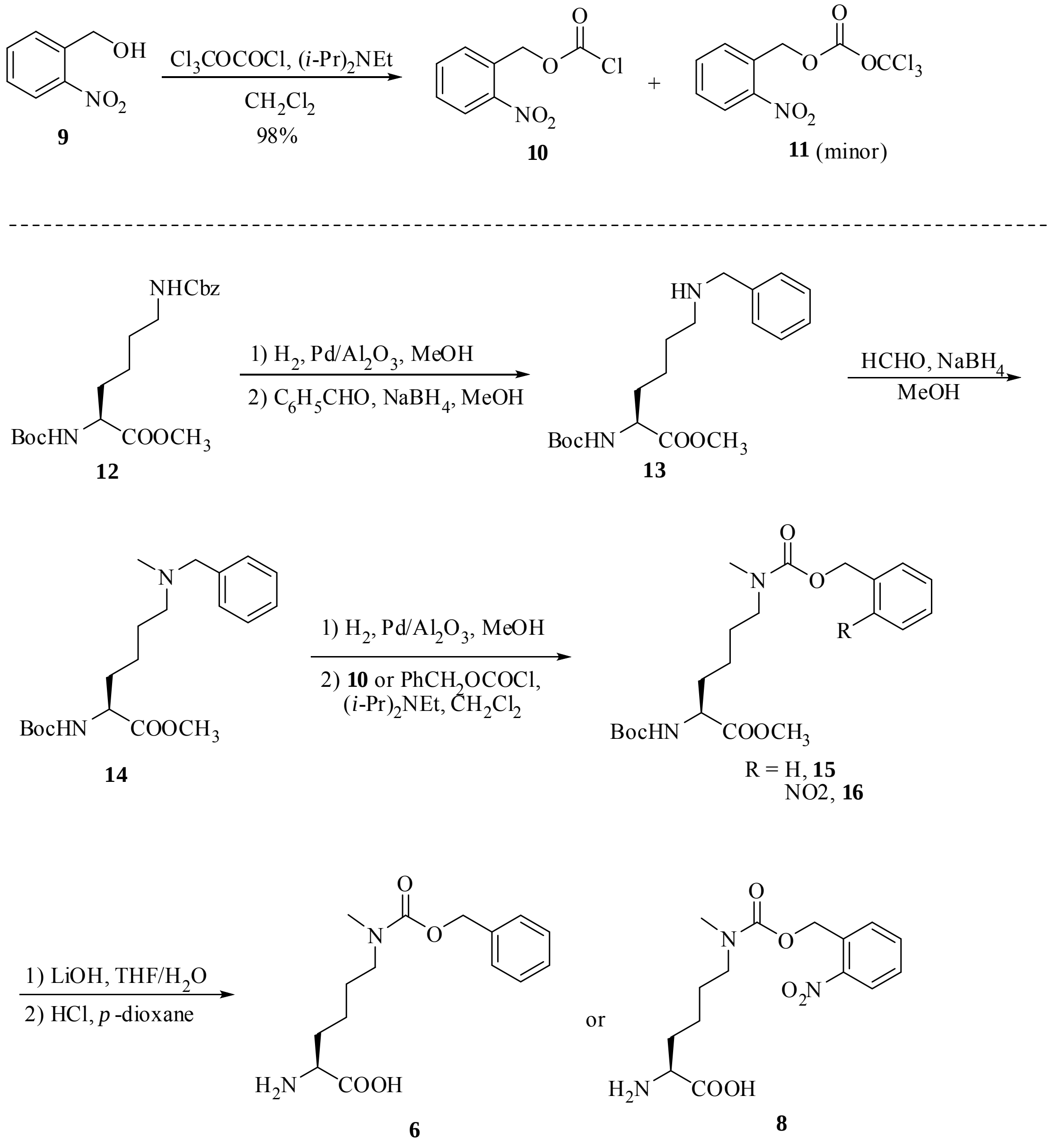
H-Lys(Z)-OH (**5**) and H-Lys(Me)-OH·HCl (**17**) was obtained from Chem-Impex International, Inc. (Wood Dale, IL). Compound **7** was synthesized as reported.[87] All other reagents were obtained from commercial suppliers and used as received.

Chemical Synthesis

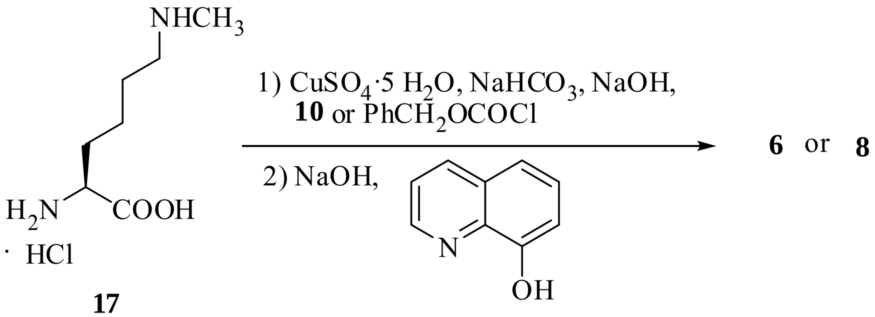
Compounds **6** and **8** were synthesized from **12**[88] in a scalable route (**Scheme II-1**). For comparison, a shorter synthesis from the relatively expensive N^{ε}-methyl-L-lysine hydrochloride (**17**) following the standard protocol of copper complexation, reaction with appropriate chloroformate, and decomplexation with 8-hydroxyquinoline[89] was also developed. (**Scheme II-2**)

2-Nitrobenzyl chloroformate (10) and 2-nitrobenzyl trichloromethyl carbonate (11).

To a solution of 2-nitrobenzyl alcohol (**9**, 1.97 g, 12.9 mmol) in anhydrous dichloromethane (55 mL) cooled in an ice bath was added diphosgene (1.71 mL, 14.2 mmol) in dichloromethane (10 mL) dropwise over 10 min followed by diisopropylethylamine (2.25 mL, 12.9 mmol) in dichloromethane (10 mL) dropwise over 10 min. The reaction mixture was then stirred at room temperature for 2 h, and sodium hydroxide (1 *N*, 20 mL) was added and stirred further at room temperature for 30 min. The mixture was washed with water (30 mL), saturated sodium bicarbonate (30 mL x 2)



Scheme II-1. Longer synthesis of **6** and **8**.



17

Scheme II-2. Shorter synthesis of **6** and **8**.

and brine (30 mL), dried (Na_2SO_4), and flash chromatographed (EtOAc/hexanes, 1:20) to give a mixture of **10** and **11** (2.71 g, 98%) as a yellow oil. A minor fraction of impurity, presumably **11**, was evident from NMR analysis but did not interfere with the next step reaction. No further purification was performed. For **10**^{90}: ^{1}H NMR ($CDCl_3$, 500 MHz) δ 8.19 (d, 1 H, J = 8.0 Hz), 7.74 (t, 1 H, J = 7.8 Hz), 7.65 (d, 1 H, J = 8.0 Hz), 7.58 (t, 1 H, J = 7.8 Hz), 5.75 (s, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 150.6, 147.7, 134.4, 129.9, 129.8, 129.2, 125.6, 69.7. For **11**: ^{1}H NMR ($CDCl_3$, 500 MHz) δ 8.20 (d, 1 H, J = 6.5 Hz), 7.74 (t, 1 H, J = 7.7 Hz), 7.68 (d, 1 H, J = 8.0 Hz), 7.58 (t, 1 H, J = 7.7 Hz), 5.75 (s, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 147.3, 147.2, 134.5, 130.2, 129.8, 129.0, 108.0, 68.0.

N$^{\alpha}$-Boc-*N*$^{\varepsilon}$-benzyl-L-lysine methyl ester (13)

A solution of **12** (4.20 g, 10.6 mmol) in methanol (100 mL) was hydrogenated under a H_2 balloon in the presence of palladium on alumina (10 wt.% Pd, 0.71 g, 0.67 mmol) at room temperature for 3 h, and TLC analysis showed a complete conversion.

The mixture was then filtered over a pad of Celite and the solution was directly used for the next step reaction. The material should be immediately used without purification since prolonged storage at room temperature or flash chromatography would contribute to lactam formation.

To a solution of the above amine (~10.6 mmol) in methanol was added benzaldehyde (4.00 mL, 39.4 mmol), and the reaction mixture was stirred at room temperature for 30 min. The mixture was then cooled in an ice bath, and sodium borohydride (0.75 g, 19.8 mmol) was added portionwise. The mixture was then stirred at room temperature overnight, and water (10 mL) was added dropwise to quench the reaction. Most of the methanol was evaporated under a reduced pressure, and the residue was dissolved in ethyl acetate (100 mL), washed with water (30 mL), saturated sodium bicarbonate (30 mL) and brine (30 mL), dried (Na_2SO_4), and evaporated to give the crude **13** as a yellow oil, which was used in the next step reaction without further purification. A small fraction of pure **13** was obtained by flash chromatography (10% methanol with 5% triethylamine in dichloromethane) for characterization. $[\alpha]_D^{19}$ +7.4 (*c* 4.85, CH_2Cl_2); ^{1}H NMR ($CDCl_3$, 500 MHz) δ 7.35-7.30 (m, 4 H), 7.27-7.24 (m, 1 H), 5.05 (d, 1 H, *J* = 7.0 Hz), 4.32-4.28 (m, 1 H), 3.78 (s, 2 H), 3.74 (s, 3 H), 2.63 (t, 2 H, *J* =7.0 Hz), 1.83-1.78 (m, 1 H), 1.67-1.60 (m, 1 H), 1.58-1.48 (m, 3 H), 1.44 (s, 9 H), 1.41-1.36 (m, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.5, 155.5, 140.5, 128.5, 128.2, 127.0, 79.9, 54.1, 53.5, 52.3, 49.1, 32.7, 29.7, 28.4, 23.2; HRMS (ESI) calcd for $C_{19}H_{31}N_2O_4$ ($[M + H]^+$) 351.2284, found 351.2282.

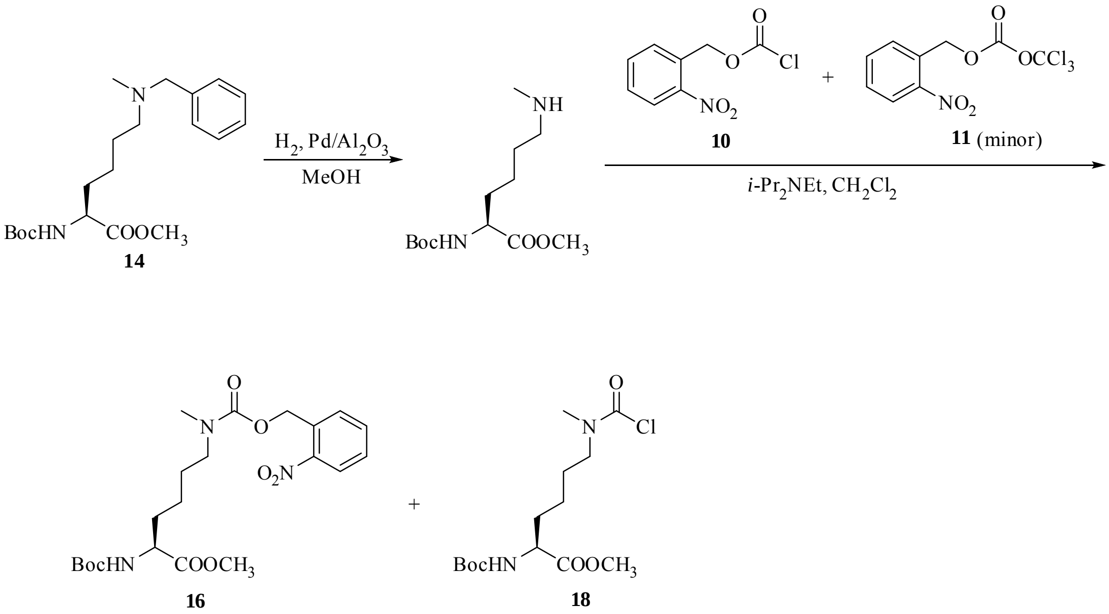
N^{α}-Boc-N^{ε}-benzyl-N^{ε}-methyl-L-lysine methyl ester (14)

To a solution of crude **13** (~10.6 mmol) in methanol (100 mL) in methanol was added formaldehyde (37% aqueous solution, 3.00 mL, 40.3 mmol), and the reaction mixture was stirred at room temperature for 30 min. The mixture was then cooled in an ice bath, and sodium borohydride (0.77 g, 20.4 mmol) was added portionwise. The mixture was then stirred further at room temperature for 4 h, and water (30 mL) was added dropwise to quench the reaction. Most of the methanol was evaporated under a reduced pressure, and the residue was dissolved in ethyl acetate (100 mL), washed with water (30 mL), hydrochloric acid (1 *N*, 30 mL), sodium hydroxide (1 *N*, 30 mL) and brine (30 mL), dried (Na_2SO_4), evaporated, and flash chromatographed (EtOAc/hexanes, 1:1 then 5% to 10% methanol in dichloromethane) to give **14** (3.32 g, 86% yield for three steps) as a yellow oil. $[\alpha]_D^{20}$ +7.7 (*c* 2.37, CH_2Cl_2); ^{1}H NMR ($CDCl_3$, 500 MHz) δ 7.33-7.29 (m, 4 H), 7.26-7.23 (m, 1 H), 5.08 (d, 1 H, J = 8.5 Hz), 4.31-4.27 (m, 1 H), 3.73 (s, 3 H), 3.46 (s, 2 H), 2.35 (t, 2 H, J =7.2 Hz), 2.17 (s, 3 H), 1.82-1.76 (m, 1 H), 1.66-1.47 (m, 3 H), 1.44 (s, 9 H), 1.40-1.32 (m, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.6, 155.6, 139.3, 129.2, 128.4, 127.1, 80.0, 62.5, 57.1, 53.6, 52.4, 42.4, 32.7, 28.5, 27.0, 23.3; HRMS (ESI) calcd for $C_{20}H_{33}N_2O_4$ ($[M + H]^+$) 365.2440, found 365.2437.

N^{α}-Boc-N^{ε}-Cbz-N^{ε}-methyl-L-lysine methyl ester (15)

A solution of **14** (2.58 g, 7.07 mmol) in methanol (50 mL) was hydrogenated under a H_2 balloon in the presence of palladium on alumina (10 wt.% Pd, 0.50 g, 0.47 mmol) at room temperature for 5 h. The mixture was then filtered over a pad of Celite and evaporated to give the crude amine (Boc-Lys(Me)-OMe) as a grey oil. A small fraction of pure amine was obtained by flash chromatography (10% methanol with 5% triethylamine in dichloromethane) for characterization. $[\alpha]_D^{19}$ +6.9 (*c* 1.90, CH_2Cl_2); ^1H NMR ($CDCl_3$, 500 MHz) δ 5.38 (d, 1 H, *J* = 7.0 Hz), 4.14-4.12 (m, 1 H), 3.59 (s, 3 H), 2.95 (s, 1 H), 2.46 (t, 2 H, *J* =7.2 Hz), 2.29 (s, 3 H), 1.69-1.64 (m, 1 H), 1.55-1.48 (m, 1 H), 1.43-1.33 (m, 2 H), 1.30 (s, 9 H), 1.27-1.22 (m, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.4, 155.5, 79.6, 53.3, 52.1, 51.2, 35.8, 32.2, 28.8, 28.2, 23.0; HRMS (ESI) calcd for $C_{13}H_{27}N_2O_4$ ($[M + H]^+$) 275.1971, found 275.1968.

To a solution of the above amine (~7.07 mmol) and diisopropylethylamine (2.00 mL, 11.5 mmol) in anhydrous dichloromethane (40 mL) cooled in an ice bath was added benzyl chloroformate (95%, 1.50 mL, 10.5 mmol) dropwise over 10 min, and the mixture was stirred at room temperature for 12 h. The mixture was then diluted in ethyl acetate (100 mL), washed with sodium hydroxide (0.5 *N*, 40 mL) and brine (40 mL), dried (Na_2SO_4), evaporated, and flash chromatographed (EtOAc/hexanes, 1:3) to give **15** (2.59 g, 90% for two steps) as a colorless oil. $[\alpha]_D^{19}$ +1.2 (*c* 1.51, CH_2Cl_2); ^1H NMR analysis showed a 1.5:1 mixture of rotamers at room temperature. Major rotamer: ^1H NMR ($CDCl_3$, 500 MHz) δ 7.33-7.32 (m, 4 H), 7.29-7.26 (m, 1 H), 5.26 (d, 1 H, *J* = 6.0 Hz), 5.10 (s, 2 H), 4.24 (m, 1 H), 3.69 (s, 3 H), 3.32-3.19 (m, 2 H), 2.87 (s, 3 H), 1.78-1.66 (m, 2 H), 1.57-1.51 (m, 2 H), 1.41 (s, 9 H), 1.36-1.28 (m, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.4, 156.5, 155.6, 137.0, 128.5, 128.0, 127.9, 79.7, 67.0, 53.4, 52.2, 48.4, 33.9, 32.0, 28.4, 26.9, 22.2; Characteristic peaks of the minor rotamer: ^1H NMR ($CDCl_3$, 500 MHz) δ 5.05 (d, 1 H, *J* = 6.0 Hz); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 156.3, 155.4, 79.9, 53.3, 34.6, 32.4, 27.5, 22.5; HRMS (ESI) calcd for $C_{21}H_{33}N_2O_6$ ($[M + H]^+$) 409.2339, found 409.2332.



N^{α}-Boc-N^{ε}-(2-nitrobenzyl)oxycarbonyl-N^{ε}-methyl-L-lysine methyl ester (16) and N^{α}-Boc-N^{ε}-chlorocarbonyl-N^{ε}-methyl-L-lysine methyl ester (18)

Compound **14** (1.06 g, 2.91 mmol) was converted into the corresponding amine by hydrogenolysis, which was then treated with crude **10** (0.94 g, 4.38 mmol) according to the procedure for **15** to give **16** (0.91g, 69% for two steps) as a yellow oil. A small amount of **18** (yield not determined), the structure of which was assigned based on NMR and MS analysis data, was obtained as a colorless oil. Presumably Boc-Lys(Me)-OMe reacts with **11** to give **16** and generates one molecule of phosgene at the same time, which then acylates the residual Boc-Lys(Me)-OMe to afford **18.**

For **16**: $[\alpha]_D^{20}$ +8.9 (*c* 1.70, CH_2Cl_2); R_f = 0.46 (EtOAc/hexanes, 1:1); ^1H NMR analysis showed a 1.1:1 mixture of rotamers at room temperature. Major rotamer: ^1H NMR ($CDCl_3$, 500 MHz) δ 7.96 (d, 1 H, *J* = 8.0 Hz), 7.58 (t, 1 H, *J* = 7.5 Hz), 7.50-7.47 (m, 1 H), 7.39 (t, 1 H, *J* = 8.2 Hz), 5.42 (s, 2 H), 5.21 (d, 1 H, *J* = 7.5 Hz), 4.20-4.16 (m, 1 H), 3.63 (s, 3 H), 3.21 (appar. nonet, 2 H, *J* = 7.1 Hz), 2.86 (s, 3 H), 1.74-1.44 (m, 4 H), 1.34 (s, 9 H), 1.30-1.24 (m, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.2, 155.6, 155.4, 147.5, 133.6, 133.1, 128.7, 128.5, 124.8, 79.6, 63.7, 53.2, 48.5, 33.8, 31.9, 28.2, 26.7, 22.2; Characteristic peaks of the minor rotamer: ^1H NMR ($CDCl_3$, 500 MHz) δ 5.13 (d, 1 H, *J* = 7.5 Hz), 2.83 (s, 3 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.1, 155.5, 155.4, 147.5, 128.9, 128.5, 79.7, 52.1, 48.4, 34.6, 32.3, 27.4, 22.4.

For **18**: $[\alpha]_D^{20}$ +6.6 (*c* 2.20, CH_2Cl_2); R_f = 0.61 (EtOAc/hexanes, 1:1); ^1H NMR analysis showed a 1.2:1 mixture of rotamers at room temperature. Major rotamer: ^1H NMR ($CDCl_3$, 500 MHz) δ 5.06 (d, 1 H, *J* = 8.0 Hz), 4.30 (m, 1 H), 4.20-3.73 (s, 3 H), 3.44 (t, 1 H, *J* = 7.5 Hz), 3.40-3.36 (m, 1 H), 3.10 (s, 3 H), 1.85-1.56 (m, 4 H), 1.43 (s, 9 H), 1.40-1.31 (m, 2 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 173.3, 155.5, 149.8, 80.1, 53.2, 52.8, 51.2, 38.6, 32.5, 28.5, 26.7, 22.4; Characteristic peaks of the minor rotamer: ^1H NMR ($CDCl_3$, 500 MHz) δ 3.74 (s, 3 H), 3.02 (s, 3 H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 149.3, 80.2, 52.5, 36.8, 32.7, 27.4, 22.4; HRMS (ESI) calcd for $C_{14}H_{25}ClN_2O_5Na$ ([M + Na]$^+$) 361.1320 (^{37}Cl)/359.1350 (^{35}Cl), found 361.1348/ 359.1359; calcd for $C_{14}H_{26}ClN_2O_5$ ([M + H]$^+$) 339.1505/337.1530, found 339.1585/ 337.1557.

N^{ε}-Benzyloxycarbonyl-N^{ε}-methyl-L-lysine (6)

To a solution of **15** (2.59 g, 6.34 mmol) in THF (20 mL) was added lithium hydroxide solution (0.5 M, 25.0 mL, 12.5 mmol), and the mixture was stirred at room temperature for 3 h. The mixture was diluted in water (20 mL) and extracted with ether (30 mL x 2). The ether extracts were discarded, and the remaining aqueous solution was adjusted to pH 3 with hydrochloric acid (3 *N*), with the concomitant formation of white precipitate. The suspension was extracted with ethyl acetate (50 mL x 2), and the combined organic phases were washed once with brine (30 mL), dried (Na_2SO_4), and evaporated to give the crude carboxylic acid as a colorless oil, which was directly used without further purification.

The above crude acid (~6.34 mmol) was dissolved in 1,4-dioxane (15 mL), and hydrogen chloride in 1,4-dioxane (4.0 M, 5.0 mL, 20.0 mmol) was added. The resulting white suspension was stirred at room temperature for 12 h, evaporated, redissolved in a minimal amount of water, and loaded onto an ion-exchange column made from Dowex 50WX4-400 cation-exchange resin (~14 mL bed volume). The column was washed with excessive water (300 mL) and then eluted with pyridine (1 M, 450 mL) to give **6** (1.51 g, 81% for two steps) as a white powder. $[\alpha]_D^{20}$ +14.1 (*c* 1.07, 3 *N* HCl) (lit.[91] $[\alpha]_D^{25}$ +14.0 (*c* 0.5, acetic acid)); ^{1}H NMR analysis showed a 1:1 mixture of rotamers at room temperature. Major rotamer: ^{1}H NMR (D_2O, 500 MHz) δ 7.46-7.42 (m, 5 H), 5.16 (s, 2 H), 3.68 (m, 1 H), 3.34 (m, 2 H), 2.90 (s, 3 H), 1.83 (m, 2 H), 1.60 (quintet, 2 H, J = 7.3 Hz), 1.34 (m, 2 H); ^{13}C NMR (D_2O, 75 MHz) δ 184.1, 158.4, 137.1, 129.4, 129.0, 128.5, 68.0, 56.5, 49.0, 35.1, 34.8, 27.6, 22.8; Characteristic peaks of the minor rotamer: ^{1}H

NMR (D_2O, 500 MHz, pH = 14) δ 2.95 (s, 3 H); ^{13}C NMR (D_2O, 75 MHz) δ 128.2, 67.2, 49.2, 34.2, 27.3. HRMS (ESI) calcd for $C_{15}H_{23}N_2O_4$ ([M + H]$^+$) 295.1658, found 295.1656.

***N*$^\varepsilon$-(2-Nitrobenzyl)oxycarbonyl-*N*$^\varepsilon$-methyl-L-lysine (8)**

According to the same procedure for **6**, **16** (0.914 g, 2.02 mmol) afforded **8** (0.514 g, 75% for two steps) as a pale yellow solid. $[\alpha]_D^{20}$ +14.2 (*c* 1.16, 3 *N* HCl); ^{1}H NMR analysis showed a 1.1:1 mixture of rotamers at room temperature. Major rotamer: ^{1}H NMR (D_2O, 500 MHz, pH = 14) δ 8.00 (d, 1 H, *J* = 8.0 Hz), 7.64 (t, 1 H, *J* = 7.7 Hz), 7.50 (d, 1 H, *J* = 8.5 Hz), 7.47 (t, 1 H, *J* = 8.2 Hz), 5.32 (s, 2 H), 3.95 (m, 1 H), 3.20 (m, 2 H), 2.77 (s, 3 H), 1.84 (m, 2 H), 1.49 (m, 2 H), 1.31-1.27 (m, 2 H); ^{13}C NMR (D_2O, 75 MHz) δ 184.2, 158.0, 147.6, 135.1, 132.7, 129.8, 129.4, 125.7, 65.0, 56.5, 49.2, 35.1, 34.9, 27.6, 22.8; Characteristic peaks of the minor rotamer: ^{1}H NMR (D_2O, 500 MHz, pH = 14) δ 2.82 (s, 3 H); ^{13}C NMR (D_2O, 75 MHz) δ 133.0, 129.7, 64.9, 49.3, 34.2, 27.2; HRMS (ESI) calcd for $C_{15}H_{22}N_3O_6$ ([M + H]$^+$) 340.1509, found 340.1513.

Synthesis of 6 from the shorter pathway

To a solution of **17** (1.50 g, 7.63 mmol) in water (30 mL) was added cupric sulfate pentahydrate (1.00 g, 4.00 mmol), followed by sodium bicarbonate (1.42 g, 16.90 mmol) in small portions to prevent excessive bubble formation. Benzyl chloroformate (2.53 g, 13.36 mmol) in dioxane (5 mL) was then added dropwise in 5 min, followed by sodium hydroxide (0.49 g, 12.25 mmol) in one portion. The reaction mixture was stirred at room temperature for 16 h, filtered, washed with water (100 mL), ethanol (50 mL) and diethyl ether (50 mL), and dried in the open air for 1 h to give the crude copper complex (2.40 g, 97%) as a blue solid.

All the above copper complex (2.40 g, ~7.40 mmol) was suspended in sodium hydroxide solution (0.2 *N*, 100 mL, 20 mmol), and a solution of 8-hydroxylquinoline (1.40 g, 9.64 mmol) in 1,4-dioxane (10 mL). The resulting green suspension was stirred at room temperature overnight and filtered. The filtrate was adjusted to pH 3 with hydrochloric acid (3 *N*) and extracted with ethyl acetate (40 mL x 2). The organic extracts were discarded, and the aqueous phase was concentrated to about 20 mL and loaded onto an ion-exchange column made from Dowex 50WX4-400 cation-exchange resin (~14 mL bed volume). The column was washed with excessive water (300 mL) and then eluted with pyridine (1 M, 450 mL) to give a yellow solid upon evaporation, which was suspended in ethanol, filtered, washed ethyl acetate dried to give **6** (1.46 g, 65% for two steps) as a white solid. $[\alpha]_D^{22}$ +15.3 (*c* 1.02, 3 *N* HCl). All other characterization data were identical to that of **6** from the longer route.

Synthesis of 8 from the shorter pathway

According to the same procedure for **6**, **17** (0.50 g, 2.54 mmol) afforded **8** (0.39 g, 45% yield for two steps). The compound was identical to **8** from the longer route in all aspects.

DNA and Protein Sequences

DNA Sequences

Z Domain:

atgactagtgtagacaactagatcaacaaagaacaacaaaacgccttctatgagatcttacatttacctaacctgaatgaggagc

agcgtgatgccttcatccaaagtttaaaagatgacccaagccaaagcgctaaccttttagcagaagctaaaaagctaaatgatgc
tcaggcgcctaagggatctgagctccatcaccatcaccatcactaa

GFP$_{UV}$:

Atgagtaaaggagaagaacttttcactggagttgtcccaattcttgttgaattagatggtgatgttaatgggcacaaattttctgtca
gtggagagggtgaaggtgatgcaacatacggaaaacttacccttaaatttatttgcactactggaaaactacctgttccatggcca
acacttgtcactactttctcttatggtgttcaatgcttttcccgttatccggatcacatgaaacggcatgactttttcaagagtgccatg
cccgaaggttatgtacaggaacgcactatatctttcaaagatgacgggaactacaagacgcgtgctgaagtcaagtttgaaggt
gatacccttgttaatcgtatcgagttaaaaggtattgattttaaagaagatggaaacattctcggacacaaactcgaatacaactat
aactcacacaatgtatacatcacggcagacaaacaaaagaatggaatcaaagctaacttcaaaattcgccacaacattgaagat
ggatccgttcaactagcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaaccattacctgtc
gacatagtctgccctttcgaaagatcccaacgaaaagcgtgaccacatggtccttcttgagtttgtaactgctgctgggattacac
atggcatggatgaactctacaaagagctccatcaccatcaccatcactaa

pylT:

ggaaacctgatcatgtagatcgaatggactctaaatccgttcagccgggttagattcccggggtttccgcca

Methanosarcina mazei PylRS:

atggataaaaaaccactaaacactctgatatctgcaaccgggctctggatgtccaggaccggaacaattcataaaataaaacac
cacgaagtctctcgaagcaaaatctatattgaaatggcatgcggagaccaccttgttgtaaacaactccaggagcagcaggact
gcaagagcgctcaggcaccacaaatacaggaagacctgcaaacgctgcagggtttcggatgaggatctcaataagttcctcac
aaaggcaaacgaagaccagacaagcgtaaaagtcaaggtcgtttctgcccctaccagaacgaaaaaggcaatgccaaaatcc
gttgcgagagccccgaaacctcttgagaatacagaagcggcacaggctcaaccttctggatctaaattttcacctgcgataccg
gtttccacccaagagtcagtttctgtcccggcatctgtttcaacatcaatatcaagcatttctacaggagcaactgcatccgcactg

gtaaaagggaatacgaaccccattacatccatgtctgcccctgttcaggcaagtgcccccgcacttacgaagagccagactga
caggcttgaagtcctgttaaacccaaaagatgagatttccctgaattccggcaagcctttcagggagcttgagtccgaattgctct
ctcgcagaaaaaaagacctgcagcagatctacgcggaagaaagggagaattatctggggaaactcgagcgtgaaattaccag
gttctttgtggacaggggttttctggaaataaaatccccgatcctgatccctcttgagtatatcgaaaggatgggcattgataatgat
accgaactttcaaaacagatcttcagggttgacaagaacttctgcctgagacccatgcttgctccaaacctttacaactacctgcg
caagcttgacagggccctgcctgatccaataaaaatttttgaaataggcccatgctacagaaaagagtccgacggcaaagaac
acctcgaagagtttaccatgctgaacttctgccagatgggatcgggatgcacacgggaaaatcttgaaagcataattacggactt
cctgaaccacctgggaattgatttcaagatcgtaggcgattcctgcatggtctatggggatacccttgatgtaatgcacggagac
ctggaactttcctctgcagtagtcggacccataccgcttgaccgggaatggggtattgataaaccctggataggggcaggtttc
gggctcgaacgccttctaaaggttaaacacgactttaaaaatatcaagagagctgcaaggtccgagtcttactataacgggattt
ctaccaacctgtaa

Proteins Sequences

Z Domain:

MTSVDNXINKEQQNAFYEILHLPNLNEEQRDAFIQSLKDDPSQSANLLAEAKKL
NDAQAPKGSELHHHHHH X represents a noncanonical amino acid.

GFP$_{UV}$:

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLP
VPWPTLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGNYK
TRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYITADKQKNGI
KANFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTXSALSKDPNEKR
DHMVLLEFVTAAGITHGMDELYKELHHHHHH

X represents a noncanonical amino acid.

Methanosarcina mazei PylRS:

MDKKPLNTLISATGLWMSRTGTIHKIKHHEVSRSKIYIEMACGDHLVVNNSRSS
RTARALRHHKYRKTCKRCRVSDEDLNKFLTKANEDQTSVKVKVVSAPTRTKK
AMPKSVARAPKPLENTEAAQAQPSGSKFSPAIPVSTQESVSVPASVSTSISSISTG
ATASALVKGNTNPITSMSAPVQASAPALTKSQTDRLEVLLNPKDEISLNSGKPFR
ELESELLSRRKKDLQQIYAEERENYLGKLEREITRFFVDRGFLEIKSPILIPLEYIER
MGIDNDTELSKQIFRVDKNFCLRPMLAPNLYNYLRKLDRALPDPIKIFEIGPCYR
KESDGKEHLEEFTMLNFCQMGSGCTRENLESIITDFLNHLGIDFKIVGDSCMVYG
DTLDVMHGDLELSSAVVGPIPLDREWGIDKPWIGAGFGLERLLKVKHDFKNIKR
AARSESYYNGISTNL

Construction of Plasmids

Constructions of pY+ and pY-

The plasmid pY+ was derived from the pRep plasmid by replacing the suppressor tRNA in pRep by pylT.[92] The gene of pylT flanked by the lpp promoter at the 5’ end and the *rrnC* terminator at the 3’ end was amplified from pBK-AcKRS-pylT.[55, 56] The plasmid pY- was derived from the pNeg plasmid by replacing the suppressor tRNA with pylT. Similarly, the gene of pylT flanked by the lpp promoter at the 5’ end and the *rrnC* terminator at the 3’ end was amplified from pBK-AcKRS-pylT. pY+ has a tetracycline selection marker, a chloramphenicol acetyltransferase gene with an amber mutation at D112. pY- has an ampicillin selection marker and a barnase gene with two amber mutations at Q2 and D44. The barnase gene is under control of a pBad promoter.

Construction of pET-pylT-GFP

Plasmid pET-pylT-GFP was derived from the plasmid pAcKRS-pylT-GFP1Amber in which GFP$_{UV}$ has an amber mutation at Q204. [55, 56] The restriction enzyme *BglII* was used to cut off the ACKRS gene. The digested pAcKRS-pylT-GFP1Amber plasmid was ligated to form pET-pylT-GFP.

Construction of pET-pylT-Z

The pET-pylT-Z plasmid was derived from pET-pylT-GFP. This gene was amplified from the pLeiZ plasmid.[93] Two restriction sites, *NdeI* at the 5' end and *SacI* at the 3' end, were introduced in the PCR product which was subsequently digested and used to replace GFP$_{UV}$ in pET-pylT-GFP.

Construction of the pRS1 Library

The plasmid pBK-mmPylRS that encodes wild-type Methanosarcina mazei PylRS was derived from a pBK plasmid containing *p*-iodophenylalanyl-tRNA synthetase.[94] The pylRS gene is under the control of *E. coli glnS* promoter and terminator. It was amplified from genomic DNA of *Methanosarcina mazei* strain DSM 3647 (ATCC) by flanking primers, pBK-mmPylRS-NdeI-F and pBK-mmPylRS-PstI~NsiI-R. To construct the pRS1 library, NNK (N=A or C or G or T, K=G or T) mutations were introduced at six sites by overlap extension PCR.[95] The following pairs of primers were used to generate a PylRS gene library with randomization at six sites: (1) pBK-mmPylRS-NdeI-F (5'-gaatcccatatggataaaaaaccactaaacactctg-3') and mmPylRS-Mutlib01-R (5'-ggccctgtcaagcttgcgmnngtagttmnnmnngtttggagcaagca tggg-3'); (2) mmPylRS-Mutlib02-F (5'-cgcaagcttgacagggccctgcctgatcc-3') and mmPylRS-Mutlib03-R

(5'-gcatcccgatcccatctgmnngaamnncagcatggtaaactcttc-3'); (3) mmPylRS-Mutlib04-F (5'-cagatgggatcgggatgcacacg-3') and mmPylRS-Mutlib05-R (5'-ccgaaacctgcccctatmnngggtttatcaatacccca-3'); (4) mmPylRS-Mutlib06-F (5'-ataggggcaggtttcgggctcgaacgcc-3') and pBK-mmPylRS-PstI~NsiI-R (5'-gtttgaaaatgcatttacaggttggtagaaatccc-3'). The gene library was digested with the restriction enzymes *NdeI* and *NsiI*, gel-purified, and ligated back into the pBK vector digested by *NdeI* and *PstI* to afford plasmid the pRS1 library. 1 μg of the ligation products were then electroporated into *E. coli* Top10 cells. Electroporated cells were recovered in SOC medium for 60 min at 37 °C, transferred into a 2 L 2YT medium with kanamycin (25 μg/mL) and were incubated at 37°C to OD_{600} at 1.0. To calculate the library size, 1 μL recovered SOC culture was subjected to serial dilutions in 2YT, then plated on LB agar plates with kanamycin (25 μg/mL), and grown overnight in a 37°C incubator. Based on the colony numbers on these plates, the pRS1 library contains approximately 1.01×10^9 independent transformants. Sequencing pylRS variants in 20 clones did not reveal any significant bias at the randomization sites.

Selection Procedure for Evolving Pyrrolysyl-tRNA Synthetase

The selections followed the scheme shown in **Scheme II-4**. For the positive selection, the pRS1 library was used to transform *E. coli* TOP10 competent cells harboring pY+ to yield a cell library greater than 1×10^9 cfu, ensuring complete coverage of the pRS1 library. Cells were plated on LB agar plates containing 12 μg/mL tetracycline (Tet), 25 μg/mL kanamycin (Kan), 68 μg/mL chloramphenicol (Cm) and 1 mM **5**. After incubation at 37°C for 72 h, colonies on the plates were collected. Total plasmids were isolated and separated by 1 % agarose gel electrophoresis. pRS1 plasmids were extracted using a Gel-extraction kit (QIAGEN). The extracted pRS1 plasmids from the positive selection were used to transform *E. coli* TOP10 harboring pY− for the negative selection. After electroporation, the cells were allowed to recover for 1 h at 37°C in SOC media before being plated on LB agar plates containing 50 µg/mL Kan, 200 µg/mL ampicillin (Amp) and 0.2% arabinose. The plates were incubated for 16 h at 37°C. Survived cells were then pooled and pRS1 plasmids were extracted. The selection power to exclude out the mutants that also took endogenous amino acids was tested on LB agar plates containing 50 µg/mL Kan, 200 µg/mL Amp, 0.2% arabinose and 1mM **5**. The plate contains 1 mM **5** showed much fewer colony numbers as times of negative selections increased. Five alternative selections (three positive + two negative) finally yielded many colonies. 22 single colonies after the third positive selection were selected and the plasmids were isolated for sequencing. 96 single colonies from the third positive selection were also chosen for testing their ability to grow on plates with 102 µg/mL chloramphenicol, 25 µg/mL Kan, 12 µg/mL Tet, and 1 mM of **5**, **6**, **7** or **8**. A plate without NAA supplementary was used as a control. Images of colonies growing on different plates were shown in **Figure II-1**. Sequences of PylRS variants that charge pylT with different NAAs are presented in **Table II-1&2**.

Protein Purification and Photolysis to Form the Monomethylated Protein

To express GFP$_{UV}$ incorporated with a NAA, we cotransformed *E. Coli* BL21(DE3) cells with pBK-mKRS1 and pET-pylT-GFP. Cells were recovered in 1 mL of LB medium for 1 h at 37 ºC before being plated on LB agar plate containing Kan (25

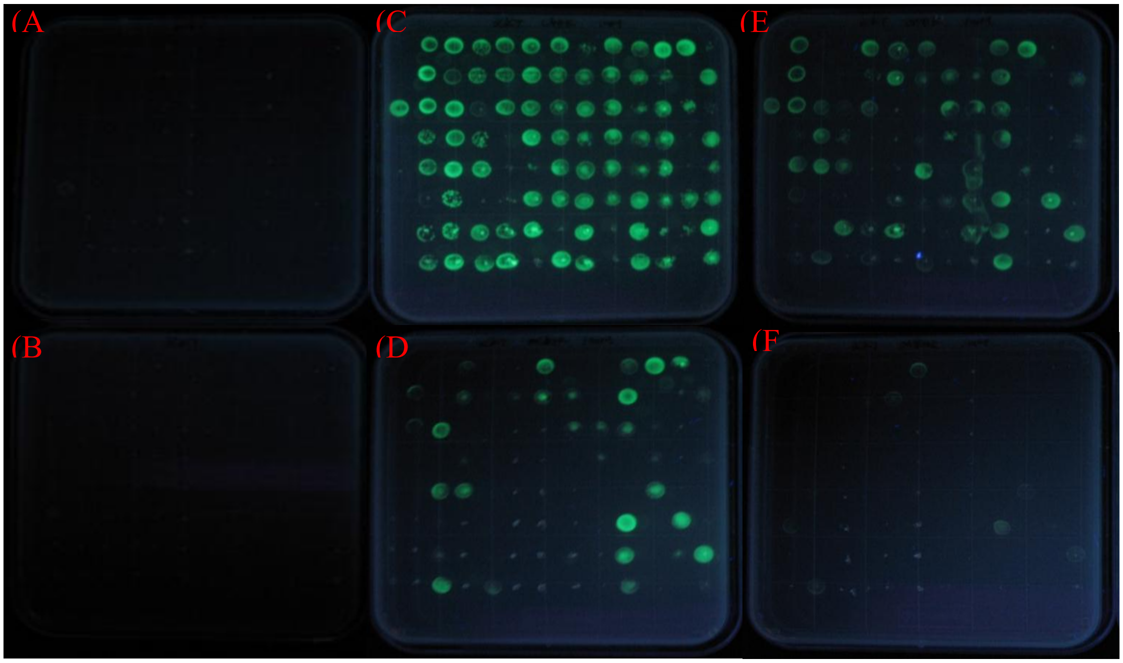


Figure II-1. Growth of 96 single colonies from the third positive selection of **5** on LB plates with different supplements. (**A**) 68 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet; (**B**) 102 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet; (**C**) 1 mM **5**, 102 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet; (**D**) 1 mM **6**, 102 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet; (**E**1 mM **7**, 102 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet; (**F**) 1 mM **8**, 102 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet. The pY+ plasmid has a GFP$_{UV}$ gene under control of a T7 promoter. Its expression is promoted by the suppression of two amber mutations at positions 1 and 107 of a T7 RNA polymerase gene in pREP. The fluorescent intensity of the expression of GFP$_{UV}$ roughly represents the suppression efficiency at amber codons.

μg/mL) and Amp (100 μg/mL). A single colony was then selected and grown overnight in a 10 mL culture. This overnight culture was used to inoculate 100 mL of M9 minimal media supplemented with 1% glycerol, 300 μM leucine, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.2% NaCl, 25 μg/mL Kan and 100 μg/mL Amp. Cells were grown at 37ºC in an incubator (300 r.p.m.) and protein expression was induced when OD_{600} was 0.7 by adding IPTG to a final concentration of 1 mM and **5** to a final concentration of 1 mM. After 6 h induction, cells were harvested, resuspended in a lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0) and sonicated. The cell lysate was clarified by centrifugation (60 min, 11,000 g, 4ºC). The supernatant was injected into a 30 mL Ni^{2+}-NTA column (Qiagen) on FPLC (ÄKTApurifier^{TM}, GE Healthcare Bio-Sciences Corp) and washed with 45 mL lysis buffer and 45 mL wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 40 mM imidazole, pH 8.0). Protein was finally eluted out by running an imidazole gradient from 40 mM to 250 mM in lysis buffer. Pure fractions were collected and concentrated. The buffer was later changed to 1 mM ammonium bicarbonate using an Amicon Ultra -15 Centrifugal Filter Devices (10,000 MWCO cut) (Millopore). The purified proteins were analyzed by 15% SDS-PAGE. GFP_{UV} proteins incorporated with other NAAs were expressed and purified similarly except the supplemented NAA was changed. For all NAAs, 1 mM final concentration was used.

Z domain proteins incorporated with different NAAs were expressed and purified as same as the expression of GFP_{UV} proteins except pET-pylT-Z was used to cotransform *E. coli* BL21(DE3) together with pBK-mKRS1. 1 mM final concentration was used for all four NAAs. The buffer of the finally purified Z domain proteins was changed to 1 mM ammonium bicarbonate using an Amicon centriplus YM-3 (3,000 MWCO cut) (Millopore). The purified proteins were analyzed by 15% SDS-PAGE. **Z-7**, **Z-8**, and **GFP-8** (1 mg/mL) in 1 mM ammonium bicarbonate solution were treated with 365 nm UV light form hand-held UV light source for one hour.

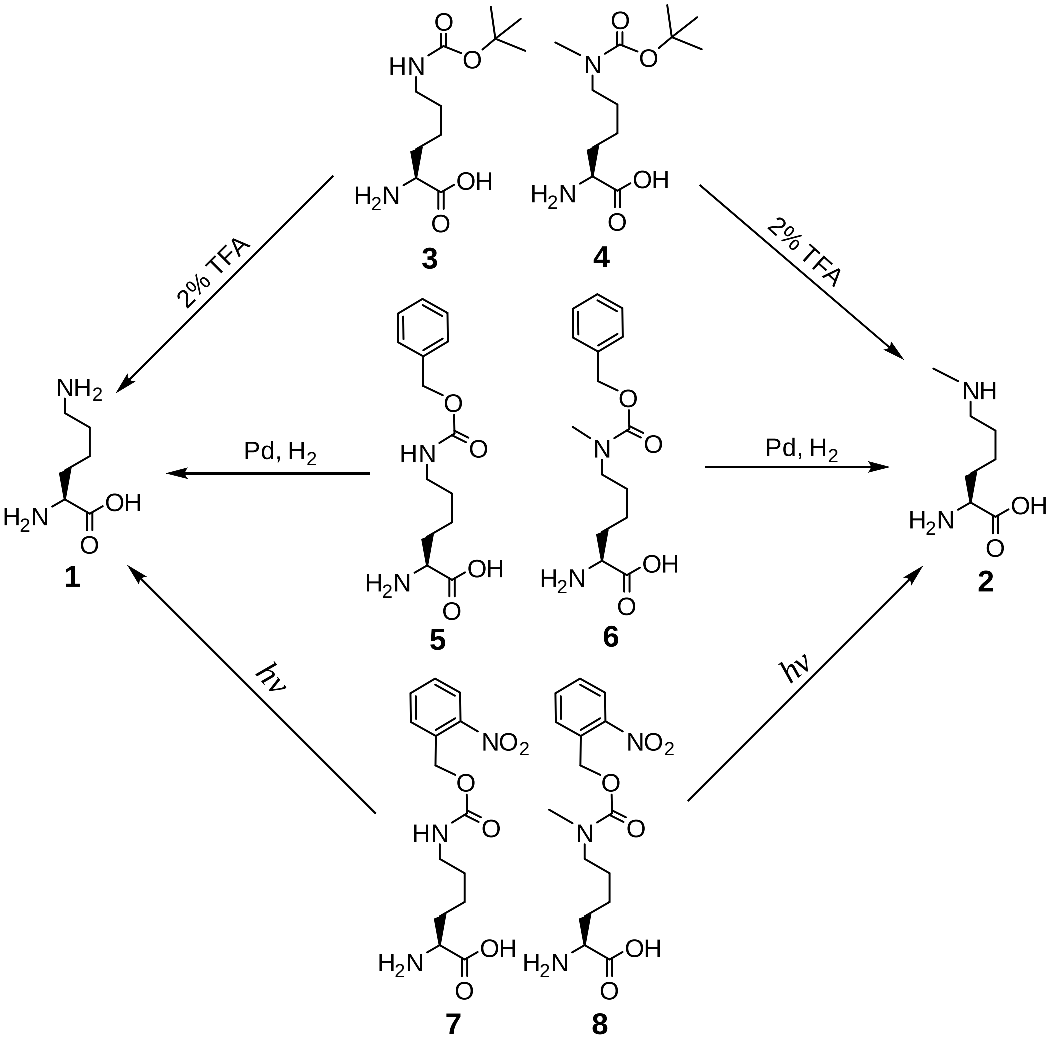
Protein LC-ESI-MS Analysis

An Agilent (Santa Clara, CA) 1200 capillary HPLC system was interfaced to an API QSTAR Pulsar Hybrid QTOF mass spectrometer (Applied Biosystems/MDS Sciex, Framingham, MA) equipped with an electrospray ionization (ESI) source. Liquid chromatography (LC) separation was achieved using a Phenomenex Jupiter C4 microbore column (150 × 0.50 mm, 300 Å) (Torrance, CA) at a flow rate of 10 µL per min. The proteins were eluted using a gradient of (A) 0.1% formic acid versus (B) 0.1% formic acid in acetonitrile. The gradient timetable was as follows: 2% B for 5 min, 2-30% in 3 min, 30-60% in 44 min, 60-95% in 8 min, followed by holding the gradient at 95% for 5 min, for a total run time of 65 min. The MS data were acquired in positive ion mode (500-1800 Da) using spray voltage of +5000 V. BioAnalyst software (Applied Biosystems) was used for spectral deconvolution. For the GFPuv protein analysis, a mass range of m/z 500-1800 was used for deconvolution and the output range was 10000-50000 Da using a step mass of 0.1 Da and a S/N threshold of 20. For the Z-Domain protein analysis, a mass range of m/z 500-2000 was used for deconvolution and the output range was 5000-15000 Da for Z-domain-His6X using a step mass of 0.1 Da and a S/N threshold of 20.

Results and Discussion

We chose to work on the genetic incorporation of protected N^{ε}-methyl-L-lysines instead of N^{ε}-methyl-L-lysine itself because it would be difficult to identify an evolved PylRS that specifically recognizes N^{ε}-methyl-L-lysine but not cellularly abundant lysine. Three protected N^{ε}-methyl-L-lysines (**4**, **6** and **8** in **Scheme II-3**) were initially considered but eventually we chose **6** and **8**. The cleavage of the Boc protection group from **4** to recover N^{ε}-methyl-L-lysine needs to be carried out under a strong acidic condition. Although it has been demonstrated that wild type PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair can genetically incorporate **3**[85] that is structurally close to **4** into proteins at amber codons and it is highly possible the same pair will also incorporate **4** into proteins, we thought the harsh condition for the deprotection is not suitable for many proteins (the incorporation of **4** into proteins using wild type PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair in *E. coli* was published by Chin *et al.*[66]). The strong acidic condition for the deprotection denatures proteins that have to be refolded later. However, refolding is problematic for most large size proteins. On the contrary, the deprotection of **6** by catalytic hydrogenation and **8**, a photocaged N^{ε}-methyl-L-lysine by UV photolysis can be achievable under mild conditions that are suitable for most proteins. In addition, the photolysis of **8** after its incorporation into proteins may also be carried out in living cells. This may allow the synthesis of methylated proteins directly in cells for their functional investigations.

The synthesis of both **6** and **8** started from lysine derivative[96] (**Scheme II-3**) and finished in gram quantities. To expand the substrate scope of PylRS to accommodate **6**



Scheme II-3: L-Lysine and N^{ε}-methyl-L-lysine derivatives and their deprotection.

or **8**, we constructed an active-site mutant library of the *Methanosarcina mazei* PylRS gene with randomization at six active site residues (L305, Y306, L309, N346, C348, and W417) (**Figure II-2**) according to a standard protocol.[97] This gene library was then cloned into a pBK plasmid to form a pRS1 plasmid library in which mutant PylRS variants are under control of a constitutive *glnS* promoter.[97] Together with two selection plasmids, pY+ for positive selection and pY- for negative selection, this plasmid library was then subjected to alternative positive and negative selections to identify PylRS variants specific for **6** or **8** (**Scheme II-4**).[52, 96] The positive selection plasmid, pY+ contains genes encoding pylT and type I chloramphenicol (Cm) acetyltransferase with an amber mutation at D112. Cotransforming *E. coli* with pY+ and pRS1 and then growing cells in Cm and NAA-containing plates conferred the selection of PylRS variants that charge pylT with native amino acids or the supplied NAA. The negative selection plasmid, pY- contains genes encoding pylT and toxic barnase that has two amber mutations at Q2 and D44. Cotransforming *E. coli* with pY- and pRS1 and then growing cells in plates without NAA conferred the selection of PylRS variants that do not charge pylT with a native amino acid. Only PylRS variants that charge pylT with the provided NAA but not any native amino acid will survive from both positive and negative selections. However, a series of selections (three positive selections + two negative selections) yielded no viable clones. The selected mutants of N^{ε}-acetyl-L-lysyl-tRNA synthetase (AcKRS) from Söll amd Chin Groups were reported in high K_m vales around 7 to 35 mM recently.[98] It suggested the selection system is able to selected low affinity mutant amd the possible colnes charging **6** and **8** have too high K_m to be stabilized in the

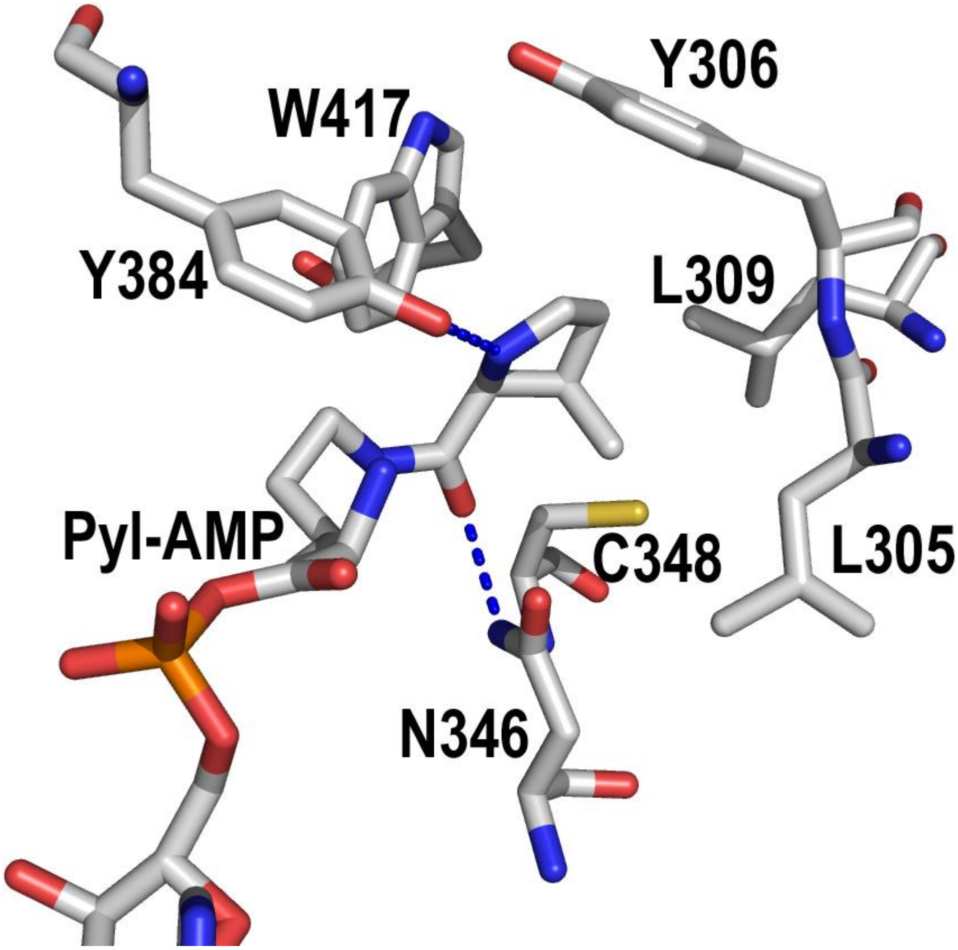
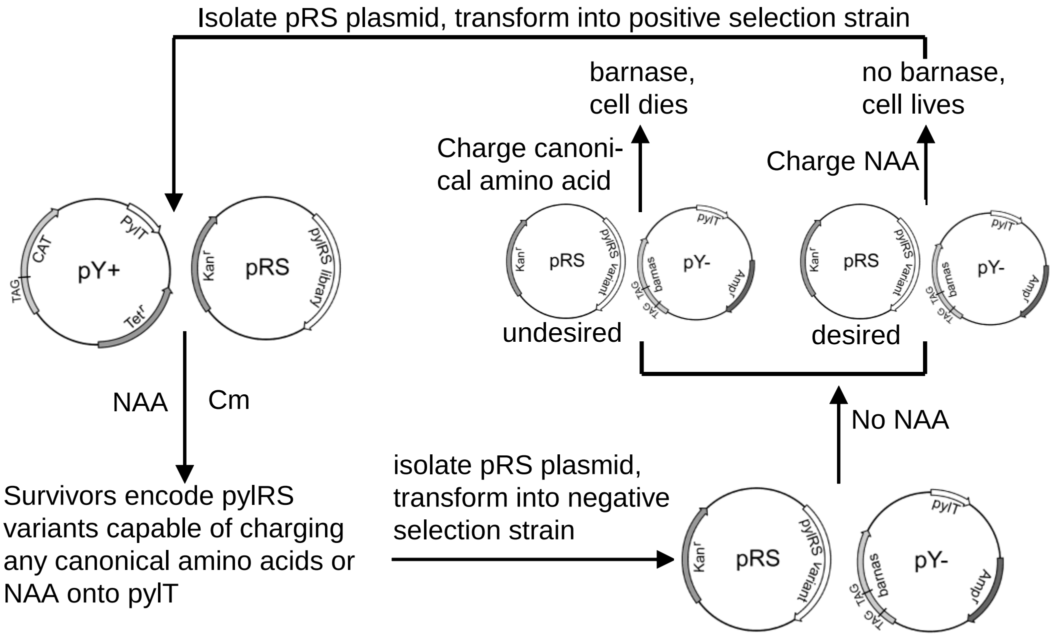


Figure II-2: The active site of PylRS. The structure was derived from pdb entry: 2Q7E.



Scheme II-4. The selection scheme to identify PylRS variants specific for a noncanonical amino acid.

active site of PylRS. We then chose an indirect route to identify clones specific for **6** or **8**. Since **5** and **7** are structurally close to **6** and **8**, respectively. We thought PylRS variants selected for **5** or **7** might also charge pylT with **6** or **8**, respectively. We carried out the selections of the pRS1 library to identify PylRS variants that are specific for **5**. After a series of selections, many colonies survived and most of them converge to two specific clones, in which mKRS1 shows the highest suppression efficiency and has mutations Y306M/L309A/C348T/T364K (**Table II-1&2**). We then tested the efficiency of the evolved mKRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair to suppress an amber mutation at Q204 of $\mathrm{GFP}_{\mathrm{UV}}$ in *E. coli*. A plasmid pET-pylT-GFP was constructed. It contains genes encoding $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ and $\mathrm{GFP}_{\mathrm{UV}}$ with an amber mutation at Q204. The $\mathrm{GFP}_{\mathrm{UV}}$ gene is under control of a T7 promoter. Cotransforming *E. coli* BL21(DE3) cells with the selected pBK-mKRS1 and pET-pylT-GFP and growing cells in minimal medium supplemented with **5** afforded full-length $\mathrm{GFP}_{\mathrm{UV}}$. No full-length $\mathrm{GFP}_{\mathrm{UV}}$ was expressed when **5** was excluded from the medium (**Figure II-3**). As what we expected, the evolved mKRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair could also incorporate **6** at an amber codon position. When **6** instead of **5** was provided in the medium, full-length $\mathrm{GFP}_{\mathrm{UV}}$ was also expressed (**Figure II-3**). To our surprise, the evolved mKRS1 can also charges $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ with **7** and **8**. When **7** or **8** was provided in the medium, growing cells transformed with pBK-mKRS1 and pET-pylT-GFP also afforded full-length $\mathrm{GFP}_{\mathrm{UV}}$ (**Figure II-3**). We also tested the efficiency of the mKRS1-pylT pair to suppress an amber mutation at K7 of Z domain. A plasmid pET-pylT-Z was constructed by replacing the $\mathrm{GFP}_{\mathrm{UV}}$ gene in pET-pylT-GFP with Z domain that contains an amber mutation at K7. Cotransforming *E. coli* BL21(DE3) cells

Table II-1: Sequences of selected PylRS variants that charge pylT with all four NAAs[a]

PylRS	Frequency	L305	Y306	L309	N346	C348	W417
mKRS1[b]	7/22	L	M	A	N	T	W
mKRS3	6/22	L	M	A	N	C	W
mKRS5	1/22	L	M	P	N	C	W

[a]Other PylRS clones are presented in **Table II-2**. [b]This clone has an additional mutation T364K.

Table II-2. Evolved PylRS variants that charge pylT with different NAAs.

Position	305	306	309	346	348	477	Remark[1]
WT	L	Y	L	N	C	W	
mKRS1[2]	L	M	A	N	T	W	**5, 7,6, 8**
mKRS2	L	V	A	N	A	W	**5, 7, 6**
mKRS3[3]	L	M	A	N	C	W	**5, 7,6, 8**
mKRS4	L	A	A	H	L	W	**5, 6**
mKRS5	L	M	P	N	C	W	**5, 6,7, 8**
mKRS6	L	M	A	N	S	W	**5**
mKRS7	L	Y	A	N	A	W	**5**
mKRS8[4]	L	M	T	N	A	W	**5**
mKRS9	L	A	A	N	A	W	**5**
mKRS10	L	A	L	N	A	W	**5, 7**
mKRS11[5]	L	A	L	N	C	W	**5, 7**

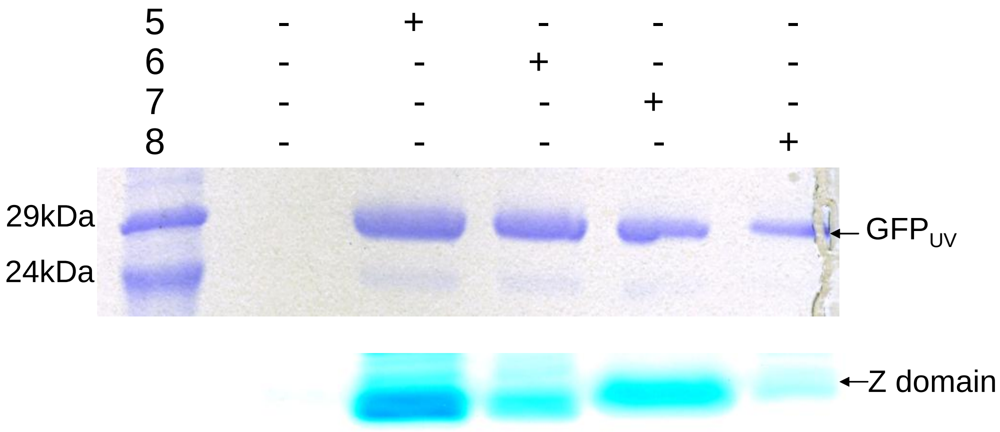
[1] This column represents the NAAs that can be taken by mutant PylRS variants. The order of compounds also indicates the decreasing encoding efficiency based on the screening results.

[2] The mutant was found seven times from 22 sequenced mutants and has extra mutation on T364K.

[3] The mutant was found six times from 22 sequenced mutants.

[4] The mutant has extra mutation on P297S.

[5] The mutant was found twice from 22 sequenced mutants.


5 - + - - -
6 - - + - -
7 - - - + -
8 - - - - +
29kDa
24kDa
GFP$_{UV}$
Z domain

with pBK-mKRS1 and pET-pylT-Z and then growing cells in the medium supplemented with either of **5**, **6**, **7**, and **8** afforded full-length Z domain. A trace amount of Z domain was expressed when no NAA was provided in the medium (**Figure II-3**).

To prove the incorporation of NAAs, purified Z domain proteins were then analyzed by electrospray ionization mass spectrometry (ESI-MS). (**Figure II-4-8**) Before analysis, Z domain containing **7** (Z-**7**) and Z domain containing **8** (Z-**8**) were photolysed under 365 nm UV light for an hour. As demonstrated previously, this treatment should efficiently cleave the photocaging group. The ESI-MS analysis confirmed the expected mass for all four Z domain proteins (**Table II-3**). For both Z-**7** and Z-**8**, the photocaging group was efficiently cleaved off. Only peaks corresponding to Z domain that contains lysine or N^{ε}-methyl-L-lysine at K7 could be detected. The ESI-MS analysis of the deprotected **Z-8** also revealed two additional peaks at 8005 Da and 8136 Da. These two peaks match mass of Z domain that contains lysine, glutamate or glutamine at K7. Since mKRS1 was evolved against endogenous native amino acids and these two additional peaks are significant in the ESI-MS spectrum, direct incorporation of lysine, glutamate, or glutamine at K7 is not likely. We thought the additional mass peak was due to the incorporation of **7** that happened to be a contaminant in **8**. this is highly possible since we synthesized **8** from lysine. A very small amount of **7** might end up in **8**. We carefully examined all the NMR and MS data of **8** and concluded the contaminant was lower than 1% and insignificant. However, **7** could become significant after its incorporation into Z domain because the evolved mKRS1 has a higher affinity to **7** than **8**. Similarly, two additional mass peaks at 8139 Da and 8181 Da corresponding to

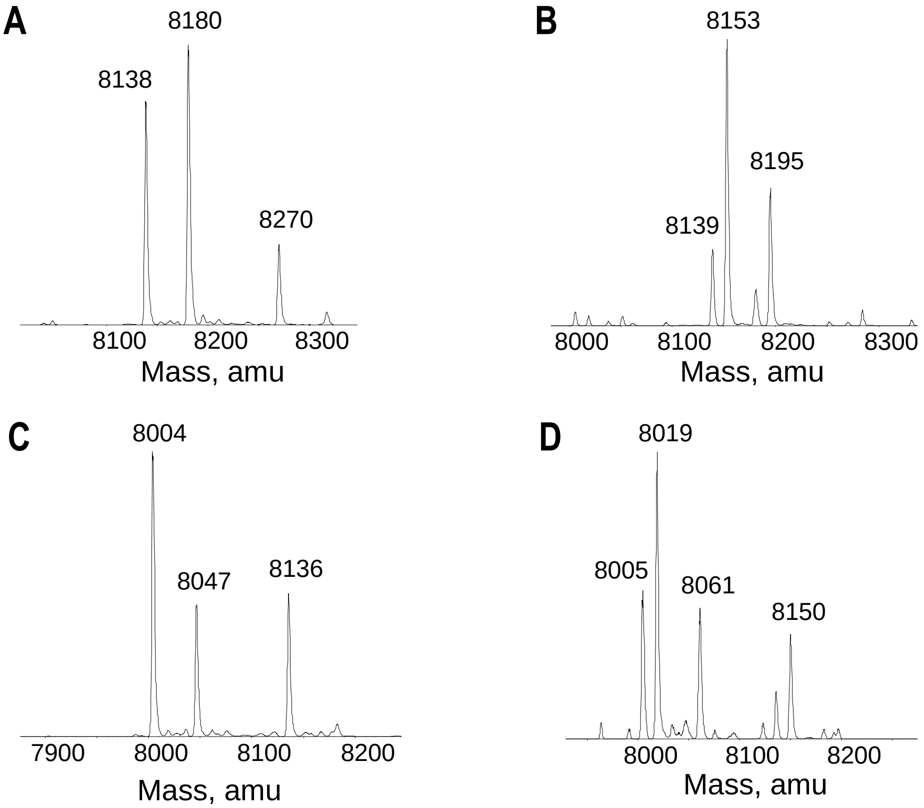
Table II-3: Z domain expression yields and MS characterization

Proteins	Yield[a] (mg/L)	Calculated Mass (Da)	Detected Mass (Da)
Z-5	9.3	8270[b]	8270
		8181[c]	8180
		8139[d]	8138
Z-6	4.0	8284[b]	
		8195[c]	8195
		8153[d]	8153
Z-7[e]	5.3	8136[b]	8136
		8047[c]	8047
		8005[d]	8004
Z-8[e]	1.2	8150[b]	8150
		8061[c]	8061
		8019[d]	8019

[a] Proteins were expressed in minimal media supplemented with 1% glycerol and 1 mM NAA. [b] Full-legnth Z domain proteins. [c] Full-length Z domain without N-terminal methionine but with a N-terminal acetylation. [d] Full-length protein without N-terminal methionine. [e] Both proteins were deprotected by UV irradiation.

Z domain containing **5** (Z-**5**) was also present in the ESI-MS spectrum of Z domain incorporated with **6** (Z-**6**). These two peaks are clearly the mass peaks of Z-**5**. This indicates the existance of **5** as a contaminant in **6**, though the contaminant was also lower than 1%. To eliminate **7** from **8**, we synthesized **8** from commerically available N^{ε}-methyl-L-lysine following a route presented in **Scheme II-2**. The purified **8** was then used to express $\mathrm{GFP_{UV}}$ containing **8** at Q204 (GFP-**8**). The purified GFP-**8** before and after 1 hour photolysis was then analyzed by ESI-MS (**Figure II-9&10**). The spectrum of GFP-**8** before photolysis showed one major peak at 27904 Da corresponding to GFP-**8** without N-terminal methionine (calculated mass: 27903 Da). No mass peak corresponding to $\mathrm{GFP_{UV}}$ incorporated with **7** (GFP-**7**) was identified. Similarly, the spectrum of GFP-**8** after photolysis showed one major peak at 27725 Da. It matches the calculated mass (27724 Da) of $\mathrm{GFP_{UV}}$ incorporated with N^{ε}-methyl-L-lysine at Q204. No mass peak corresponding to $\mathrm{GFP_{UV}}$ incorporated with lysine at Q204 was present.

From the ESI-MS spectral data of deprotected Z-**7**, Z-**8**, and GFP-**8**, it is clear that deprotecting the photocaging group under UV light is very efficient. No additional treatment is necessary. We have also tried palladium black catalyzed hydrogenation to deprotect Z-**6**. However, the protein either aggregated or did not show any detectable deprotection. We are currently searching for homogenous hydrogenation catalysts that can efficient deprotect Cbz group from Z-**6.**


A
8180
8138
8270
8100
8200
8300
Mass, amu
B
8153
8195
8139
8000
8100
8200
8300
Mass, amu
C
8004
8047
8136
7900
8000
8100
8200
Mass, amu
D
8019
8005
8061
8150
8000
8100
8200
Mass, amu

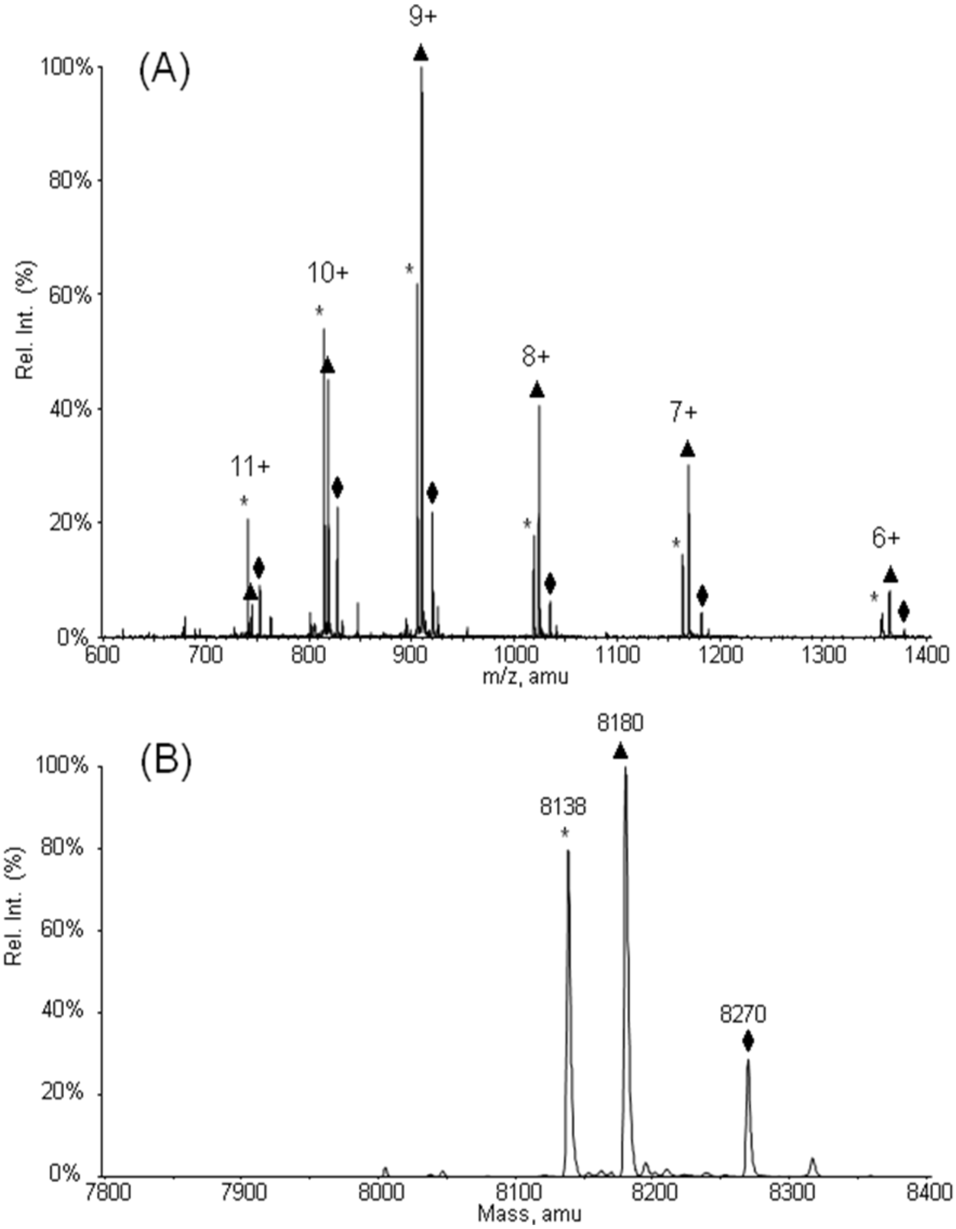


Figure II-5: Mass determination of Protein Z-**5** (A) ESI-MS spectrum of Z-**5** (B) The deconvoluted ESI-MS spectrum of Z-**5**.

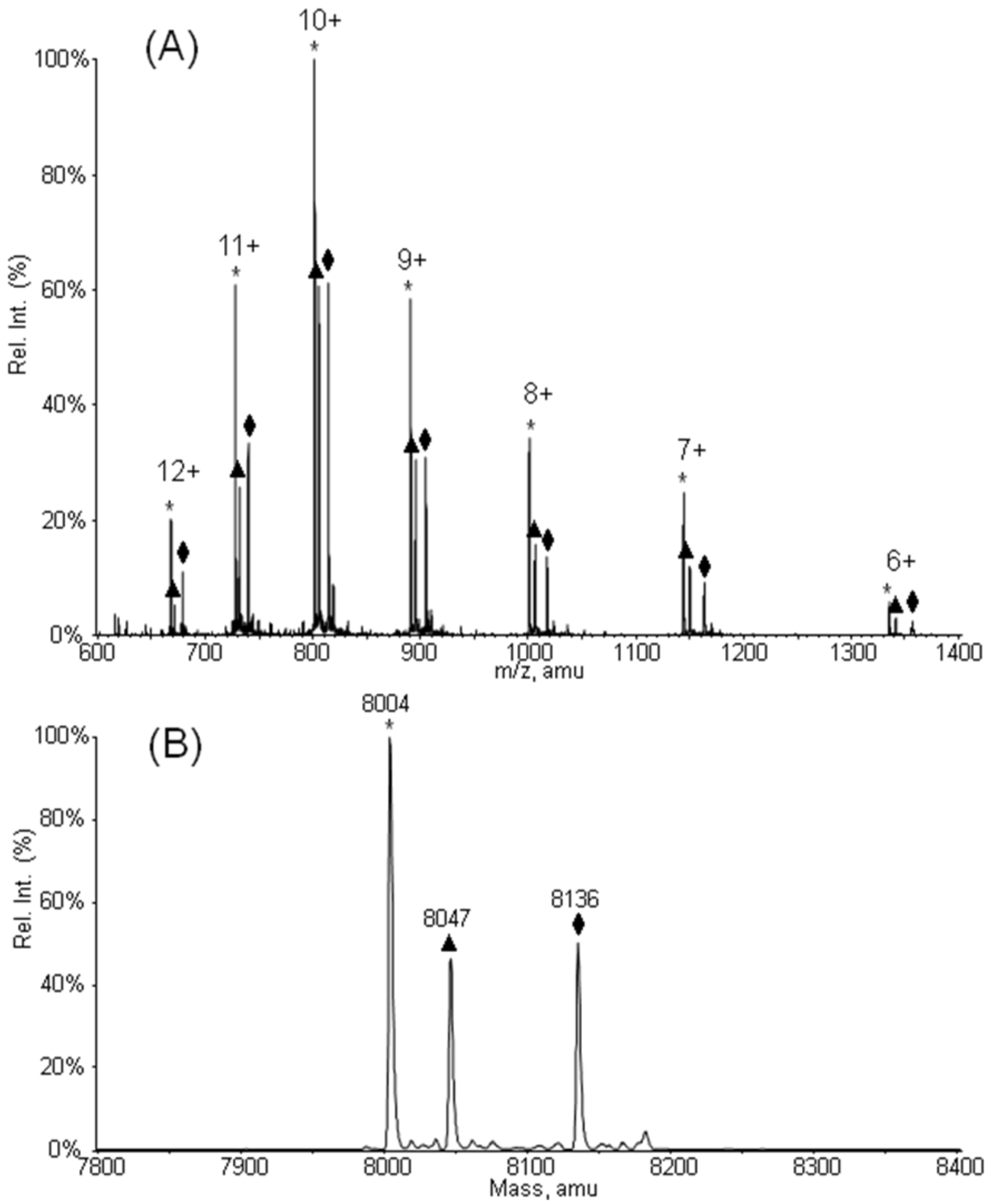


Figure II-6: Mass determination of Protein Z-**6** (A) ESI-MS spectrum of Z-**6** (B) The deconvoluted ESI-MS spectrum of Z-**6**.

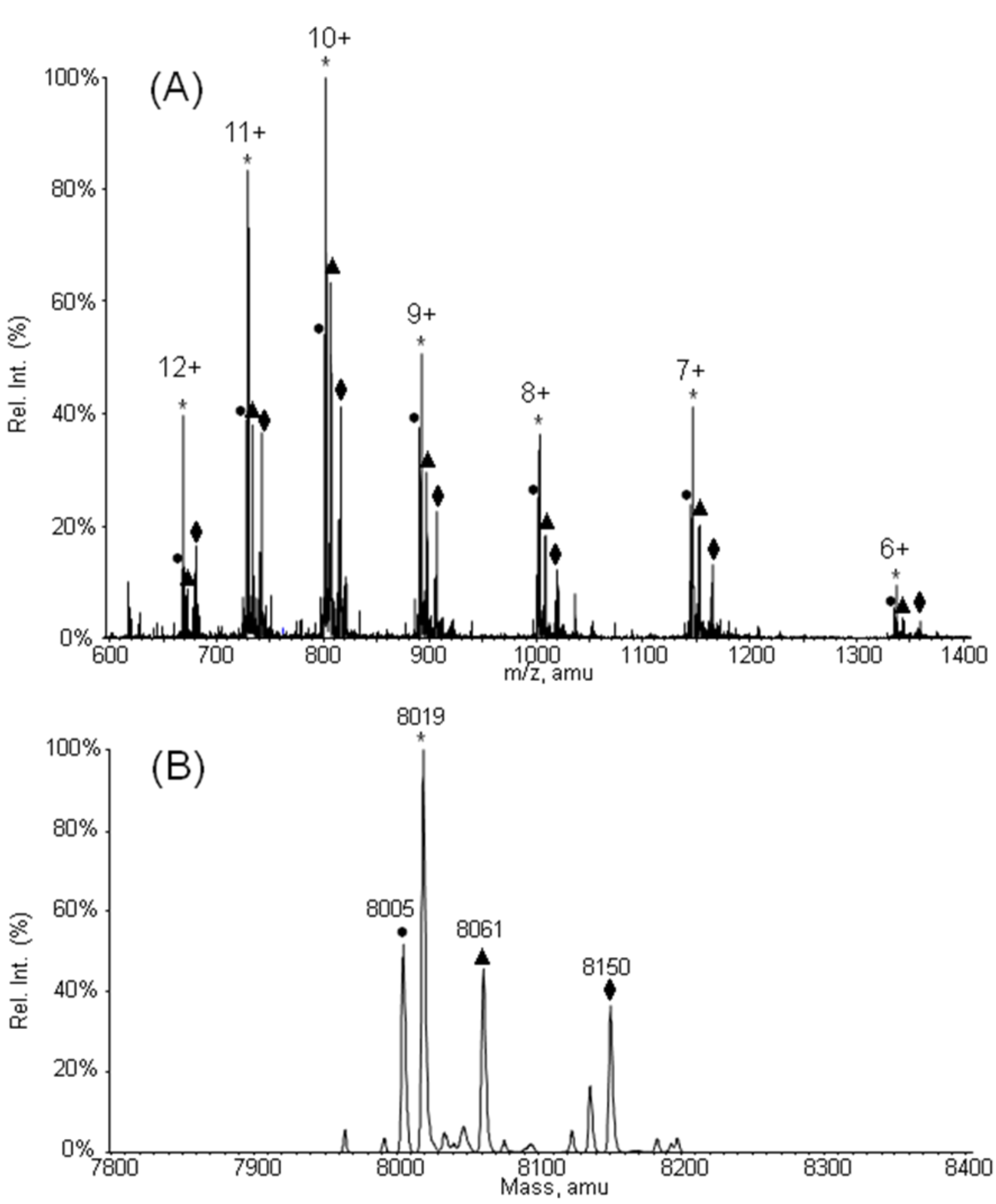


Figure II-7. Mass determination of Protein Z-**7** (A) ESI-MS spectrum of Z-**7** (B) The deconvoluted ESI-MS spectrum of Z-**7**.

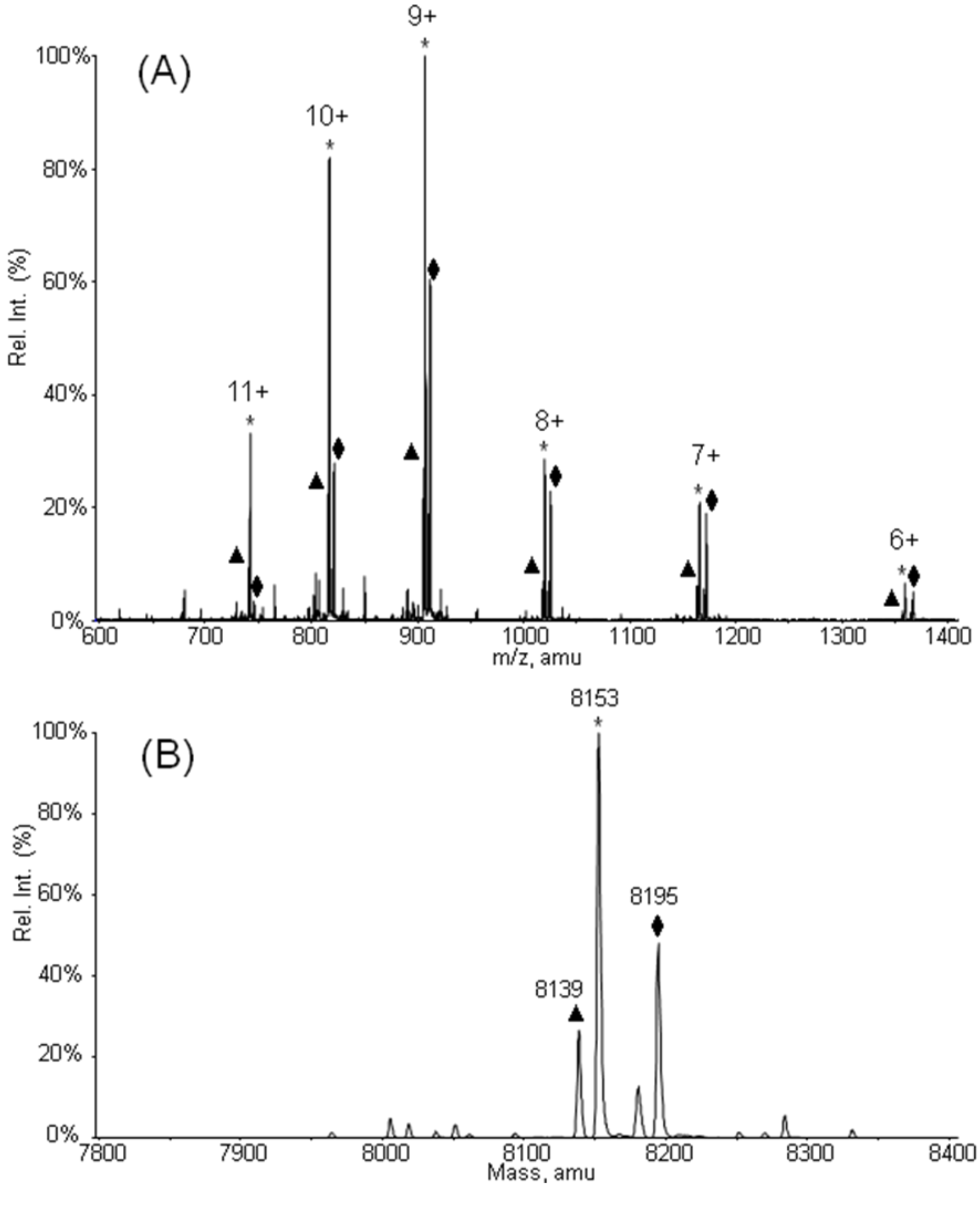


Figure II-8: Mass determination of Protein Z-**8** (A) ESI-MS spectrum of Z-**8** (B) The deconvoluted ESI-MS spectrum of Z-**8**.

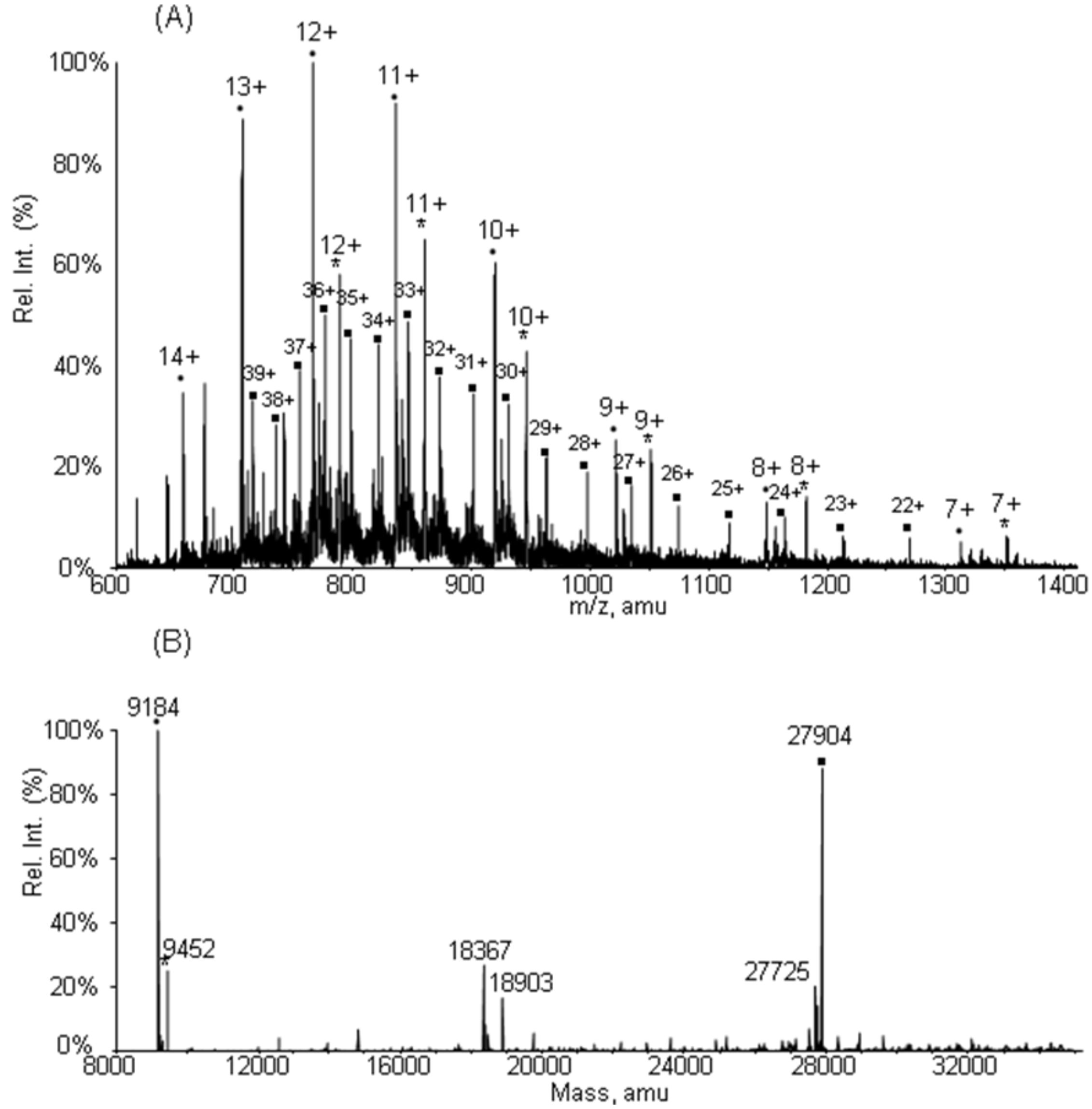


Figure II-9. Mass determination of Protein GFP-**8** (A) ESI-MS spectrum of GFP-**8** (B) The deconvoluted ESI-MS spectrum of GFP-**8**.

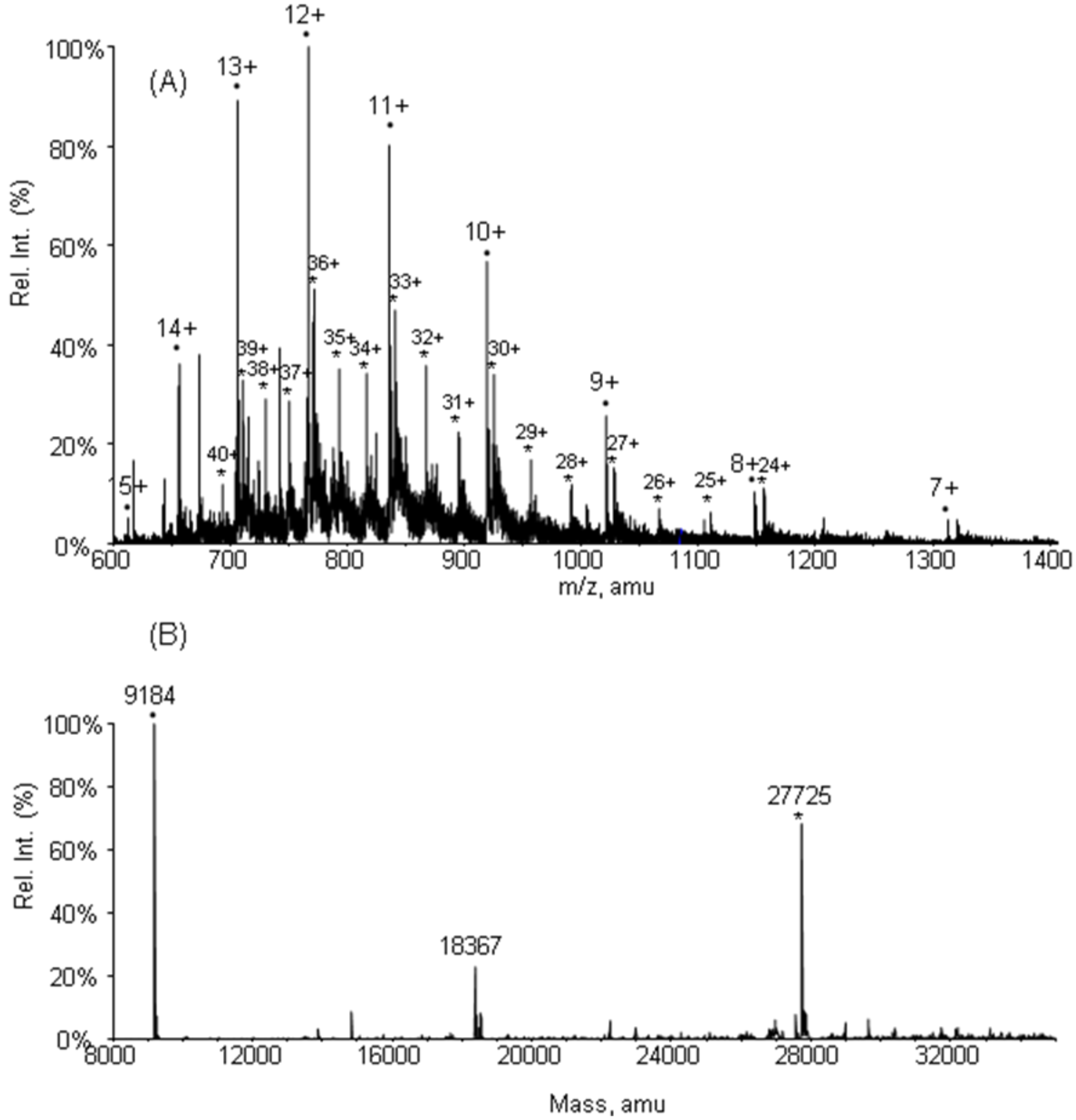


Figure II-10. Mass determination of Protein GFP-**8** after photolysis. (A) ESI-MS spectrum of GFP-**8** (B) The deconvoluted ESI-MS spectrum of GFP-**8**.

Conclusion

In summary, we have demonstrated the genetic incorporation of a photocaged N^{ε}-methyl-L-lysine into proteins in *E. coli*. Since deprotecting the photocaging group to recover N^{ε}-methyl-L-lysine only requires UV irradiation, this method is suitable to directly synthesize many proteins with monomethylated lysines. Given the fact that protein lysine methylation has a fundmental role in regulating functions of chromatin and many transcription factors, a broad application of this developed method is anticipated. Since wild type and evovled PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pairs have been directly used to incorproate NAAs into proteins in mammalian cells,[85, 96] the evolved mKRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair might also be applied to synthesize proteins with monomethylated lysines directly in mamalian cells. This will allow the functional analysis of lysine monomethylations directly *in vivo.*

CHAPTER III

A FACILE METHOD TO SYNTHESIZE HISTONES WITH POSTTRANSLATIONAL MODIFICATION MIMICS

Introduction

Epigenetic changes are crucial for the development and differentiation of the various cell types in an organism and typically involve postreplicational modifications to DNA[99-102] and posttranslational modifications to proteins that are closely associated with DNA.[5, 6, 81, 82, 103, 104] Among posttranslational modifications of proteins, those on histones are probably the most diversified. These include several types of methylation on lysine and arginine, phosphorylation on serine and threonine, lysine acetylation, ubiquitination, glycosylation, etc.[81, 105-107] Modifications of histones are central to the regulation of chromatin dynamics, and therefore, many biological processes involving chromatin, such as replication, repair, transcription, and genome stability, are regulated by histone modifications. The importance and complexity of histone modifications has also led to the coin of the term “the histone code”. In order to decipher “the histone code”, recently several methods have been introduced to synthesize histones with posttranslational modifications. It has been demonstrated that native chemical ligation and its derivative, expressed protein ligation, can be combined together with solid phase peptide synthesis to synthesize acetylated and ubiquitinated histones.[19, 20, 108-110] Although elegant, the general practice of this method is hard. Another method to synthesize histones with posttranslational modifications is to site-specifically incorporate modified amino acids directly into histones at amber mutation sites during protein translation. Using evolved PylRS- $\text{tRNA}^{\text{Pyl}}_{\text{CUA}}$ pairs,[26, 111, 112] histones with lysine acetylation and lysine monomethylation have been recombinantly synthesized in *E. coli*.[52, 53, 55, 66-69] Alternatively, dehydroalanine can be genetically installed at designated sites followed by reactions with thiol-containing molecules to install histones with different posttranslational mimics.[70, 71, 113] Using the genetic incorporation of phenylselenocysteine followed by oxidative elimination to generate dehydroalanine that then undertook Michael additions with a series of cysteamine derivatives, Schultz and coworkers showed that histones with acetylated and methylated lysine mimics could be synthesized.[71]

We have been primarily focusing on the dehydroalanine-based approach to synthesize histones with posttranslational modification mimics for its easy access to multiple types of modifications. In addition, the methods allowing the genetic incorporation of N^{ε},N^{ε}-dimethyllysine and N^{ε},N^{ε},N^{ε}-trimethyllysine directly into histones are not available so far. Although oxidative elimination of phenylselenocysteine in a histone protein to install dehydroalanine is straightforward, the genetic incorporation of phenylselenocysteine has relatively low efficiency. For wild type histone H3, its expression in *E. coli* can reach to 300 mg/L in LB medium. However, H3 with phenylselenocysteine incorporated could only reach to 15 mg/L in LB medium as shown in a previous publication.[71] The low solubility of phenylselenocysteine at the physiological pH also excludes the possibility of boosting its incorporation level by increasing its concentration in the growth medium. For these reasons, we have been searching alternative methods to incorporate dehydroalanine into histones for the site-specific installation of posttranslational mimics on them.

Experimental Section

General Experimental

All reactions involving moisture sensitive reagents were conducted in oven-dried glassware under an argon atmosphere. Anhydrous solvents were obtained through standard laboratory protocols. Analytical thin-layer chromatography (TLC) was performed on Whatman SiO_2 60 F-254 plates. Visualization was accomplished by UV irradiation at 254 nm or by staining with ninhydrin (0.3% w/v in glacial acetic acid/n-butyl alcohol 3:97). Flash column chromatography was performed with flash silica gel (particle size 32-63 μm) from Dynamic Adsorbents Inc (Atlanta, GA).

Specific rotations of chiral compounds were obtained at the designated concentration and temperature on a Rudolph Research Analytical Autopol II polarimeter using a 0.5 dm cell. Proton and carbon NMR spectra were obtained on Varian 300 and 500 MHz NMR spectrometers. Chemical shifts are reported as δ values in parts per million (ppm) as referenced to the residual solvents: chloroform (7.27 ppm for ^{1}H and 77.23 ppm for ^{13}C) or water (4.80 ppm for ^{1}H). A minimal amount of 1,4-dioxane was added as the reference standard (67.19 ppm for ^{13}C) for carbon NMR spectra in deuterium oxide, and a minimal amount of sodium hydroxide pellet or concentrated hydrochloric acid was added to the NMR sample to aid in the solvation of amino acids which have low solubility in deuterium oxide under neutral conditions. ^{1}H NMR spectra are tabulated as follows: chemical shift, multiplicity (s = singlet, bs = broad singlet, d = doublet, t = triplet, q = quartet, m = multiplet), number of protons, and coupling constant(s). Mass spectra were obtained at the Laboratory for Biological Mass Spectrometry at the Department of Chemistry, Texas A&M University.

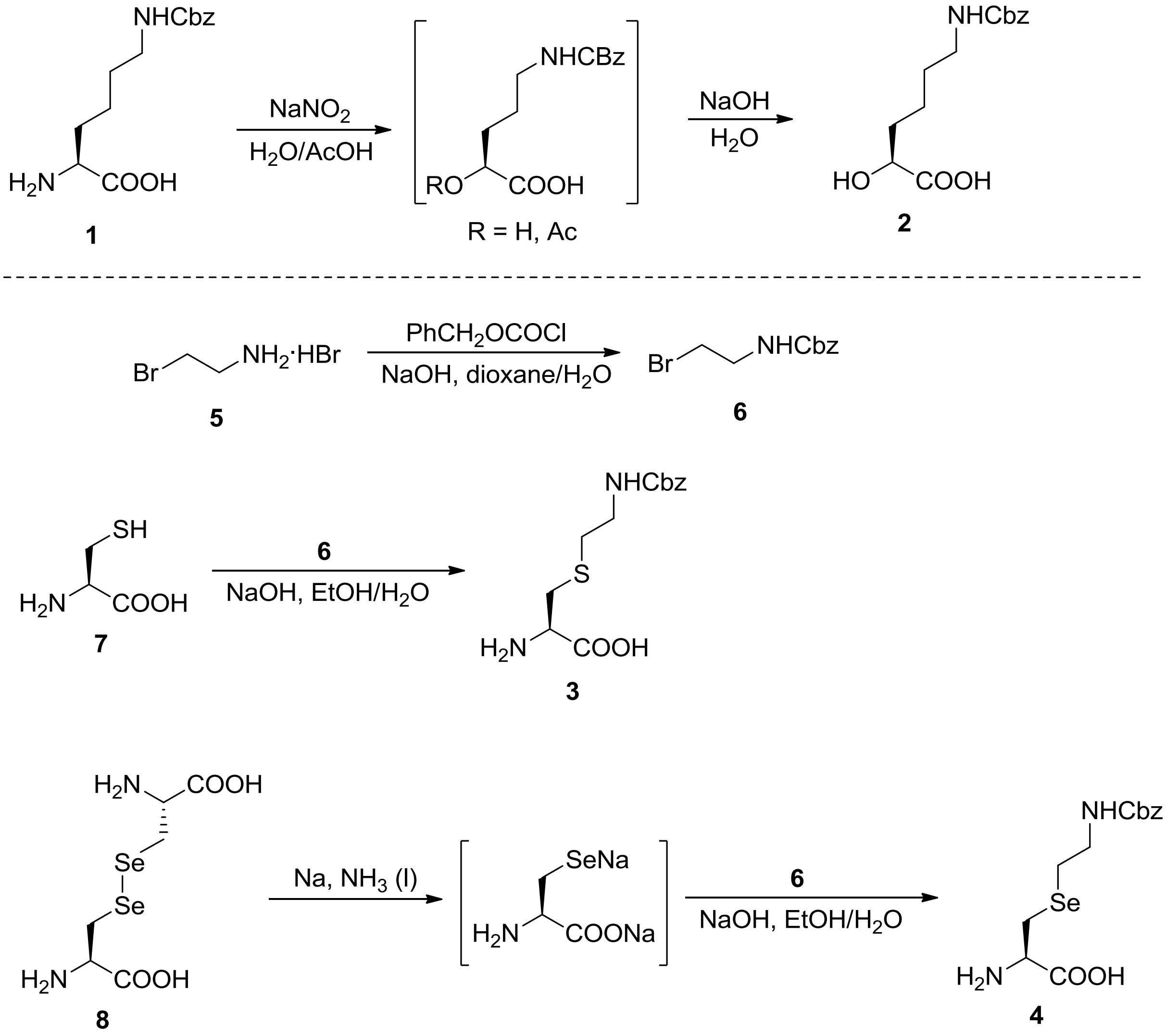
H-Lys(Z)-OH (**1**) was obtained from Chem-Impex International, Inc. (Wood Dale, IL). *O*-Mesitylsulfonylhydroxylamine (MSH), 2-(methylamino)ethanethiol hydrochloride (monomethyllysine mimic precursor), 2-(dimethylamino)ethanethiol hydrochloride (dimethyllysine mimic precursor), and 2-(mercaptoethyl)trimethyl-ammonium chloride (trimethyllysine mimic precursor) were prepared according to procudres by Bernardes and coworkers.[113] All other reagents were obtained from commercial suppliers and used as received.

Chemical Synthesis

Compounds **2** was synthesized from **1** on multi-gram scale[114, 115] (**Scheme III-1**). Compounds **3** and **4** were synthesized by nucleophilic substitution of **6** with appropriate precursors. Reductive cleavage of the diselenide bond in **8** with sodium borohydride[116] turned out to be incomplete, and was best effected with sodium in liquid ammonia.

(*S*)-6-(((Benzyloxy)carbonyl)amino)-2-hydroxyhexanoic acid (2)[114]

Compound **1** (5.0 g, 17.8 mmol) was dissolved in a mixture of acetic acid and water (1:2, 445 mL) with heating and then cooled to room temperature. Sodium nitrite (3.1 g, 44.9 mmol) in water (20 mL) was added dropwise over 10 min. After about 30 min gas evolution ceased and a yellow clear solution was formed. The solution was stirred at room temperature for 24 h and extracted with ethyl acetate (150 mL x 2). The



Scheme III-1. Synthesis of **2**, **3**, and **4**.

combined organic layers were washed with brine, dried (Na_2SO_4), evaporated, redissolved in sodium hydroxide (2 *N*, ~ 30 mL) and stirred at room temperature overnight. The suspension was adjusted to pH 1with hydrochloric acid (6 *N*), extracted with ethyl acetate (150 mL x 2), washed with brine, dried (Na_2SO_4), evaporated, and recrystallized in ethyl acetate/hexanes (1:1) to give **2** (3.5 g, 70%) as a white solid. ^{1}H NMR (DMSO-d6, 500 MHz) δ 7.38-7.30 (m, 5 H), 7.25 (t, 1 H, *J* = 5.7 Hz), 5.00 (s, 2 H), 3.90 (dd, 1 H, *J* = 8.0, 4.5 Hz), 3.36 (bs, 1 H), 2.97 (dt, 2 H, *J* = 6.5, 6.5 Hz), 1.73-1.44 (m, 2 H), 1.42-1.29 (m, 4 H); ^{13}C NMR (DMSO-d6, 75 MHz) δ 175.8, 156.0, 137.3, 128.3, 127.7, 69.5, 65.1, 40.2, 33.6, 29.2, 22.1; HRMS (ESI) calcd for $C_{14}H_{20}NO_5$ ($[M+H]^+$) 282.3123, found 282.2140.

Benzyl (2-bromoethyl)carbamate (6)[117, 118]

To a solution of **5** (13.5 g, 64.6 mmol) in a mixture of sodium hydroxide (2.0 *N*, 70 mL, 0.14 mol) and 1,4-dioxane (50 mL) cooled in an ice bath was added a solution of benzyl chloroformate (11.6 g, 64.6 mmol) in 1,4-dioxane (20 mL) dropwise over 15 min. The mixture was stirred at room temperature overnight, and most of the dioxane was evaporated. The residue was adjusted to pH 5 with hydrochloric acid (2 *N*) and extracted with ethyl acetate (100 mL x 2). The combined organic layers were washed with brine, dried (Na_2SO_4), evaporated, chromatographed (EtOAc/hexanes, 1:9), and crystallized in hexanes to give **6** (8.9 g, 53%) as a white solid. ^{1}H NMR ($CDCl_3$, 500 MHz) δ 7.40-7.33 (m, 5 H), 5.29 (bs, 1 H), 5.12 (s, 2 H), 3.60 (appar. q, 2 H, *J* = 6.0 Hz), 3.47 (t, 2 H, *J* = 5.7 Hz); ^{13}C NMR ($CDCl_3$, 75 MHz) δ 156.3, 136.4, 128.8, 128.4, 128.3, 67.2, 43.0, 32.7.

(*R*)-2-Amino-3-((2-(((benzyloxy)carbonyl)amino)ethyl)thio)propanoic acid (3)

To a degassed solution of L-cysteine (**7**, 1.25 g, 10.0 mmol) in sodium hydroxide (2 *N*, 15.0 mL, 30.0 mmol) cooled in an ice bath was added **6** (2.85 g, 11.0 mmol) in degassed ethanol (10 mL) dropwise over 5 min. The mixture was stirred at room temperature overnight, and hydrochloric acid (3 *N*, 8.0 mL, 24.0 mmol) was added to give a white suspension. Filtration followed by washing with excessive water, ethanol, and dichloromethane and drying under vacuum afforded **3** (2.4 g, 81%) as a white solid. $[\alpha]_D^{21}$ +2.4 (*c* 1.20, 1 *N* NaOH); ^{1}H NMR (D_2O, 500 MHz, pH > 10) δ 7.45-7.39 (m, 5 H), 5.11 (s, 2 H), 3.38 (t, 1 H, *J* = 5.7 Hz), 3.36-3.28 (m, 2 H), 2.84 (dd, 1 H, *J* = 13.5, 4.5 Hz), 2.76 (dd, 1 H, *J* = 13.2, 6.7 Hz), 2.68 (t, 2 H, *J* = 6.5 Hz); ^{13}C NMR (D_2O, 125 MHz, pH > 10) 181.6, 159.0, 137.1, 129.4, 128.3, 67.5, 55.8, 40.5, 37.5, 32.2; HRMS (ESI) calcd for $C_{13}H_{19}N_2O_4S$ ($[M+H]^+$) 299.1066, found 299.1074.

(*R*)-2-Amino-3-((2-(((benzyloxy)carbonyl)amino)ethyl)selanyl)propanoic acid (4)

To an argon-protected solution of seleno-L-cystine (**8**, 3.75 g, 11.0 mmol) in liquid ammonia (~ 80 mL) cooled in a dry ice/acetone bath was added sodium metal (1.3 g, 56.5 mmol) in small pieces (*CAUTION!*) over 2 h, and a yellow suspension resulted in the end. The bath temperature was gradually raised up to room temperature and excessive ammonia was blown away with a gentle stream of argon *inside a well-ventilated fume hood*. Residual ammonia was removed on a rotorvap *inside a well-ventilated fume hood*, and the solid was cooled in an ice bath and carefully dissolved with degassed ice-cold water (25 mL). Compound **6** (5.85 g, 22.7 mmol) in degassed ethanol (15 mL) was added dropwise over 5 min, and the mixture was stirred at room temperature for 20 h and then filtered to give a red solution. Hydrochloric acid (3 *N*, 38.0 mL, 0.11 mol) was added to give a pink suspension, which was filtered, washed with plenty of water, ethanol, and dichloromethane, and dried under vacuum to afford **4** (6.6 g, 87%) as a slightly pink solid. $[\alpha]_D^{21}$ +14.8 (*c* 1.42, 1 *N* NaOH); ^{1}H NMR (D_2O, 500 MHz, pH < 1) δ 6.94-6.91 (m, 5 H), 4.62 (s, 2 H), 3.84 (m, 1 H), 2.91 (dt, 2 H, *J* = 6.7, 1.5 Hz), 2.67-2.65 (m, 1 H), 2.57 (m, 1 H), 2.28 (t, 2 H, *J* = 6.5 Hz); ^{13}C NMR (D_2O, 125 MHz, pH ~ 14) δ 181.4, 158.6, 136.8, 129.1, 128.7, 128.0, 66.9, 55.9, 41.0, 29.7, 24.1; HRMS (ESI) calcd for $C_{13}H_{19}N_2O_4Se$ ($[M+H]^+$) 349.0512/347.0510/345.0518 (major Se isotopes), found 349.0521/347.0501/345.0506.

DNA and Protein Sequences

Gene and protein sequences of pylT, *Methanosarcina mazei* PylRS and GFP_{UV} were listed in the DNA sequence of **Experimental Section** in Chapter II.

DNA Sequences

sfGFP:

atgtagaaaggagaagaacttttcactggagttgtcccaattcttgttgaattagatggtgatgttaatgggcacaaattttctgtccgtggagagggtgaaggtgatgctacaaacggaaaactcacccttaaatttatttgcactactggaaaactacctgttccgtggccaacacttgtcactactctgacctatggtgttcaatgcttttcccgttatccggatcacatgaaacggcatgactttttcaagagtgccatgcccgaaggttatgtacaggaacgcactatatctttcaaagatgacgggacctacaagacgcgtgctgaagtcaagtttgaaggtgatacccttgttaatcgtatcgagttaaagggtattgattttaaagaagatggaaacattcttggacacaaactcgagtacaactttaactcacacaatgtatacatcacggcagacaaacaaaagaatggaatcaaagctaacttcaaaattcgccacaacgttgaagatggttccgttcaactagcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaaccattacctgt

cgacacaatctgtcctttcgaaagatcccaacgaaaagcgtgaccacatggtccttcttgagtttgtaactgctgctgggattaca
catggcatggatgagctctacaaaggatcccatcaccatcaccatcactaa

mkRS1:

atggataaaaaaccactaaacactctgatatctgcaaccgggctctggatgtccaggaccggaacaattcataaaataaaacac
cacgaagtctctcgaagcaaaatctatattgaaatggcatgcggagaccaccttgttgtaaacaactccaggagcagcaggact
gcaagagcgctcaggcaccacaaatacaggaagacctgcaaacgctgcagggtttcggatgaggatctcaataagttcctcac
aaaggcaaacgaagaccagacaagcgtaaaagtcaaggtcgtttctgcccctaccagaacgaaaaaggcaatgccaaaatcc
gttgcgagagccccgaaacctcttgagaatacagaagcggcacaggctcaaccttctggatctaaattttcacctgcgataccg
gtttccacccaagagtcagtttctgtcccggcatctgtttcaacatcaatatcaagcatttctacaggagcaactgcatccgcactg
gtaaaagggaatacgaaccccattacatccatgtctgcccctgttcaggcaagtgcccccgcacttacgaagagccagactga
caggcttgaagtcctgttaaacccaaaagatgagatttccctgaattccggcaagcctttcagggagcttgagtccgaattgctct
ctcgcagaaaaaaagacctgcagcagatctacgcggaagaaagggagaattatctggggaaactcgagcgtgaaattaccag
gttctttgtggacaggggttttctggaaataaaatccccgatcctgatccctcttgagtatatcgaaaggatgggcattgataatgat
accgaactttcaaaacagatcttcagggttgacaagaacttctgcctgagacccatgcttgctccaaaccttatgaactacgcgc
gcaagcttgacagggccctgcctgatccaataaaaatttttgaaataggcccatgctacagaaaagagtccgacggcaaagaa
cacctcgaagagtttaccatgctgaacttcacgcagatgggatcgggatgcacacgggaaaatcttgaaagcataattaaggac
ttcctgaaccacctgggaattgatttcaagatcgtaggcgattcctgcatggtctttggggataccettgatgtaatgcacggaga
cctggaactttcctctgcagtagtcggacccataccgcttgaccgggaatggggtattgataaaccctggataggggcaggtttc
gggctcgaacgccttctaaaggttaaacacgactttaaaaatatcaagagagctgcaaggtccgagtcttactataacgggattt
ctaccaacctgtaa

Xenopus laevis H3/C110A:

Atggctcgtactaagcagaccgcccgttagtccaccggagggaaggctccccgcaaacagctggccaccaaggcagccag
gaagagcgctccggccacaggcggagtcaagaaacctcaccgttaccggcccggcacagtcgctctccgcgagatccgcc
gctaccagaaatccaccgagctgctcatccgcaaactgcctttccagcgcctggtccgggagatcgctcaggacttcaagacc
gacctgcgcttccagagctcggccgtcatggctctgcaggaggccagcgaggcttatctggtcggtttgtttgaggacaccaac
ctggccgccatccacgccaagagggtcaccatcatgcccaaggacatccagctggcccgcaggatccggggcgagagggc
tgagctccatcaccatcaccatcactaa

Proteins Sequences

sfGFP:

MXKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFI**C**TTGKLP
VPWPTLVTTLTYGVQ**C**FSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYK
TRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIK
ANFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRD
HMVLLEFVTAAGITHGMDELYKGSHHHHHH

X represents a noncanonical amino acid.

mkRS1:

MDKKPLNTLISATGLWMSRTGTIHKIKHHEVSRSKIYIEMACGDHLVVNNSRSS
RTARALRHHKYRKTCKRCRVSDEDLNKFLTKANEDQTSVKVKVVSAPTRTKK
AMPKSVARAPKPLENTEAAQAQPSGSKFSPAIPVSTQESVSVPASVSTSISSISTG
ATASALVKGNTNPITSMSAPVQASAPALTKSQTDRLEVLLNPKDEISLNSGKPFR
ELESELLSRRKKDLQQIYAEERENYLGKLEREITRFFVDRGFLEIKSPILIPLEYIER
MGIDNDTELSKQIFRVDKNFCLRPMLAPNLMNYARKLDRALPDPIKIFEIGPCYR

KESDGKEHLEEFTMLNFTQMGSGCTRENLESIIKDFLNHLGIDFKIVGDSCMVFG DTLDVMHGDLELSSAVVGPIPLDREWGIDKPWIGAGFGLERLLKVKHDFKNIKR AARSESYYNGISTNL

Xenopus laevis H3/C110A :

MARTKQTARXSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIR RYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAV**M**ALQEASEAYLVGLFEDT NLAAIHAKRVTI**M**PKDIQLARRIRGERAELHHHHHH

Plasmid Constructions

Construction of pET-sfGFP2TAG

The plasmid pET-sfGFP2TAG was derived from the plasmid pET-pylT-Z[68] in which sfGFP has an amber mutation at S2. This gene was amplified from the Superfolder GFP plasmid (Theranostech®). Two restriction sites, *NdeI* at the 5' end and *SacI* at the 3' end, were introduced in the PCR product which was subsequently digested and used to replace Z domain gene in pET-sfGFP2TAG.

Construction of pBK-mKRS1 with the Y384F mutation

The pBK-mKRS1 plasmid in which mKRS1 has the Y384 mutation was derived from a pBK plasmid containing *p*-iodo-L-phenylalanyl-tRNA synthetase[94] that was initially evolved from *M. jannaschii* tyrosyl-tRNA synthetase. The PylRS gene is under the control of *E. coli glnS* promoter and terminator. This gene was amplified from the pBK-mKRS1 (Y306M/L309A/C348T/T364K) plasmid by flanking primers, pBK-mmPylRS-NdeI-F and pBK-mmPylRS-PstI~NsiI-R. To construct the pBK-mkRS1F plasmid Y384F mutation was introduced by overlap extension PCR. The following pairs of primers were used to generate an mKRS1 gene with the Y384F mutation: (1) pBK-mmPylRS-NdeI-F (5'-gaatcccatatggataaaaaaccactaaacactctg-3') and mmPylRS-Y384F-R (5'-tacatcaagggtatccccaaagaccatgcaggaatcgcctacg-3'); (2) mmPylRS-Y384F-F (5'-cgtaggcgattcctgcatggtctttggggataccottgatgta-3') and pBK-mmPylRS-PstI~NsiI-R (5'-gtttgaaaatgcatttacaggttggtagaaatccc-3'). The amplified gene was digested with the restriction enzymes *NdeI* and *NsiI*, gel-purified, and ligated back into the pBK vector digested by *NdeI* and *PstI* to afford plasmid pBK-mKRS1.

Construction of pEVOL-pylT

The pEVOL-pylT plasmid was derived from a pEVOL plasmid.[119] The pylT gene with *prok* promoter and terminator was synthesized by overlap PCR with eight primers ((1) pEVOL-PylT-ApaLI-F-1 (5'-gatatgatcagtgcacggctaactaagcggcctgctgactttctcg-3'); (2) pEVOL-PylT-R-2 (5'-caatcccttaatagcaaaatgccttttgatcggcgagaaagtcagcag-3'); (3) pEVOL-pylT-F-3 (5'-gctattaagggattgacgagggcgtatctgcgcagtaagatgcgcccc-3'); (4) pEVOL-pylT-R-4 (5'-agtccattcgatctacatgatcaggtttccaatgcggggcgcatcttac-3'); (5) pEVOL-pylT-F-5 (5'-gtagatcgaatggactctaaatccgttcagccgggttagattcccgggg-3'); (6) pEVOL-pylT-R-6 (5'-ggcttttcgaatttggcggaaaccccgggaatctaac-3'); (7) pEVOL-pylT-F-7 (5'-caaattcgaaaagcctgctcaacgagcaggctttttg-3'); (8) pEVOL-pylT-Xho1-R-8 (5'-ctgagctgctcgagcatgcaaaaagcctgctc-3') and introduce by two restriction sites, ApaLI at the 5' end and XhoI at the 3' end. Two pairs of restriction sites (SpeI and SalI sites between the *pBAD* promoter and terminator; NdeI and NotI between the *glnS* promoter and terminator) were introduced using two pairs of primers ((1) pEVOL-SpeI-R (5'-ttactagtaattcctcctgttagccc-3') and pEVOL-SalI-F (5'-ccgtcgaccatcatcatcatcatc-3'); (2) 5'-pEVOL-NdeI-R (5'-atcatatgggattcctcaaagcgtaaac-3') and pEVOL-NotI-F (5'-acgcggccgctttcaaacgctaaattgc-3'). These restriction sites in pEVOL-pylT were constructed for further installations of two copies of mKRS1.

Construction of pEVOL-mKRS1-pylT

The pEVOL-mKRS1-pylT plasmid was derived from the pEVOL-pylT plasmid with sequential insertion of two copies of the mKRS1 gene. The first copy of mKRS1 gene was amplified from the pBK-mKRS1 plasmid by flanking primers, pEVOL-PylRS-SpeI-F and pEVOL-PylRS-SalI-R, digested by SpeI and SalI restriction enzymes, and ligated to a precut pEVOL-pylT plasmid. The resulted plasmid was digested by NdeI and NotI enzymes and used to insert the second copy of the mKRS1 gene that was amplified using primers pEVOL-PylRS-NdeI-F and pEVOL-PylRS-NotI-R and digested by NdeI and NotI restriction enzymes. The resulted plasmid is pEVOL-mKRS1-pylT.

Construction of pET-H3K9TAG

The plasmid pET-H3K9TAG was derived from the plasmid pET-pylT-GFP in which H3 has an amber mutation at K9. This gene was amplified from the plasmid pET22b-xlH3.[71] Two restriction sites, *NdeI* at the 5' end and *SacI* at the 3' end, were introduced in the PCR product which was subsequently digested and used to replace GFP gene in pET-H3K9TAG.

Construction of pET-H3

The plasmid pET-H3 with wild type H3 gene was derived from the commercial plasmid pET-Duet1. This gene was amplified from the plasmid pET22b-xlH3/C110A. Two restriction sites, *NdeI* at the 5' end and *KpnI* at the 3' end, were introduced in the PCR product which was subsequently digested and used to install H3 gene in pET-Duet1 to afford pET-H3.

Screening Procedures for mKRS and mKRS1 Strains

The initial mKRS1 strain (designated as mKRS) and the mKRS1 strain with the Y384F mutation (designated as mKRS1) in pBK plasmid were cotransformed with pY+ to *E. Coli* TOP10 cells and tested for their ability to grow on plates with 102 μg/mL chloramphenicol (Cm), 25 μg/mL kanamycin (Kan), 12 μg/mL tetracycline (Tet), and 1 mM of **1**, **3**, or **4**. A plate without any noncanonical amino acid (NAA) supplement was used as the negative control. *E. Coli* cells containing the two plasmids were separately placed on LB agar plates with decreasing cell number by serial dilution from 3×10^6 to 1. The original cell solution was prepared with OD_{600} = 1.0 (~1×10^9 cells/mL). The cells on the plate were grown at 37 ^{o}C for 48 h. Images of colonies growing on different plates were shown in **Figure III-1.**

GFP_{UV} and sfGFP and H3 Protein Expression and Purification

GFP_{UV} and sfGFP expressions and purifications are performed by the method in Chapter II. To express H3 incorporated with a NAA, *E. Coli* BL21(DE3) cells were cotransformed with pEVOL-mKRS1-pylT and pET-H3K9TAG. Cells were recovered in

1 mL of LB medium for 1 h at 37 ºC before being plated on LB agar plate containing Cm (34 μg/mL) and Amp (100 μg/mL). A single colony was then selected and grown overnight in a 10 mL culture. The culture was then used to inoculate 500 mL of 2YT media supplemented with 34 μg/mL Cm and 100 μg/mL Amp. Cells were grown at 37 ºC in an incubator (300 r.p.m.) and protein expression was induced when OD_{600} reached 0.7 - 1.0 by adding 1 mM IPTG, 0.2% arabinose and 2 mM **1**. After 6-8 h induction, cells were harvested, resuspended in a lysis buffer (50 mM Tris-HCl, 100 mM NaCl, 5 mM EDTA, 0.1% NaN_3, 0.5% Triton-X100, 0.1 mM PMSF and 1mM DTT, pH 8.0) and sonicated. $MgSO_4$ (final concentration 10 mM) was added to chelate EDTA, and 0.01 mg/ml DNase and 0.1 mg/ml lysozyme were then added to the solution. The mixture was incubated at room temperature for 20 min. The cell lysate was clarified by centrifugation (20 min, 6,000 r.p.m., 4 ºC). The centrifuged pellet was crushed with a spatula, then resuspended by sonication in the lysing buffer. Another portion of DNase and lysozyme was added at this point to improve the purity of the pellet. After centrifugation, the inclusion body was washed twice with the washing buffer (50 mM Tris-HCl, 100 mM NaCl, 5 mM EDTA, 0.1% NaN3, pH 8.0). The semi-purified H3 protein inclusion body was dissolved by a dissolving buffer (100 mM NaH_2PO_4, 10 mM Tris HCl, 8 M urea, pH 8.0) and incubated for 1 h at 37 ^oC. The solution was then centrifuged (20 min, 10,000 r.p.m). The supernatant was incubated with 3 mL Ni^{2+}-NTA resin (Qiagen) (2 h, 4 ºC) and washed with 30 mL of washing buffer (dissolving buffer in pH = 6.2). The protein was finally eluted out by an elution buffer (dissolving buffer in pH = 4.5). The pure eluted fractions were collected and concentrated. The buffer was

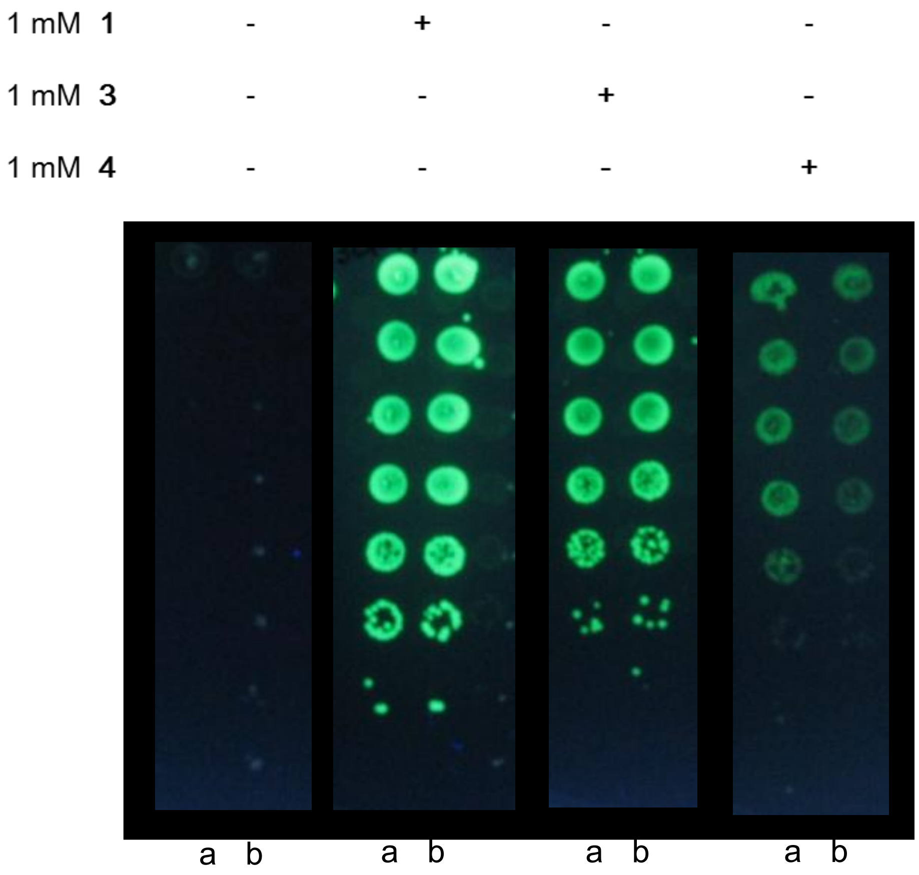


Figure III-1: Growth of two strains, mKRS1 (a) and mKRS (b), on LB plates with different supplements. The LB agar plates contain 102 μg/mL Cm, 25 μg/mL Kan and 12 μg/mL Tet. The TOP10 cells with pBK-mkRS1 (a) (Y306M/L309A/C348T/T364K/Y384F) and pY+ or pBK-mkRS (b) (Y306M/L309A/C348T/T364K) and pY+ were plated with serial dilution from 3×10^6 (1) cells to 0.3 (8) (10 fold for each dilution). The pY+ plasmid has a GFP_{UV} gene under the control of a T7 promoter. The expression is promoted by the suppression of two amber mutations at positions 1 and 107 of a T7 RNA polymerase gene in pREP. The fluorescent intensity of the expression of GFP_{UV} roughly represents the suppression efficiency at amber codons.

later changed to 10 mM ammonium bicarbonate and 8 M urea using an Amicon Ultra -15 Centrifugal Filter Devices (10,000 MWCO cut, Millopore). The purified proteins were analyzed by 15% SDS-PAGE. H3 proteins incorporated with **3** and **4** were expressed in the presence of 2 mM **2** and 1.5 mM **4**, respectively, and similarly purified.

Synthesis of H3 Mimics with Posttranslational Modifications

H3K9Dha was synthesized from H3K9-**4**. Aqueous H_2O_2 solution (100 mM, 5 μL, 100 eq, 500 nmol) was added to the protein H3K9-**4** (1.6 mg/mL, 100 μM, 50 μL, 5 nmole) in a dissolving buffer (100 mM NaH_2PO_4, 10 mM Tris HCl, 8 M urea, pH 8.0), and the mixture was periodically agitated at room temperature for 1 h. The mixture was then dialyzed by Amicon Ultra -15 Centrifugal Filter Devices (10,000 MWCO cut, Millopore) against the dissolving buffer to terminated the reaction.

To synthesize H3 mimics H3K9AcsK, H3K9mesK, H3K9m^{2}sK, H3K9m^{3}sK and H3K9pC, solutions of the corresponding thiol nucleophiles (400 mM, 12.5 μL, 5 μmole) in the dissolving buffer (100 mM NaH_2PO_4, 10 mM Tris HCl, 8 M urea, pH 8.0) were add into protein samples in the same buffer. The Michael addition reaction was performed at room temperature for 1 h and then terminated by dialysis against the dissolving buffer.

Immunoprecipitation of wt-H3, H3K9mesK, H3K9m^{2}sK and H3K9m^{3}sK by HP1β and Western Blotting Assay

HP1β (1 μM) was incubated with wt-H3, H3K9mesK, H3K9m^{2}sK and H3K9m^{3}sK, in 500 μl of binding buffer (0.5 M NaCl, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris HCl, pH 8.0). A portion of this sample (20 µl) was removed to check the total protein level (input). The remaining supernatant was incubated for 4 h at 4 ºC with 1 µg of a goat polyclonal antibody to CBX1/HP1 beta (Abcam, ab40828). After 1 h of incubation, 30 µl of protein A-agarose (Sigma) was added. The beads were pelleted, washed 5 times with 700 µl of RIPA buffer, and the bound protein was released by boiling in SDS-sample buffer. A Rabbit polyclonal antibody to C-terminus of H3 (9715, Cell Signaling Technology) was used to detect H3 proteins immunoprecipitated by HP1β.

A Rabbit polyclonal antibody for histone H3 (acetyl K9) (Abcam, ab10812) was used to detect wtH3 or H3K9-AcsK. Further ECL test was carried out by treatment with a donkey polyclonal secondary antibody to rabbit IgG-H&L (HRP) (Abcam, ab16284).

ESI-MS Analysis of Intact Proteins

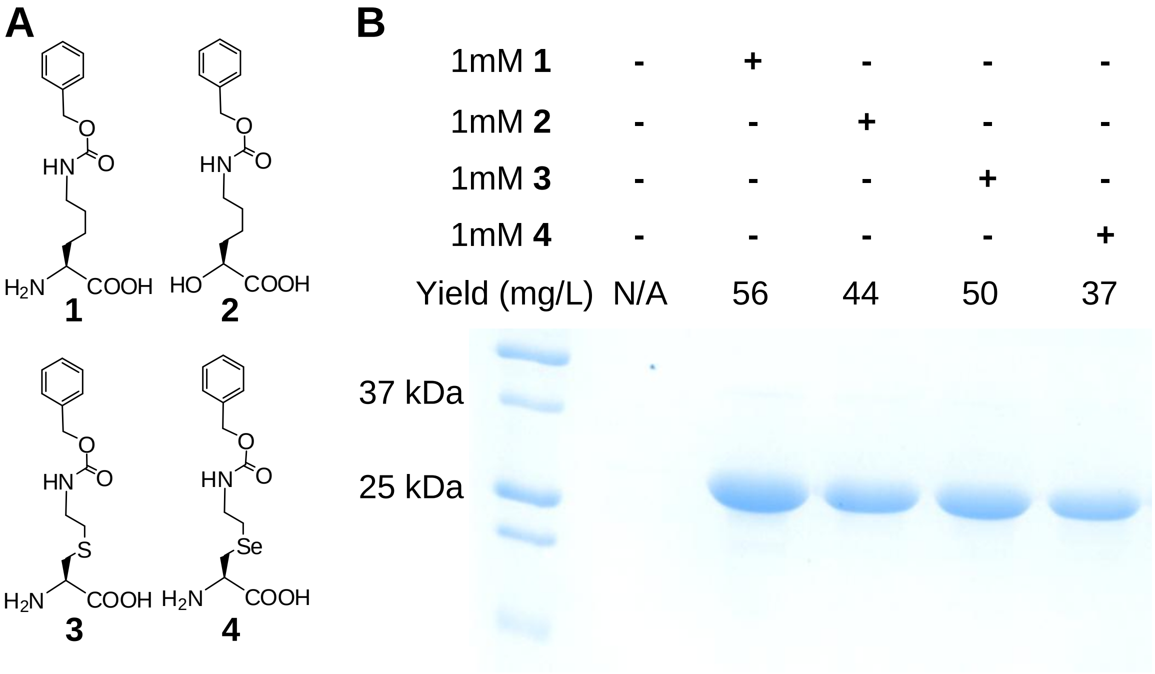
Protein samples were prepared by desalting the protein using C18 ZipTip^{TM} (Millipore) following the manufacturer's protocol and eluted with 60% acetonitrile containing 0.1% formic acid. The resulting solution was diluted to 1 µM with 50% methanol containing 0.1% formic acid, and then used for electrospray mass spectrometry (ESI-MS) analysis. ESI ion-mobility (IM) MS experiments were performed on a SYNAPT G2 HDMS mass spectrometer (Waters Corp., Milford, MA) equipped with a nano-ESI source. The IM-MS data were acquired in positive ion mode (400-2500 Da) using spray voltage of +1800 V. Data analysis and protein signal extraction were performed using the MassLynx^{TM} and DriftScope^{TM} software packages (Waters Corp., Milford, MA). For the histone H3 series, a mass range of m/z 500-1200 was used for spectral deconvolution and the output range was 15000 to 19000 Da using a resolution of 0.1 Da per channel. For the green fluorescent proteins, a mass range of m/z 700-1300 was used for spectral deconvolution and the output range was 27000 to 29000 Da using a resolution of 0.1 Da per channel. For top-down analysis of H3K9-AcsK by CID-IM-MS, the precursor ion of m/z 655.54, corresponding to the $^{25+}$ charged state of H3K9-AcsK, was selected to perform collision-induced dissociation (CID) experiment using argon as collision gas and 25 V of collision energy in the Trap region. High-resolution MS experiments were performed on a SolariX 9.4 T: hybrid quadrupole-FTICR mass spectrometer (Burker Daltonik GmbH, Bremen, Germany) equipped with a nano-ESI source and acquired in positive ion mode (m/z 300-3000) using electrospray voltage of +1600 V. All MS spectra were obtained by quadrupole mass selection of m/z 800 to 1000 and accumulation of 100 spectra. Data analysis and protein signal extraction were performed using the DataAnalysis^{TM} software packages (Burker Daltonik GmbH, Bremen, Germany). For spectral deconvolution to a singly charged spectrum, the output range was 5000 to 100000 m/z using an abundance cutoff of 0 %.

Tandem Mass Spectrometry analysis

GFP-**4** from the SDS-PAGE gel was excised and digested with endoproteinase Asp-N (Roche Diagnostics Co., Indianapolis, IN) or trypsin (Promega, Madison, WI) at 37 ^{0}C overnight using the following protocol: the gel slice was washed with 25 mM ammonium bicarbonate (ABC, pH 8) and dehydrated with a solution mixture of acetonitrile (ACN) and 50mM ABC (v/v, 2/1). The washing and dehydrating steps were repeated for another two times. Supernatant was removed and the gel slice was dried in a vacuum centrifuge (SpeedVac Concentrator, Savant, Farmingdale, NY). 10 µL of 20 ng/µL Asp-N or trypsin in 25 mM ABC was added to the dried gel slice. After the gel slice was completely rehydrated, 20 µL of 25mM ABC was added to cover gel slice and incubated at 37 °C overnight. Peptides resulting from the Asp-N or trypsin digestion were mixed 1:1 (v/v) with matrix (5 mg/mL α-cyano-4-hydroxycinnamic acid, 50% (v/v) acetonitrile, 10 mM ammonium dihydrogen phosphate, 0.1% TFA) and 1 µL of the resulting mixture was spotted onto a stainless steel target plate. Mass spectra and tandem MS spectra were collected using an Applied Biosystems 4800 TOF/TOF^{TM} Analyzer (Framingham, MA). Collision induced dissociation tandem MS spectra were acquired using air at the medium pressure setting and at 2 kV of collision energy. Tandem MS data was manually interpreted using the Data Explorer™ software package (Applied Biosystems, Framingham, MA).

Results and Discussion

We previously showed that an evolved PylRS, mKRS1 together with $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ allows the specific incorporation of N^{ε}-Cbz-lysine (**1** in **Figure III-2A**) at an amber mutation site of a protein in *E. coli*. This mutant enzyme also shows high substrate promiscuity and is able to charge $\mathrm{tRNA}^{\mathrm{Pyl}}$ with N^{ε}-Cbz-N^{ε}-methyllysine (75% protein yield of **1** incorporation), N^{ε}-(*o*-nitrobenzyloxycarbonyl)-lysine (75% protein yield of **1** incorporation) , and N^{ε}-(*o*-nitrobenzyloxycarbonyl)-N^{ε}-methyllysine (63% protein yield of **1** incorporation).[68] To test the substrate scope of mKRS1, we recently synthesized three N^{ε}-Cbz-N^{ε}-methyllysine analogues whose structures are shown in **Figure III-2A** as **2-4** and analyzed their uptake by the mKRS1- $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair to incorporate at amber mutation sites in *E. coli*. Two plasmids were constructed for the test. One plasmid pEVOL-mKRS1-pylT[119] contains the $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ gene under control of the *proK* promoter and the *proK* terminator, one copy of the mKRS1 gene under control of the constitutive *glnS* promoter, and one copy of the mKRS1 gene under control of the *pBAD* promoter; the other plasmid pET-sfGFP2TAG contains a superfolder GFP (sfGFP) gene[120, 121] that is under control of the *T7* promoter and has an amber mutation at S2. One addition mutation Y384F was also introduced to mKRS1 given the fact that this mutation can increase the binding of PylRS or its mutants to substrates with hydrophobic side chains.[122] When growing *E. coli* BL21 cells transformed with both plasmids in LB medium supplemented with **1**, **2**, **3**, or **4**, overexpression of sfGFP was observed. The expression levels for all four conditions are comparable. On the contrary,



Proteins were expressed in BL21 cells transformed with pEVOL-mKRS1-pylT and pET-sfGFP2TAG in LB medium supplemented with 1 mM **1**, **2**, **3**, or **4**.

only a trace amount of sfGFP was detected when growing same cells in LB medium without providing a noncanonical amino acid. As shown **in Table III-1**, the purified proteins all showed the expected molecular weight (MW) when analyzed by electrospray ionization mass spectrometry (ESI-MS). (**Figure III-3-6**) For sfGFP with **2** incorporated (sfGFP-**2**), the full-length protein was not the major component as detected by ESI-MS. **2** is an α-hydroxy acid instead of an α-amino acid. Its incorporation into sfGFP at the S2 position generates an ester bond between the first methionine (M1) and the incorporated **2** which is highly susceptible to hydrolysis catalyzed by methionine aminopeptidase, an essential gene in *E. coli*. The mass peak for the major component of the purified sfGFP-**2** did match that of the protein without M1. To independently confirm the incorporation of **4**, **4** was genetically incorporated at the Q204 position of $\mathrm{GFP_{UV}}$. The purified protein was digested by Asp-N protease and the target fragment N-DNHYLSTXSALSK-C (X denotes **4**) was analysed and confirmed by tandem mass spectrometry. (**Figure III-7**). Importantly, the isotope pattern observed for the selenium containing peptide fragment matches the theoretical isotope pattern. (**Figure III-8**)

With our initial success, we next moved on to synthesize histone H3 with **3** or **4** incorporated at its K9 position. Both **3** and **4** are susceptible to oxidative elimination to generate dehydroalanine after their incorporation into H3 (**Scheme III-2**). **3** is an *S*-alkylcysteine that may be converted to dehydroalanine using *O*-mesitylenesulfonylhydroxyl amine (MSH);[113] **4** is a *Se*-alkylselenocysteine that could be specifically oxidized by H_2O_2 to form dehydroalanine. To express H3 incorporated with

Table III-1: Theoretical and detected average molecular weight of sfGFP with different NAAs incorporated at its S2 position.

Proteins[a]	Theoretic MW_{avg} (Da)	Detected MW_{avg} (Da)
sfGFP-**1**	27903	27903
sfGFP-**2**	27904[b]	27904[b]
	27773[c]	27774[c]
sfGFP-**3**	27921	27922
sfGFP-**4**	27968	27968

[a] sfGFP-**1** contains **1** at S2 position; sfGFP-**3** contains **3** at its S2 position; sfGFP-**4** contains **4** at its S2 position. [b] Full length sfGFP-**2**. [c] sfGFP-**2** without M1.

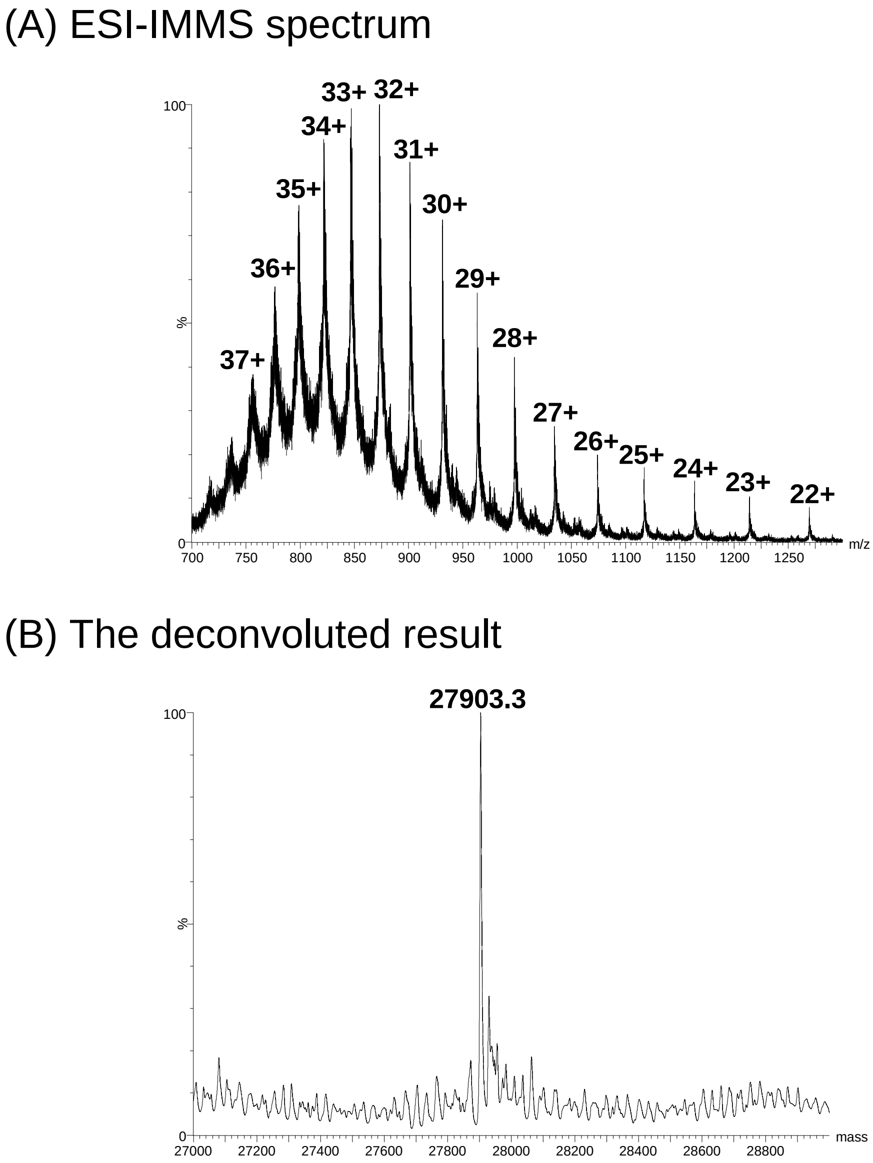
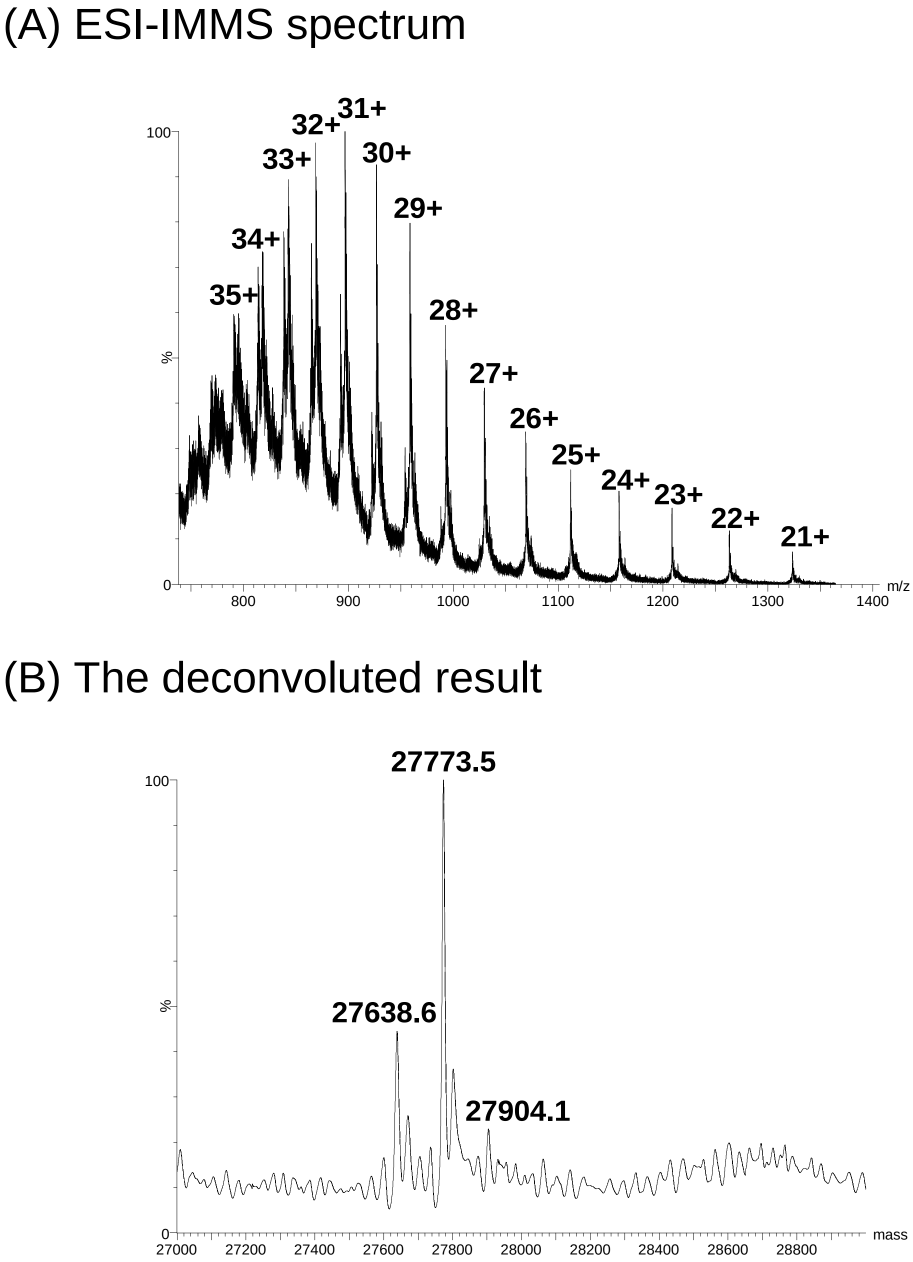


Figure III-3: Mass determination of the protein sfGFP-**1**: (A) ESI-IM-MS spectrum of sfGFP-**1** and (B) the deconvoluted ESI-IM-MS spectrum of sfGFP-**1**.



(A) ESI-IMMS spectrum

(B) The deconvoluted result

Figure III-4: Mass determination of the protein sfGFP-**2**: (A) ESI-IM-MS spectrum of sfGFP-**2** and (B) the deconvoluted ESI-IM-MS spectrum of sfGFP-**2**.

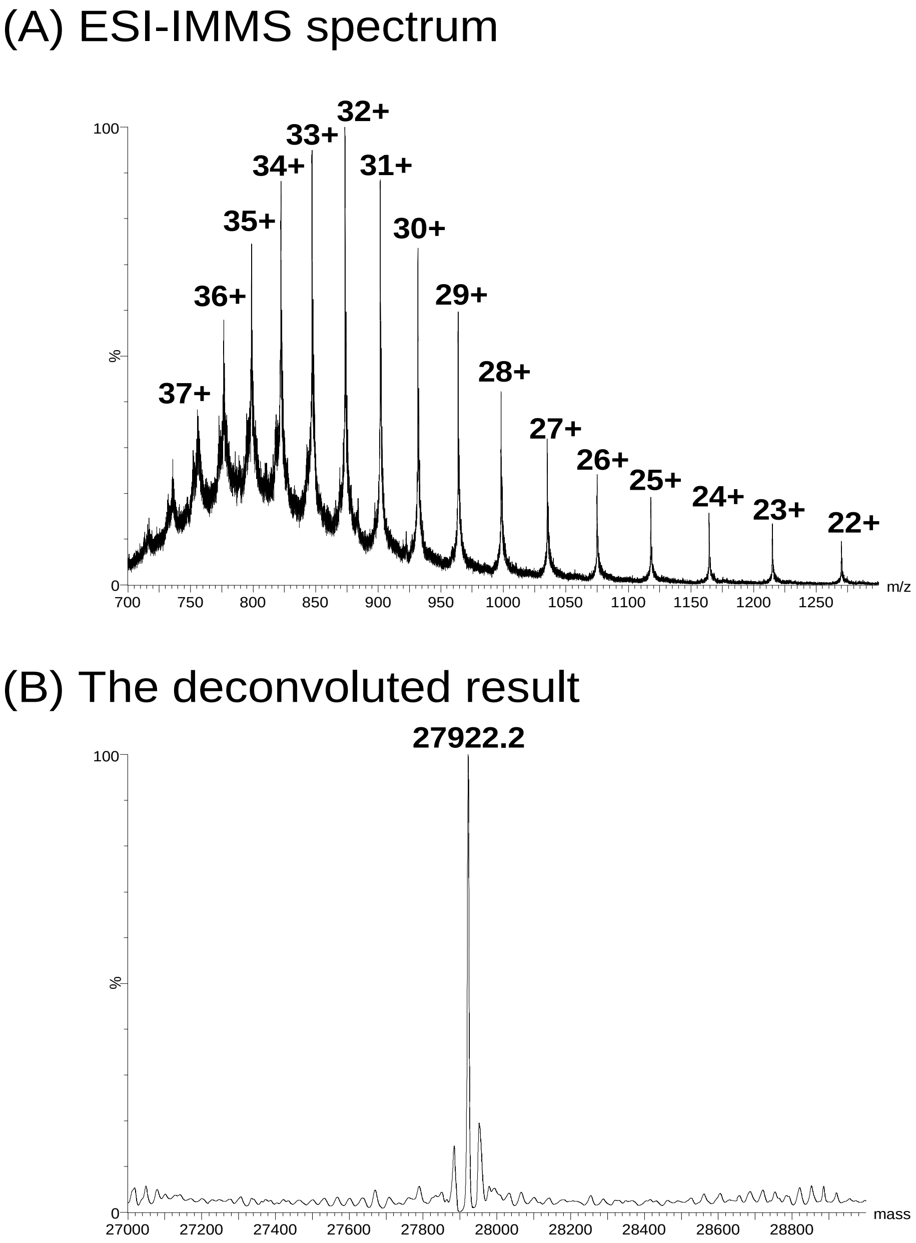


Figure III-5: Mass determination of the protein sfGFP-**3**: (A) ESI-IM-MS spectrum of sfGFP-**3** and (B) the deconvoluted ESI-IM-MS spectrum of sfGFP-**3.**

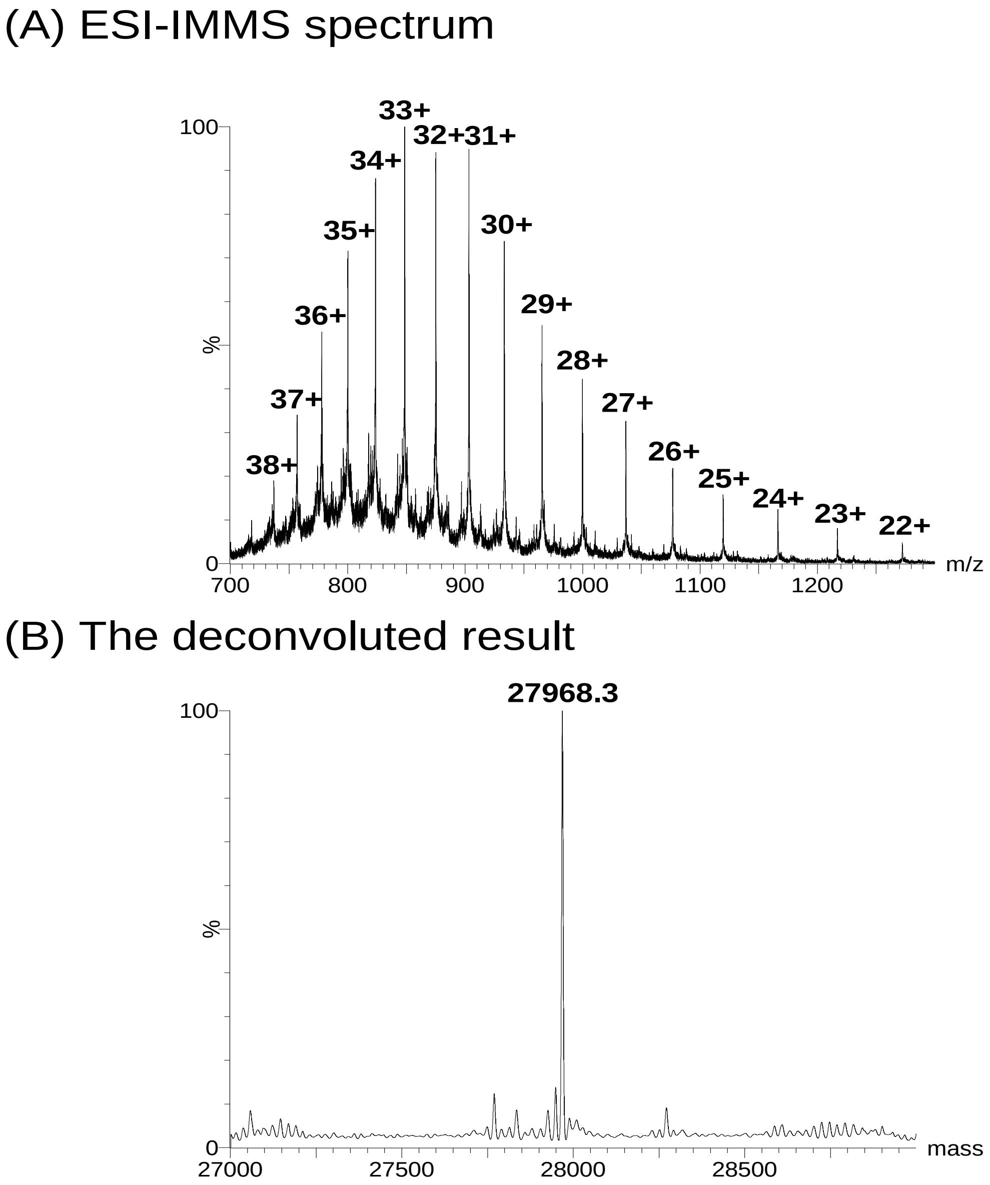


Figure III-6: Mass determination of the protein sfGFP-**4**: (A) ESI-IM-MS spectrum of sfGFP-**4** and (B) the deconvoluted ESI-IM-MS spectrum of sfGFP-**4**.

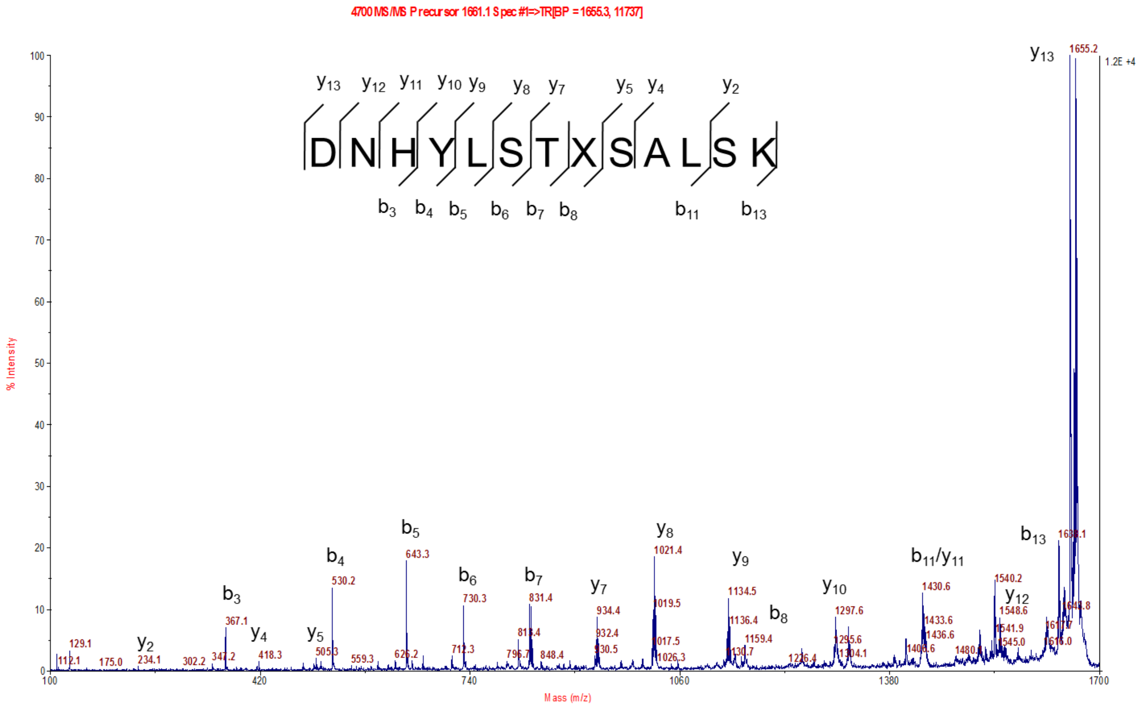


Figure III-7: GFP-**4** MALDI-MS/MS analysis on fragment DNHYLSTXALSK. X represents **4**.

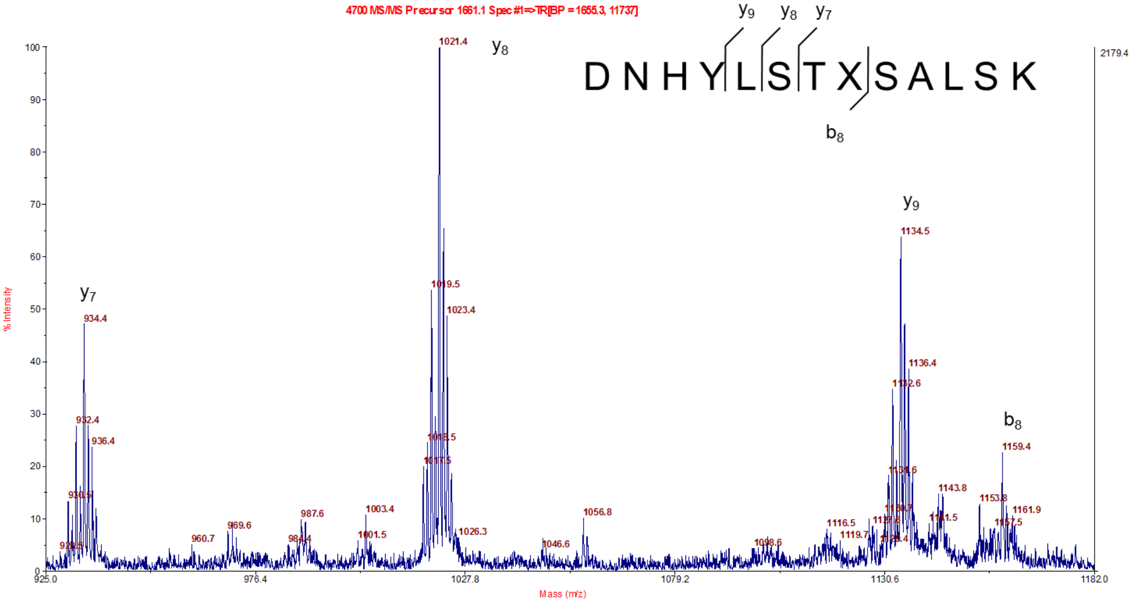
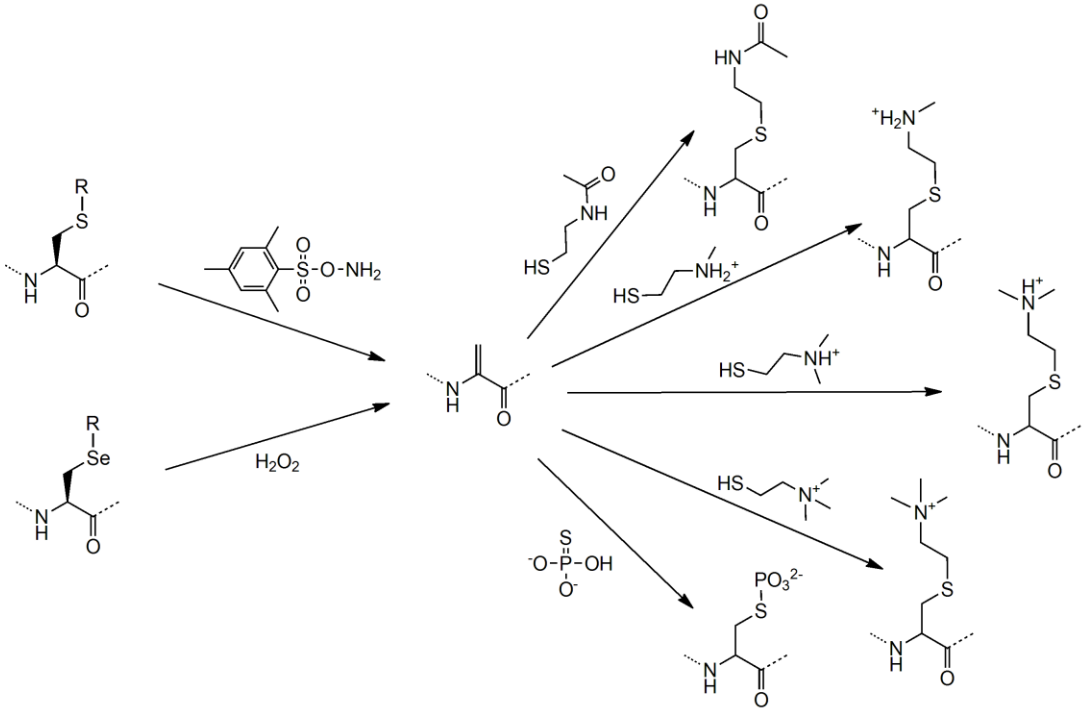


Figure III-8: GFP-**4** MALDI-MS/MS analysis on fragment DNHYLSTXALSK. X represents **4**. The *Se* isotopic patterns were clear in fragments y_7, y_8, y_9 and b_8.



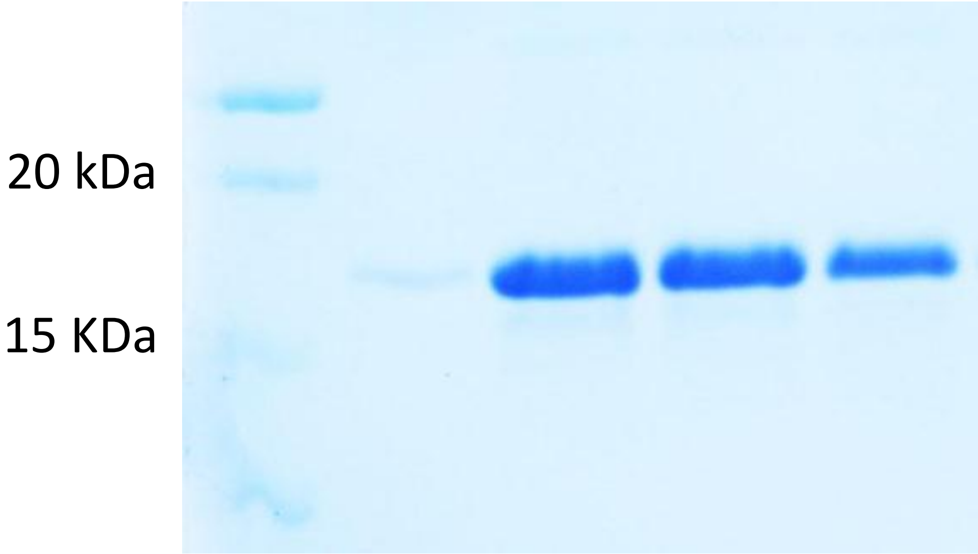
Scheme III-2: Oxidative elimination reactions to generate Dha that undergoes Michael addition reactions to form different posttranslational modification mimics.

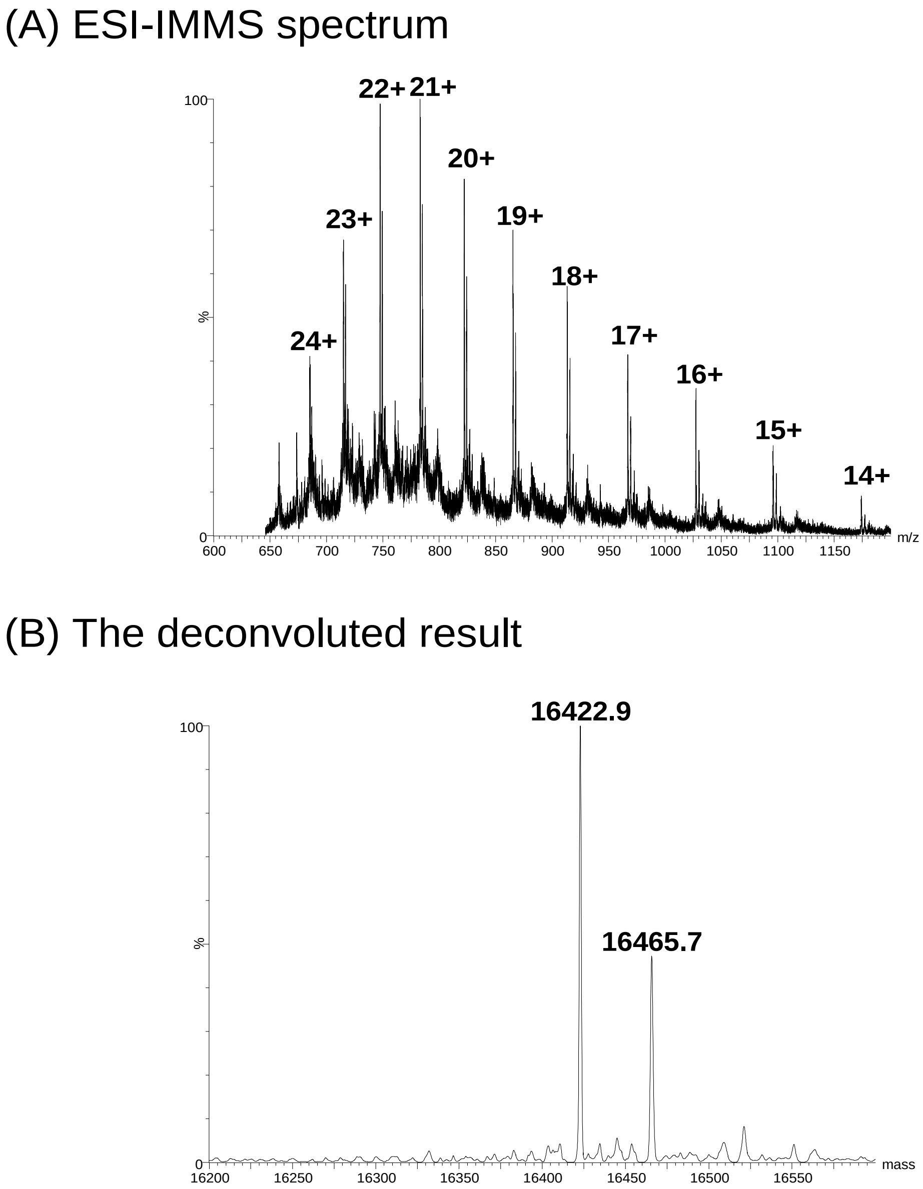
3 or **4**, the *Xenopus laevis* H3 gene with an amber mutation at its K9 position was used to replace the sfGFP gene in pET-sfGFP2TAG to afford pET-H3K9TAG. This plasmid together with pEVOL-mKRS1-pylT was used to cotransform *E. coli* BL21 cells that were subsequently grown in LB medium supplemented with either **3** or **4**. Both conditions led to high H3 expression levels with 185 mg/L for **3** and 126 mg/L for **4** (**Figure III-9**). Given that the expression level of wild type H3 at the same condition was 300 mg/L, these expression levels represented more than 50% of that of wild type H3 and are significantly better than the incorporation level of phenylselenocysteine. When both **3** and **4** were not provided in the medium, only a trace amount of H3 could be observed. The purified proteins both showed the expected mass when analyzed by ESI-MS. The observed mass peaks (H3 with **3** incorporated (H3K9-**3**): 16423 Da and 16466 Da; H3 with **4** incorporated (H3K9-**4**): 16470 Da and 16513 Da) agreed well with the calculated mass (H3K9-**3**: 16423 Da and 16465 Da; H3K9-**4**: 16470 Da and 16512 Da) (**Figure III-10-11**).

Reaction of H3K9-**3** with 8 M urea at pH 8.0 with 1000 eq. of MSH for 0.5 h failed to convert **3** to dehydroalanine. When the reaction time was prolonged, unexpected products were observed that did not correspond to the elimination product, indicating that MSH reacted with other amino acids in H3K9-**3**. Since H3K9-**3** is fully denatured, the structure rigidity is not a factor that may influence the conversion.[123] At this stage, we suspect the reaction itself is not favorable in 8 M urea.

To convert **4** in H3K9-**4** to dehydroalanine, we tried two oxidative reagents, sodium periodate and H_2O_2. Although it has been shown that sodium periodate can

2 mM **1**	-	+	-	-
2 mM **3**	-	-	+	-
1.5 mM **4**	-	-	-	+
Yield (mg/L)	N/A	205	185	126

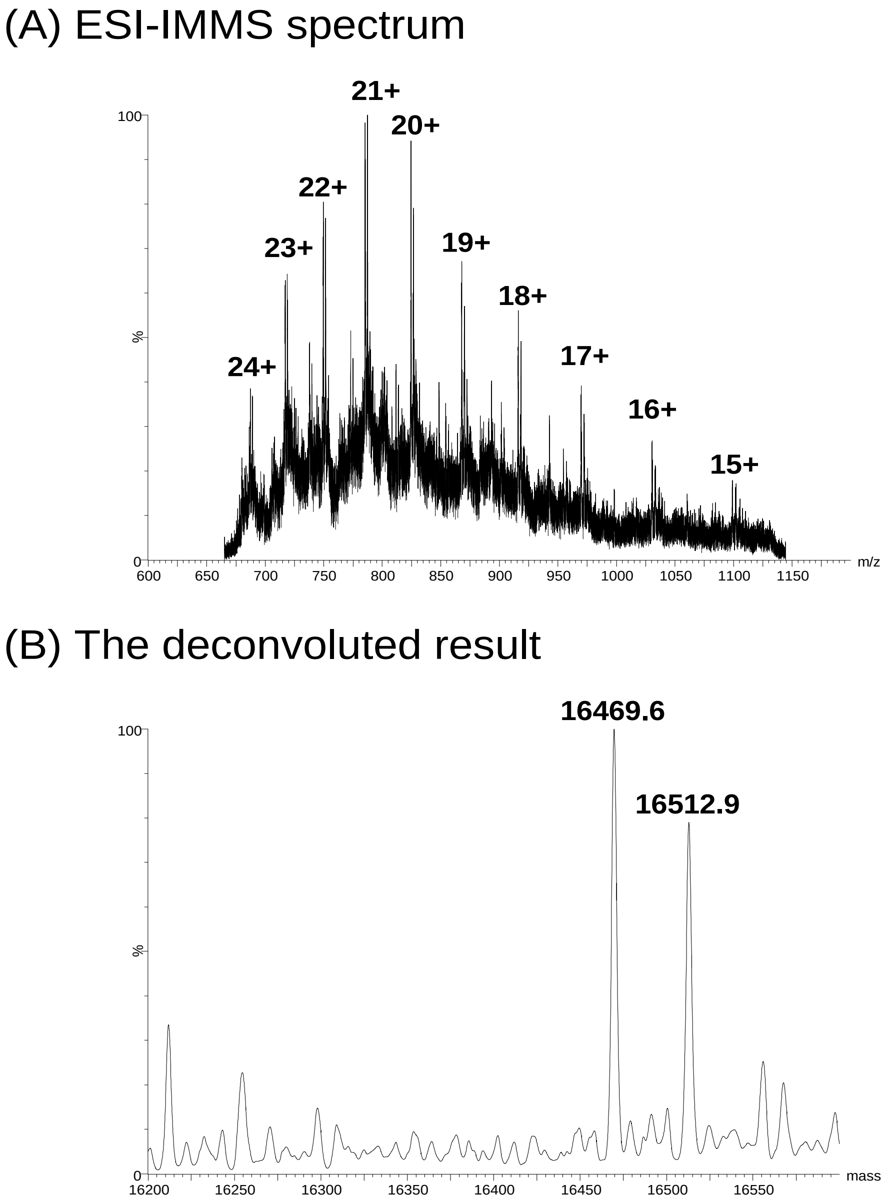




(A) ESI-IMMS spectrum

(B) The deconvoluted result

Figure III-10. Mass determination of the protein H3K9-**3**: (A) ESI-IM-MS spectrum of H3K9-**3** and (B) the deconvoluted ESI-IM-MS spectrum of H3K9-**3**.


(A) ESI-IMMS spectrum

(B) The deconvoluted result

Figure III-11. Mass determination of the protein H3K9-**4**: (A) ESI-IM-MS spectrum of H3K9-**4** and (B) the deconvoluted ESI-IM-MS spectrum of H3K9-**4**.

efficiently convert a derivatized selenocysteine to dehydroalanine in peptides, it apparently didn't work on H3K9-**4**. Adducts that clearly indicated single and multiple oxygen atom(s) added to the protein were detected. On the contrary, oxidative elimination of H3K9-**4** using H_2O_2 processed smoothly. In the presence of 100 eq. of H_2O_2, **4** in H3K9-**4** was efficiently converted to dehydroalanine in 1 h. The ESI-MS analysis of the final product also indicated the oxidation of the two methionine residues in H3K9-**4** to methionine sulfoxide (**Figure III-12-14**). The detected molecular mass (16244 Da and 16287 Da) agreed well with the calculated mass of H3 with dehydroalanine at its K9 position (H3K9Dha) that also has two oxidized methionine residues (16244 Da and 16286 Da). There is also one peak at 12229 Da that matched the mass of H3K9Dha with only one oxidized methionine residue (calculated mass: 16228 Da). To convert dehydroalanine to N^{ε}-acetylthialysine, a H3K9Dha was incubated together with 1500 eq. of *N*-acetylcysteamine at 25 degree for 1 h as shown in **Scheme III-2**. ESI-MS analysis of the final product indicated the quantitative conversion of dehydroalanine to N^{ε}-acetylthialysine (**Table III-2**) (**Figure III-15**). Top down tandem MS analysis of the final product confirmed the site-specific installation of N^{ε}-acetylthialysine at the K9 position (**Figure III-16-19**).[124-126] This top down tandem MS analysis also confirmed that an additional oxidation at M120 (**Figure III-16-19**). We performed similar reactions to convert dehydroalanine in H3K9Dha to N^{ε}-methylthialysine, N^{ε},N^{ε}-dimethylthialysine, and N^{ε},N^{ε},N^{ε}-trimethylthialysine by reacting H3K9Dha with corresponding cysteamine derivatives. (**Figure III-20-22**) The final

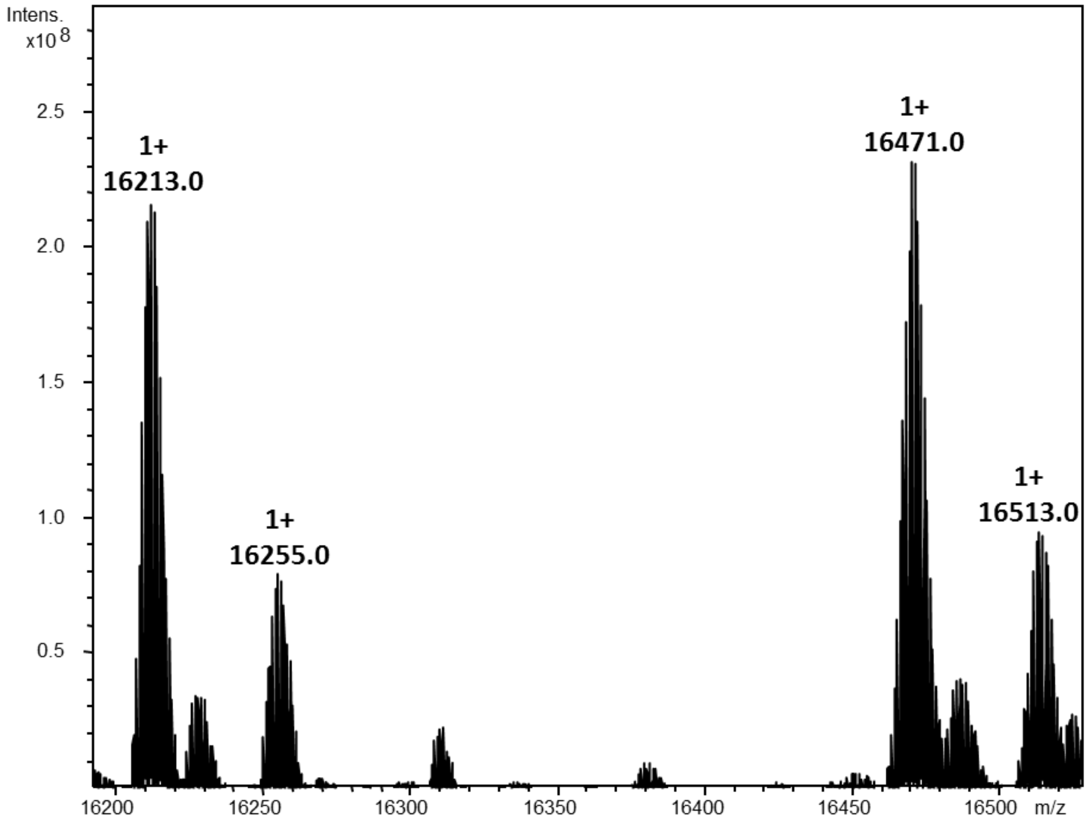


Figure III-12. Molecular weight determination of H3K9-**4** after its treatment with 5eq H_2O_2. The deconvoluted singly charged ESI-MS spectrum by FT-ICR MS. Both H3K9-**4** and H3K9Dha were found.

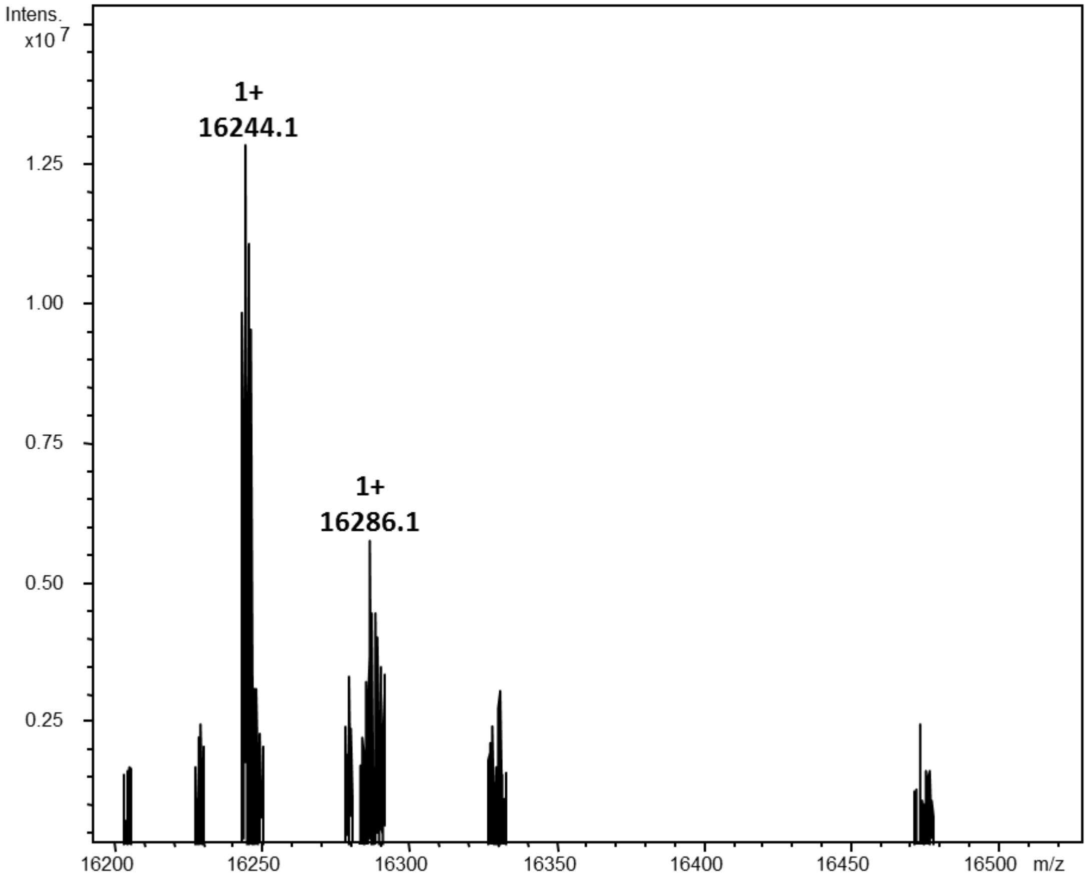
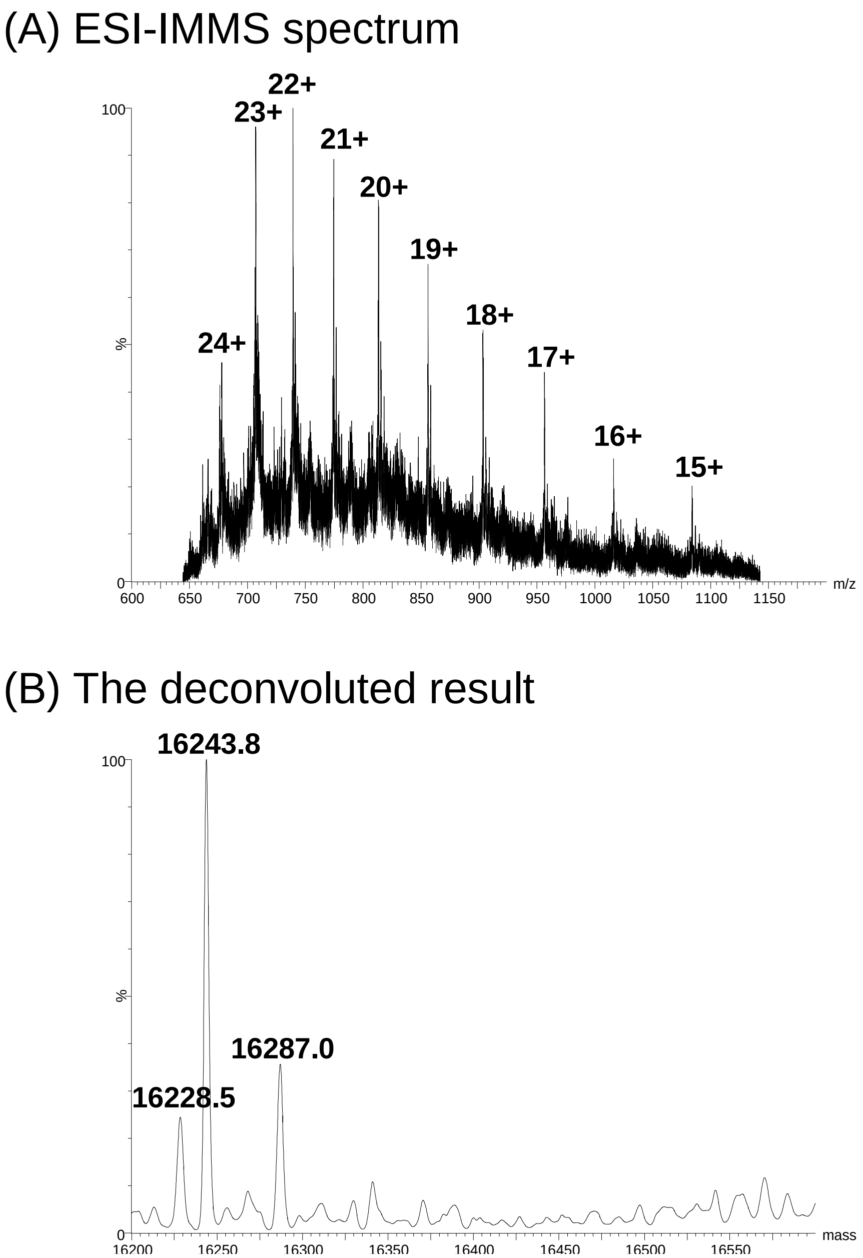


Figure III-13. Molecular weight determination of H3K9Dha. The deconvoluted singly charged ESI-MS spectrum of H3K9Dha by FT-ICR MS. H3K9Dha was synthesized from H3K9-**4** by treatment with 100 eq. H_2O_2 at room temperature for 1 h.



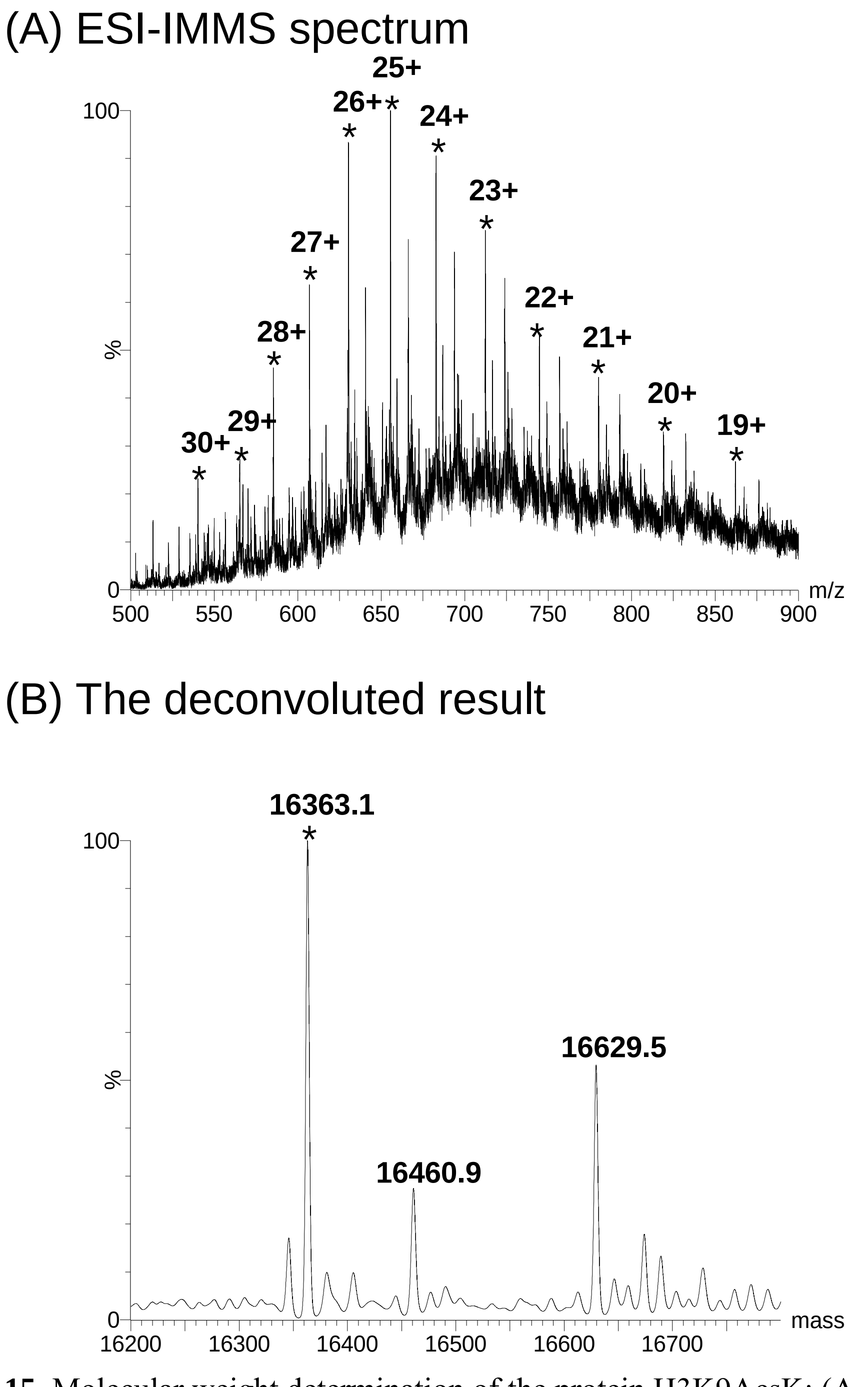
(A) ESI-IMMS spectrum

(B) The deconvoluted result

FigureIII- 14. Molecular weight determination of H3K9Dha: (A) ESI-IM-MS and (B) the deconvoluted spectra of H3K9Dha. H3K9Dha was synthesized from H3K9-**4** by treatment with 100 eq. H_2O_2 at room temperature for 1 h.

products were analysed by Fourier transform ion cyclotron resonance (FT-ICR) MS. As shown in **Table III-2**, the detected mass agreed well with the calculated mass. A similar reaction was also carried out to install phosphocysteine at the K9 position of H3 by reacting out to install phosphocysteine at the K9 position of H3 by reacting H3K9Dha with 1000 eq. of thiophosphate for 1 h at room temperature. The FT-ICR-MS analysis of the final product indicated a quantitative conversion. (**Figure III-23)** This chemically installed phosphocysteine was stable at the physiological pH. Storing the final product in a 4 degree fridge for a week did not show significant degradation.

To demonstrate that N^{ε}-methylthialysine, N^{ε},N^{ε}-dimethylthialysine, and N^{ε},N^{ε},N^{ε}-trimethylthialysine can closely mimic their naturally counterparts, H3 with N^{ε}-methylthialysine installed at K9 (H3K9msK), H3 with N^{ε},N^{ε}-dimethylthialysine installed at K9 (H3K9m^{2}sK), and H3 with N^{ε},N^{ε},N^{ε}-trimethylthialysine incorporated at K9 (H3K9m^{3}sK) were immunoprecipitated by heterochromatin protein 1β (HP-1β) that specifically binds to H3 with mono-, di-, or trimethylated lysine at K9 but not wild type H3. The immunoprecipitates were then probed by anti-H3 antibody that specifically recognizes the C-terminal domain of H3. All three proteins were pulled out by HP-1β, showing intense bands when detected by anti-H3 antibody. Importantly, wild type H3 was not immunoprecipitated by HP-1β at all. To demonstrate N^{ε}-acetylthialysine closely mimics N^{ε}-acetyllysine, H3 with N^{ε}-acetylthialysine incorporated at K9 (H3K9AcsK) was probed by anti-H3K9Ac antibody that was raised against H3 with N^{ε}-acetyllysine at K9 in a Western blot analysis. The intense detected band for H3K9AcsK proved the specific interaction between H3K9AcsK and anti-H3K9Ac antibody. A similar assay for


(A) ESI-IMMS spectrum

(B) The deconvoluted result

Figure III-15. Molecular weight determination of the protein H3K9AcsK: (A) ESI-IM-MS and (B) the deconvoluted spectra of H3K9AcsK.

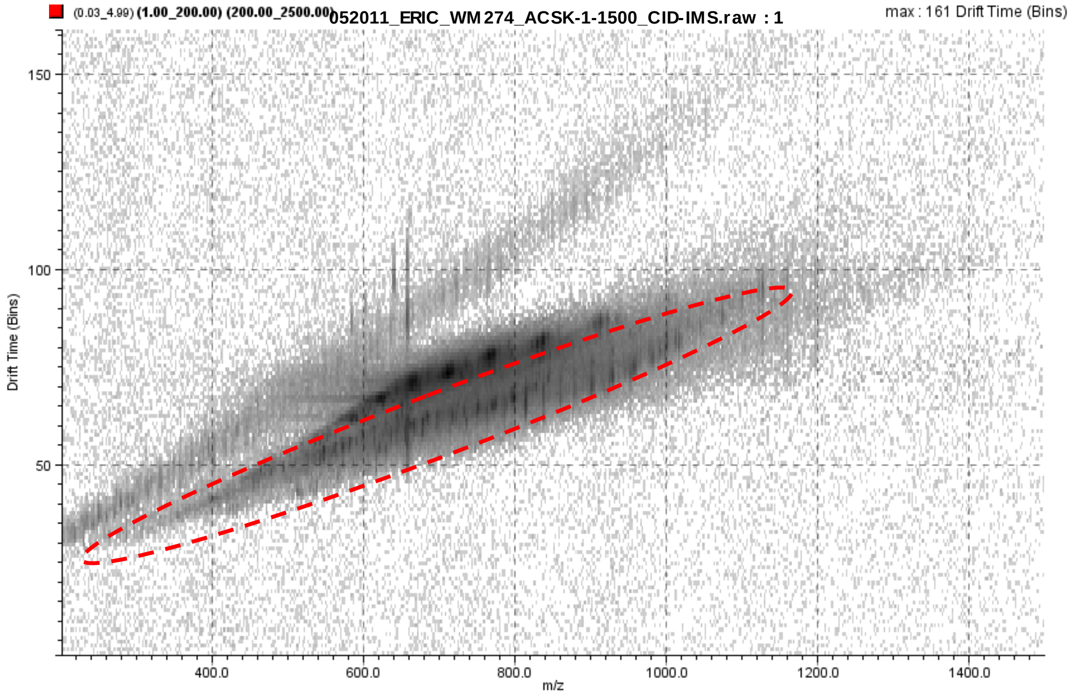


Figure III-16. Top-down analysis of H3K9AcsK by CID-IM-MS.

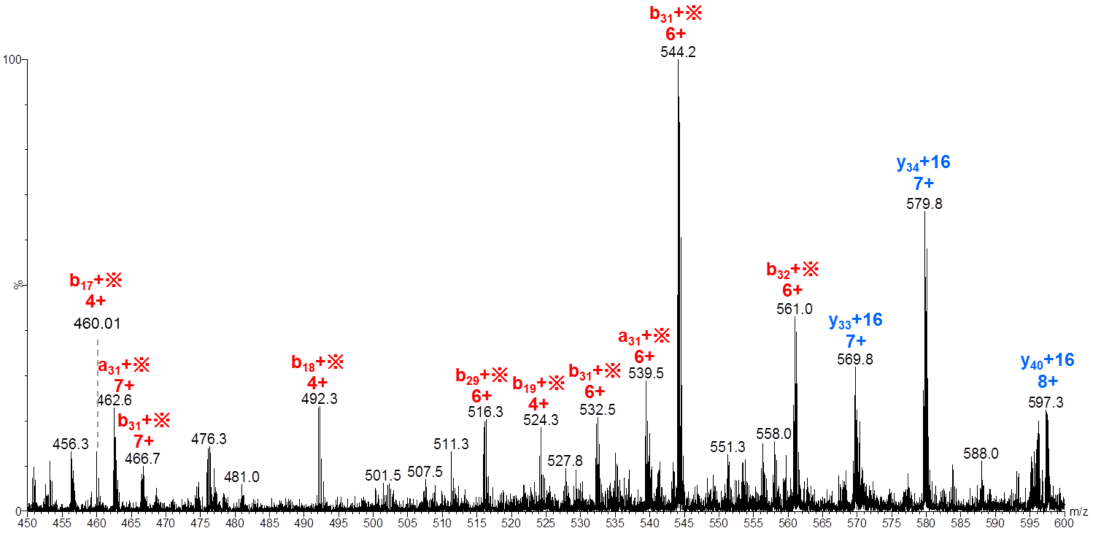


Figure III-17. The extracted MS/MS spectrum for top-down analysis of H3K9AcsK by CID-IM-MS. The extracted MS/MS region was represented as red dotted line in Figure III-16. The spectrum was shown in zoom-in region of m/z 450-600.

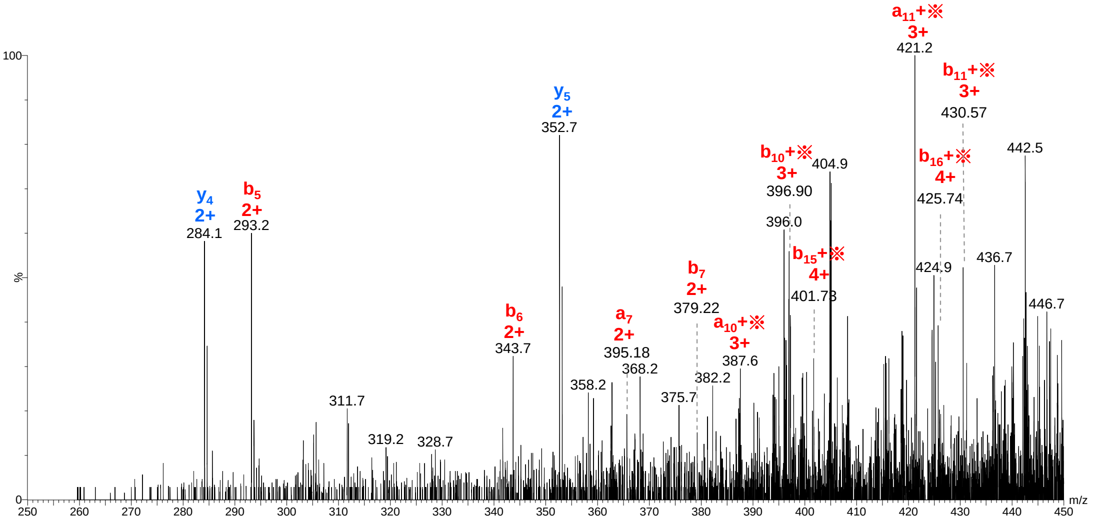


Figure III-18. The extracted MS/MS spectrum for top-down analysis of H3K9AcsK by CID-IM-MS. The spectrum was shown in zoom-in region of m/z 250-450.

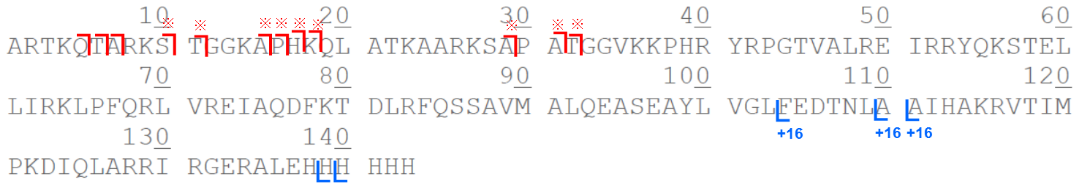

10
20
30
40
50
60
ARTKQTARKS TGGKAPHKQL ATKAARKSAP ATGGVKKPHR YRPGTVALRE IRRYQKSTEL
70
80
90
100
110
120
LIRKLPFQRL VREIAQDFKT DLRFQSSAVM ALQEASEAYL VGLFEDTNLA AIHAKRVTIM
+16
+16
+16
130
140
PKDIQLARRI RGERALEHHH HHH

Figure III-19. The summary of top-down analysis for H3K9AcsK by CID-IM-MS. This result indicates the AcsK modification site is between the protein sequences of R8 to S10. ※ denotes the AcSK modification (M+60); +16 denotes methionine oxidation (M+16). The b fragment ions were represented in red and y in blue.

Table III-2. Theoretical and detected molecular weight of various H3 proteins.

Proteins	Theoretic MW_{avg} (Da)	Detected MW_{avg} (Da)
H3K9-**3**	16423[b]	16423[b]
	16465[c]	16466[c]
H3K9-**4**	16470[b]	16470[b]
	16512[c]	16513[c]
H3K9Dha	16228[d]	16229[d]
	16244[e]	16244[e]
	16270[f]	16287[g]
	16286[g]	
H3K9AcsK	16347[d]	16363[e]
	16363[e]	
	16389[f]	
	16405[g]	
H3K9mesK	16319[d]	16318[d]
	16335[e]	16334[e]
	16361[f]	16361[f]
	16377[g]	16377[g]
H3K9m^{2}sK	16333[d]	16332[d]
	16349[e]	16348[e]
	16375[f]	16391[g]
	16391[g]	
H3K9m^{3}sK	16348[d]	16347[d]
	16364[e]	16363[e]
	16390[f]	
	16406[g]	
H3K9pC[a]	16342[d]	16341[d]
	16358[e]	16357[e]
	16384[f]	
	16400[g]	

[a] H3K9pC contains phosphocysteine at its K9 position. [b] A full length protein without M1. [c] A full length protein without M1 but containing the N-terminal acetylation. [d] A full length protein without M1 but containing one oxidized methionine residue. [e] A full length protein without M1 but containing two oxidized methionine residues. [f] A full length protein without M1 but containing one oxidized methionine residues and the N-terminal acetylation. [g]. A full length protein without M1 but containing two oxidized methionine residues and the N-terminal acetylation.

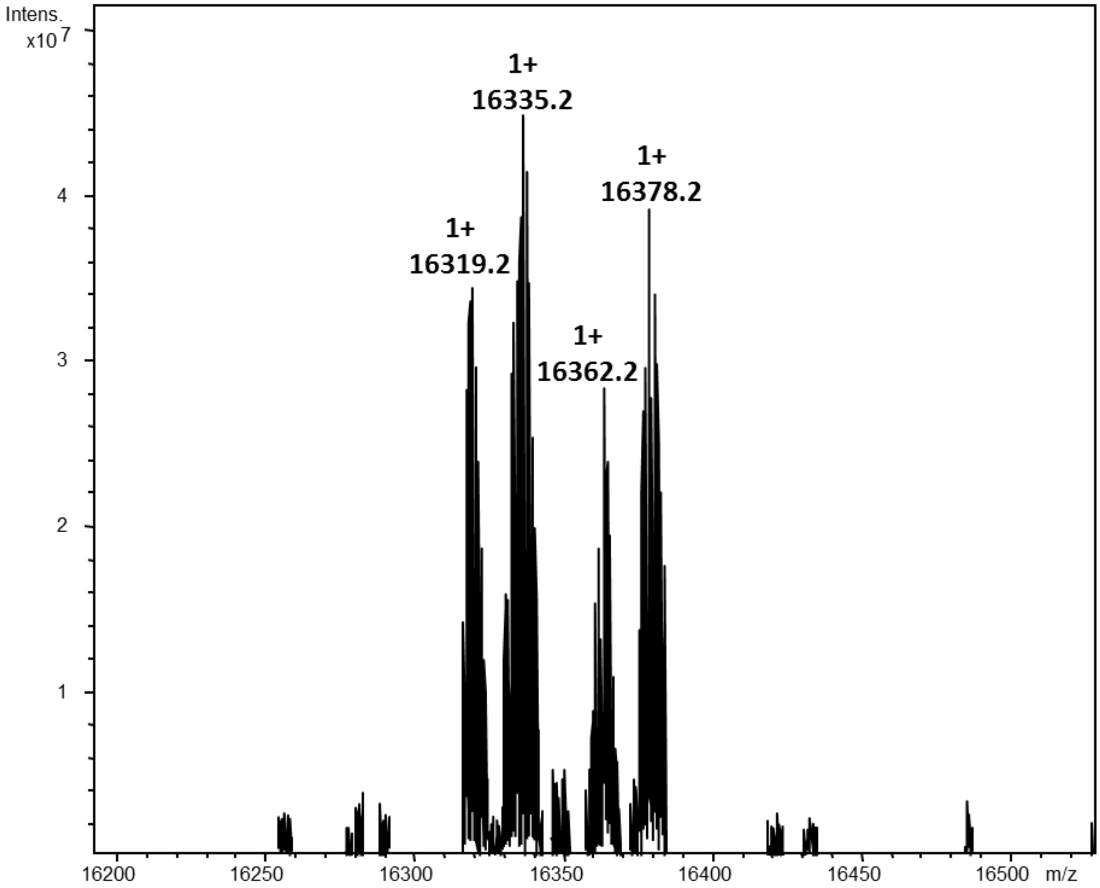


Figure III-20. Molecular weight determination of Protein H3K9mesK. The deconvoluted singly charged ESI-MS spectrum of H3K9 H3K9mesK by FT-ICR MS.

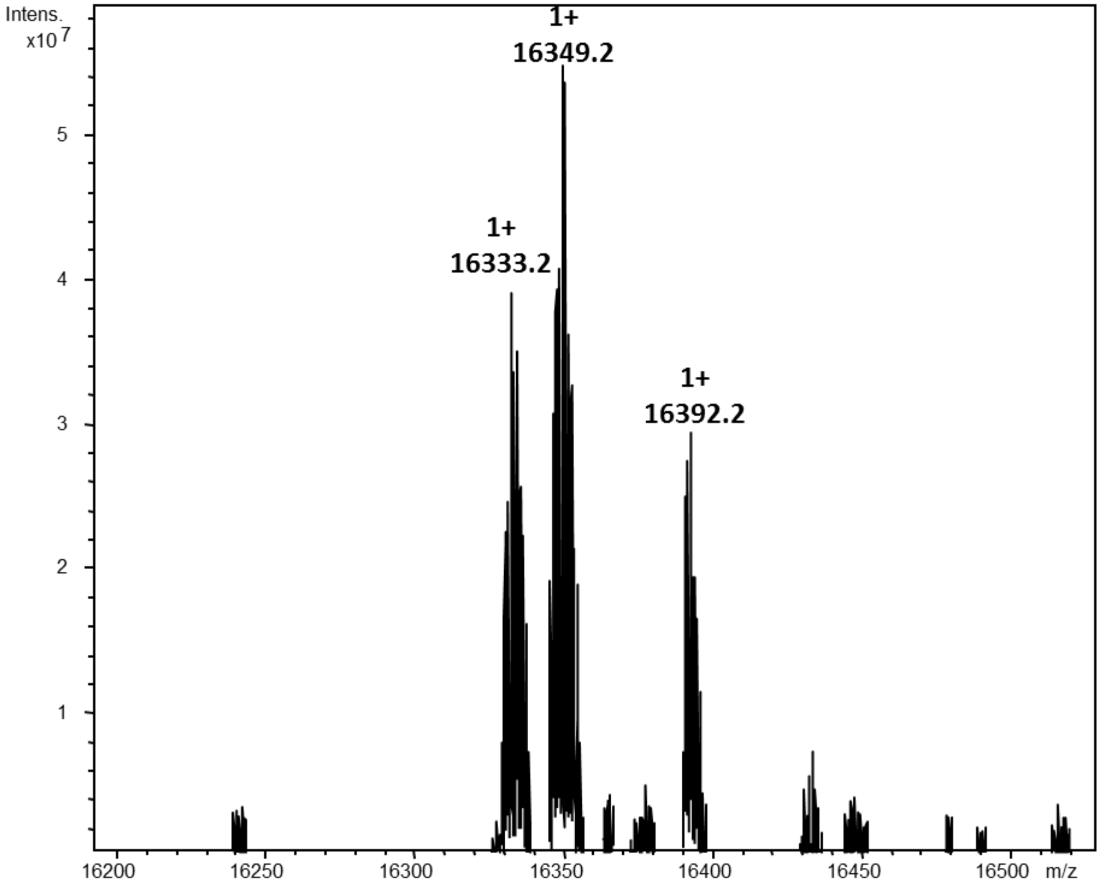


Figure III-21. Molecular weight determination of Protein H3K9m^{2}sK. The deconvoluted singly charged ESI-MS spectrum of H3K9 H3K9m^{2}sK by FT-ICR MS.

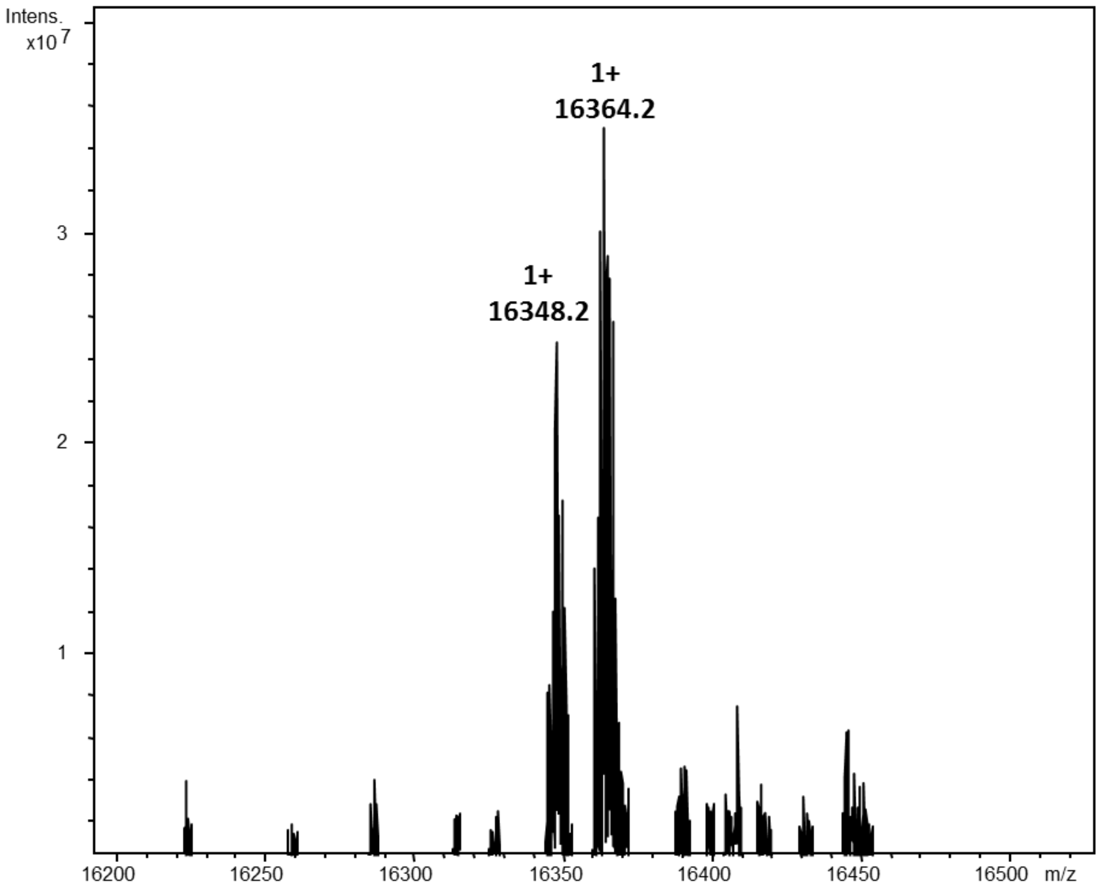


Figure III-22. Molecular weight determination of Protein H3K9m^{3}sK. The deconvoluted singly charged ESI-MS spectrum of H3K9 H3K9m^{3}sK by FT-ICR MS.

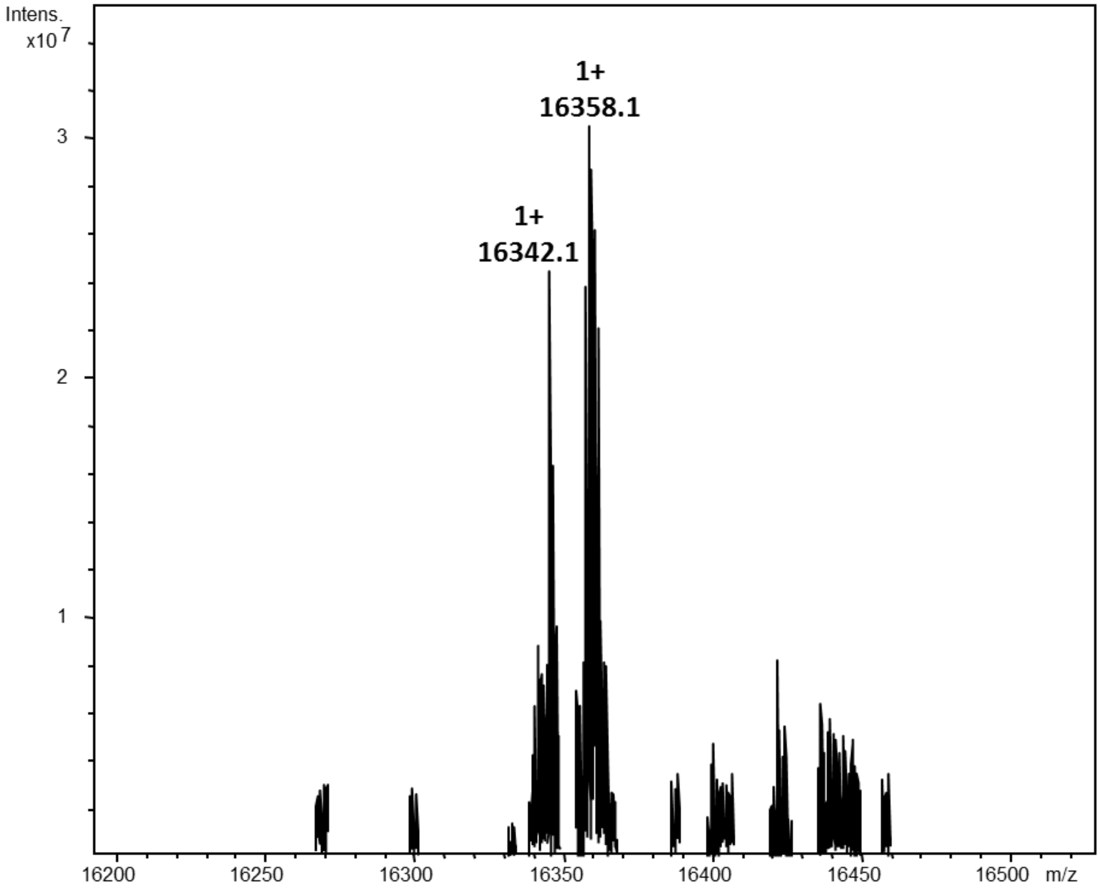


Figure III-23. Molecular weight determination of Protein H3K9pC. The deconvoluted singly charged ESI-MS spectrum of H3K9 H3K9pC by FT-ICR MS.

wild type H3 did not give detectable signal. All these experiments clearly demonstrated that N^{ε}-methylthialysine, N^{ε},N^{ε}-dimethylthialysine, N^{ε},N^{ε},N^{ε}-trimethylthialysine, and N^{ε}-acetylthialysine closely mimic their natural modified lysine counterparts in biological interactions. Given that all thialysine derivatives can be obtained from Michael addition reactions of dehydroalanine and the genetic incorporation of **3** into histones followed by oxidative elimination can lead to a large quantity of dehydroalanine-containing histones, the method we report here is an optimal choice to synthesize histones with different lysine modifications for their functional analysis. Although we did not do biochemistry analysis of H3 with phosphocysteine incorporated at K9, we believe phosphocysteine closely mimics phosphoserine and may mimic phosphothreonine. Since the reported method can be generalized to synthesize histones with phosphocysteine incorporated at sites that naturally have phosphoserine or phosphothreonine, this method can definitely be applied to study histone phosphorylation that is related to transcription regulation, DNA repair and cell division.

Conclusion

In conclusion, we have developed a method to genetically incorporate a *Se*-alkylselenocysteine into a histone protein in *E. coli*. The yield of the expressed histone can reach to 50% of its wild type counterpart. The selective oxidative elimination of *Se*-alkylselenocysteine to form dehydroalanine and its following reaction with thiol-containing small molecules allow the access of histones with several posttranslational modification mimics, including lysine methylation, lysine acetylation, and serine phosphorylation. The application of the reported method in the histone biology research area will facilitate addressing the questions such as how posttranslational modifications affect the chromatin structure, how these modifications regulate the interactions between chromatin and other non-chromatin regulatory proteins, and how modifications on chromatin cross regulate each other.

CHAPTER IV

THE DE NOVO ENGINEERING OF PYRROLYSYL-TRNA SYNTHETASE FOR GENETIC INCORPORATION OF L-PHENYLALANINE AND ITS DERIVATIVES*

Introduction

Using orthogonal aminoacyl-tRNA synthetase (aaRS)-amber suppressor tRNA ($tRNA_{CUA}$) pairs, more than seventy NAAs have been site-specifically incorporated into proteins at amber mutation sites in *E. coli*, yeast, and mammalian cells.[21, 23, 127-129] These NAAs provide new opportunities to generate proteins with novel properties and amend proteins with biochemical or biophysical probes for their structural and functional analysis. There are two broadly used aaRS-$tRNA_{CUA}$ pairs for the NAA incorporation in *E. coli*, namely the *Mj*TyrRS-$tRNA^{Tyr}_{CUA}$ pair derived from *Methanococcus jannaschii* tyrosyl-tRNA synthetase-$tRNA^{Tyr}$ pair and the pyrrolysyl-tRNA synthetase (PylRS)-$tRNA^{Pyl}_{CUA}$ pair.[26, 52, 112] Both pairs have their limitations. Although the *Mj*TyrRS-$tRNA^{Tyr}_{CUA}$ pair has been proved powerful for the genetic code expansion in *E. coli*, NAAs that were genetically encoded using evolved *Mj*TyrRS-$tRNA^{Tyr}_{CUA}$ pairs must contain a β- or γ-aromatic side chain[24] and the *Mj*TyrRS-$tRNA^{Tyr}_{CUA}$ pair can not be directly used in

*Reprinted in part from “The *de novo* engineering of pyrrolysyl-tRNA synthetase for genetic incorporation of L-phenylalanine and its derivatives”, Yane-Shih Wang, William K. Russell, Zhiyong Wang, Wei Wan, Lindsey E. Dodd, Pei-Jing Pai, David H. Russell and Wenshe, R. Liu. *Mol. Biosyst.* **2011**, 7, 714 - 717. Copyright 2011, with permission from The Royal Society of Chemistry.

yeast and mammalian cells (Liu W. R., unpublished data). It has been demonstrated that the PylRS-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair is orthogonal in *E. coli*, yeast and mammalian cells.[85, 96, 130, 131] However, NAAs that were genetically encoded using the wild-type and evolved PylRS-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pairs either are *N*$^{\varepsilon}$-acylated lysine derivatives or contain a long aliphatic side chain.[56, 66, 68, 85, 86, 96, 130, 131] An ideal aaRS-tRNA$_{\mathrm{CUA}}$ pair that resolves the aforementioned limitations would fulfill the two following requirements. First, it should be orthogonal in *E. coli*, yeast and mammalian cells. Therefore, the pair could be evolved simply and quickly in *E. coli* and subsequently transferred directly to yeast and mammalian cells, avoiding identification and evolution of every aaRS-tRNA$_{\mathrm{CUA}}$ pair in different cell lines. Second, the pair could be evolved for genetic incorporation of NAAs with diversified side chains that could be either short aromatic, long aromatic, or aliphatic. This would relieve the burden to identify unique orthogonal aaRS-tRNA$_{\mathrm{CUA}}$ pairs for NAAs with different side chains. Herein, we demonstrate that the PylRS-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair fulfills both requirements and can be engineered for genetic incorporation of NAAs with short aromatic side chains into proteins. The radical change of substrate specificity of PylRS from pyrrolysine (**1** in **Scheme IV-1**) to L-phenylalanine, *p*-iodo-L-phenylalanine, and *p*-bromo-L-phenylalanine (**2**, **3**, **4** in **Scheme IV-1**) indicates powerful potentials of engineering PylRS for genetic incorporation of other NAAs with short aromatic side chains. This development is expected to greatly expand the inventory of NAAs that can be genetically incorporated into proteins in *E. coli*, yeast, and mammalian cells. Given that derivatives of an evolved *Mj*TyrRS-tRNA$^{\mathrm{Tyr}}_{\mathrm{CUA}}$ pair and a wild-type or evolved PylRS-$\text{tRNA}^{\text{Pyl}}_{\text{CUA}}$ pair can be coupled together for genetic incorporation of two different NAAs into one protein in *E. coli*,[88, 132] this development makes it possible to incorporate two L-phenylalanine derivatives into one protein and will have potential applications in protein folding dynamics and enzyme mechanism studies.[47, 133]

Experimental Section

General Experimental

The *p*-iodo-L-phenylalanine and *p*-bromo-L-phenylalanine were purchased form ChemImpex. 3-(Dansylamino)phenylbornic acid (DaFBA) was purchased from Sigma-Aldrich. Z Domain, GFP_{UV}, pylT and *Methanosarcina mazei* PylRS gene and protein sequences are the same as described in Chapter II, as well as plasmids pET-pylT-GFP, pET-pylT-GFP, pY+, pY- and pRS1 library used in the chapter.

Plasmid Constructions

The construction methods of plasmids pET-pylT-Z, pET-pylT-GFP, pY+ and pY- used in the work was described in the **General Experimental** section of Chapter II.

Construction of pET-Z. The pET-Z plasmid was derived from pETDuet-1 (Novagen$^{\text{TM}}$). Wild type Z domain gene was amplified from the pLeiZ plasmid. Two restriction sites, *NdeI* at the 5' end and *SacI* at the 3' end, were introduced in the PCR product which was subsequently digested and used to replace GFP_{UV} in pET-pylT-GFP.

Selection Procedure for Evolving Pyrrolysyl-tRNA Synthetase

The selections followed the scheme shown in **Scheme II-2** by 1 mM **3**. Five alternative selections (three positive (P) + two negative (N) with P-N-P-N-P order) finally yielded many colonies. 16 single colonies after the third positive selection were selected and the plasmids were isolated for sequencing. 4 single colonies from the third positive selection were also chosen for testing their ability to grow on plates with 102 µg/mL chloramphenicol, 25 µg/mL Kan, 12 µg/mL Tet, and 1 mM of **3** or 5 mM of **4**. A plate without NAA supplementary was used as a control. Images of colonies growing on different plates were shown in **Figure IV-1**. The positive selection of PylRS mutants specific for L-phenylalanine was carried out similarly as that for *p*-iodo-L-phenylalanine except that no noncanonical amino acid was supplemented into the medium.

GFP$_{UV}$ and Z domain Protein Expression and Purification

To express GFP$_{UV}$ and Z domain incorporated with a NAA were performed by the method in the **Experimental Section** of Chapter II. To express wild-type Z-domain proteins (**Z-wt**), we transformed *E. Coli* BL21(DE3) cells with pET-Z. Cells were recovered in 1 mL LB medium for 1 h at 37 ºC before being plated on a LB agar plate containing Kan (50 µg/mL) and Amp (100 µg/mL). A single colony was then selected and grown overnight in a 10 mL culture. This overnight culture was used to inoculate 100 mL LB medium supplemented with 100 µg/mL Amp. Cells were grown at 37ºC in an incubating shaker and protein expression was induced when OD$_{600}$ reached 0.7 by adding IPTG to a final concentration of 1 mM. After 6 h induction, cells were harvested

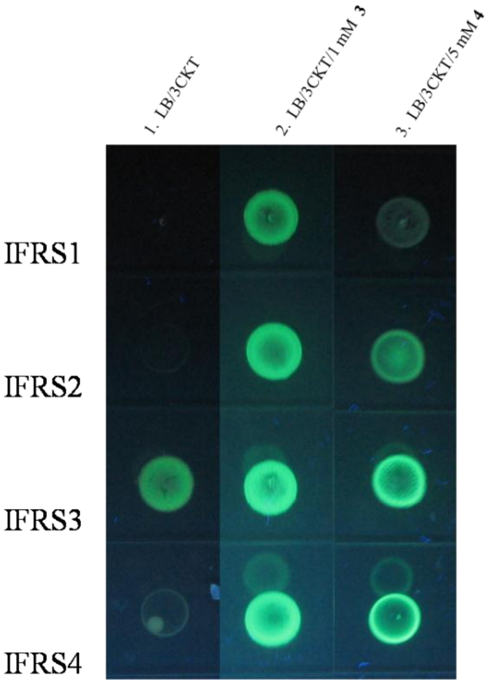


Figure IV-1. Growth of 4 selected IFRS mutants from the third positive selection of **3** on LB plates with different supplements. (**1**) Growth on LB/3CKT (LB agar plates containing 102 μg/mL Cm, 25 μg/mL Kan, and 12 μg/mL Tet); (**2**) Growth on plates containing 1 mM **3**, 102 μg/mL Cm, 25 μg/mL Kan, and 12 μg/mL Tet; (**3**) Growth on plates containing 5 mM **4**, 102 μg/mL Cm, 25 μg/mL Kan, and 12 μg/mL Tet. All the colonies were cultured at 37°C for 24 hours. Images were taken under UV 365 nm radiation. The pY+ plasmid has a GFP_{UV} gene under control of a T7 promoter and a T7 RNA polymerase gene that contains two amber mutations at positions 1 and 107. The expression of GFP_{UV} is promoted by the suppression of two amber mutations in the T7 RNA polymerase. The fluorescent intensity of the expression of GFP_{UV} roughly represents the suppression efficiency at amber codons.

and resuspended in PBS buffer. Further protein purification was the same as that for **Z-3**. The purified protein was analyzed by 15% SDS-PAGE.

Suzuki Coupling on the Z-domain Proteins

A water-souble palladium catalyst using 2-amino-4,6-dihydroxyprimidine as its ligand was prepared according to the literature protocol, and the coupling of **Z-3** with 3-(dansylamino)phenylboronic acid (DaFBA) was carried out accordingly with minor modification. Z-wt was used as a control. To a protein sample of The buffer of 50 μL of **Z-3** (0.5 mg/mL, 80 μL, 4.9 nmol) or **Z-wt** (0.4 mg/mL, 80 μL, 3.9 nmol) in 0.1× PBS solution (1 mM Na_2HPO_4, 0.18 mM KH_2PO_4, 13.7 mM NaCl, 0.27 mM KCl, pH 7.4) was added the palladium catalyst in water (10 mM, 7 μL, 70 nmol), aqueous formic acid (20 mM, 3 μL, 60 nmol) and DaFBA in DMSO (20 mM, 20 μL, 400 nmol). The mixture was vortexed and then heated in a 35 ^{o}C water bath for 6 h. After dialysis against 0.1× PBS overnight (2L × 2, 8 h each time) to remove excessive dye and catalyst, the protein samples were lyophilized, redissolved in 10 μL of 8 M urea, and analyzed by the SDS-PAGE (15%) electrophoresis.

Protein LC-ESI-MS Analysis

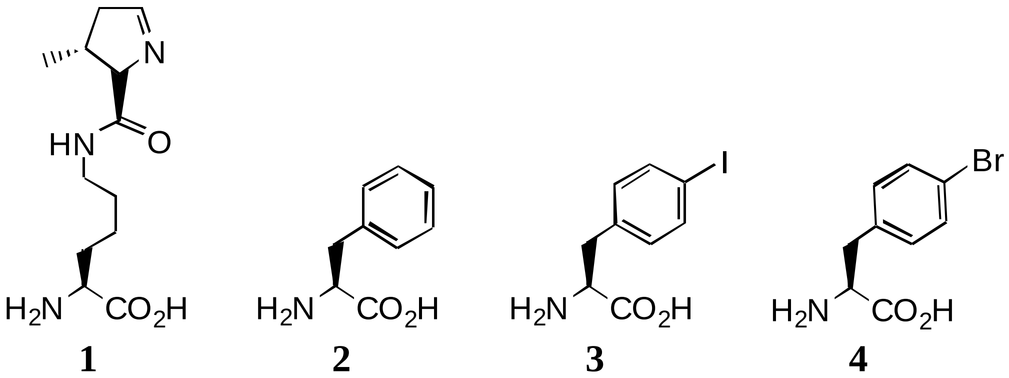
An Agilent (Santa Clara, CA) 1200 capillary HPLC system was interfaced to an API QSTAR Pulsar Hybrid QTOF mass spectrometer (Applied Biosystems/MDS Sciex, Framingham, MA) equipped with an electrospray ionization (ESI) source. Liquid chromatography (LC) separation was achieved using a Phenomenex Jupiter C4 microbore column (150 × 0.50 mm, 300 Å) (Torrance, CA) at a flow rate of 10 μL per min. The proteins were eluted using a gradient of (A) 0.1% formic acid versus (B) 0.1% formic acid in acetonitrile. The gradient timetable was as follows: 2% B for 5 min, 2-30% in 3 min, 30-60% in 44 min, 60-95% in 8 min, followed by holding the gradient at 95% for 5 min, for a total run time of 65 min. The MS data were acquired in positive ion mode (500-1800 Da) using spray voltage of +5000 V. BioAnalyst software (Applied Biosystems) was used for spectral deconvolution. For the GFPuv protein analysis, a mass range of m/z 500-1800 was used for deconvolution and the output range was 10000-50000 Da using a step mass of 0.1 Da and a S/N threshold of 20. For the Z-Domain protein analysis, a mass range of m/z 500-2000 was used for deconvolution and the output range was 5000-15000 Da for Z-domain-His6X using a step mass of 0.1 Da and a S/N threshold of 20.

Tandem Mass Spectrametry analysis

GFP$_{UV}$ variants from the SDS-PAGE gels was cut, dissolved in 25 mM Ammonia bicarbonate, and denatured at 90 degree for 15 min. Proteinase Asp-N (Roche) was dissolved in 0.01% TFA (pH 3). Proteinase Asp-N solution was added to the substrate protein solution (w/w=1:50), and incubated at 37 degree overnight. Peptides resulting from the proteinase Asp-N digestion were mixed 1:1 (v/v) with matrix (5 mg mL-1 α-cyano-4-hydroxycinnamic acid, 50% (v/v) acetonitrile, 10 mM ammonium dihydrogen phosphate, 1% TFA) and 1 μL of the resulting mixture was spotted onto a stainless steel target plate. Mass spectra and tandem MS spectra were collected using an Applied Biosystems 4800 Tof/Tof (Framingham, MA). Collision induced dissociation tandem MS spectra were acquired using air at the medium pressure setting and at 2 kV of collision energy. Tandem MS data was manually interpreted using the Data Explorer™ software package (Applied Biosystems, Framingham, MA).

Results and Discussion

In order to evolve a PylRS mutant that specifically acylates $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ with L-phenylalanine, we carried out a standard positive selection of the pRS1 plasmid library in *E. coli*. The pRS1 plasmid library contains the *Methanosarcina mazei* PylRS gene with randomization at six active-site residues (L305, Y306, L309, N346, C348, and W417; **Figure II-1**).[68] The selection was based on the resistance to chloramphenicol, which was conferred by suppressing a permissive amber mutation in the chloramphenicol acetyltransferase gene in the minimal medium.[23, 52] Cells containing PylRS mutants that acylate $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ with natural amino acids survive the selection. Screening the survived clones revealed two PylRS mutants (FRS1 and FRS2 in **Table IV-1**) that are specific for L-phenylalanine. Both clones contain the mutation N346A. This mutation apparently removes the steric clashes between the β-amide of N346 and the aromatic side chain of L-phenylalanine that prevent the binding of L-phenylalanine to the wild-type PylRS. C348 in both FRS1 and FRS2 is mutated to a larger amino acid that reduces the size of the active site to better accommodate L-phenylalanine. Site-specific incorporation of L- phenylalanine at an amber mutation at Q204 of $\mathrm{GFP}_{\mathrm{UV}}$ using the FRS1-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair or the FRS2-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair in the GMML medium (liquid glycerol minimal medium containing 0. 3 mM L-leucine) supplemented with 1 mM L-



Scheme IV-1. Structures of pyrrolysine, L-phenylalanine, *p*-iodo-L-phenylalanine, and *p*-bromo-L-phenylalanine

Table VI-1. Selected PylRS mutants.

PylRS	L305	Y306	L309	N346	C348	W417
FRS1	L	Y	L	A	L	W
FRS2	L	Y	L	A	K	W
IFRS1 (9/16)[a]	M	L	S	S	M	W
IFRS2 (3/16) [a]	L	Y	L	A	M	L
IFRS3 (3/16) [a]	L	T	A	S	M	W
IFRS4 (1/16) [a]	L	T	R	S	M	W

[a] The occurring frequency of a selected mutant in all sequenced clones.

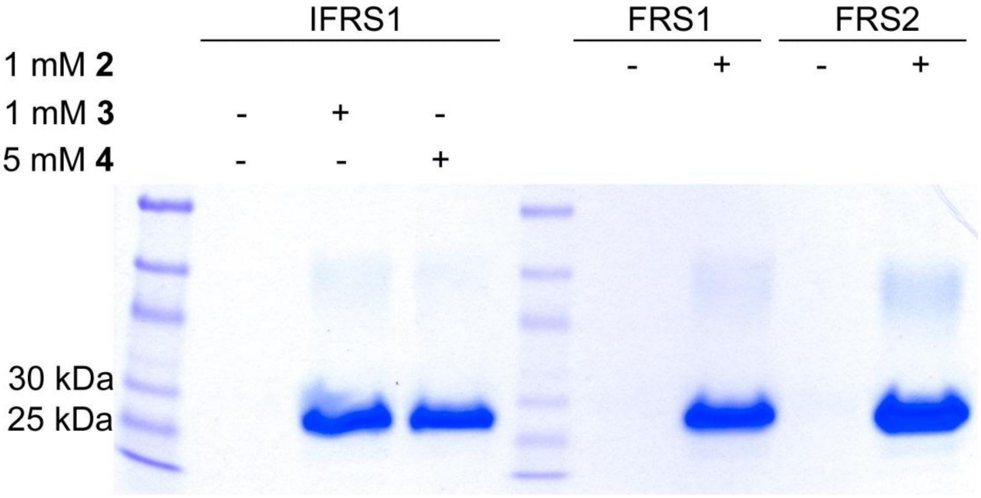


Figure IV-2: Suppression of the amber mutation at Q204 of GFP$_{UV}$ by the FRS1-, FRS2-, and IFRS1-tRNA$^{Pyl}_{CUA}$ pairs at different conditions. The gel was stained by Gelcode blue. All GFP$_{UV}$ proteins are expressed in 500 mL GMML, purified and concentrated to 250 μL. 10 μL of each sample was loaded and analyzed by the SDS-PAGE.

phenylalanine afforded the full-length GFP$_{\mathrm{UV}}$ (**Figure IV-2**). The yields of the full-length GFP$_{\mathrm{UV}}$ expression were ~1.6 mg/L for the FRS1-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair and ~2.9 mg/L for the FRS2-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair. In the absence of L-phenylalanine in the GmmL medium, only a trace amount of the full-length GFP$_{\mathrm{UV}}$ could be detected. The electrospray ionization mass spectrometry (ESI-MS) analysis of the expressed full-length GFP$_{\mathrm{UV}}$ proteins show molecular weights (27,730 Da for the FRS1-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair and 27,729 Da for the FRS2-tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair) that agree well with the calculated mass (27,729 Da) of the full-length GFP$_{\mathrm{UV}}$ with L-phenylalanine incorporated at Q204 but without the N-terminal methionine (**Figures IV-3** & **4**). The hydrolysis of the N-terminal methionine from GFP$_{\mathrm{UV}}$ has been observed in several similar studies.[55, 68, 96, 134] The incorporation of L-phenylalanine at Q204 was also independently confirmed by the tandem mass spectral (MS-MS) analysis of endoproteinase Asp-N-digested fragments of the purified full-length GFP$_{\mathrm{UV}}$ proteins. The tandem mass spectra of the fragment of DNHYLSTF*SALSK (F* denotes the designated L-phenylalanine) validated the incorporation of L-phenylalanine at Q204. The observed F*-containing ions (y_6 to y_{13} and b_8 to b_{13}) all had expected mass (**Figures IV-5A** & **B**).

Since L-phenylalanine is a canonical aromatic amino acid, we further tested whether PylRS could be evolved to specifically acylate tRNA$^{\mathrm{Pyl}}_{\mathrm{CUA}}$ with a noncanonical short aromatic amino acid. We chose to work on *p*-iodo-L-phenylalanine because its genetic incorporation into proteins may facilitate protein structure determination and serve as an anchor to label proteins through the Suzuki-Miyaura cross-coupling

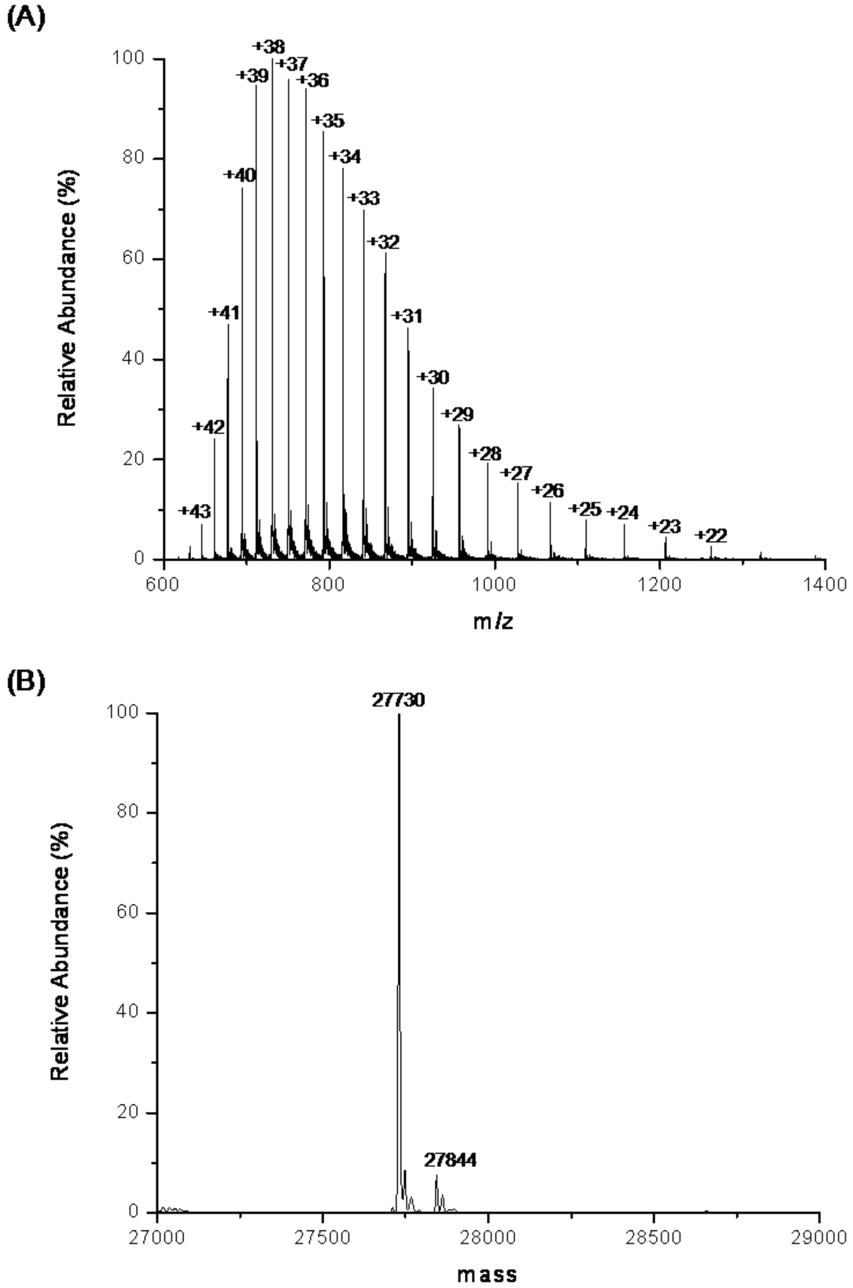


Figure IV-3. Mass determination of GFP$_{UV}$ incorporated with L-phenylalanine at Q204 using the FRS1-tRNA$^{Pyl}_{CUA}$ pair. (**A**) The ESI-MS spectrum and (**B**) the deconvoluted ESI-MS spectrum.

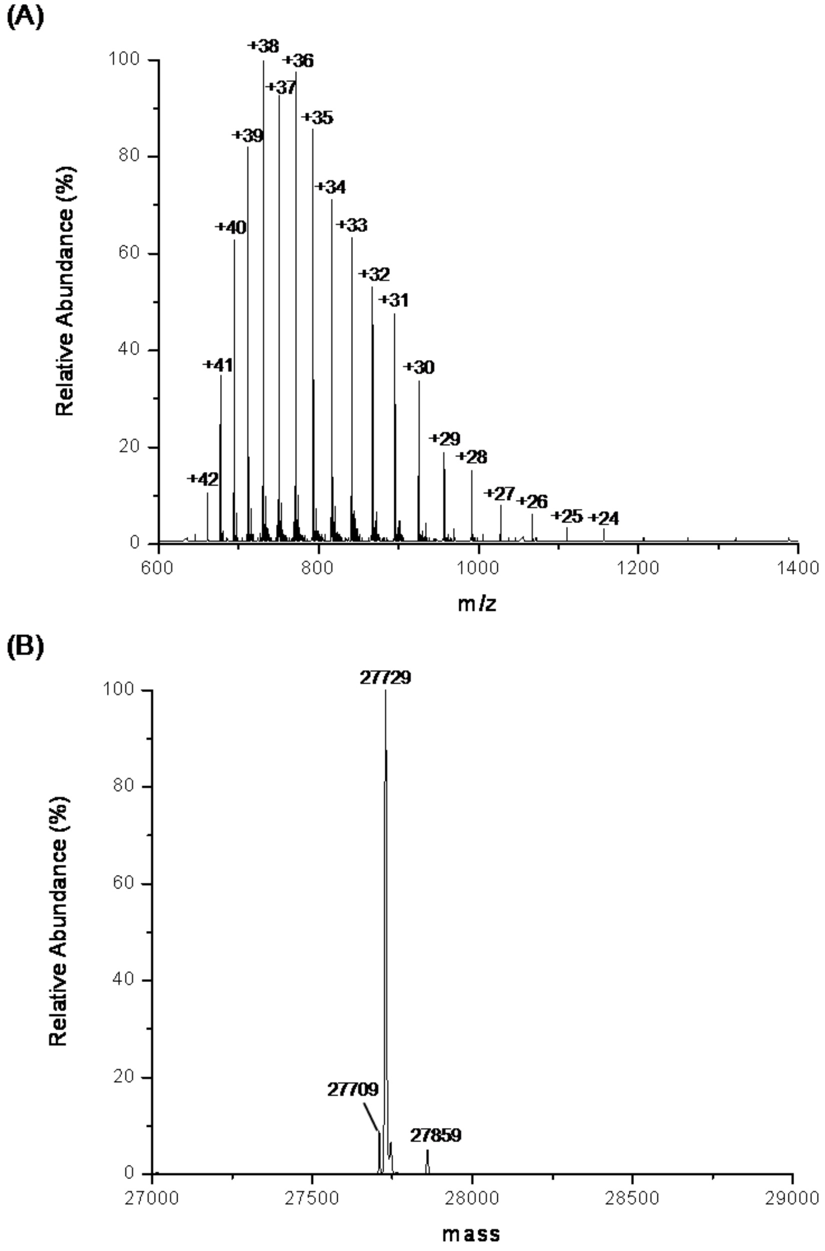


Figure IV-4. Mass determination of $\mathrm{GFP_{UV}}$ incorporated with L-phenylalanine at Q204 using the FRS2-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair. (**A**) The ESI-MS spectrum and (**B**) the deconvoluted ESI-MS spectrum.

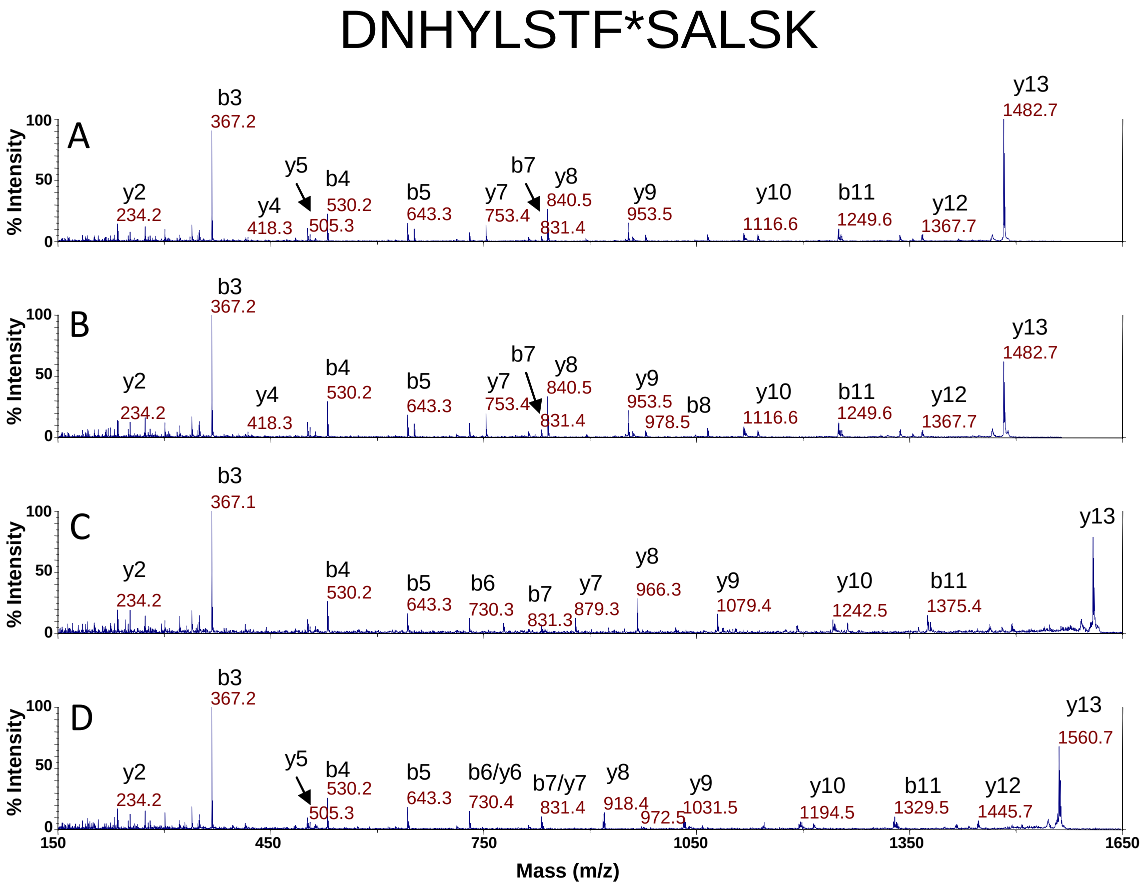


Figure [illegible] Tandem mass spectra of the DNHYLSTF*SALSK (F* denotes L-phenylalanine or its derivatives) fragments from the full-length GFP_{UV} protein. (**A**) The full-length GFP_{UV} was expressed using the evolved FRS1-$tRNA^{Pyl}_{CUA}$ pair in the presence of 1 mM L-phenylalanine. (**B**) The full-length GFP_{UV} was expressed using the evolved FRS2-$tRNA^{Pyl}_{CUA}$ pair in the presence of 1 mM L-phenylalanine. (**C**) The full-length GFP_{UV} was expressed using the evolved IFRS1-$tRNA^{Pyl}_{CUA}$ pair in the presence of 1 mM *p*-iodo-L-phenylalanine. (**D**) The full-length GFP_{UV} was expressed using the evolved IFRS1-$tRNA^{Pyl}_{CUA}$ pair in the presence of 5 mM *p*-bromo-L-phenylalanine.

reaction.[94, 135] The pRS1 plasmid library was passed through rounds of alternative positive and negative selections according to a standard protocol[23] to identify PylRS mutants that are specific for *p*-iodo-L-phenylalanine. The positive selection was carried out similarly to the selection of L-phenylalanine-specific PylRS mutants except that 1 mM *p*-iodo-L-phenylalanine was supplemented in the medium. The negative selection utilized the toxic barnase gene with amber mutations at two permissive sites and was carried out in the LB medium with the absence of *p*-iodo-L-phenylalanine. PylRS mutants that acylate $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ selectively with *p*-iodo-L-phenylalanine survive both positive and negative selections. After five rounds of selections, sixteen survived clones were validated and sequenced. They converged to four unique mutants (IFRS1-IFRS4 in **Table IV-1**). The most abundant mutant, IFRS1, was further characterized. When the IFRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair was used to suppress the amber mutation at Q204 of $\mathrm{GFP_{UV}}$ in the presence of 1 mM *p*-iodo-L-phenylalanine, full-length $\mathrm{GFP_{UV}}$ was produced. The expression yield of the full-length $\mathrm{GFP_{UV}}$ in the GMML medium was ~1.0 mg/L (**Figure IV-2**). The ESI-MS detected mass (27,855 Da) of the purified protein matched perfectly well with the calculated mass (27,855 Da) of the full-length $\mathrm{GFP_{UV}}$ with *p*-iodo-L-phenylalanine incorporated at Q204 but without the N-terminal methionine (**Figure IV-6**). In contrast, only a negligible amount of full-length $\mathrm{GFP_{UV}}$ was expressed when *p*-iodo-L-phenylalanine was absent. These results indicate that the IFRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$pair specifies *p*-iodo-L-phenylalanine but not any canonical amino acids. Since *p*-bromo-L-phenylalanine and *p*-iodo-L-phenylalanine are structurally similar, we suspected that

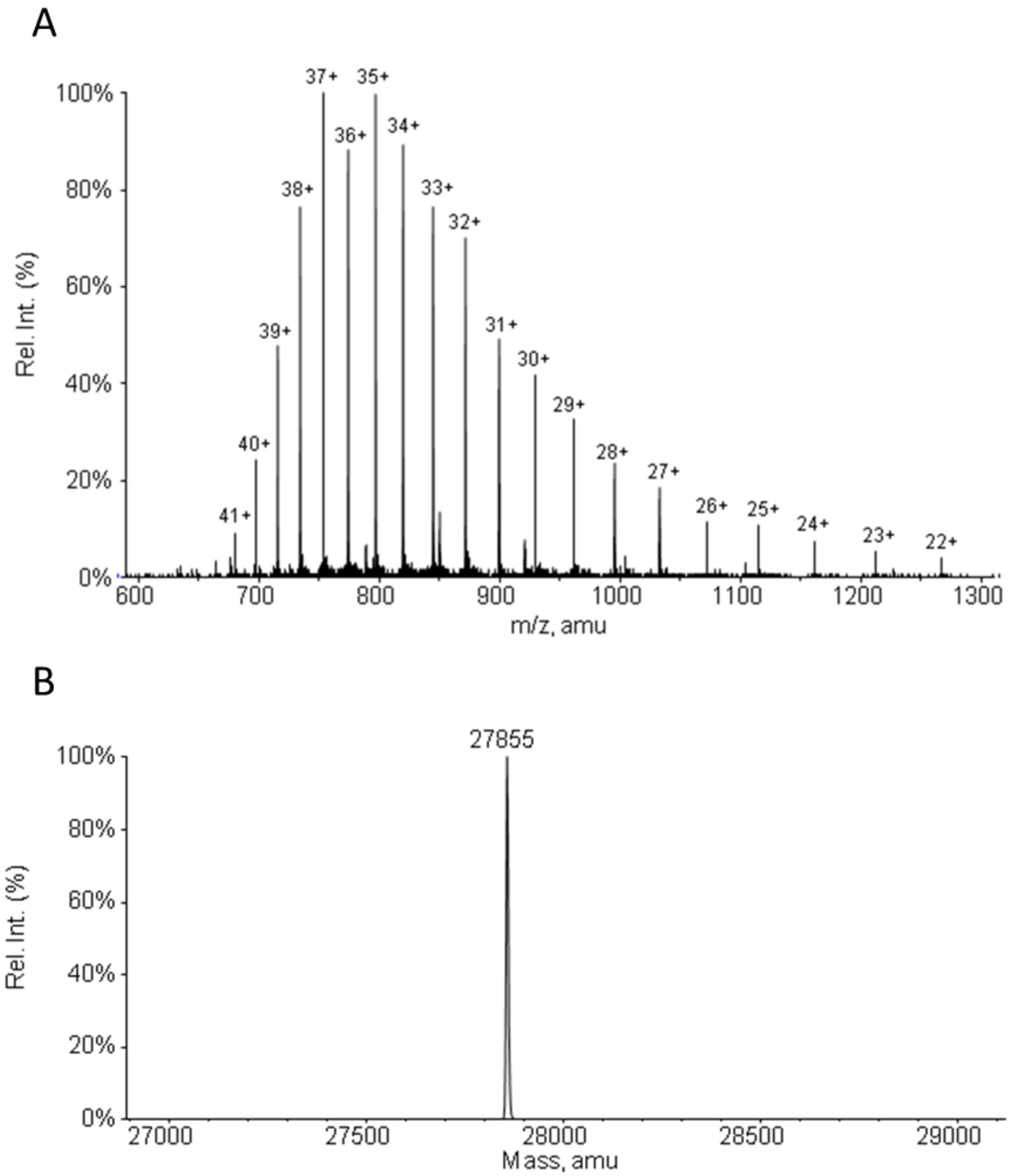


Figure IV-6. Mass characterization of **GFP-3**. (**A**) The ESI-MS spectrum and (**B**) the deconvoluted ESI-MS spectrum.

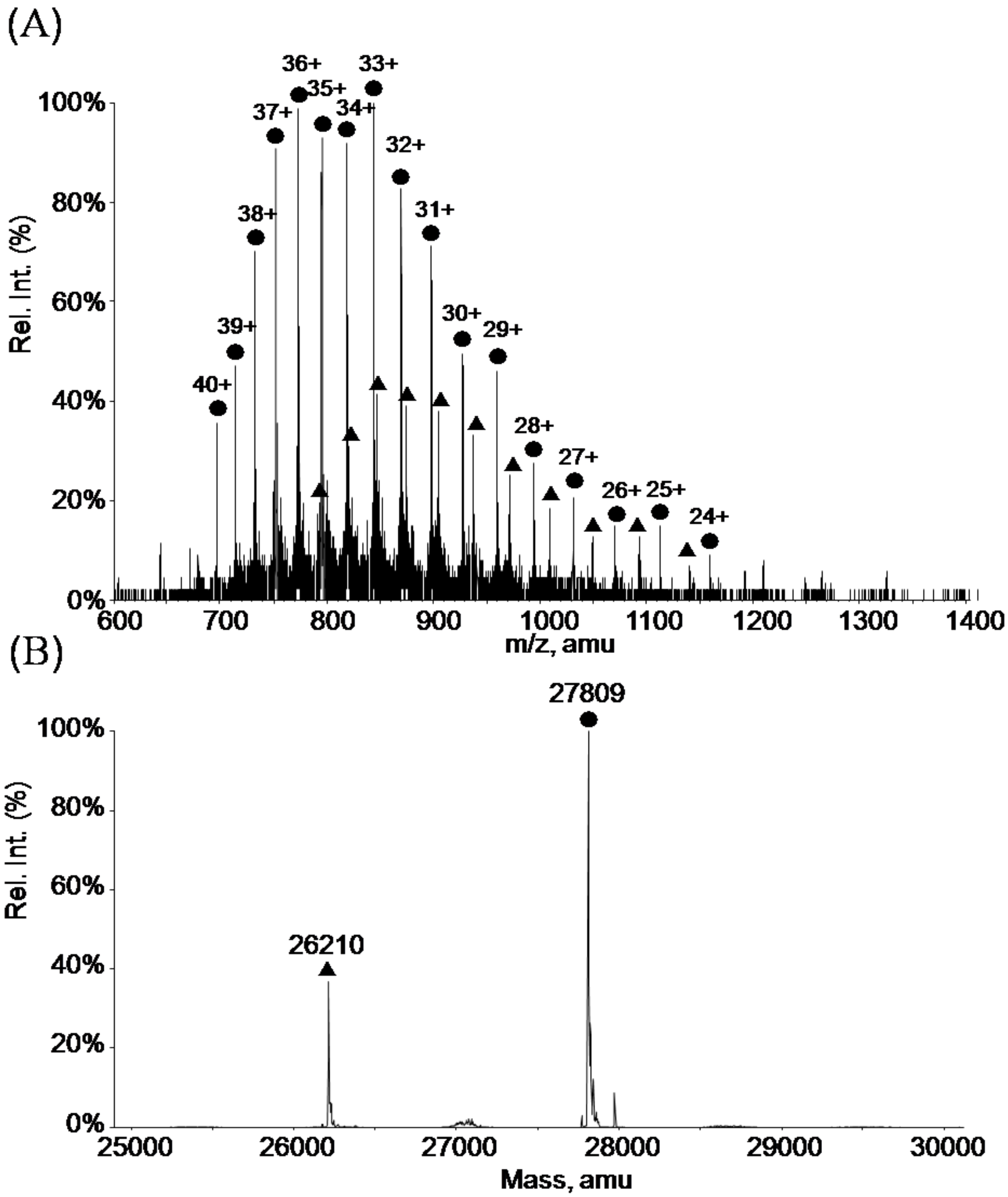


Figure IV-7. Mass characterization of **GFP-4**. (**A**) The ESI-MS spectrum and (**B**) the deconvoluted ESI-MS spectrum.

IFRS1 might also acylate $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ with *p*-bromo-L- phenylalanine. When the IFRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$pair was used to suppress the amber mutation at Q204 of $\mathrm{GFP}_{\mathrm{UV}}$ in the presence of 5 mM *p*-bromo-L-phenylalanine, the full-length $\mathrm{GFP}_{\mathrm{UV}}$ was also produced. The expression yield in GMML medium was ~0.8 mg/L. The ESI-MS detected mass (27,809 Da) of the purified protein agreed well with the calculated mass (27,808 Da) of the full-length $\mathrm{GFP}_{\mathrm{UV}}$ with *p*-bromo-L-phenylalanine incorporated at Q204 but without the N-terminal methionine (**Figure IV-7**). The observed small peak at 26,210 Da may due to a protein contaminant. To independently confirm the incorporation of *p*-iodo-L-phenylalanine and *p*-bromo-L-phenylalanine at Q204 of the expressed $\mathrm{GFP}_{\mathrm{UV}}$ proteins, the purified proteins were digested by endoproteinase Asp-N. The digested fragments were then subjected to the MS-MS analysis. As **Figure IV-5C** shows, the tandem mass spectrum of the *p*-iodo-L-phenylalanine-containing fragment of DNHYLSTF*SALSK (F* denotes *p*-iodo-L-phenylalanine) validated the incorporation of L-phenylalanine at Q204. The observed F*-containing ions (y_6 to y_{13} and b_8 to b_{13}) all had expected mass. Similarly, the observed F*-containing ions of the *p*-bromo-L-phenylalanine-containing fragment of DNHYLSTF*SALSK (F* denotes *p*-bromo-L-phenylalanine) all had expected mass (**Figure IV-5D**). The mass peaks for two bromine isotopes were clearly observed, supporting the site-specific incorporation of *p*-bromo-L-phenylalanine at Q204 (**Figure IV-8-9**).

Davis et al. recently demonstrated that a palladium complex with 2-amino-4,6-dihydroxypyrimidine could efficiently catalyze the cross-coupling between an arylboronic acid and an aryl iodide that was covalently installed on a protein.[135] We

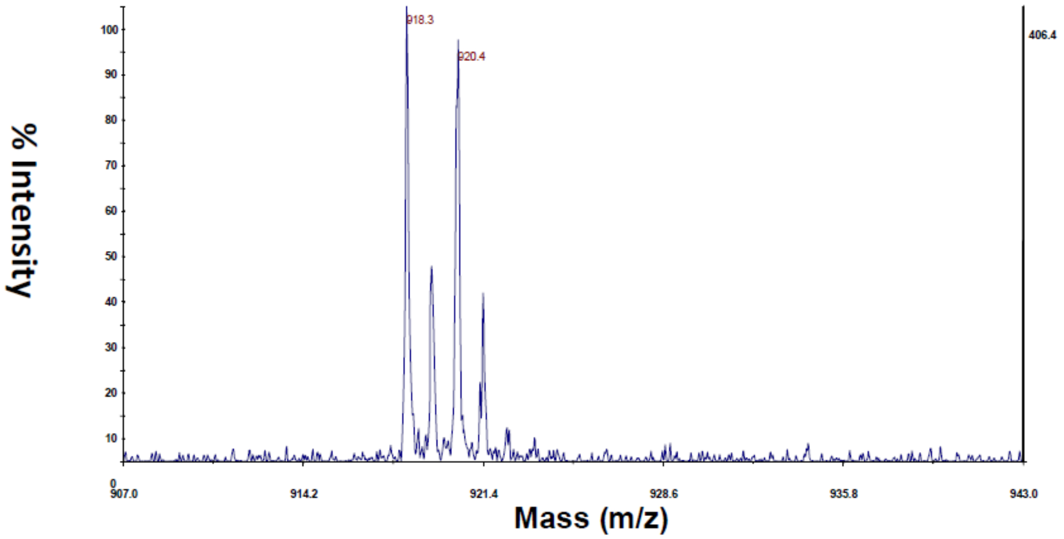


Figure IV-8. Tandem mass spectrametry analysis predicted two bromine isotopes is observed in the y_8 fragment of *p*-bromo-L-phenylalanine-containing DNHYLSTF*SALSK (F* denotes *p*-bromo-L-phenylalanine).

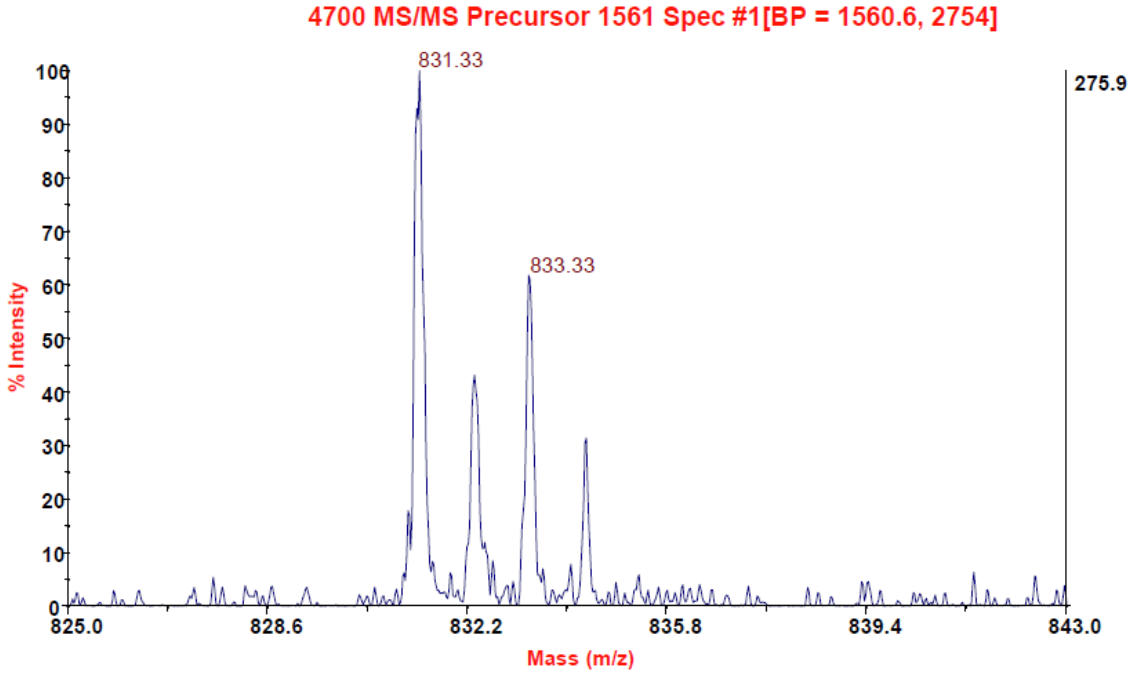


Figure IV-9. Tandem mass spectrametry analysis predicted two bromine isotopes is observed in the y_7 fragment of *p*-bromo-L-phenylalanine-containing DNHYLSTF*SALSK (F* denotes *p*-bromo-L-phenylalanine).

suspected that similar reactions might allow the labeling of proteins containing *p*-iodo-L-phenylalanine. The palladium complex was synthesized according to the literature procedure.[135] This catalyst was subsequently used to catalyze the reaction between the Z-domain protein containing *p*-iodo-L-phenylalanine and a dye, 3-(dansylamino)phenylboronic acid. The incorporation of *p*-iodo-L-phenylalanine at an amber mutation at K7 of Z-domain was achieved in *E. coli* using the IFRS1-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair. The ESI-MS analysis of the expressed Z-domain indicated three forms: the full-length Z domain without the N-terminal methionine, the full-length Z-domain, and the full-length Z-domain without the N-terminal methionine but with an N-terminal acetylation (**Table IV-2** & **Figure IV-10-11**), which have also been observed in other similar studies.[68, 93, 94, 97] This mutant Z-domain protein does not contain any cysteine that can potentially toxify the synthesized palladium catalyst. The 5h reaction between the mutant Z-domain and 3-(dansylamino)phenylboronic acid yielded a labeled Z-domain protein. When the labeled protein was denatured and analyzed in a SDS-PAGE gel, it showed a strong yellow fluorescence band under long wavelength UV light (365 nm) (**Figure IV-12**). The same labeling reaction between the wild-type Z-domain protein and the dye did not give any detectable yellow fluorescent protein. These results indicate that the labeling reaction was site-specific at *p*-iodo-L-phenylalanine in the mutant Z-domain protein.

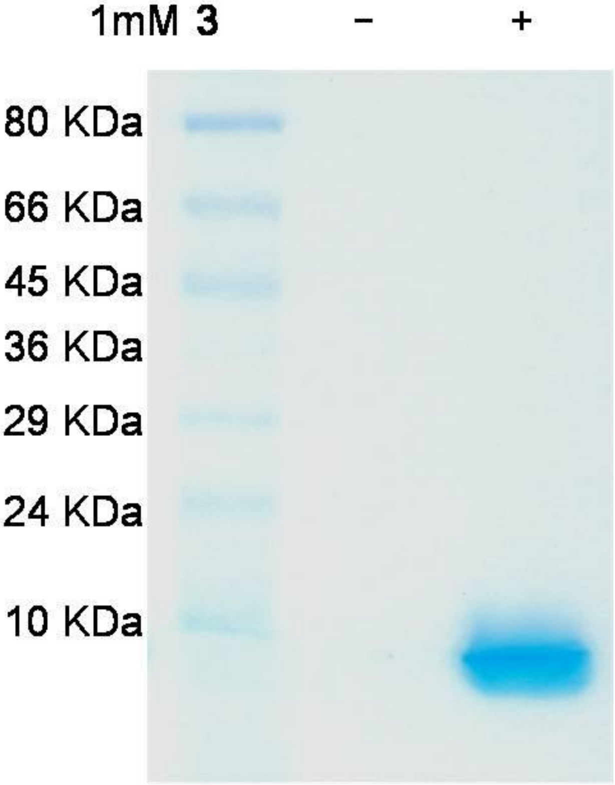


Figure IV-10. The expression of Z-domain containing an amber mutation at K7. Proteins were expressed in BL21(DE3) cells that grew in minimal media supplemented with 1% glycerol and 1 mM **3**. The proteins were analyzed by SDS-PAGE (15%) gel electrophoresis with gelcole blue staining.

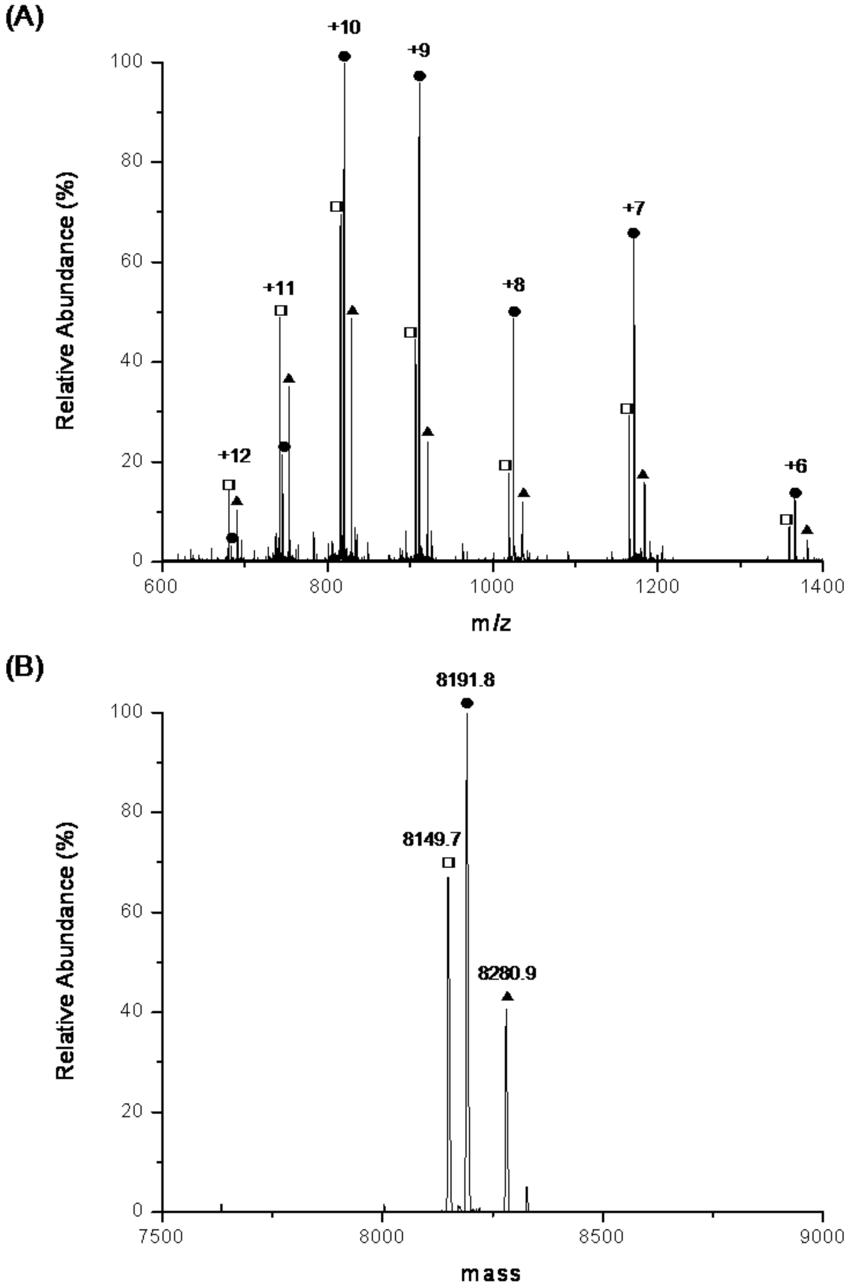


Figure IV-11. Mass characterization of **Z-3**. (**A**) The ESI-MS spectrum and (**B**) the deconvoluted ESI-MS spectrum.

Table IV-2. GFP_{UV} and Z-domain expression yields and MS characterizations

Proteins	Yield (mg/L)[j]	Calculated Mass (Da)	Detected Mass (Da)
GFP-3 [a]	1.0	27855[f]	27855
GFP-4 [b]	0.8	27808[f]	27810
Z-3 [c]	1.5	8281[g]	8280
		8192[h]	8191
		8150[i]	8149
GFP-F[d]	1.6	27729[f]	27730
GFP-F[e]	2.9	27729[f]	27729

[a] GFP_{UV} incorporated with *p*-iodo-L-phenylalanine at Q204.

[b] GFP_{UV} incorporated with *p*-bromo-L-phenylalanine at Q204.

[c] Z-domain incorporated with *p*-iodo-L-phenylalanine at K7.

[d] GFP_{UV} incorporated with L-phenylalanine at Q204 that was expressed in the LB medium using the FRS1- $\text{tRNA}^{\text{Pyl}}_{\text{CUA}}$ pair.

[e] GFP_{UV} incorporated with l-phenylalanine at Q204 that was expressed in the LB medium using the FRS2- $\text{tRNA}^{\text{Pyl}}_{\text{CUA}}$ pair.

[f] Full-legnth GFP_{UV} proteins without N-terminal methionine.

[g] Full-legnth Z domain proteins.

[h] Full-length Z domain without N-terminal methionine but with an N-terminal acetylation.

[i] Full-length Z-domain without N-terminal methionine.

[j] The yields were determined by the BCA protein assay.

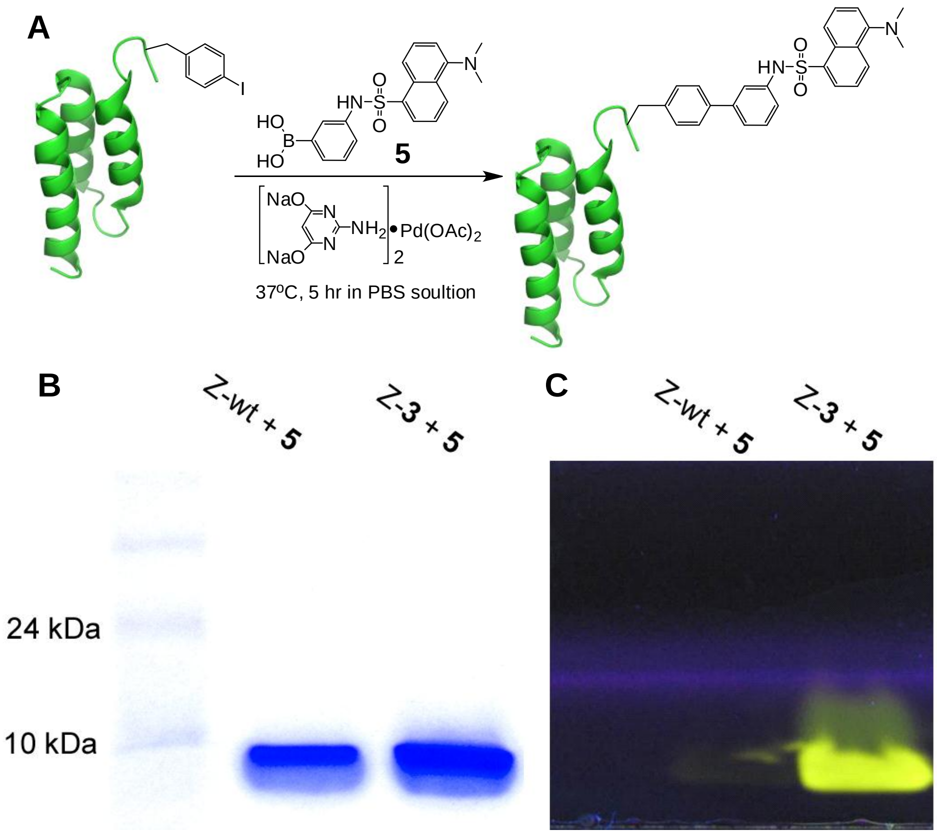


Figure IV-12: Site-selective labeling of Z-domain incorporated with *p*-iodo-L-phenylalanine (Z-**3**). (**A**) The Suzuki-Miyaura cross-coupling reaction between a *p*-iodo-L-phenylalanine-containing protein and 3-(dansylamino)phenylboronic acid (**5**). (**B**) The Gelcode blue staining of Z-**3** and the wild-type Z-domain protein (Z-wt) that were labeled with **5**. (**C**) Thhe fluorescent imaging of the same gel when excited by 365 nm UV light. The fluorescent image shows real colors captured by a regular camera.

Conclusion

In summary, we have demonstrated that PylRS can be mutated to acylated $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ with L-phenylalanine and its derivatives whose side-chain structures are drastically different from pyrrolysine. We expect the structural information obtained from these PylRS mutants will further facilitate the engineering of PylRS for genetic incorporation of other NAAs with short aromatic side chains, including 3-nitro-L-tyrosine, *O*-sulfo-L-tyrosine, 3,4-dihydro-L-phenylalanine, 3-amino-L-tyrosine, *p*-cyano-L-phenylalanine, etc. Given the fact that the PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair has been used to genetically incorporate NAAs into proteins in *E. coli*, yeast, and mammalian cells, this breakthrough will greatly expand the inventory of genetically encoded NAAs and our abilities to do protein engineering in these cells. This development also makes it possible to incorporate two different L-phenylalanine derivatives into one protein in *E. coli* using an evolved *Mj*TyrRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pair and an evolved PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair and has potential applications in enzyme mechanism and protein dynamics analysis. For example, two tyrosine derivatives that have different redox potentials or different pKa's can be incorporated at two catalytic tyrosine sites in the ribonucleotide reductase to clarify the catalytic roles of these tyrosine residues in the enzyme.

CHAPTER V

A RATIONALLY DESIGNED PYRROLYSYL-TRNA SYNTHETASE MUTANT HAS A BROAD SUBSTRATE SPECIFICITY*

Introduction

In the past decade, we have seen the genetic incorporation of many noncanonical amino acids (NAAs) into proteins at nonsense mutation sites in *Escherichia coli*, *Saccharomyces cerevisiae*, and mammalian cells using unique aminoacyl-tRNA synthetase (aaRS)-nonsense suppressing tRNA pairs that do not cross interact with endogenous aaRS-tRNA pairs.[21, 23, 24, 52, 128, 129, 136] Except the wild-type pyrrolysyl-tRNA synthetase (PylRS) that has been directly used for the genetic incorporation >10 NAAs[26, 83, 85, 112, 131], most of these unique aaRSs were evolved via complicated positive and negative selection systems.[21, 23, 92, 98, 137, 138] So far, evolution is still the primary approach to identify NAA-specific aaRSs. The prerequisites of the referred aaRS evolution are the construction of a large mutant aaRS gene library and a readily available selection system.[23] Although several easily accessible selection systems have been developed for *E. coli* and *S. cerevisiae* cells, constructing a large mutant aaRS gene library that most of time needs to cover random variations of 5-6 aaRS active site residues is not straightforward and needs an experienced molecular biologist to practice

*Reprinted in part from “A rationally designed pyrrolysyl-tRNA synthetase mutant with a broad substrate spectrum”, Yane-Shih Wang, Xinqiang Fang, Ashley L. Wallace, Bo Wu, and Wenshe, R. Liu. *J. Am. Chem. Soc.* **2012**, 134, 2950 - 2953. Copyright 2012, with permission from American Chemical Society.

many times to achieve close to the coverage of the library variations.[97] Statistically, full coverage is not likely. One factor that also significantly influences the aaRS evolution is the selection pressure. Positive and negative selections used in the evolution are based on either antibiotic resistance or toxic gene expression. Varying concentrations of antibiotics and inducers that are used to trigger toxic gene expression could lead to dramatic different evolution results. In comparison to the evolution approach, rational design of aaRSs is relatively more straightforward and easier to be carried out by following standardized site-directed mutagenesis protocols. However, attempts to rationally design unique NAA-specific aaRSs often led to mutant aaRSs with nonexclusive recognition of endogenous canonical amino acids. [139, 140 139, 140 138, 139, 140, 89] In this work, we show that a rationally designed PylRS mutant N346A/N348A displays exclusive binding toward NAAs and has a broad substrate spectrum.

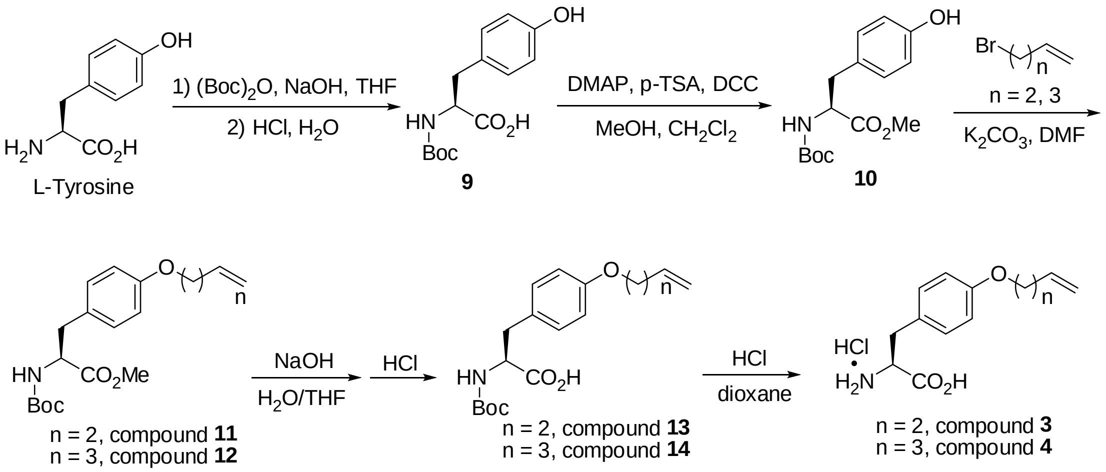
Experimental Sections

Chemical Synthesis

Compounds **1** and **2** were synthesized according to literature procedures. Compounds **3** and **4** were synthesized via the route showed in Scheme 1. Compounds **5-7**, **9-15** are commercial available from Chem-Impex International Inc.

Typically, to a 500 mL of round-bottom-flask was added L-Tyrosine (18.2 g, 0.1 mol), followed by THF (100 mL) and NaOH aqueous (1.0 M, 110 mL, 0.11 mol). The reaction mixture was cooled to 0^{o}C. Di-*tert*-butyl dicarbonate (24.2 g, 0.11 mol) was dissolved in THF (100 mL) and added to the above mixture dropwise over 40 min. Then the resulting mixture was stirred at room temperature for 20 h. THF was removed under reduced pressure. The residue was diluted with water (100 mL), and extracted with EtOAc (100 mL). The aqueous layer was collected and acidified to pH = 3 with HCl (1.0 M in water), then extracted with EtOAc (100 mL × 3). The combined organic layers were dried over anhydrous Na_2SO_4 and concentrated under reduced pressure to afford compound **9** (22.7 g, 81% yield). Compound **9** was pure enough for the next step without further purification.

To a 500 mL of round-bottom-flask was added compound **9** (22.7 g, 0.081 mol), 4-(dimethyl- amino)pyridine (DMAP, 4.9 g, 0.04 mol), *p*-toluenesulfonic acid monohydrate (p-TSA, 7.7g, 0.04 mol) and MeOH (250 mL). *N*,*N'*-dicyclohexylcarbodiimide (DCC, 16.7 g, 0.081 mol) was dissolved in CH_2Cl_2 (50 mL) and added dropwise to the above solution over 1 h. Then the resulting mixture was stirred at room temperature for 20 h. The solvents were removed under reduced pressure



Scheme V-1. Synthesis of compounds **3** and **4**.

and the residue was diluted with EtOAc (150 mL). Then the reaction mixture was filtered through Celite and washed with EtOAc (100 mL × 3). The filtrate was dried over anhydrous Na_2SO_4 and then concentrated under reduced pressure to afford product **10** (18.7 g, 78% yield). Compound **10** was pure enough fot the next step without further purification.

To a 100 mL of round-bottom-flask was added sequentially compound **10** (2.9 g, 0.0098 mol), K_2CO_3 (1.63g, 0.0118 mol), 4-bromo-1-butene (1.19 mL, 0.118 mol) and anhydrous DMF (10 mL). The resulting mixture was stirred at room temperature for 24 h. Then diluted with EtOAc (30 mL), and washed with HCl (1.0 M in water, 10 mL × 3). The organic layer was collected and dried over anhydrous Na_2SO_4 and concentrated under reduced pressure. The residue was purified via column chromatography on silica gel with hexanes/EtOAc (10:1 v/v) as eluent to afford the pure product **11** (1.3g, 38% yield).

Compound **11** (1.3g, 0.00372 mol) was dissolved in THF (5 mL), and NaOH aqueous (1.0 M, 4.8 mL) was added dropwise. The mixture was stirred at room temperature for 2 h, and then diluted with water (10 mL), extracted with Et_2O (20 mL). The aqueous layer was collected, and adjusted to pH = 1 with HCl (3.0 M in water). Then the resulting mixture was extracted with EtOAc (10 mL × 2) and the combined organic layers were dried over anhydrous Na_2SO_4 and concentrated under reduced pressure to get compound **13**, which was used directly in the next step without further purification.

Compound **13** obtained from the above step was all dissolved in anhydrous dioxane (5 mL), and followed by an injection of HCl/dioxane solution (4.0 M, 1.86 mL). The resulting mixture was stirred at room temperature for 12 h. The white solid was collected by filtration and washed sequentially with EtOAc (5 mL) and CH_2Cl_2 (5 mL), then dried under vacuum using oil pump to afford the final product **3** as a white solid (0.5g, 47% over two steps). ^{1}H NMR (CD_3OD, 300 MHz) 2.35-2.41 (m, 2 H), 2.99 (dd, 1H, J = 7.5, 14.7 Hz), 3.12 (dd, 1H, J = 5.4, 14.7 Hz), 3.88 (t, 2H, J = 6.6 Hz), 4.08 (dd, 1H, J = 5.4, 7.5 Hz), 4.92-5.06 (m, 2H), 5.75-5.84 (m, 1H), 6.78 (d, 2H, J = 8.7 Hz), 7.09 (d, 2H, J = 8.7 Hz); ^{13}C NMR (CD_3OD, 75 MHz) 34.7, 36.4, 55.2, 68.4, 116.1, 117.2, 127.3, 131.6, 135.9, 160.0, 171.3; HRMS (ESI) calcd for $C_{13}H_{18}NO_3$ (M^+-Cl) 236.1287, found 236.1282.

Compound **4** was synthesized using the similar procedure showed above, except that 5-bromo-1-butene was used instead of 4-bromo-1-butene. Product **4** was gotten as a white solid. ^{1}H NMR (CD_3OD, 300 MHz) 1.59-1.68 (m, 2H), 1.98-2.05 (m, 2H), 1.98-2.05 (m, 2H), 2.93 (dd, 1H, J = 7.5, 14.7 Hz), 3.05 (dd, 1H, J = 5.7, 14.7 Hz), 3.76 (t, 2H, J = 6.3 Hz), 4.00 (dd, 1H, J = 5.7, 7.5 Hz), 4.75-4.81 (m, 2H), 5.59-5.72 (m, 1H), 6.70 (d, 2H, J = 8.7 Hz), 7.02 (d, 2H, J = 8.7 Hz); ^{13}C NMR (CD_3OD, 75 MHz) 29.6, 31.2, 36.4, 55.2, 68.2, 115.6, 116.0, 127.1, 131.6, 139.1, 160.1, 171.2; HRMS (ESI) calcd for $C_{14}H_{20}NO_3$ (M^+-Cl) 250.1443, found 250.1438.

DNA and Protein Sequence

Gene and protein sequences of pylT, *Methanosarcina mazei* PylRS and sfGFP were listed in the DNA sequence of **Experimental Section** in Chapter III.

Methanosarcina mazei **PylRS**:

MDKKPLNTLISATGLWMSRTGTIHKIKHHEVSRSKIYIEMACGDHLVVNNSRSS

RTARALRHHKYRKTCKRCRVSDEDLNKFLTKANEDQTSVKVKVVSAPTRTKK

AMPKSVARAPKPLENTEAAQAQPSGSKFSPAIPVSTQESVSVPASVSTSISSISTG

ATASALVKGNTNPITSMSAPVQASAPALTKSQTDRLEVLLNPKDEISLNSGKPFR

ELESELLSRRKKDLQQIYAEERENYLGKLEREITRFFVDRGFLEIKSPILIPLEYIER

MGIDNDTELSKQIFRVDKNFCLRPMLAPNLYNYLRKLDRALPDPIKIFEIGPCYR

KESDGKEHLEEFTML**N**F**C**QMGSGCTRENLESIITDFLNHLGIDFKIVGDSCMVYG

DTLDVMHGDLELSSAVVGPIPLDREWGIDKPWIGAGFGLERLLKVKHDFKNIKR

AARSESYYNGISTNL

The bold and underline letters indicate the chosen amino acids, N346 and C348 for mutations.

Construction of Plasmids

The construction of pET-pylT-sfGFPS2TAG was described in Experimental Section of Chapter III.

Construction of pBK- PylRS(N346A/C348A)

To construct the pBK-PylRS(N346A/C348A) plasmid with N346A and C348A mutation was introduced by overlap extension PCR from pBK- mmPylRS plasmid. The following pairs of primers were used to generate a PylRS(N346A/C348A) gene: (1) pBK-mmPylRS-NdeI-F (5'-gaatcccatatggataaaaaaccactaaacactctg-3') and PylRS-N346A/C348A-R -R (5'- cccgtgtgcatcccgatcccatctgcgcgaacgccagcatggtaaactcttcgaggtg -

3'); (2) PylRS-N346A/C348A-F (5'- cacctcgaagagtttaccatgctggcgttcgcgcagatgggatc gggatgcacacggg -3') and pBK-mmPylRS-PstI~NsiI-R (5'-gtttgaaaatgcatttacaggt tggtagaaatccc-3'). The gene PylRS(N346A/C348A) was digested with the restriction enzymes NdeI and NsiI, gel-purified, and ligated back into the pBK vector digested by NdeI and PstI to afford plasmid pBK- PylRS(N346A/C348A).

Construction of pEVOL-pylT-PylRS(N346A/C348A)

The pEVOL-pylT-PylRS(N346A/C348A) plasmid was derived from the pEVOL-pylT plasmid with sequential insertion of two copies of the PylRS-N346A/C348A gene. The PylRS-N346A/C348A gene was amplified from the pBK-PylRS-N346A/C348A plasmid by flanking primers, pEVOL-PylRS-SpeI-F and pEVOL-PylRS-SalI-R, digested by SpeI and SalI restriction enzymes, and ligated to a precut pEVOL-pylT plasmid. The resulted plasmid is pEVOL-pylT-PylRS-N346A/C348A.

sfGFP Expression and Purification and Fluorescence Intensity Test

sfGFP proteins incorporated with **2-6, 8-15** in 2 mM and **7** in 1 mM were expressed and purified similarly as the method of $\mathrm{GFP}_{\mathrm{UV}}$ in Protein Expression and Purification of Experimental Section in Chapter II. To express sfGFP incorporated with a natural amino acid (AA), *E. Coli* BL21(DE3) cells were cotransformed with pEVOL-pylT-PylRS(N346A/C348A) and pET-sfGFP2TAG. Cells were recovered in 1 mL of LB medium for 1 h at 37 ºC before being plated on LB agar plate containing chloramphenicol (Cm) (34 μg/mL) and ampicillin (Amp) (100 μg/mL). A single colony was then selected and grown overnight in a 10 mL culture. This overnight culture was used to inoculate 500 mL of GmmL minimal medium supplemented with 34 μg/mL Cm and 100 μg/mL Amp. Cells were grown at 37 ºC in an incubator (300 r.p.m.) and protein expression was induced when OD_{600} reached 1.0 by adding 1 mM IPTG, 0.2% arabinose and 5 mM AAs for 6 h. Cell lysate was purified by 1 mL bench-top Ni^{2+}-NTA column and detected the fluorescence emission intensity at 510 nm by excitation wavelength 450 nm.

*sfGFP-**1** Protein Labeling*

To a solution of sfGFP-**1** (0.03 mM, 270 μL) in PBS (pH 7.4) buffer was added $CuSO_4$ (100 μM), $NiCl_2$ (1 mM), tris[(1-benzyl-1H-1,2,3-triazol-4-yl)methyl]amine (TBTA, in DMSO, 500 μM) and **8** (50 equiv. to the protein) sequentially, followed by sodium ascorbate (5 mM). The reaction was performed under room temperature for 3 h. Then ethylenediaminetetraacetic acid (EDTA, 0.5 M, pH 8.0, 5 μL) was added to the reaction mixture to chelate the two metals. The reaction product was added into lysis buffer (50 mM HEPES, 500 mM NaCl, 10 mM imidazole, 10 mL, pH 7.8) with 1 mL Ni^{2+}-NTA resin and incubated at 4 ^{o}C for 1 h. The resin was washed by lysis buffer (100 mL), and then labeled sfGFP-**1** was eluted out by elution buffer (50 mM HEPES, 500 mM NaCl, 250 mM imidazole, 6 mL, pH 7.8), concentrated, dialysed against PBS buffer (pH 7.4). Wild type sfGFP was used as a control and the same reaction was performed on it with the same manner. The two product were then analyzed by the SDS-PAGE (12%) electrophoresis for fluorescence image (BioRad^{TM} ChemiDoc XRS+) and further coomassie blue staining.

Results and Discussion

We previously evolved two PylRS mutants that display specific recognition of phenylalanine.[141] In both mutants, N346 is mutated to alanine and C348 is mutated to a larger amino acid, leucine or lysine. No mutation at other sites was found. **Figure V-1** shows the structure of the PylRS complex with pyrrolysyl-AMP. Phenylalanyl-AMP was also modeled into the active site of the N346A mutant of PylRS and is shown as an overlay with pyrrolysyl-AMP in **Figure V-1**. The structure in **Figure V-1** clearly explains how two mutations N346A and C348L (or C348K) instigate the substrate specificity change from pyrrolysine to phenylalanine. The side-chain amide nitrogen of N346 forms a hydrogen bond with the side-chain amide oxygen of pyrrolysine, an interaction that anchors pyrrolysine at the active site. The amide of N346 also has a steric clash with the modeled phenylalanine in the active site and excludes its binding to the wild-type PylRS. The N346A mutation not only significantly decreases the binding of PylRS to pyrrolysine and also relieves the steric hindrance that prevents the binding of phenylalanine. Since both the aromatic side chain of Y384 and two backbone amides of residues 419-421 could form π-π stacking interactions with the phenyl group of a bound phenylalanine, we think that diminish the steric hindrance between a bound phenylalanine and N346 is the major contributing factor for the direct binding of two PylRS mutants to phenylalanine. The mutation of C348 to an amino acid with a larger side chain that apparently occupies the space of pyrrolidine of pyrrolysine in the active site simply brings more van der Waals interactions with the bound phenylalanine and increases its binding potential. Since the N346A mutation relieves the steric hindrance

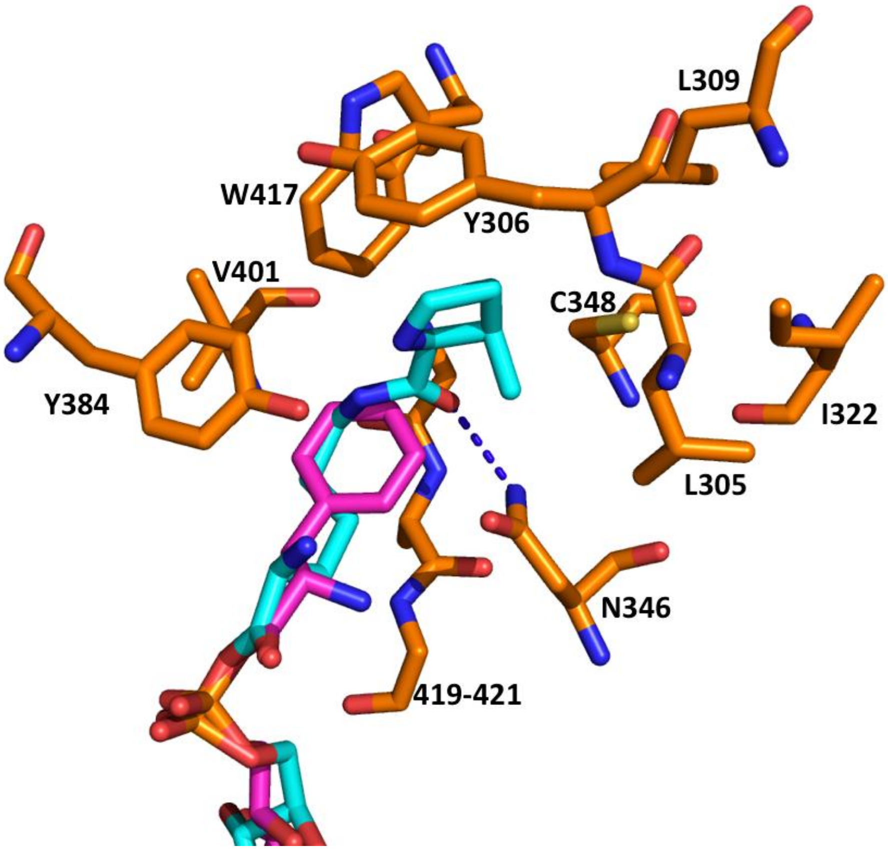


Figure V-1. Structure of the PylRS complex with pyrrolysyl-AMP. Phenylalanyl-AMP that potentially binds the N346A mutant is shown as an overlay with pyrrolysyl-AMP. The structure is based on the PDB entry: 2Q7H[142] that prevents the binding of phenylalanine and there is left a large empty space at the active site around the *para* position of phenylalanine when it binds to the N346A mutant of PylRS, we speculate that the N346A mutant could bind a phenylalanine derivate with a large *para* substituent better than phenylalanine itself. Given that mutations at C348 were prevalent in almost all evolved NAA-specific PylRS mutants and its mutation to alanine was observed for PylRS mutants specific for NAAs with large side chains,[68, 96, 130, 143] we think an extra C348A mutation to the N346A mutant will generate a much larger active site pocket that provides a *para*-substituted phenylalanine with more structural flexibility to bind to the active site.

To test the idea of large binding pocket, we constructed a plasmid pBK-N346A/C348A that carries the gene coding the PylRS mutant N346A/C348A and used it together with pET-pylT-sfGFP2TAG to transform BL21 cells. The plasmid pET-pylT-sfGFP2TAG carries genes coding $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ and an IPTG-inducible superfolder green fluorescent protein (sfGFP) with an amber mutation at the S2 position. The transformed BL21 cells were then used to examine the recognition of N346A/C348A toward twenty natural amino acids by expressing sfGFP in liquid glycerol minimal media (GMML) supplemented with 5 mM of a designated canonical amino acid. The sfGFP expression was induced by the addtion of 1 mM IPTG. Since the sfGFP expression levels at all twenty conditions were very low, we chose to detect the fluorescent emission of the expressed sfGFP at these conditions and show their relative intensities to represent the corresponding sfGFP expression levels in **Figure V-2**. The condition that contained 5 mM phenylalanine displayed the highest sfGFP expression level, confirming our initial

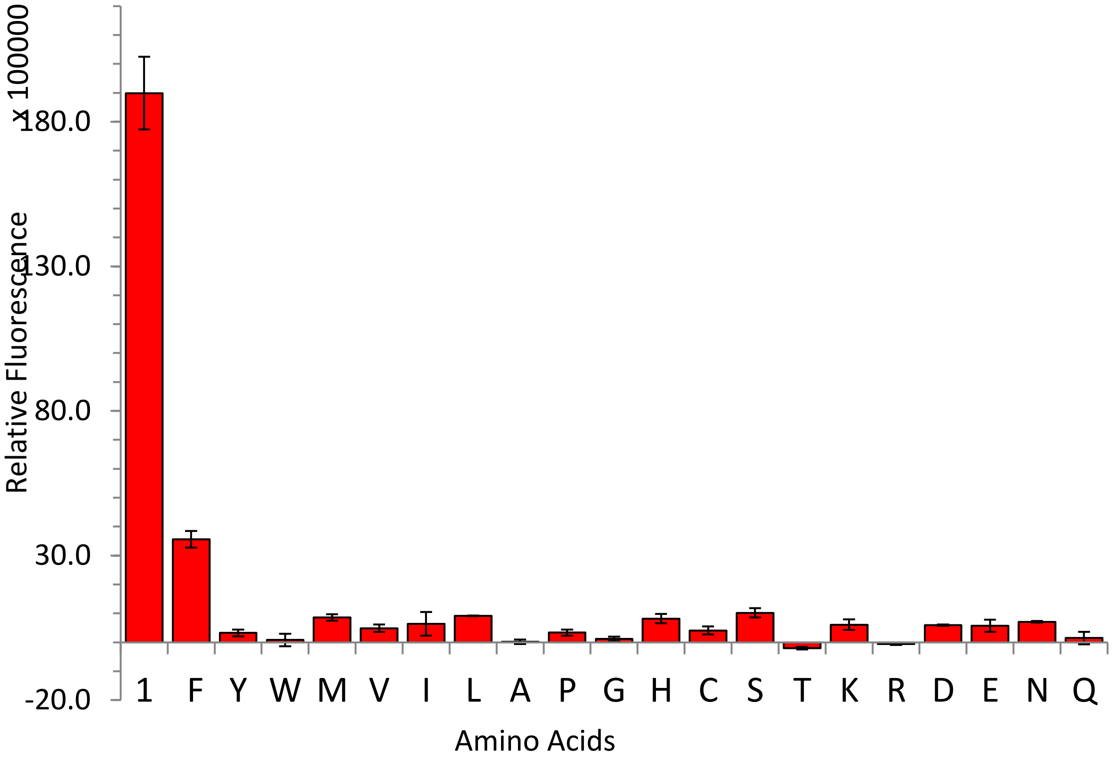


Figure V-2. Relative fluorescence emission intensities of sfGFP expressed in BL21 cells transformed with pBK-N346A/C348A and pET-pylTsfGFP2TAG and grown in GMML supplemented with different amino acids. Cell lysates were excited at 450 nm and fluorescence emission intensities were detected at 510 nm. Background emission from cell lysate of same cells grown in GMML and induced with the addition of 1 mM IPTG was subtracted from each data set. Twenty CAAs are shown as one-letter abbreviations in the x-axis labels.

speculation that the N346A mutation will result in stronger binding of PylRS toward phenylalanine. Although stronger than the wild-type PylRS, the binding of phenylalanine to N346A/C348A is still relatively low. This low binding affinity is not a surprise since phenylalanine is expected less engaged in interactions with N346A/C348A than N346A/C348L and N346A/C348K. These data show that N346A/C348A has low binding affinities toward all twenty canonical amino acids, suggesting its potential application for the genetic incorporation of NAAs. To see whether N346A/C348A can recognize a phenylalanine derivative with a large *para* substituent, the same BL21 cells were let grown in liquid glycerol minimal medium supplemented with 5 mM *p*-propargyloxy phenylalanine (**1** in **Figure V-3**). SfGFP was overexpressed 6 h after induction with 1 mM IPTG.

With our initial success with **1**, we then either purchased or synthesized several other *para*-substituted phenylalanine derivatives shown as **2-7** in **Figure V-3** and tested their recognition by N46A/C348A. To better quantify the expression levels of sfGFP incorporated with these NAAs, another plasmid pEVOL-pylT-N346A/C348A was constructed. This plasmid carries genes coding both $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ and N346A/C348A. Its $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ is under control of a strong *proK* promoter that can boost up the expression level of $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ in *E. coli*.[119] Together with pET-pylT-sfGFP2TAG, this plasmid was used to transform BL21 cells. The transformed cells were grown in GMML supplemented with 2 mM of a designated NAA. As shown in **Figure V-3**, all seven NAAs promoted sfGFP overexpression 10 h after induction with 1 mM IPTG. The molecular weights of the expressed sfGFP variants determined by electrospray

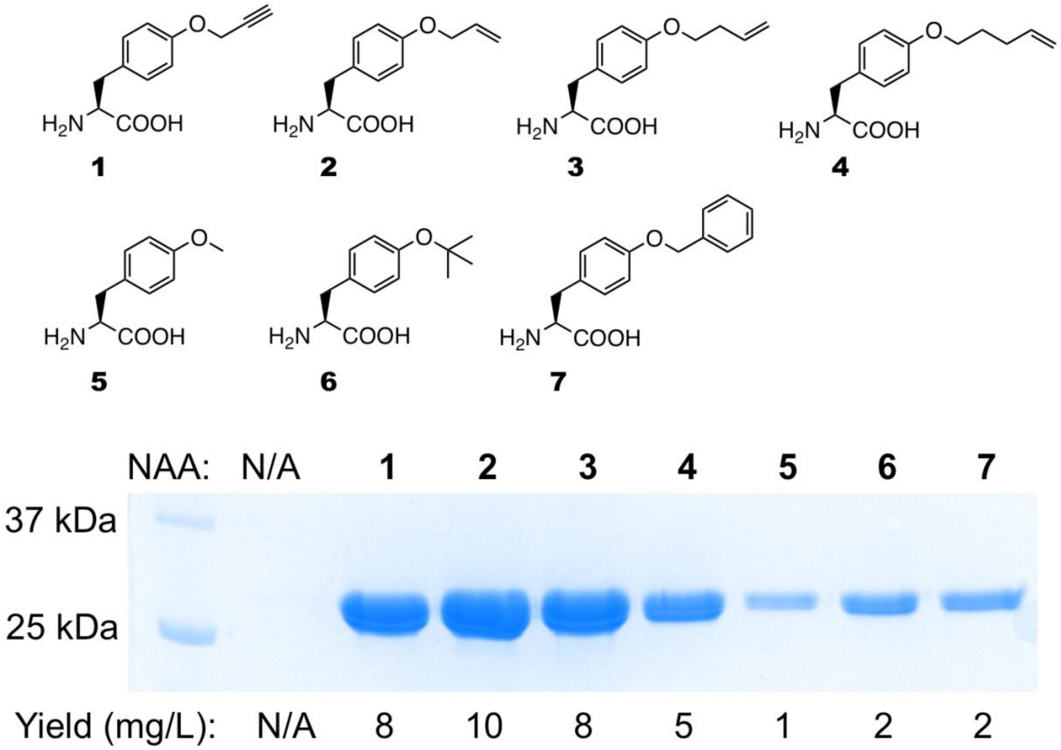


Figure V-3. Structures of **1**-**7** and their specific incorporation into sfGFP at the S2 position.

Table V-1.sfGFP protein molecular weight determination by ESI-MS.

Protein	Calculated Molecular Weight (Da)	Found Molecular Weight (Da)
sfGFP-1	27843	27845
sfGFP-2	27845	27845
sfGFP-3	27858	27857
sfGFP-4	27872	27871
sfGFP-5	27818	27818
sfGFP-6	27802	27803
sfGFP-7	27860	27861
sfGFP-8	27894	27894

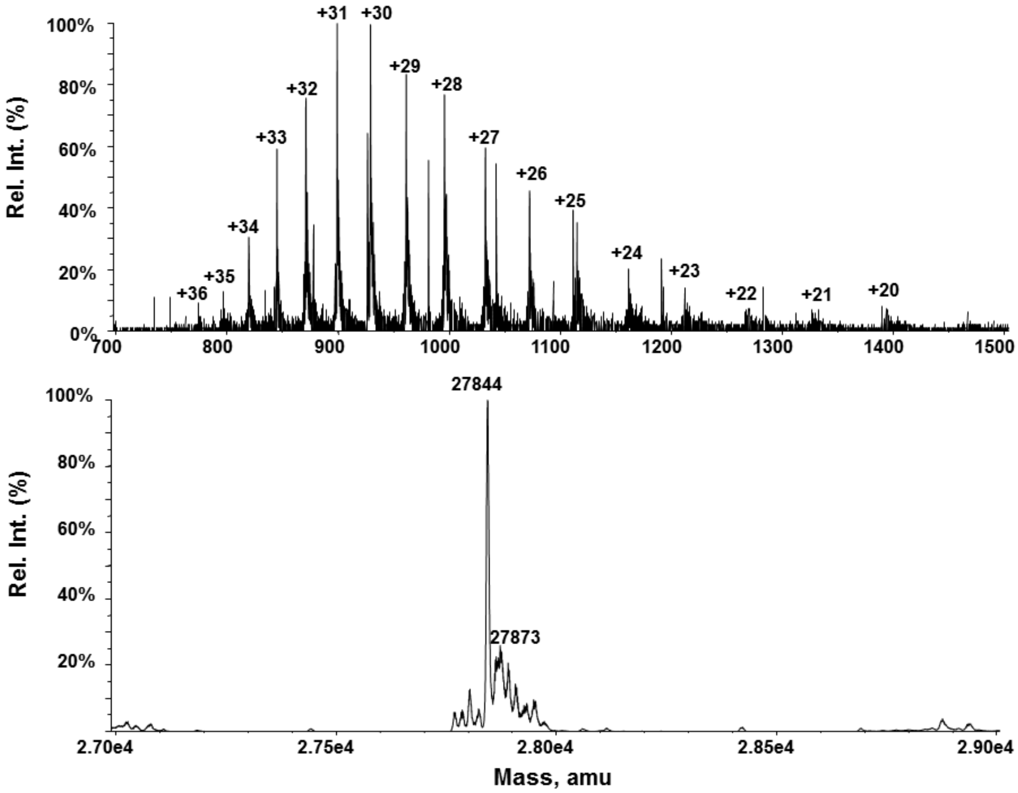


Figure V-4. Molecular weight determination of the protein sfGFP-**1**: (A) ESI-MS spectrum of sfGFP-**1** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**1**.

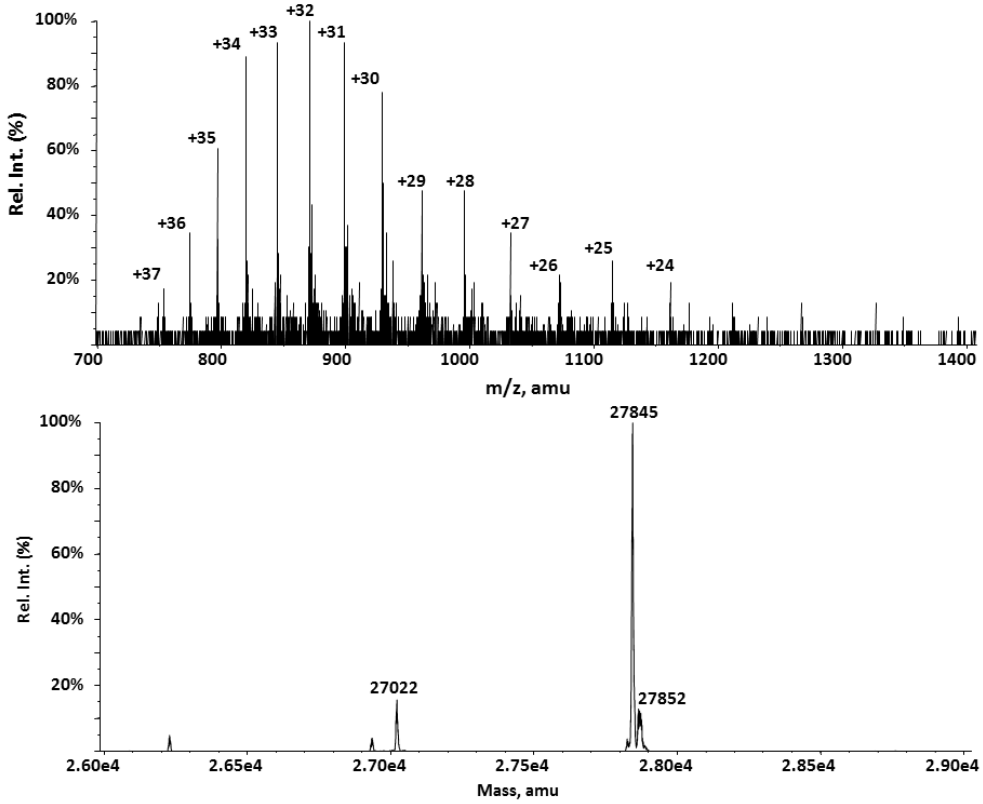


Figure V-5. Molecular weight determination of the protein sfGFP-**2**: (A) ESI-MS spectrum of sfGFP-**2** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**2**.

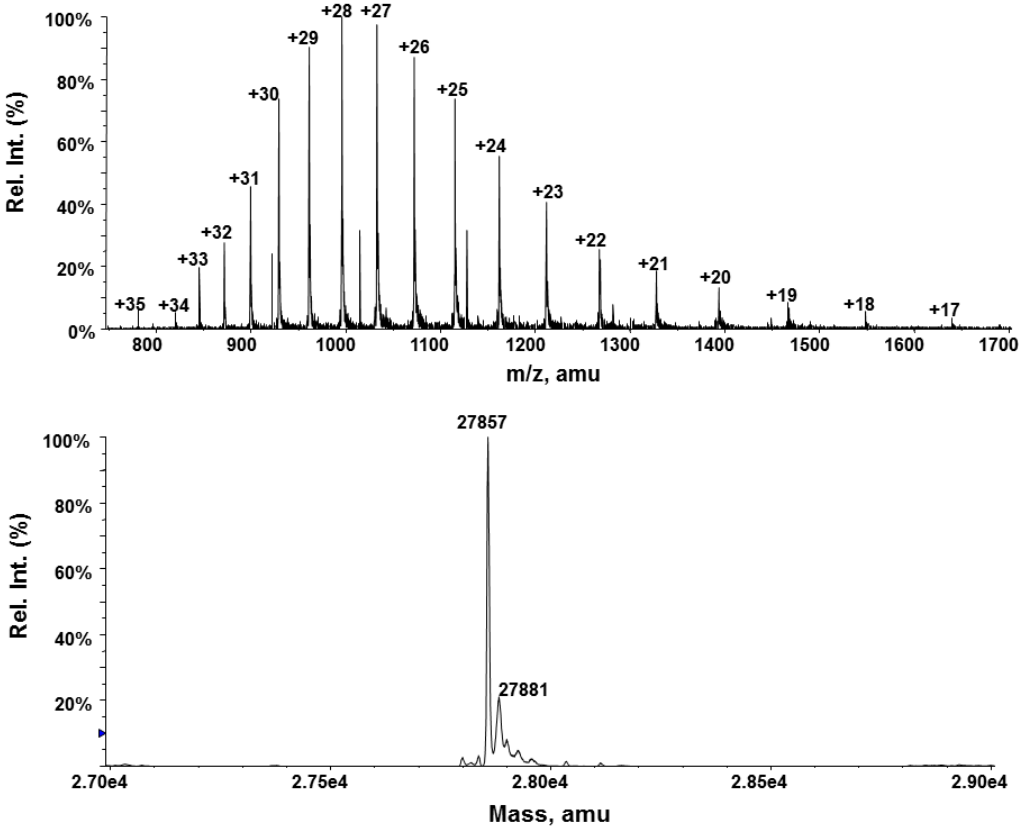


Figure V-6. Molecular weight determination of the protein sfGFP-**3**: (A) ESI-MS spectrum of sfGFP-**3** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**3**.

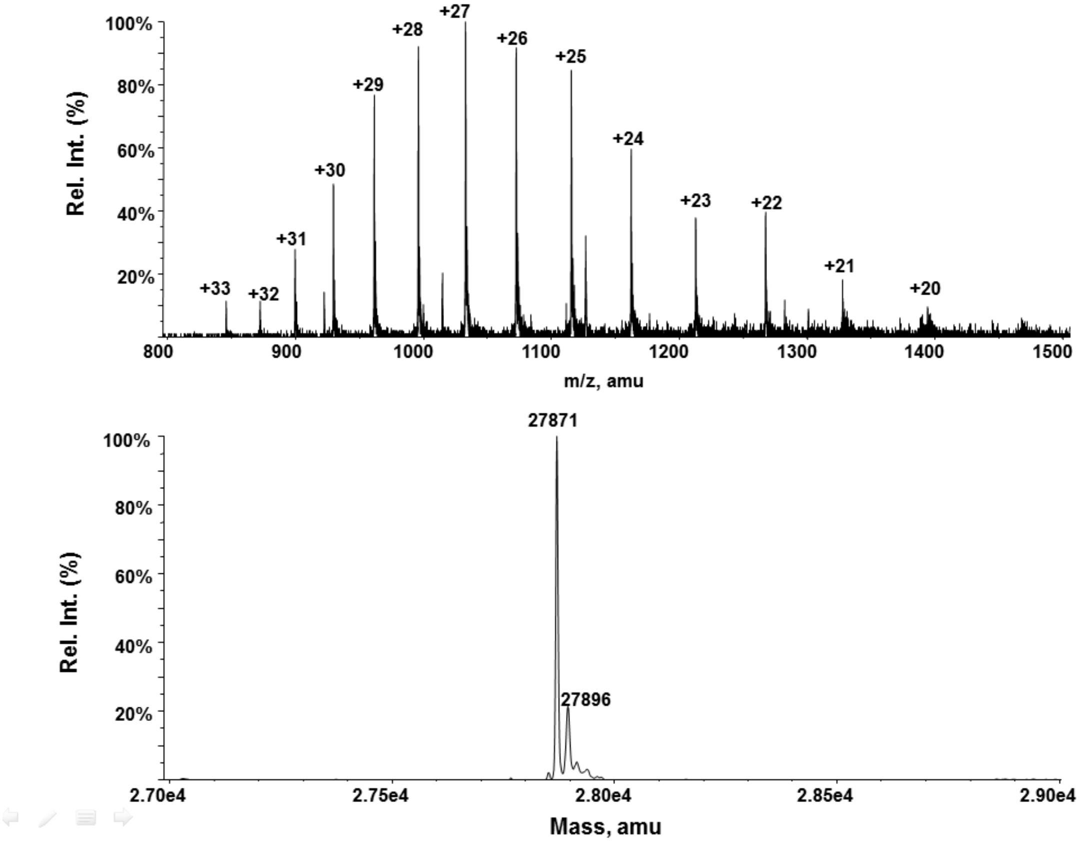


Figure V-7. Molecular weight determination of the protein sfGFP-**4**: (A) ESI-MS spectrum of sfGFP-**4** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**4**.

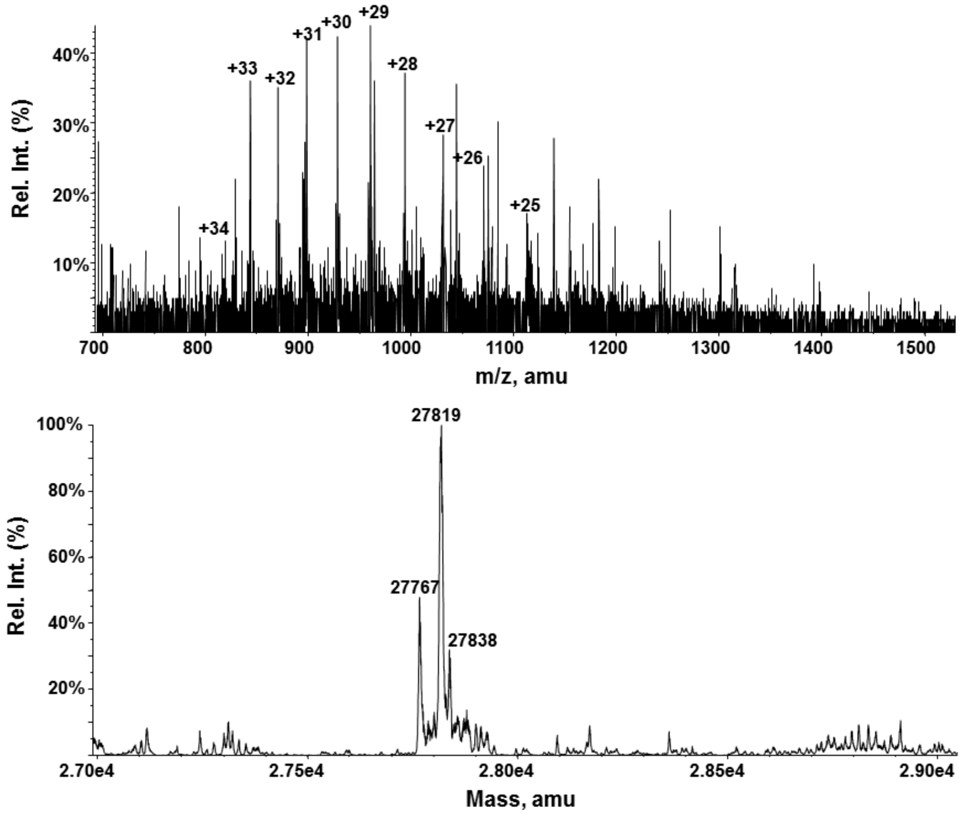


Figure V-8. Molecular weight determination of the protein sfGFP-**5**: (A) ESI-MS spectrum of sfGFP-**5** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**5**.

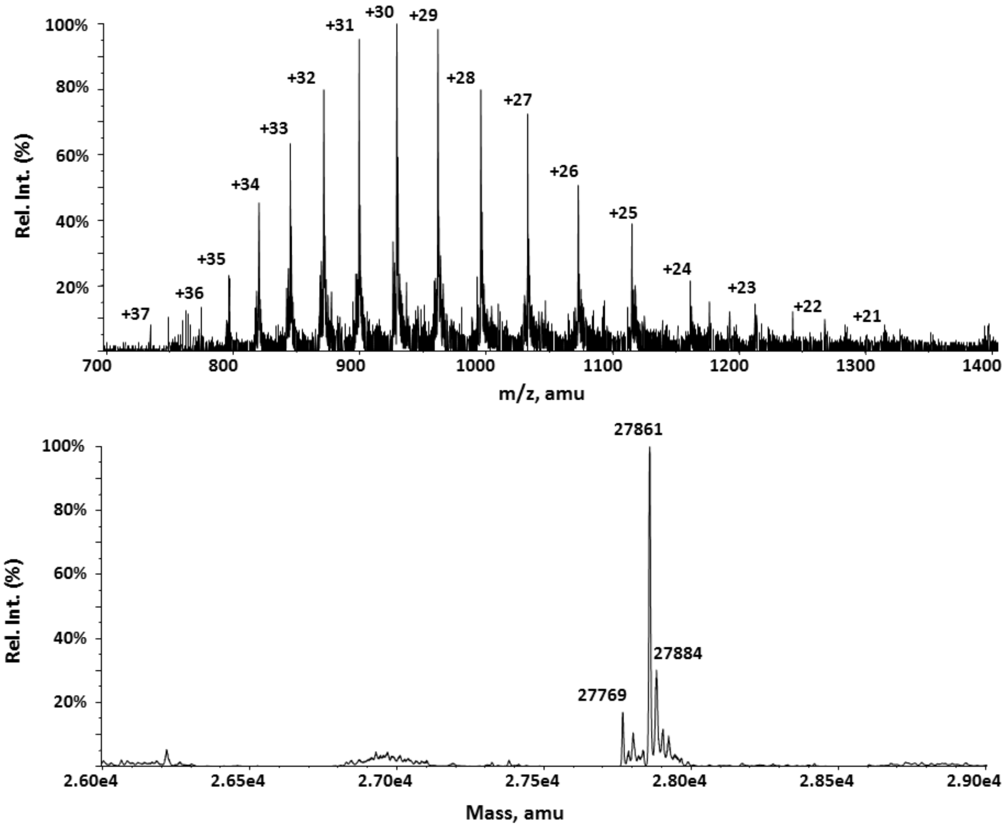


Figure V-9. Molecular weight determination of the protein sfGFP-**6**: (A) ESI-MS spectrum of sfGFP-**6** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**6**.

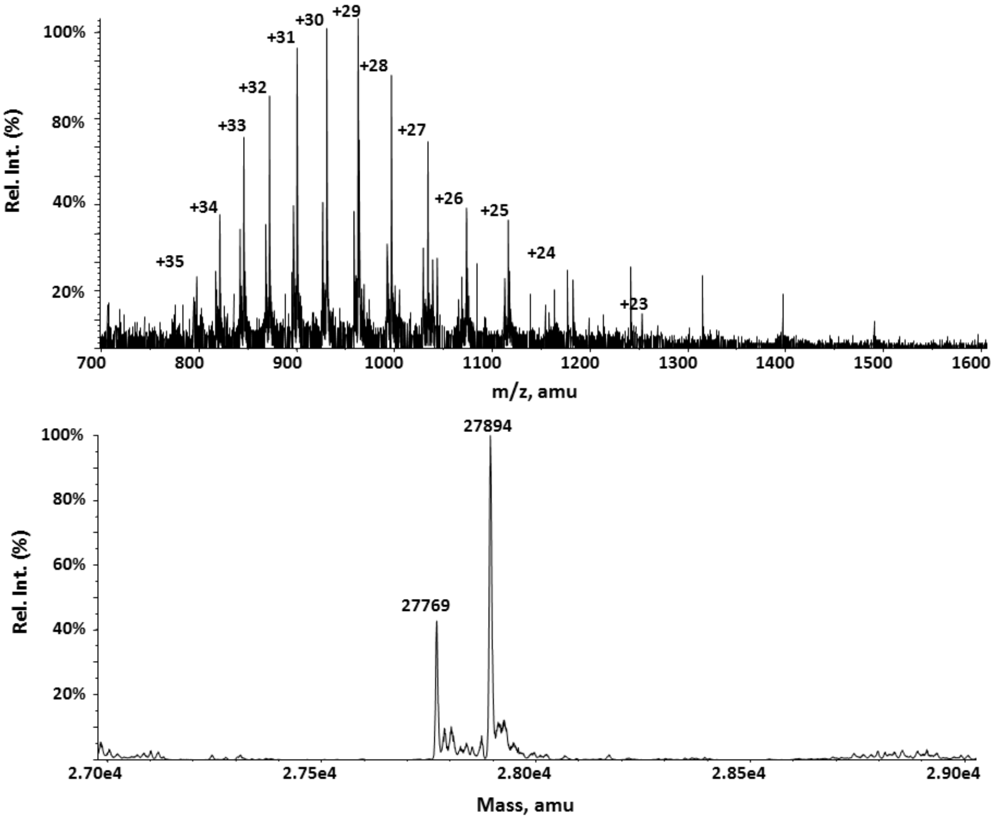


Figure V-10. Molecular weight determination of the protein sfGFP-**7**: (A) ESI-MS spectrum of sfGFP-**7** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**7**.

ionization spectrometry (ESI-MS) analysis agree well with their theoretic molecular weights (**Table V-1**)(**Figure V-4-10**). Without a NAA, no sfGFP was expressed. Since **1** contains a terminal alkyne that undergoes the Cu(I)-catalyzed azide-alkyne cyclization reaction,[144] the expressed sfGFP incorporated with **1** (sfGFP-**1**) was also used separately to label with a fluorescein azide (**8** in **Figure V-11**) in an optimized labeling condition that contained 0.1 mM Cu(I):tris[(1-benzyl-1H-1,2,3-triazol-4-yl)methyl]amine (TBTA) complex, 0.5 mM additional TBTA, 5 mM ascorbate, and 1 mM $NiCl_2$.[145] 3 h incubation led to specific labeling of sfGFP-**1**. A parallel labeling reaction of the wild-type sfGFP in the same condition gave non-detectable labeling with **8**.

One intriguing question is whether N346A/C348A could also recognize a phenylalanine derivative with a small *para*-substituent. Given that **5** has a relatively small *para*-substituent and providing **5** led to reasonable sfGFP expression, one would expect N346A/C348A also recognizes a phenylalanine derivative with a small *para*-substituent such as Cl, Br, I, CN, etc. To test this possibility, we examined the genetic incorporation of **9-15** shown in **Figure V-12** into sfGFP at S2 using the N346A/C348A-$\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair. As shown in **Figure V-12**, providing 2 mM of any of these NAAs in GMML led to sfGFP expression levels that are significantly lower than those for NAAs shown in **Figure V-3** but still higher than the background expression level in GMML in which no NAA was provided. Therefore, N346A/C348A recognizes **9-15** but with relatively lowbinding affinities. Since the expression levels of sfGFP incorporated with **9-15** are low, we did not attempt to characterize these proteins by the ESI-MS analysis. Although the current analysis suggests that it is not applicable to use the

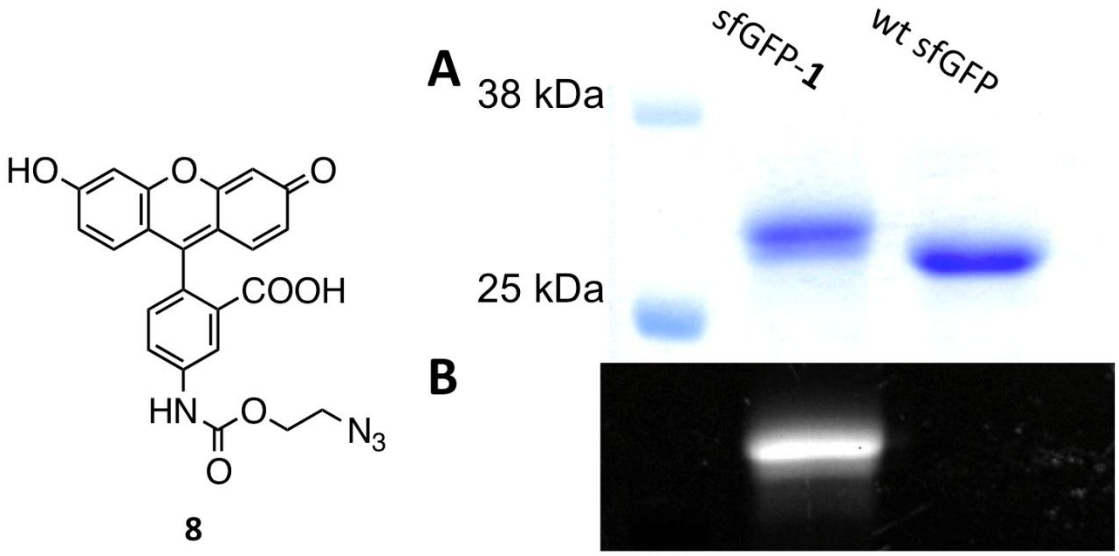


Figure V-11. Site-selective labeling of sfGFP-**1** with **8**. (**A**) SDS-PAGE analysis of sfGFP-**1** and wild-type (wt) sfGFP after their reactions with **8**. The gel was stained with Coomassie blue. **(B**) Fluorescent imaging of the same gel under 365 nm UV irradiation.

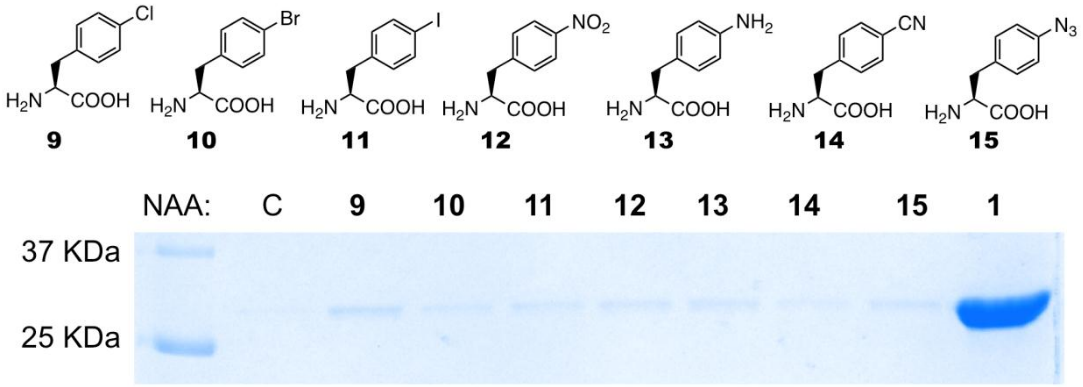


Figure V-12. Structures of **9-14** and their incorporation into sfGFP at S2.The protein expression yields are lower than 1 mg/L for all NAAs.

N346A/C348A-tRNA$^{Pyl}_{CUA}$ pair to express proteins incorporated with **9-15**, it clearly indicates PylRS can be engineered to recognize phenylalanine derivatives with small *para* substitutes. Since a PylRS mutant specific for **5** evolved by Wang and co-workers contains mutations at A302 and V401,[146] we are now introducing additional mutations to N346A/N348A to generate PylRS mutants that show high binding affinities toward phenylalanine derivatives with small *para* substituents.

Conclusion

In summary, we have rationally designed a PylRS mutant N346A/C348A that shows very low recognition toward canonical amino acids but, together with tRNA$^{Pyl}_{CUA}$, mediates efficient incorporation of NAAs **1-7** into proteins at amber mutation sites in *E. coli*. These NAAs contain functional groups such as alkyne and alkene and can be applied to install different biochemical and biophysical probes to proteins for their structural and functional analysis. Since the PylRS-tRNA$^{Pyl}_{CUA}$ pair has been successfully introduced into *S. cerevisiae*, mammalian cells, and even the multiple cellular organisms,[85, 96, 130, 147, 148] the N346A/C348A-tRNA$^{Pyl}_{CUA}$ pair could be potentially used in these systems to genetically encode **1**-**7**. Although many phenylalanine derivatives have been incorporated into proteins in *E. coli* using evolved *Methanococcus jannaschii* tyrosyl-tRNA synthetase (MjTyrRS)-tRNA$^{Tyr}_{CUA}$ pairs,[127, 149, 150] specifically evolved MjTyrRS variants for individual phenylalanine derivatives are usually required and the MjTyrRS-tRNA$^{Tyr}_{CUA}$ pair cannot be used in eukaryotic cells because of the recognition of

$\mathrm{tRNA}^{\mathrm{Tyr}}_{\mathrm{CUA}}$ by endogenous eukaryotic aaRSs (Liu & Schultz, unpublished data). Using the N346A/C348A-$\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ pair will resolve both issues. In addition, phenylalanine derivatives **3**, **4**, **6** and **7** that are taken by N346A/C348A are also genetically encoded in *E. coli* for the first time. Given that N346A/C348A has a relatively deep and big binding pocket, the current study also opens a gate to test the recognition of this mutant toward other large phenylalanine derivatives. Another potential application of N346A/C348A is to couple its pair with $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{UUA}}$ together with evolved MjTyrRS-$\mathrm{tRNA}^{\mathrm{Tyr}}_{\mathrm{CUA}}$ pairs for the genetic incorporation of two different phenylalanine derivatives into one protein.[88] This may find applications in enzyme mechanistic studies, protein FRET labeling, and phage displayed unnatural peptide library construction.

CHAPTER VI

A PYRROLYSYL-TRNA SYNTHETASE MUTANT RECOGNIZES META-SUBSTITUTED PHENYLALANINE DERIVATIVES

Introduction

Using evolved TyrRSs that were originally from *Methanococcus jannaschii* and *Escherichia coli* and their corresponding amber suppressor tRNAs, Schultz and coworkers showed that more than twenty *para*-substituted phenylalanine derivatives could be genetically incorporated into proteins at amber mutation sites in *E. coli*, *Saccharomyces cerevisiae*, and mammalian cells.[21, 23, 29, 43, 44, 92-94, 128, 151-164] These *para*-substituted phenylalanine derivatives contain functional groups that can serve as biophysical probes for structural and functional investigations of proteins and biochemical probes for protein modifications.[24, 127, 165] Compared to *para*-substituted phenylalanine derivatives, genetic incorporation of *meta*-substituted phenylalanine derivatives is far less explored. Zhang *et al.* demonstrated that a *meta*-substituted phenylalanine derivative, *meta*-acetyl-phenylalanine, could be genetically encoded in *E. coli* using an evolved *M j*TyrRS.[166] However, there has been no following study since this work was published in 2003. Developing methods for genetic incorporation of other *meta*-substituted phenylalanine derivatives will not only expand the genetically encoded NAA inventory, leaving more choices for protein engineering when subtle variations are necessary, but also help understand the NAA tolerance scope of the cellular protein translation machinery. In a previous study, we showed that a rationally designed PylRS mutant N346A/C346A together with $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$ mediates genetic incorporation of seven *para*-substituted phenylalanine derivatives into proteins at amber mutation sites in *E. coli*. Here, we reveal that the same enzyme also recognizes seven *meta*-substituted phenylalanine derivatives and when coupled with $\mathrm{tRNA}^{\mathrm{Pyl}}_{\mathrm{CUA}}$, it can be applied to genetically encode these NAAs in *E. coli*.

In our previous study, we show that N346A/C348A has an enhanced recognition of phenylalanine in comparison to the wild-type PylRS. We believe this recognition enhancement is due to the removal of the N346 amide that can potentially prevent the binding of phenylalanine to the active site of the wild-type PylRS. [141, 167-169] [141, 167-169] [140-141, 166-169] **Figure VI-1A** shows a modeled phenylalanyl-AMP at the active site of N346A/C348A. The aromatic side chain of phenylalanine is sandwiched between the Y384 phenol group and the two backbone amides of residues 419-421, forming strong π-π interactions that likely contribute to the recognition of phenylalanine by N346A/C348A. Although N346A/C348A shows an enhanced recognition of phenylalanine in comparison to the wild-type PylRS, its binding toward phenylalanine is still very weak. In addition, there is left a large empty hydrophobic pocket formed by residues A302, Y306, L309, A348 and W417 at the active site after the binding of phenylalanine. After carefully examining this pocket that is shown in **Figure VI-1B** and obviously located around the *meta* position of the bound phenylalanine side chain, we suspected that a phenylalanine derivative with a hydrophobic *meta*-substituent might also show a better binding potential toward N346A/C348A than phenylalanine.

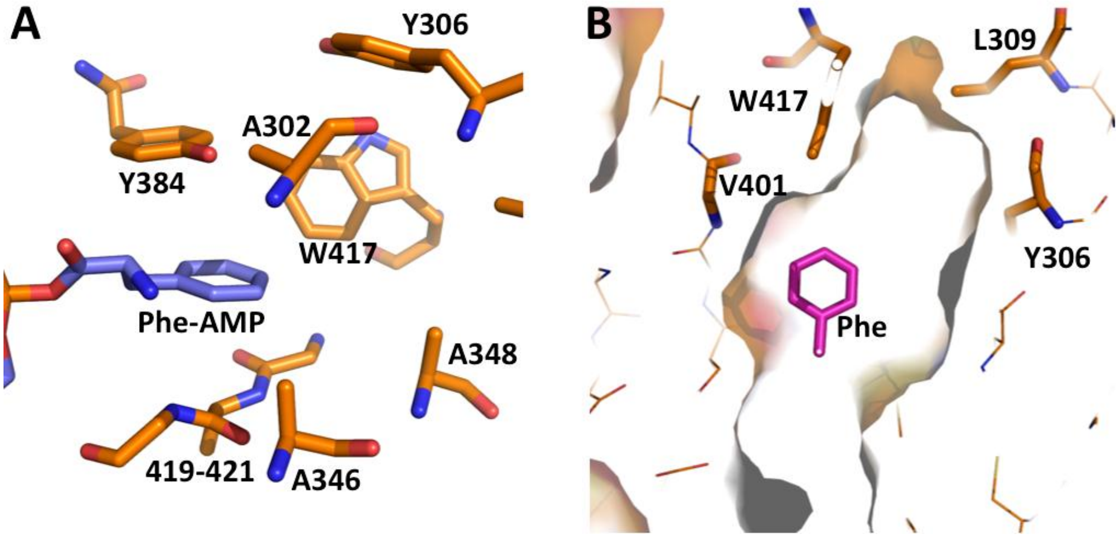


Figure VI-1. (**A**) N346A/C348A with a modeled phenylalanyl-AMP (Phe-AMP) in the active site and (**B**) the active site cavity of N346A/C348A.

Experimental Section

General Experimental

DNA and protein sequences of sfGFP, *mm*PylRS-N346A/C348A, pylT and plasmid constructions of pET-sfGFP2TAG, pEVOL-pylT-PylRS(N346A/C348A) were listed in Chapter III and V. sfGFP proteins incorporated with **1-7** in 2 mM in GmmL or 1 mM in LB were expressed and purified similarly as the method of GFP_{UV} in Protein Expression and Purification of Experimental Section in Chapter II. In the ^{15}N labeling sfGFP27-**6** protein expression, the BL21 (DE3) cells were culture in similar condition. The cells cultured in LB medium were centrifuged when OD_{600} reached 0.7 - 0.8 and replace by GmmL with ^{15}N-labeled ammonium chloride (^{15}N, 99%, Cambridge Isotope Laboratories Inc.). The protein expression was induced by adding 1 mM IPTG, 0.2% arabinose and 1 mM **6** for 6 h. Samples for folding studies by ^{19}F NMR contained 0.7 mM sfGFP27-**6**, 12 mM phosphate buffer (pH 7.0), 140 mM NaCl, 3 mM KCl. Chemical shifts were referenced against an external standard of trifluoroacetic acid at -75.2 ppm in D_2O. ^{19}F spectra were recorded on a 400 MHz spectrometer equipped with a Broadband Observe (BBO) probe (Bruker Biospin).

Construction of pBAD-sfGFP27TAG, pBAD-sfGFP8TAG, pBAD-sfGFP130TAG

The pBAD-sfGFP27TAG plasmid was derived from the pBAD-sfGFP plasmid with quickchange method. Primers pBAD-sfGFP-F27TAG-F (5'-tctgttcgtggtgaaggtgaaggtgatg-3') and pBAD-sfGFP-F27TAG-R (5'-ctatttatgaccattaacatcaccatcaagttc-3') are used for the construction of pBAD-sfGFP27TAG. Primers pBAD-sfGFP-F8TAG-F (5'- actggtgttgttcctattcttgttgaacttg -3') and pBAD-sfGFP-F8TAG-R (5'- ctaaagttcttcacctttagaaaccatggttaattcc -3') are used for pBAD-sfGFP8TAG, as well as primers pBAD-sfGFP-F130TAG-F (5'-aaagaagatggtaatattcttggtcataaacttg-3') and pBAD-sfGFP-F130TAG-R (5'-ctaatcaatacctttaagttcaatacgattaac -3') for pBAD-sfGFP130TAG.

Equilibrium Fluorescence of sfGFP

Equilibrium fluorescence values were measured by adding different concentration GndCl between 1.0 and 7.0 M to folded sfGFP variants in increments of 0.5 to 1 M GndCl, and allowing equilibrium to procee up to 12 to 48 hr in room temperature. Fluorescence values were measured using a FL600 Microplate Fluorescence Reader (450-nm excitation, 510-nm emission, 0.5-nm band pass). Midpoint recovery concentrations of GndCl C_m (recovery of 50% of the initial fluorescence) were determined from sigmoidal fits using SOLVER in EXCEL, to the scaled fluorescence value F using the equation $F_j = ¼\ a + b/(1 + (C_j/C_m)^h)$, where a, b, C_m and h are adjustable parameters, and C_j is the molarity of the GndCl in the unfolding experiment j.

ESI-MS Analysis of Intact Proteins

Nanoelectrospray ionization in positive mode was performed using an Applied Biosystems QSTAR Pulsar (Concord, ON, Canada) equipped with a nanoelectrospray ion source. Solution was flowed at 700 nL/min through a 50 µm ID fused-silica capillary that was tapered at the tip. Electrospray needle voltage was held at 2100 V.

Results and Discussion

To test the possibility of meta-substituted phenylalanine, we employed BL21 cells transformed with two plasmids pEVOL-pylT-N346A/C348A and pET-pylT-sfGFP2TAG. Plasmid pEVOL-pylT-N346A/C348A carried a $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ gene under control of a strong *proK* promoter and a N346A/C348A gene under control of a *pBAD* promoter; plasmid pET-pylT-sfGFP2TAG carried a $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ gene under control of a *lpp* promoter and a superfolder green fluorescent protein (sfGFP) gene with an amber mutation at its S2 position under control of an IPTG-inducible T7 promoter. Since N346A/C348A has low binding affinities toward twenty canonical amino acids (CAAs), growing the transformed cells in liquid glycerol minimal medium (GMML) and inducing with the addition of 1 mM IPTG and 0.2% arabinose led to no detectable sfGFP due to the premature translation termination at the S2 position. However, providing 2 mM *meta*-chloro-phenylalanine, *meta*-bromo-phenylalanine, or *meta*-iodo-phenylalanine (**2**, **3**, and **4** in **Figure VI-2**) in the medium promoted sfGFP overexpression (**Figure VI-2**). All three purified sfGFP variants had detected molecular weights determined by the electrospray ionization mass spectrometry (ESI-MS) analysis that agreed well with their theoretical molecular weights (**Table VI-1** and **Figures VI-3-6**). We also tested the uptake of *meta*-fluoro-phenylalanine (**1** in **Figure VI-2**) by N346A/C348A. Although providing **1** in GMML promoted sfGFP expression in BL21 cells transformed with pEVOL-pylT-N346A/C348A and pET-pylT-sfGFP2TAG, the yield was significantly lower than for **2-4**. Given that **1** could be misincorporated at phenylalanine sites, which leads to disruption of cellular function, we think this low sfGFP expression yield is

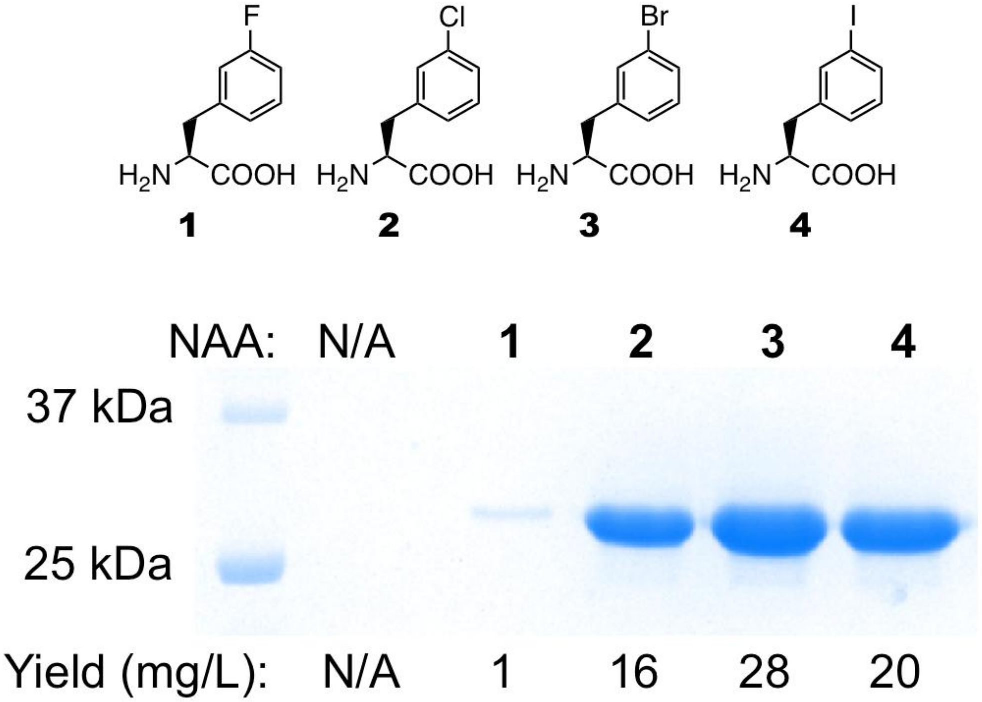


Figure VI-2. The specific incorporation of **1-4** at the S2 position of sfGFP. The expression was induced by the addition of 1 mM IPTG, 0.2% arabinose and 2 mM of a designated NAA. Cells were collected 6h after induction..

Table VI- 1. sfGFP protein molecular weight (MW) determination by ESI-MS.

Protein[a]	Calculated Molecular Weight (Da)	Found Molecular Weight (Da)
sfGFP-1	27806	27806 27826[b] 27844[b] 27788[b]
sfGFP-2	27822	27822
sfGFP-3	27867	27868
sfGFP-4	27914	27912
sfGFP-5	27802	27802
sfGFP-6	27856	27856
sfGFP-7	27818	27818

[a] Protein expression was performing in GmmL medium with 2 mM NAA

[b] There are other 13 F in sfGFP primary sequence. The MW 27826 Da indicates another NAA 1 incorporated into one of 13 F of sfGFP sequence. MW 27844 indicates other two F replaced by NAA **1**. MW 27788 indicated F incorporated at S2 position.

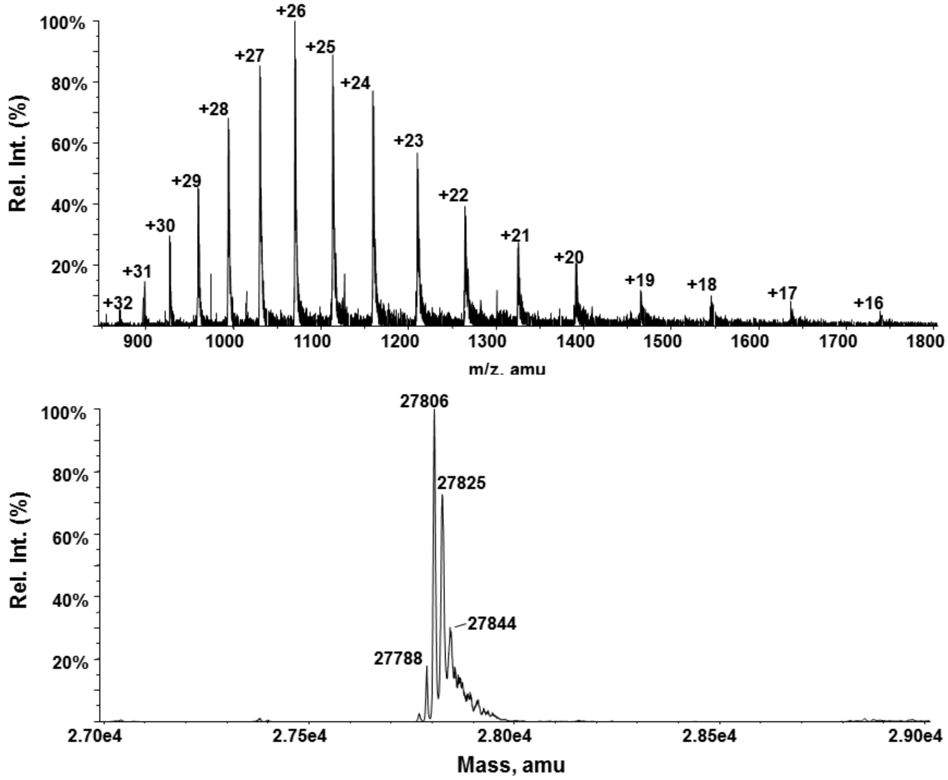


Figure VI-3. Molecular weight determination of the protein sfGFP-**1**: (A) ESI-MS spectrum of sfGFP-**1** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**1**.

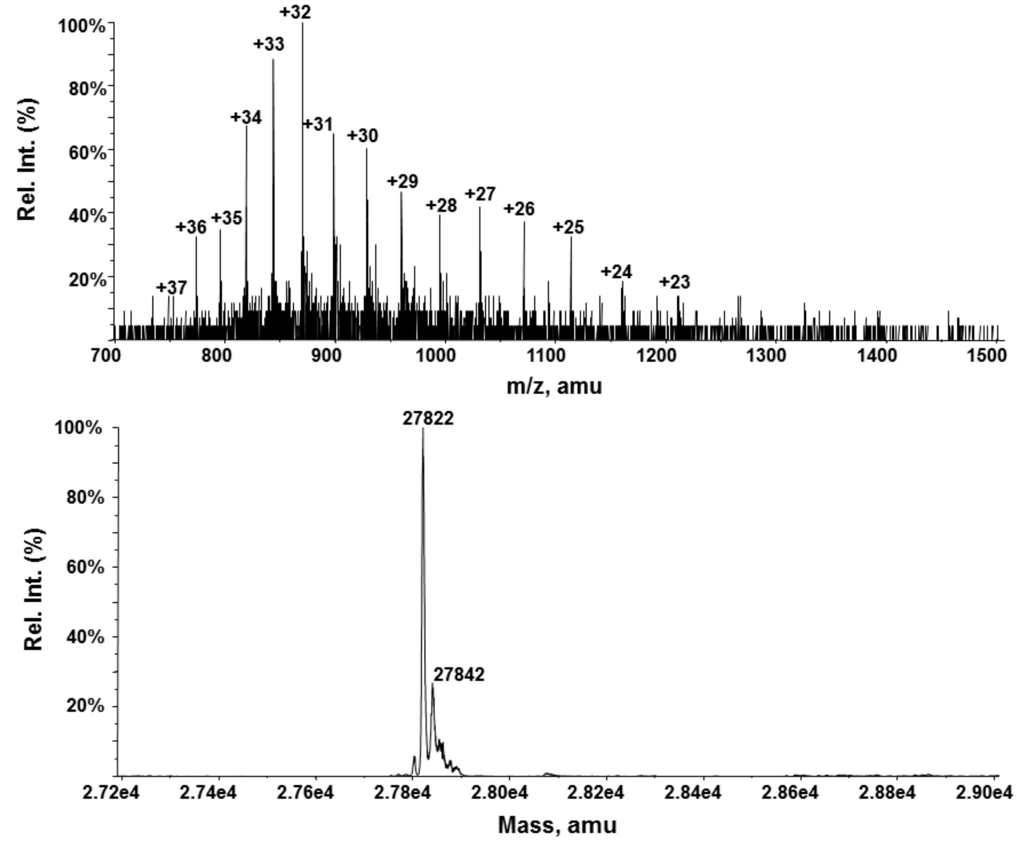


Figure VI-4. Molecular weight determination of the protein sfGFP-**2**: (A) ESI-MS spectrum of sfGFP-**2** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**2**.

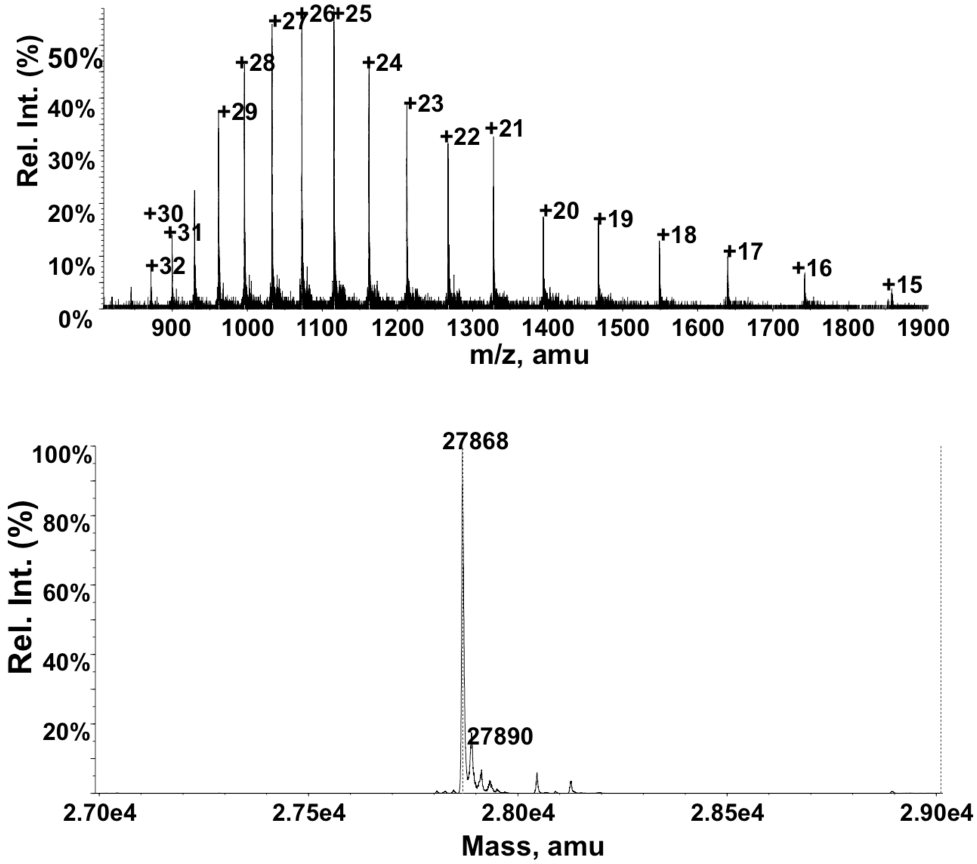


Figure VI-5. Molecular weight determination of the protein sfGFP-**3**: (A) ESI-MS spectrum of sfGFP-**3** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**3**.

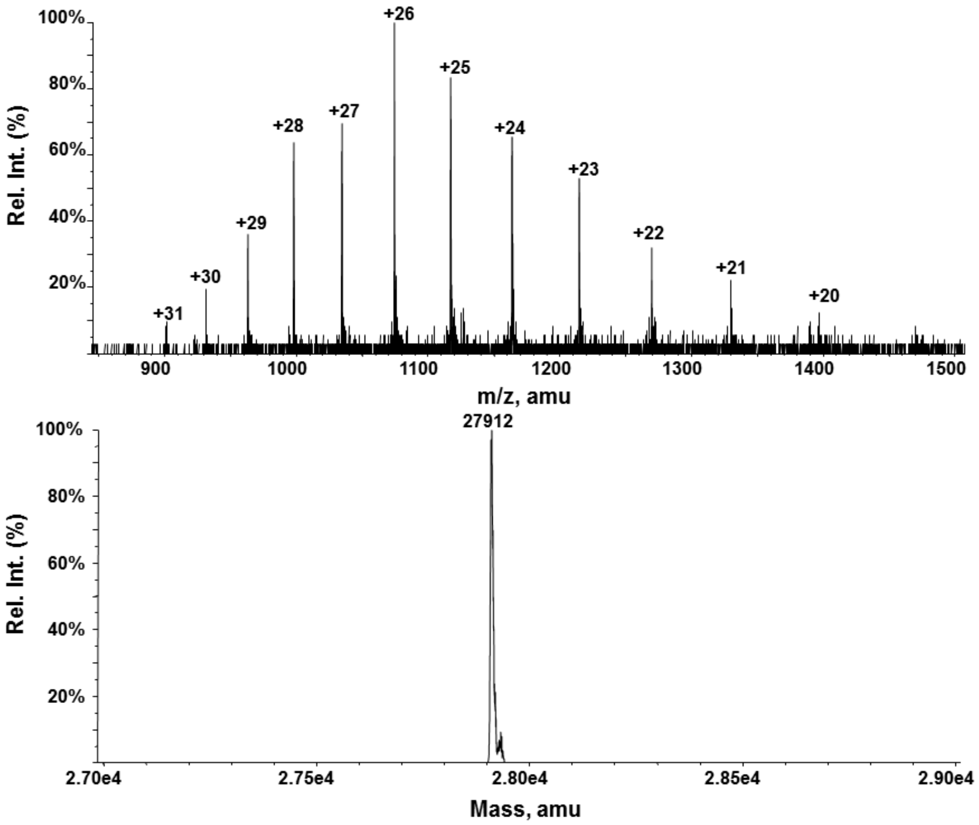


Figure VI-6. Molecular weight determination of the protein sfGFP-**4**: (A) ESI-MS spectrum of sfGFP-**4** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**4**.

partly due to the toxicity effect of **1**. Indeed, the collected cells grown in the presence of **1** were considerably fewer than those grown in GMML supplemented with either of **2-4**. In addition, except for the expected molecular weight, the ESI-MS spectrum of the purified sfGFP incorporated with **1** displayed several side peaks that clearly indicate misincorporation of **1** at phenylalanine sites (**Figure VI-3**).

Given the success with **1-4**, we anticipated that N346A/C348A would also recognize other phenylalanine derivatives with small *meta*-substituents such as **5-7** shown in **Figure VI-7**. Growing BL21 cells transformed with pEVOL-pylT-N346A/C348A and pET-pylT-sfGFP2TAG in GMML supplemented with 2 mM of either of **5-7** and inducing with the addition of 1 mM IPTG and 0.2% arabinose led to overexpression of sfGFP (**Figure VI-7**). The detected molecular weights of the purified sfGFP variants determined by the ESI-MS analysis agreed well with their theoretic molecular weights (**Table VI-1** and **Figures VI-8-10**). Among all *meta*-substituted phenylalanine derivatives we tested, **6** had the best corresponding sfGFP expression yield. **6** apparently has the most hydrophobic *meta*-substituent that likely contributes to its strong interaction with the hydrophobic active site of N346A/C348A and its high incorporation rate. In comparison to seven *para*-substituted phenylalanine derivatives we tested previously, **2-6** gave much better incorporation levels. We think this is due to the fact that the hydrophobic *meta*-substituents of these NAAs can fit well in the hydrophobic pocket shown in **Figure VI-1B**.

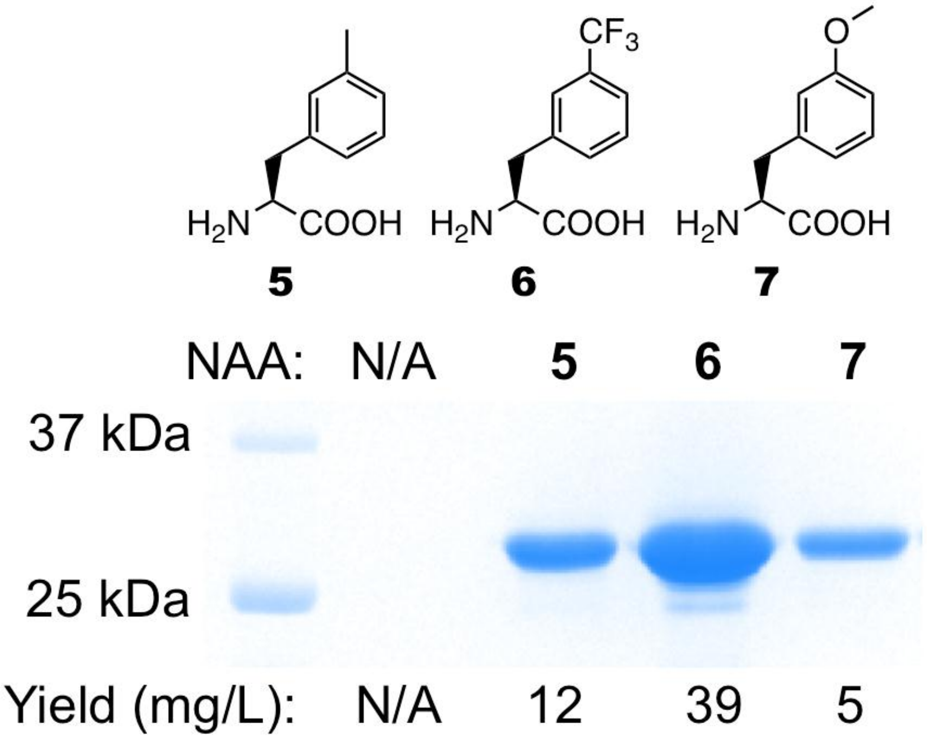


Figure VI-7. The specific incorporation of **5-7** at the S2 position of sfGFP

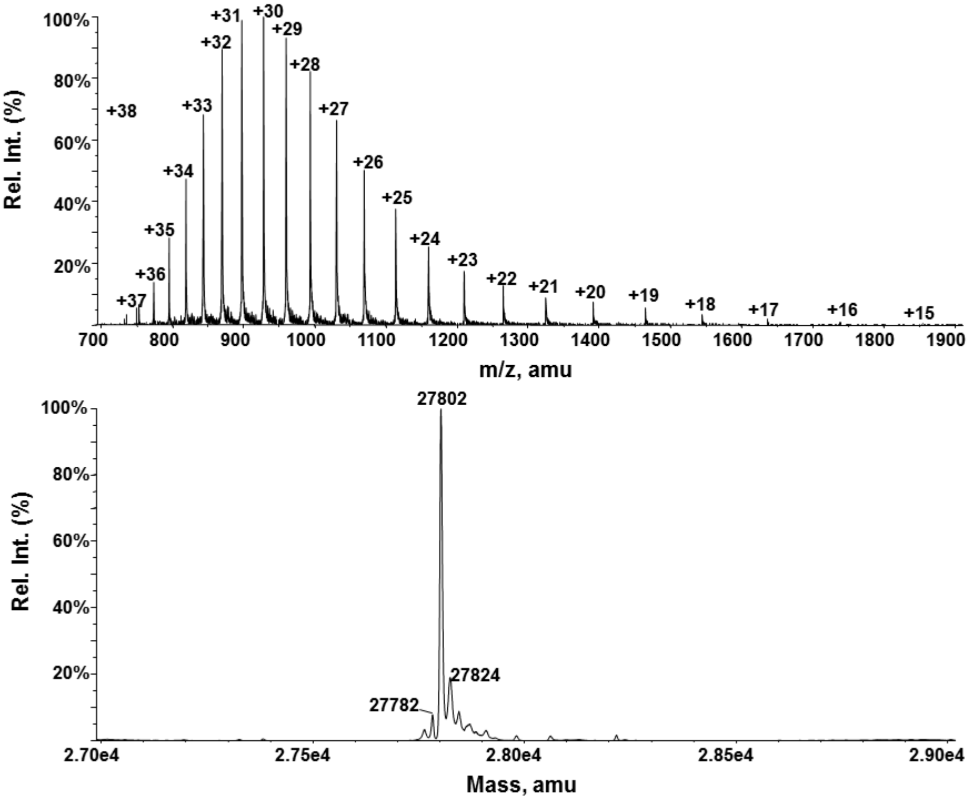


Figure VI-8. Molecular weight determination of the protein sfGFP-**5**: (A) ESI-MS spectrum of sfGFP-**5** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**5**

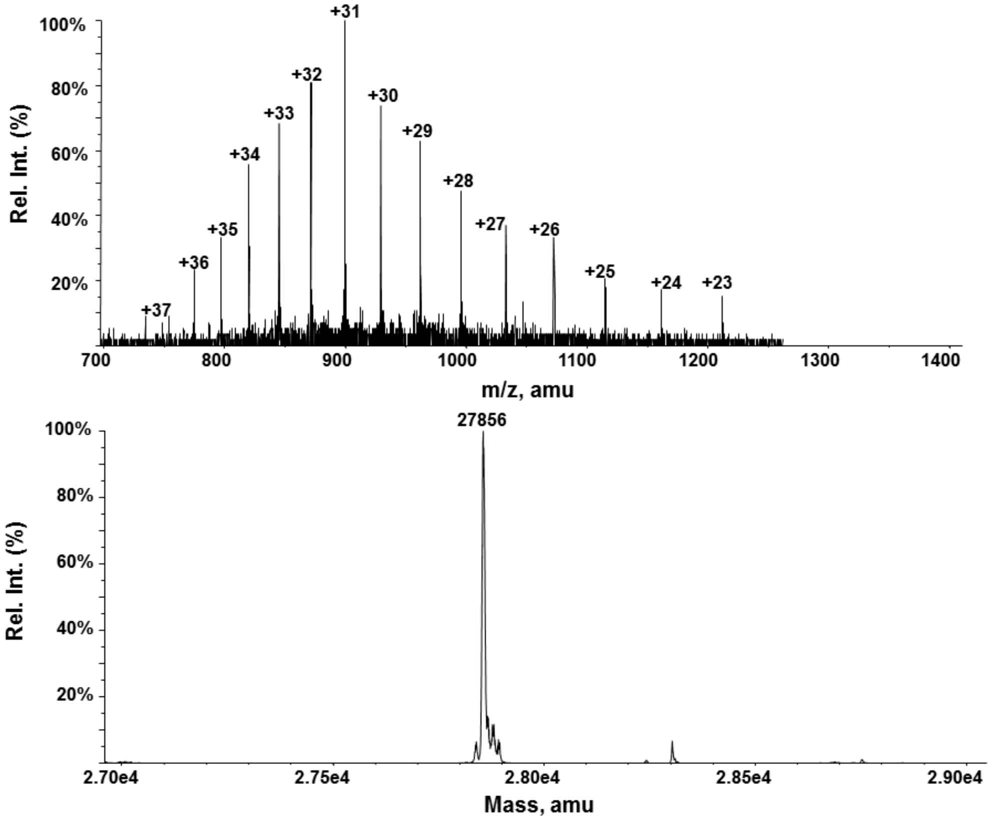


Figure VI-9. Molecular weight determination of the protein sfGFP-**6**: (A) ESI-MS spectrum of sfGFP-**6** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**6**.

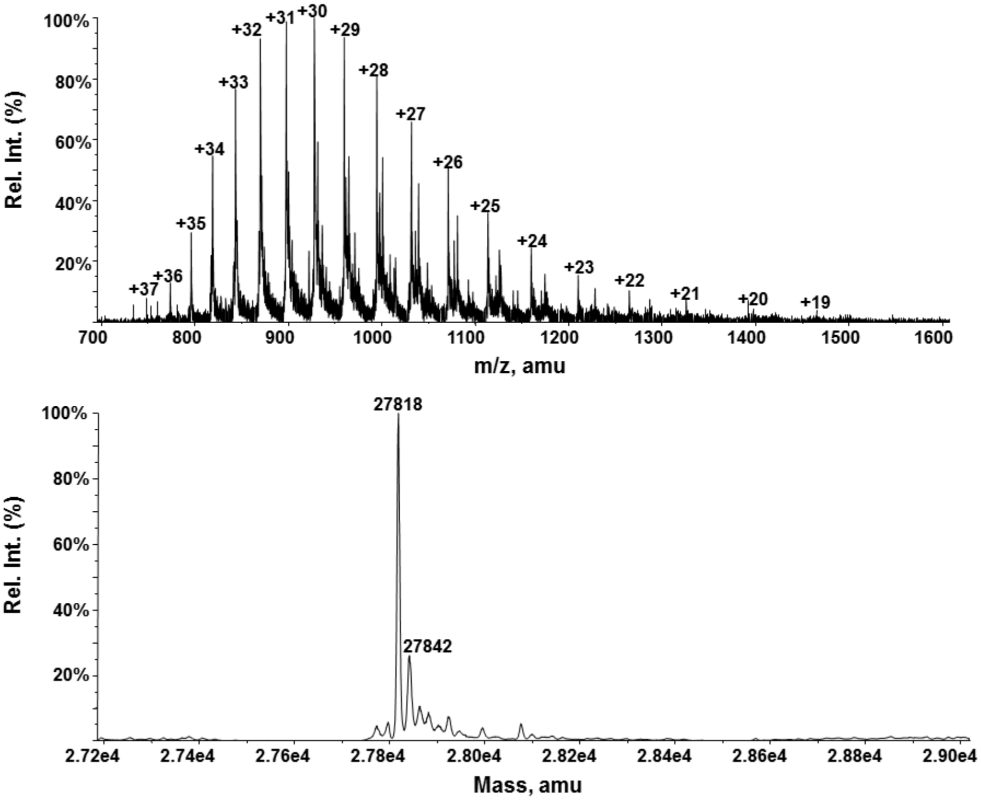


Figure VI-10. Molecular weight determination of the protein sfGFP-**7**: (A) ESI-MS spectrum of sfGFP-**7** and (B) the deconvoluted ESI-MS spectrum of sfGFP-**7**.

Among seven *meta*-substituted phenylalanine derivative substrates of N346A/C348A we tested so far, **3** and **4** can undergo Suzuki coupling reactions with organoboronic acids and Sonogashira coupling reactions with terminal alkynes using biocompatible palladium complex catalysts, and **1** and **6** contain fluorine so that they can serve as ^{19}F NMR sensors for protein folding/conformation rearrangement and protein-protein/ligand interaction investigations. Since we previously showed that a *para*-halo-phenylalanine that was incorporated into a protein using an evolved PylRS-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair could be fluorescently labeled using a Suzuki coupling reaction, here we focused to demonstrate the application of **6** in protein folding/unfolding analysis. sfGFP was chosen as a model protein for this study. A plasmid pBAD-sfGFP27TAG was constructed. This plasmid carried a sfGFP gene with an amber mutation at F27. F27 is located at the second β strand of sfGFP. Its side chain faces towards the protein interior and issurrounded by several hydrophobic residues of sfGFP. During its unfolding process, sfGFP is expected to completely expose F27 to a hydrophilic environment. We anticipated **6** at F27 of sfGFP would show a large chemical shift variation during this unfolding process. The structure of GFP also indicates there is additional space around F27 to accommodate a small substituent at the *meta* position of its side chain. Thus, replacing F27 with **6** was anticipated not to significantly perturb the sfGFP fold. Together with pEVOL-pylT-N346A/C348A, pBAD-sfGFP27TAG was used to transform *E. coli* Top10 cells. The transformed cells were then grown in LB medium supplemented with 1 mM **6** and induced by the addition of 0.2% arabinose to express sfGFP with **6** incorporated at F27 (sfGFP27-**6**). The expression level was 79 mg/L

(**Figure VI-11**). Only a minimal amount of sfGFP was expressed in LB medium without a NAA supplement. The purified sfGFP-**6** had a fluorescent spectrum and intensity similar to sfGFP (**Figure VI-12**), and a sample of ^{15}N-sfGFP-**6** also showed a similar (but not identical, with differences including small shifts in various peaks such as e.g. those near 10 ppm / 112 ppm and the appearance of a weak peak near 10.2 ppm / 128 ppm) fingerprint as sfGFP in a [^{15}N,^{1}H] correlation spectrum (**Figure VI-13**).[170] These observations suggest that replacing F27 with **6** does not significantly affect the protein folding and chromophore formation processesThe purified sfGFP27-**6** was then used to undergo unfolding analysis in the presence of guanidinium chloride (GndCl). In titrations of stepwise increasing concentrations of GndCl, ^{19}F NMR signals of sfGFP27-**6** were detected. **Figure VI-14A** shows the dependence of ^{19}F chemical shifts on the concentration of GndCl for sfGFP27-**6**. Prior to denaturation, the chemical shift of folded sfGFP27-**6** was dependent on GndCl concentration. This chemical shift change is significantly larger than that observed in a titration of the free amino acid, even though the latter may contain contributions from binding of guanidinium to the carboxylate of the amino acid, a process that cannot occur in the polypeptide (**Figure VI-15**). The GndCl concentration dependent shift in folded sfGFP27-**6** may likely be explained by a local structural change prior to denaturation. The appearance of the peak at 61.5 ppm indicates that the unfolding transition in this titration occurred at 4.6 M GndCl, which based on the time of 1 h between titrated points appears consistent with an unfolding hysteresis observed in the work of Jennings and co-workers.[170] Likewise, a stepwise titration of decreasing GndCl indicates the presence of unfolded sfGFP27-**6** to a much

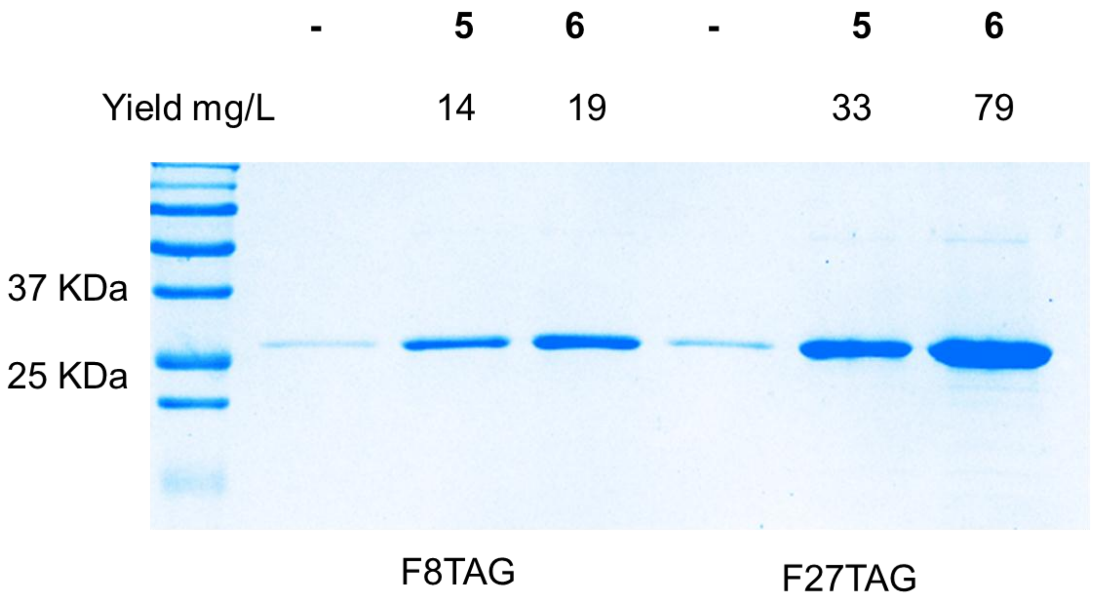


Figure VI-11. The expression of sfGFP containing an amber mutation at F8 and F27. Proteins were expressed in BL21(DE3) cells that grew in LB medium and 1 mM **5** or **6**. The proteins were analyzed by SDS-PAGE (15%) gel electrophoresis with Coomassie blue staining.

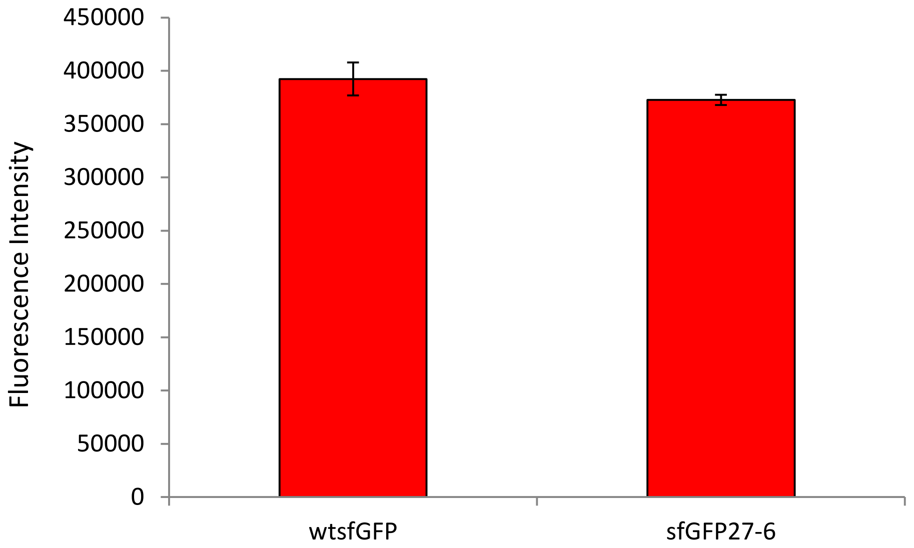


Figure VI-12. The fluorescence intensity of wt-sfGFP and sfGFP27-**6**. The same protein concentration 500 nM was prepared in PBS buffer in pH = 7.0. Excitation and emission wavelength are 450 nm and 520 nm respectively. The fluorescence intensity of sfGFP2-**6** was 95% of wt-sfGFP.

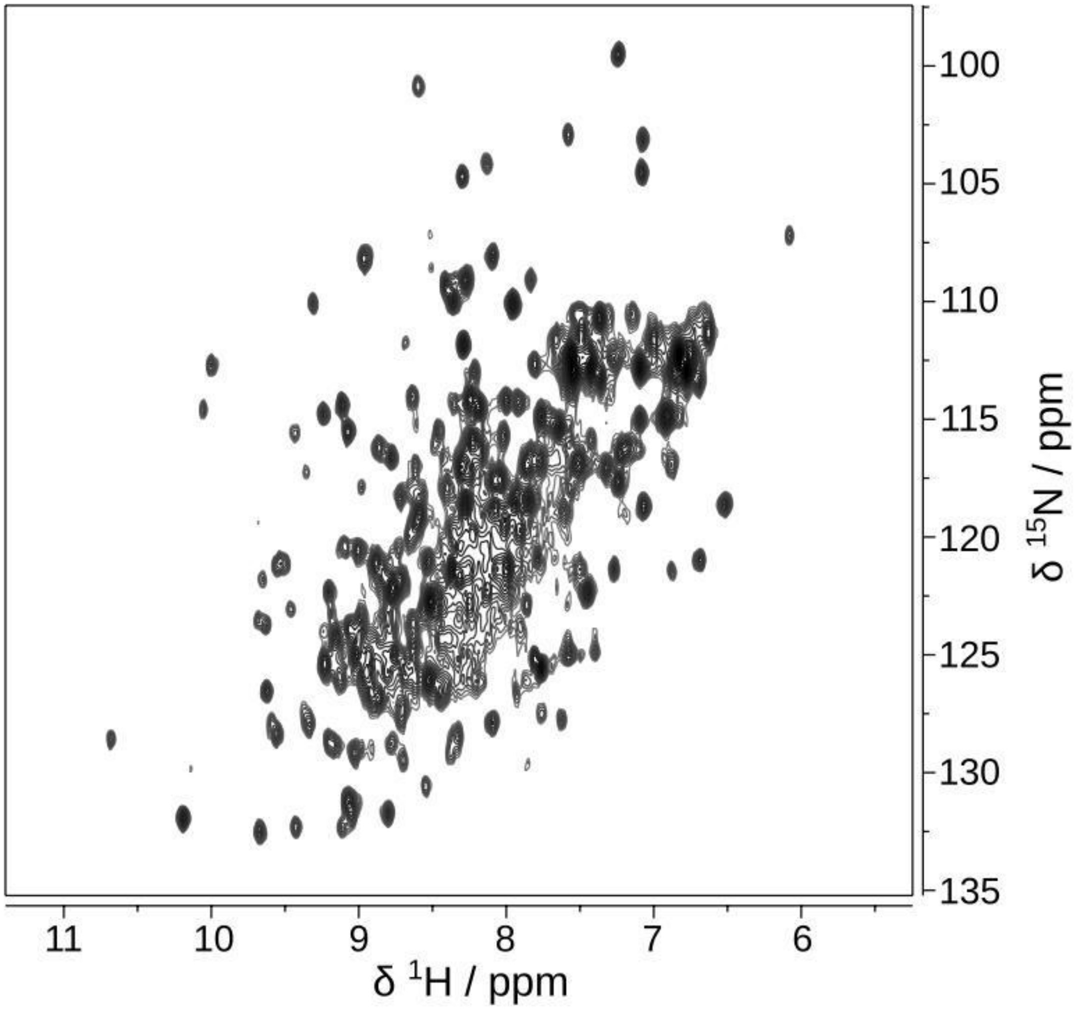


Figure VI-13. [^{15}N,^{1}H]-HMQC spectrum of ^{15}N-sfGFP27-**6** in 12 mM phosphate buffer (pH 7.0) with 140 mM NaCl and 3 mM KCl, acquired at a temperature of 35 ºC, at 500 MHz. Chemical shifts are referenced against an internal standard of 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS).

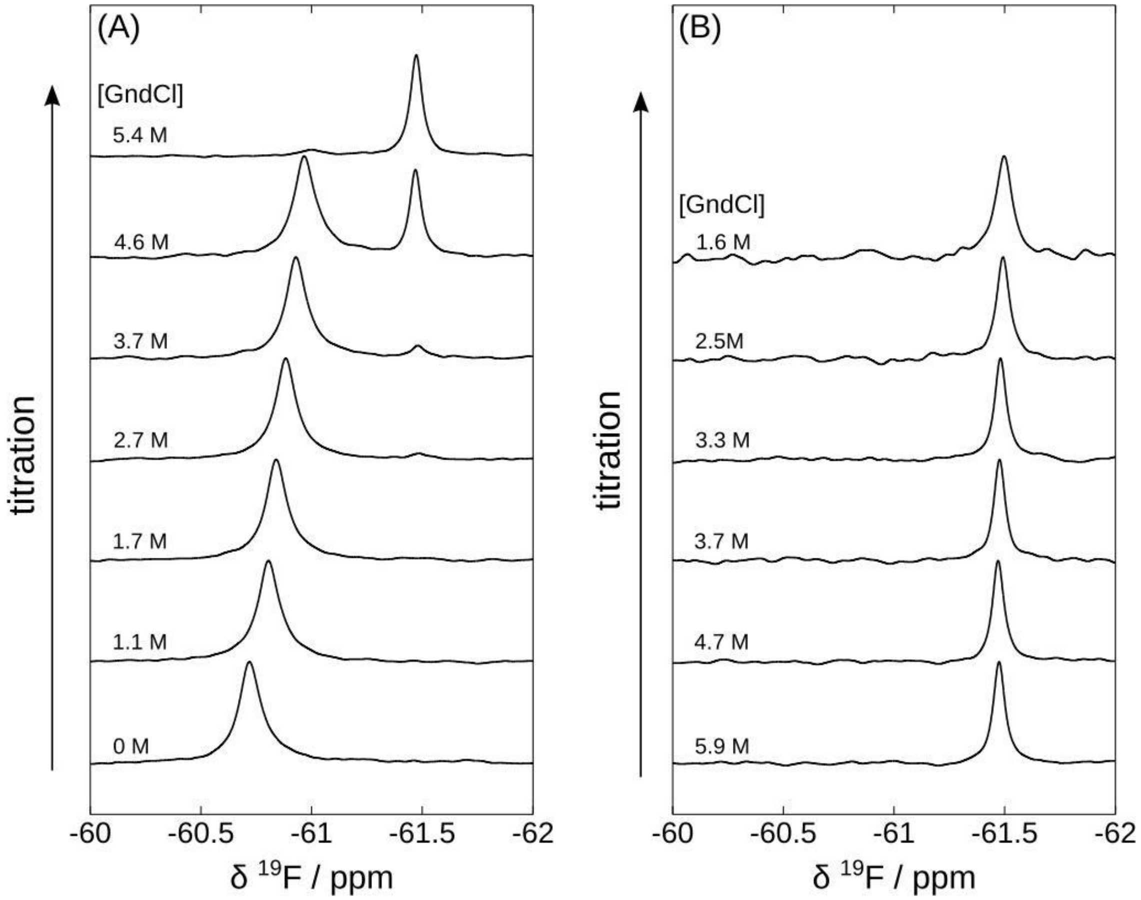


Figure VI-14. (**A**) Unfolding of sfGFP27-**6**, by stepwise addition of GndCl, characterized by ^{19}F NMR. **(B**) Spectra of unfolded sfGFP27-**6** upon stepwise reduction of GndCl. Precipitation occurred below 1.6 M GndCl. In A and B, time between each point was 1h, and measurement temperature was 298 K.

.

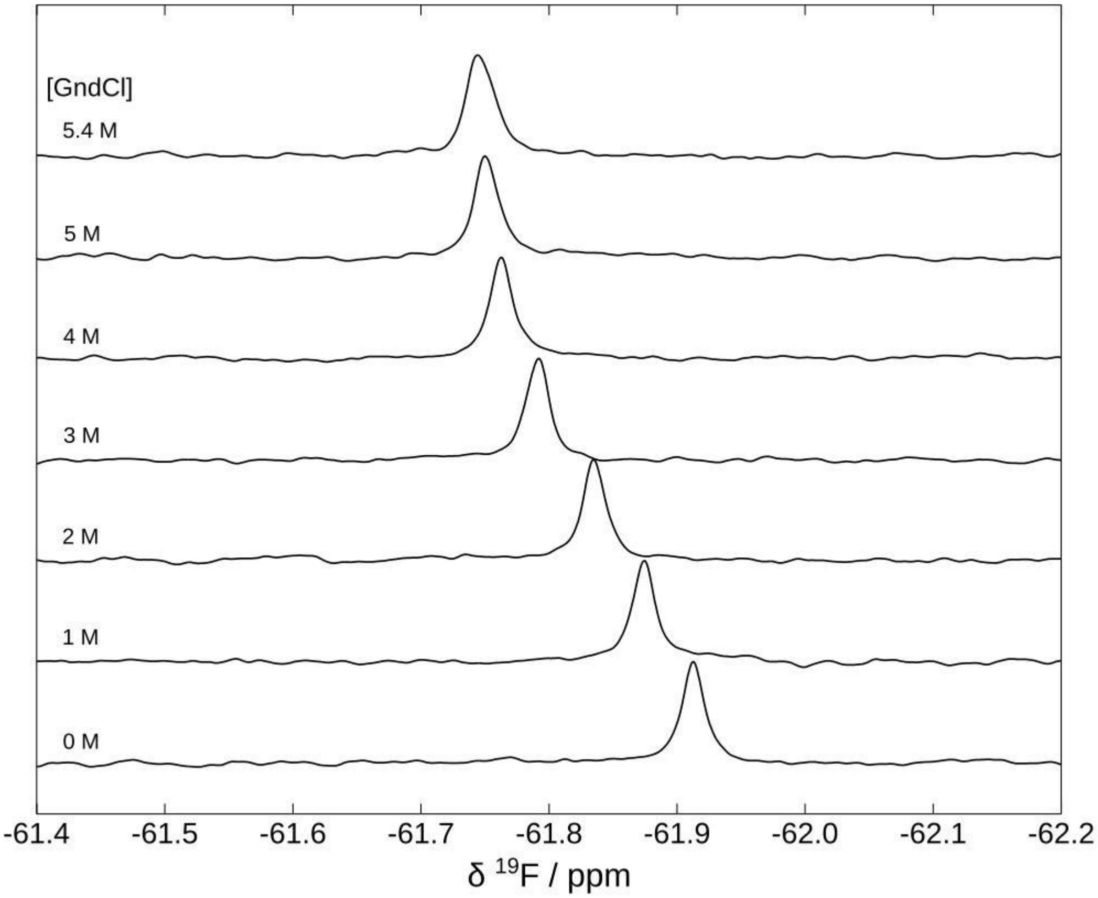


Figure VI-15. Titration of **6** in 12 mM phosphate buffer (pH 7.0) with 140 mM NaCl and 3 mM KCl with 0 M – 5.4 M GndCl. Chemical shifts are referenced against an external standard of trifluoroacetic acid at -75.2 ppm in D_2O.

lower GndCl concentration of 1.6 M (**Figure VI-14B**). The significant change in ^{19}F chemical shift observed illustrates that **6** can be used as a sensitive site specific probe for protein folding, or potentially for protein-ligand interaction. ^{19}F allows for highly receptive, background free NMR detection, and the chemical shift of this nucleus is sensitive to structural changes. Therefore, a single probe inserted at a known position will in particular be useful for measurements with large proteins, or with dilute samples, where chemical shift assignments typically cannot be obtained.

In order to study local structural change prior to denaturation of sfGFP. 5 different positions, S2, F8, F27, F130 and N134, of sfGFP were chose to study. S2 indicated non-structured amino acid, as well as F8 for first α helix from N-termini, F27 for second b sheet; F130 and N135 for loop region amino acids. (**Figure VI-16-17**) pBAD-sfGFP8TAG, pBAD-sfGFP130TAG, pBAD-sfGFP135TAG, pET-sfGFPS2TAG and pBAD-sfGFP27TAG were constructed to incorporated NAA **6** into assigned position of sfGFP. Cotransformed with pEVOL- PylRS-N346A/C348A, 5 different sfGFP variants are expressed in LB in 19-81 mg/L of sfGFP-**6** and 14-35 mg/L of sfGFP-**5**. (**Figure VI-11** and **VI-18**) In S2 and N135 sfGFP variants, sfGFP-**F** has expression by adding 5 mM L-phenylalanine in the GmmL medium during protein induction.

Unfolding equilibrium fluorescence and ^{19}F-NMR experiments are performed for all sfGFP variants. Equilibrium fluorescence profiles in S2 and F8 sfGFP variants showed high sensitivity of the hydrophobicity of local structure in first α helix cap. (**Figure VI-19-20** and **Figure VI-24**) sfGFP-S2F and sfGFP-**5** variant shows better local structure to maintain the hydrophobic chamber keeping the fluorescence of sfGFP. In sfGFP2-**6**, although C_m is better than wild type sfGFP, the result imply fluoro atoms on *m*-trifluoromethyl group may disturb the local structure by hydrogen bonding. sfGFP8-**5** and sfGFP8-**6** unfolding experiment in GndCl were similar pattern but in different chemical shift. (**Figure VI-25-26**) sfGFP-27TAG, sfGFP-130TAG, sfGFP-134TAG variants show limited change in unfolding equilibrium fluorescence pattern and midpoint recovery concentrations of GndCl C_m. (**Figure VI-21-23**) However, the unfolding profile in^{19}F-NMR experiments, it provides rich and detail informations for local structure environment.(**Figure VI-27-29**) In sfGFP27-**6** and sfGFP130-**6**, the unfolding spectrums show the interconversion between locked and the soft folding structures in proposed dual-basin folding mechanism.[170] The results in sfGFP27-**6** and sfGFP130-**6** of ^{19}F-NMR unfolding experiments, the iso-state intermediate with fast frustrated signal between native folded signal was observed in the change of chemical shift as increasing of GndCl concentration. The fast interconversion between native and iso-state shows average signals. However, the chemical shift changing in GndCl denatural experinment may also cause the GndCl concentration change. The refolding ^{19}F-NMR spectrum of sfGFP27-**6** indicated the influerence of GndCl concentration to chemical shift of unfolded sfGFP27-**6** was little.

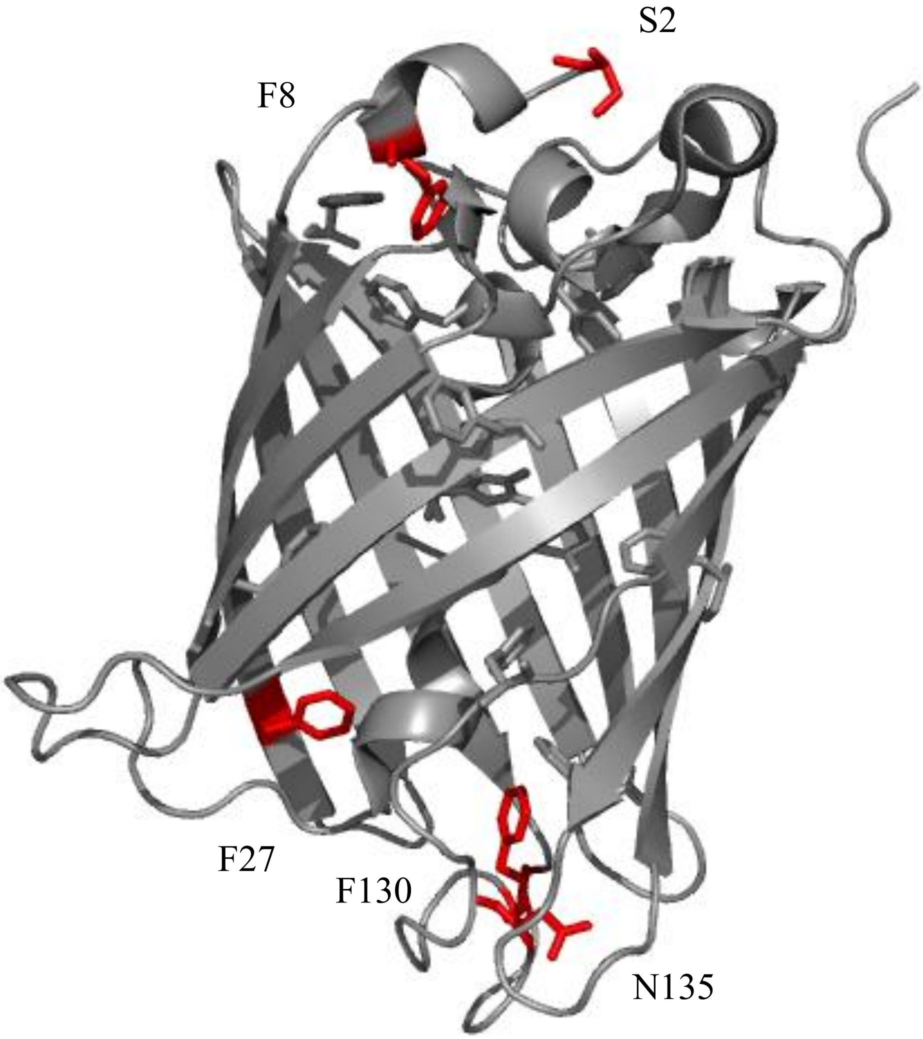


Figure VI-16. sfGFP x-ray crystal structure. The five sites, S2, F8, F27, F130 and N135 were labeled in red color in the structure. PDB code: 2B3P.

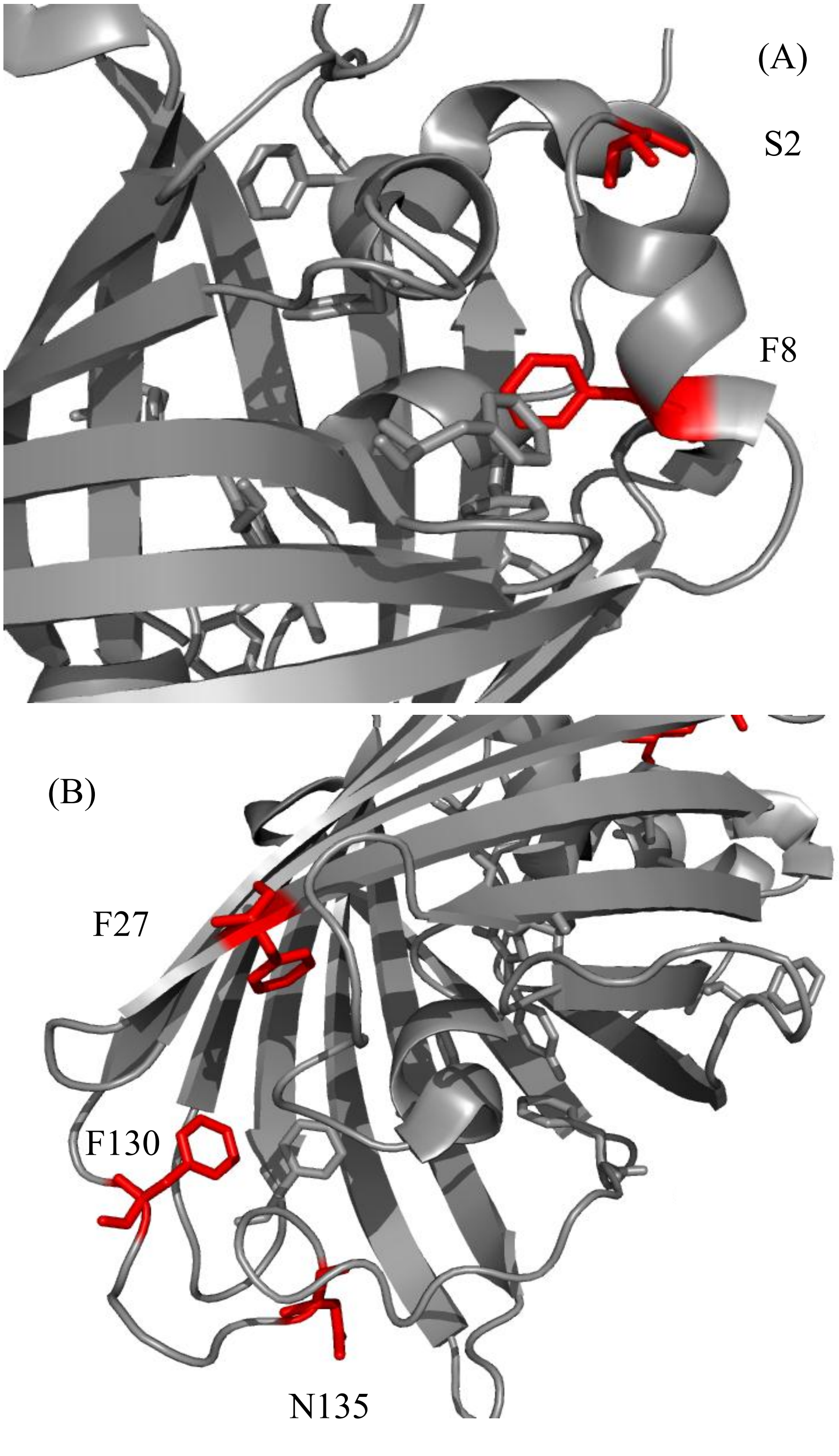


Figure VI-17. sfGFP x-ray crystal structure. (A) The local structure near S2 and F8 were labeled in red. (B) The local structure near F27, F130 and N135 were labeled in red. PDB code: 2B3P.

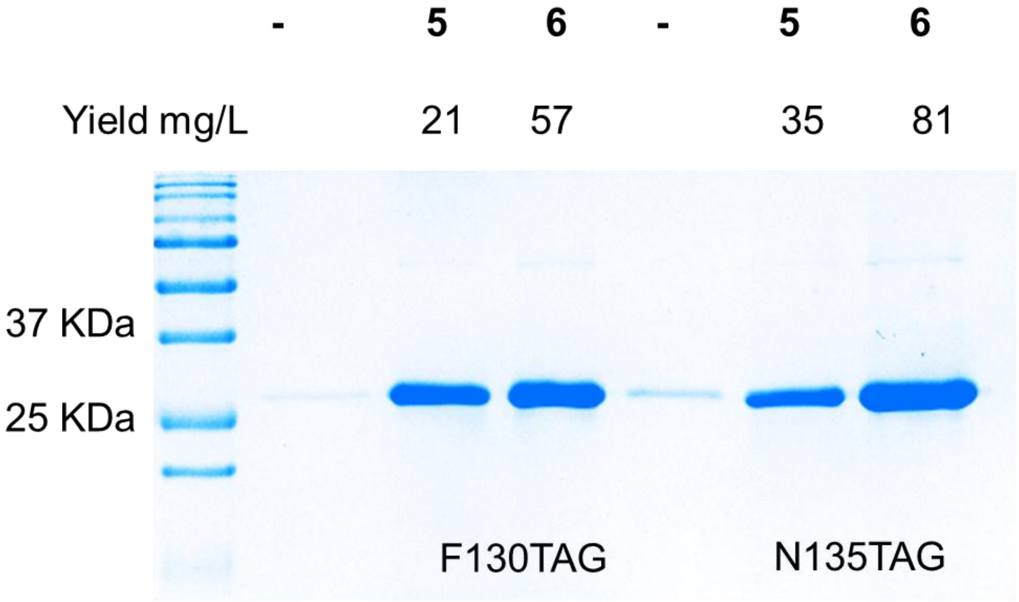


Figure VI-18. The expression of sfGFP containing an amber mutation at F130 and N135. Proteins were expressed in BL21(DE3) cells that grew in LB medium and 1 mM **5** or **6**. The proteins were analyzed by SDS-PAGE (15%) gel electrophoresis with Coomassie blue staining.

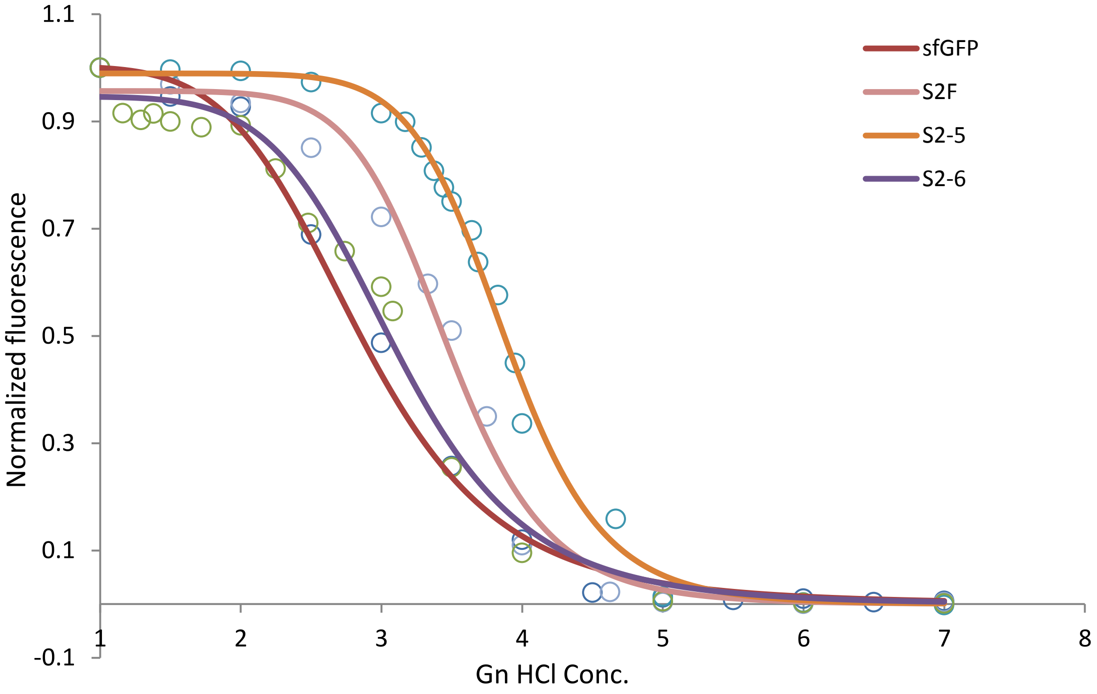


Figure VI-19. Unfolding of sfGFP, sfGFP2-**F,** sfGFP2-**5** and sfGFP2-**6**, by stepwise addition of GndCl, characterized by equilibrium fluorescence values. It was measured by adding various final GndCl concentrations to sfGFP variants into to between 1.0 and 7.0 M in increments of 0.5 M GndCl, and allowing equilibrium to process up to 48 hr in room temperature.

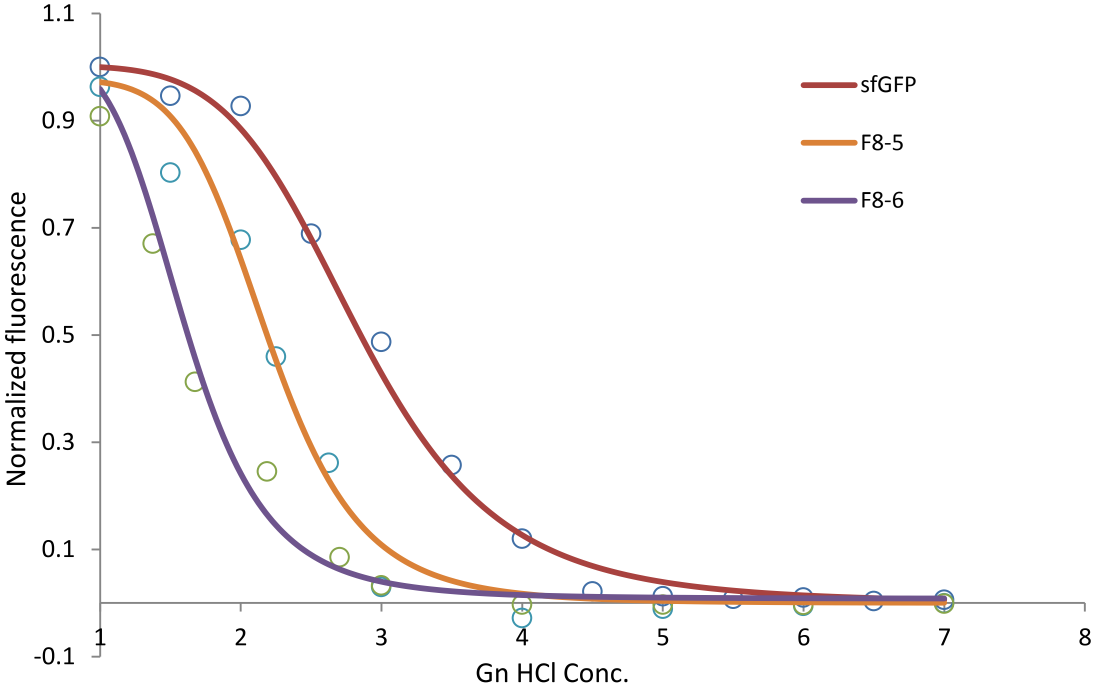


Figure VI-20. Unfolding of sfGFP, sfGFP8-**5** and sfGFP8-**6**, by stepwise addition of GndCl, characterized by equilibrium fluorescence values. It was measured by adding various final GndCl concentrations to sfGFP variants into to between 1.0 and 7.0 M in increments of 0.5 M GndCl, and allowing equilibrium to process up to 48 hr in room temperature.

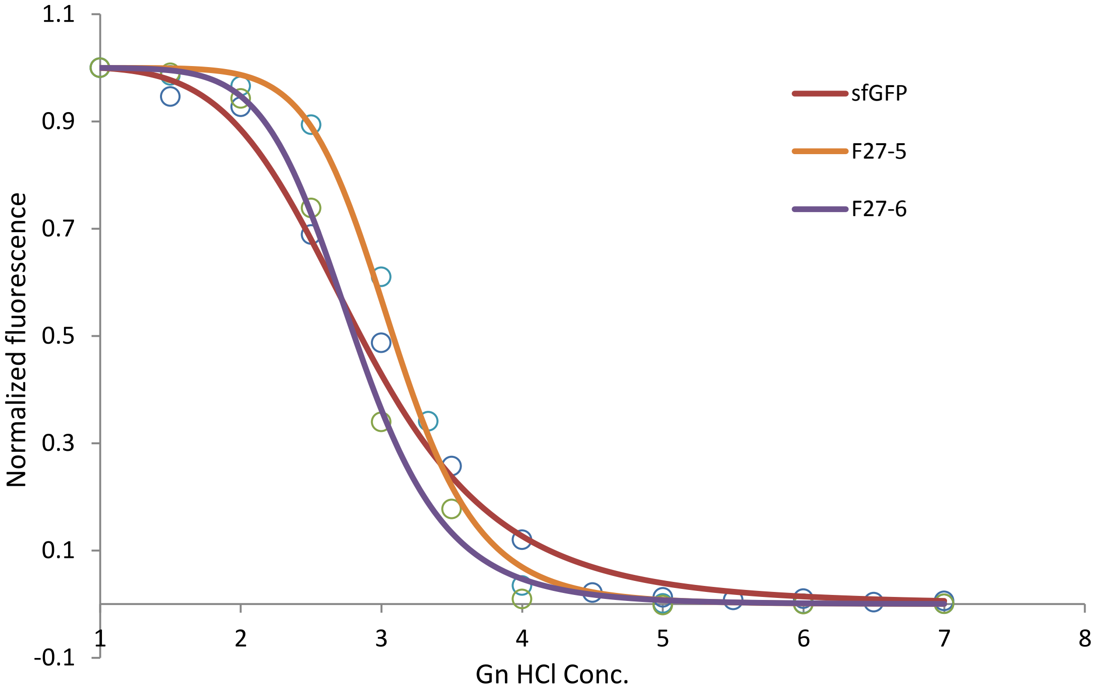


Figure VI-21. Unfolding of sfGFP, sfGFP27-**5** and sfGFP27-**6**, by stepwise addition of GndCl, characterized by equilibrium fluorescence values. It was measured by adding various final GndCl concentrations to sfGFP variants into to between 1.0 and 7.0 M in increments of 0.5 M GndCl, and allowing equilibrium to process up to 48 hr in room temperature.

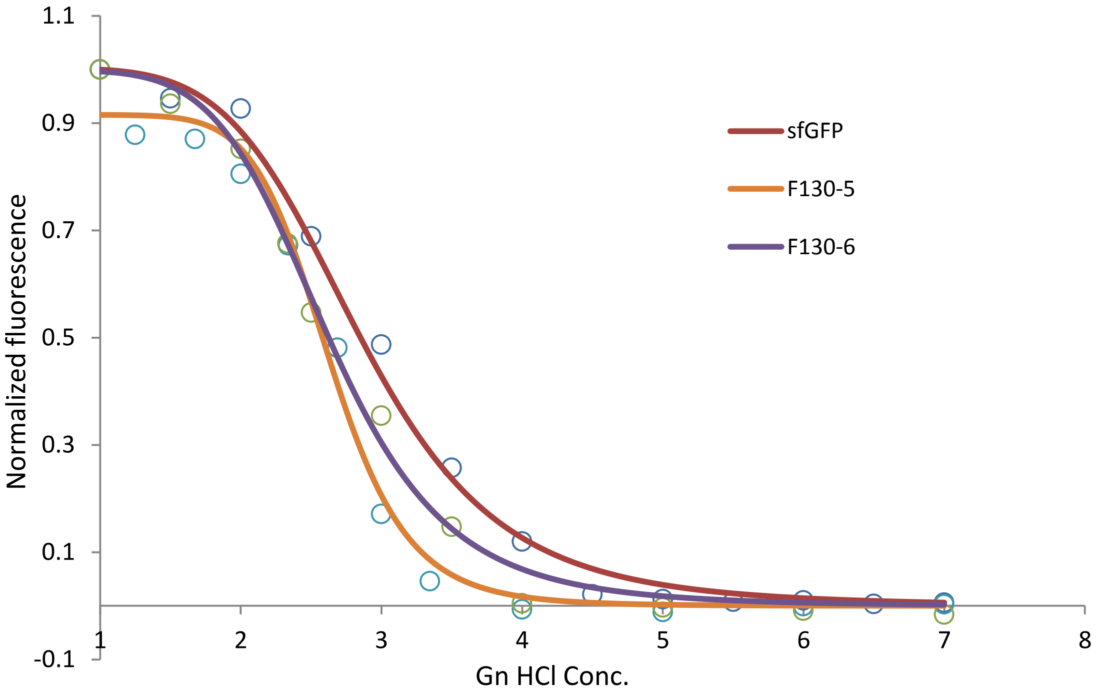


Figure VI-22. Unfolding of sfGFP, sfGFP130-**5** and sfGFP130-**6**, by stepwise addition of GndCl, characterized by equilibrium fluorescence values. It was measured by adding various final GndCl concentrations to sfGFP variants into to between 1.0 and 7.0 M in increments of 0.5 M GndCl, and allowing equilibrium to process up to 48 hr in room temperature.

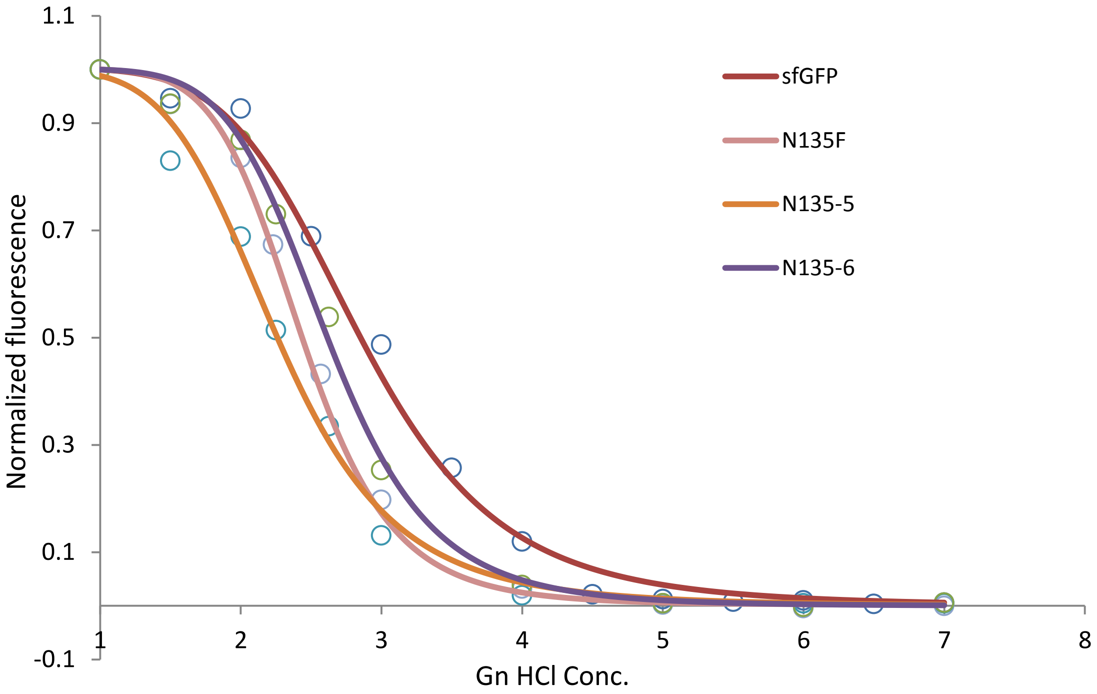


Figure VI-23. Unfolding of sfGFP, sfGFP134F, sfGFP134-**5** and sfGFP134-**6**, by stepwise addition of GndCl, characterized by equilibrium fluorescence values. It was measured by adding various final GndCl concentrations to sfGFP variants into to between 1.0 and 7.0 M in increments of 0.5 M GndCl, and allowing equilibrium to process up to 48 hr in room temperature.

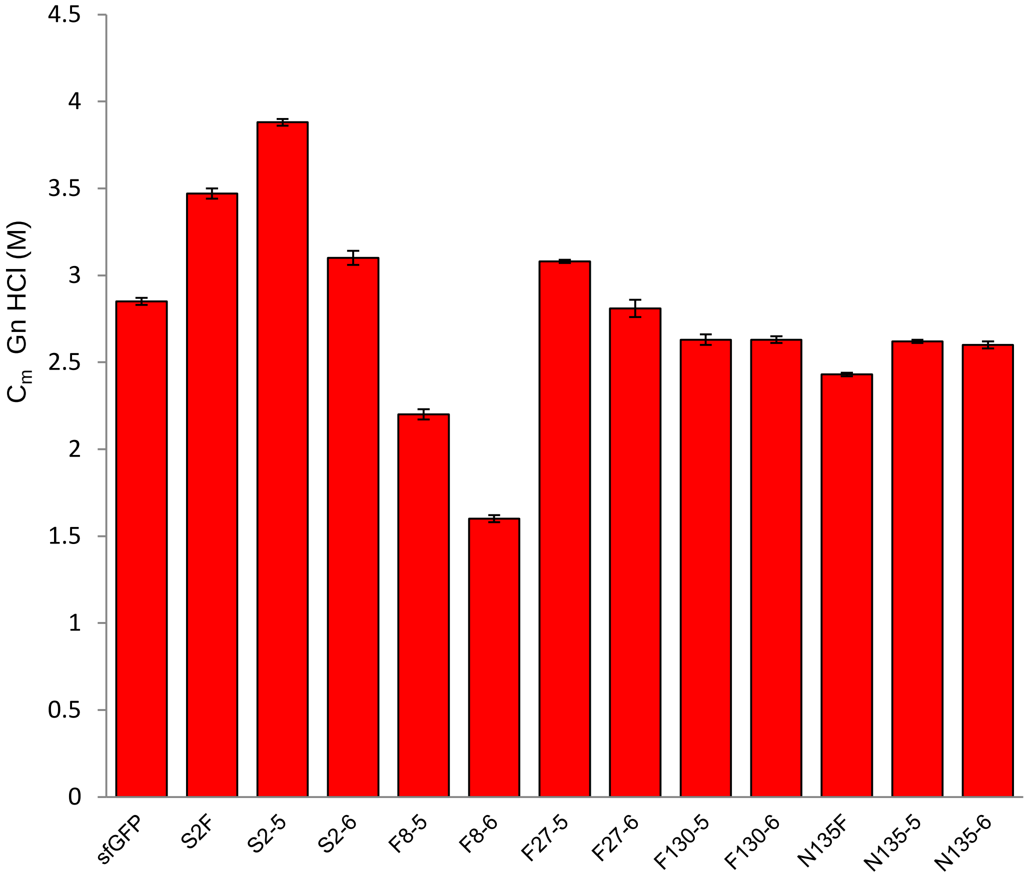


Figure VI-24. Midpoint recovery concentrations of GndCl C_m (recovery of 50% of the initial fluorescence) of sfGFP variants in unfolding equilibrium fluorescence experiments.

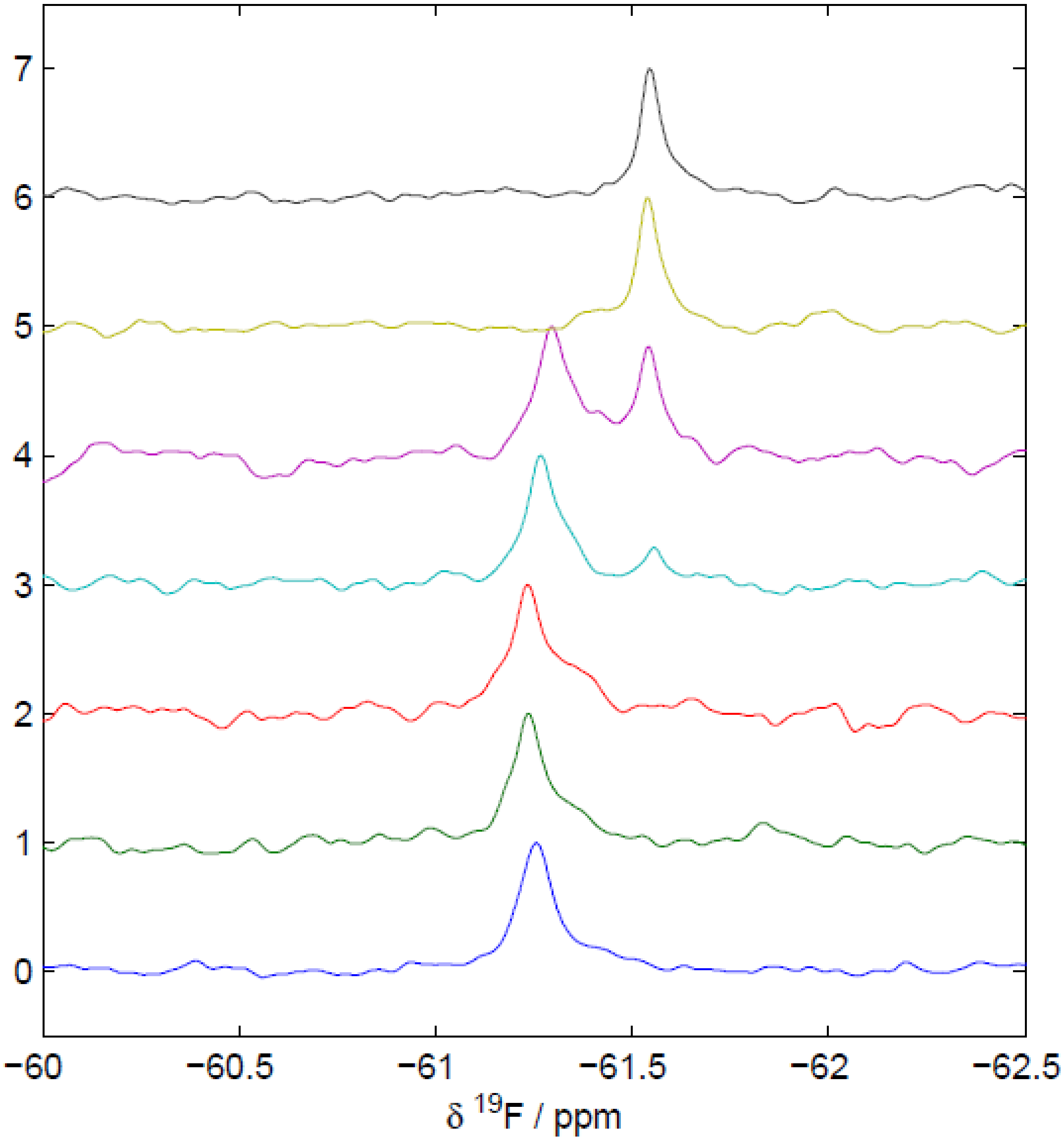


Figure VI-25. Unfolding of sfGFP2-**6**, by stepwise addition of GndCl, characterized by ^{19}F NMR. Y axis indicates the concentration of GndCl from 0 to 6 M by in increments of 1 M GndCl.

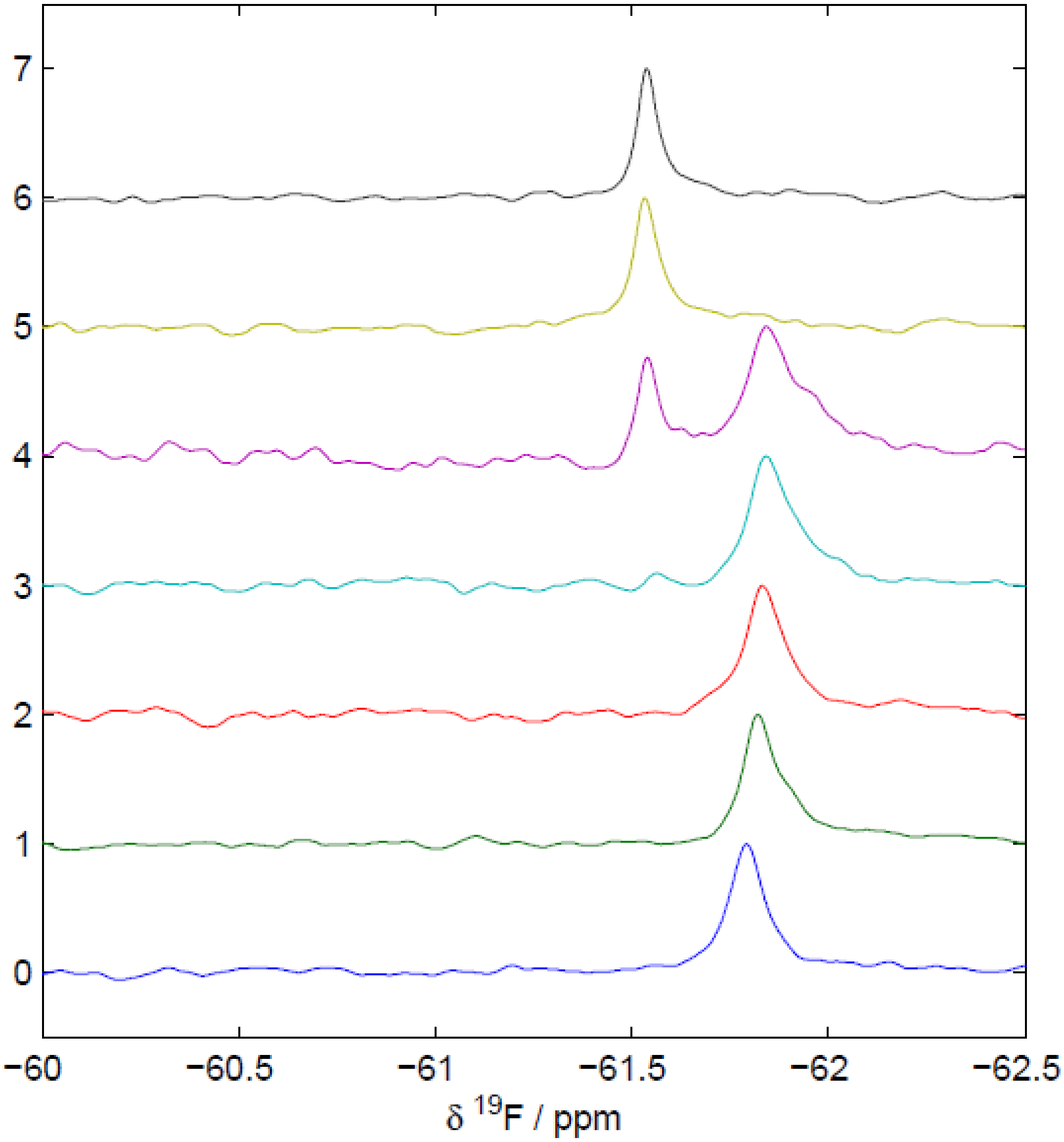


Figure VI-26. Unfolding of sfGFP8-**6**, by stepwise addition of GndCl, characterized by ^{19}F NMR. Y axis indicates the concentration of GndCl from 0 to 6 M by in increments of 1 M GndCl.

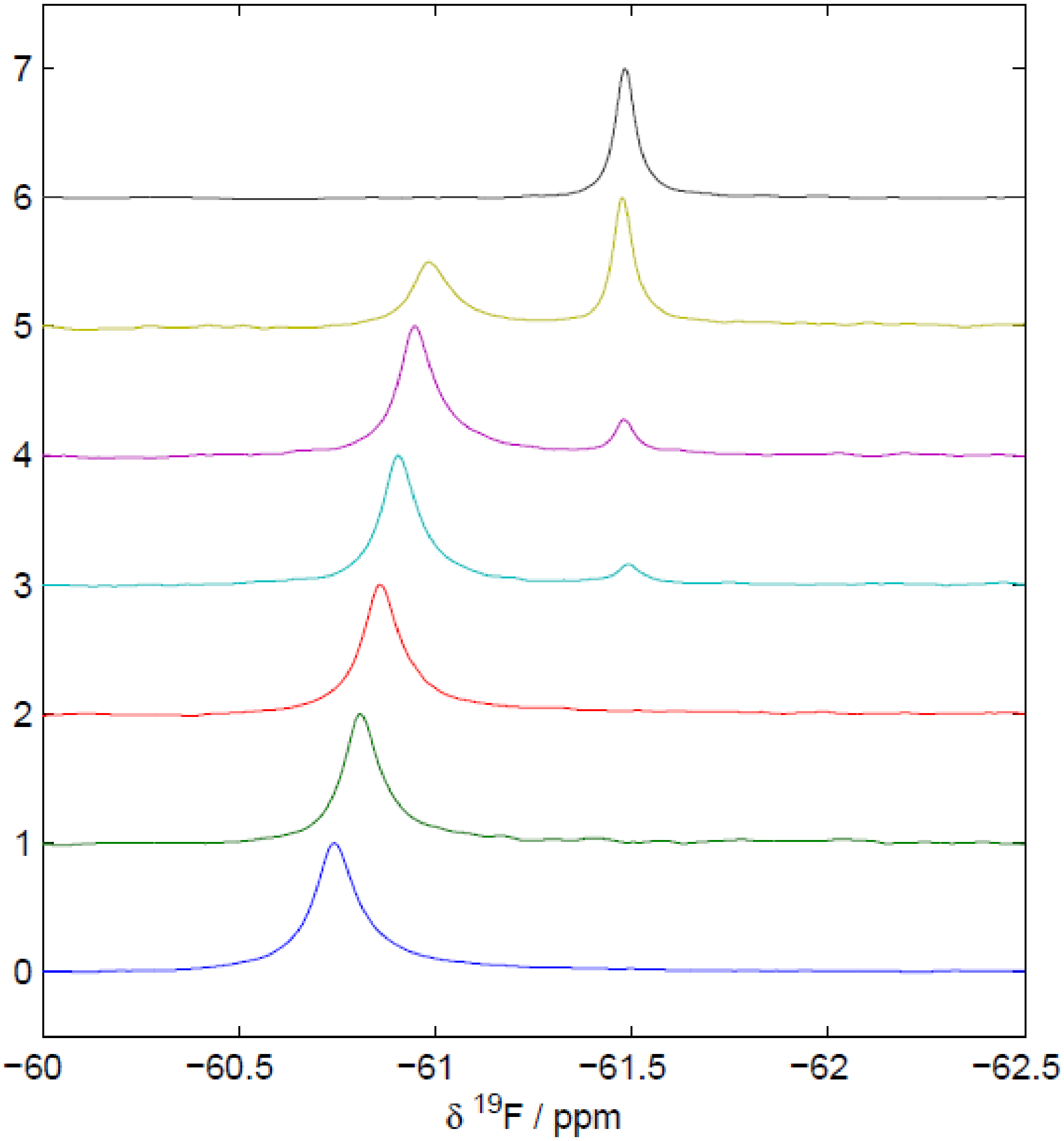


Figure VI-27. Unfolding of sfGFP27-**6**, by stepwise addition of GndCl, characterized by ^{19}F NMR. Y axis indicates the concentration of GndCl from 0 to 6 M by in increments of 1 M GndCl.

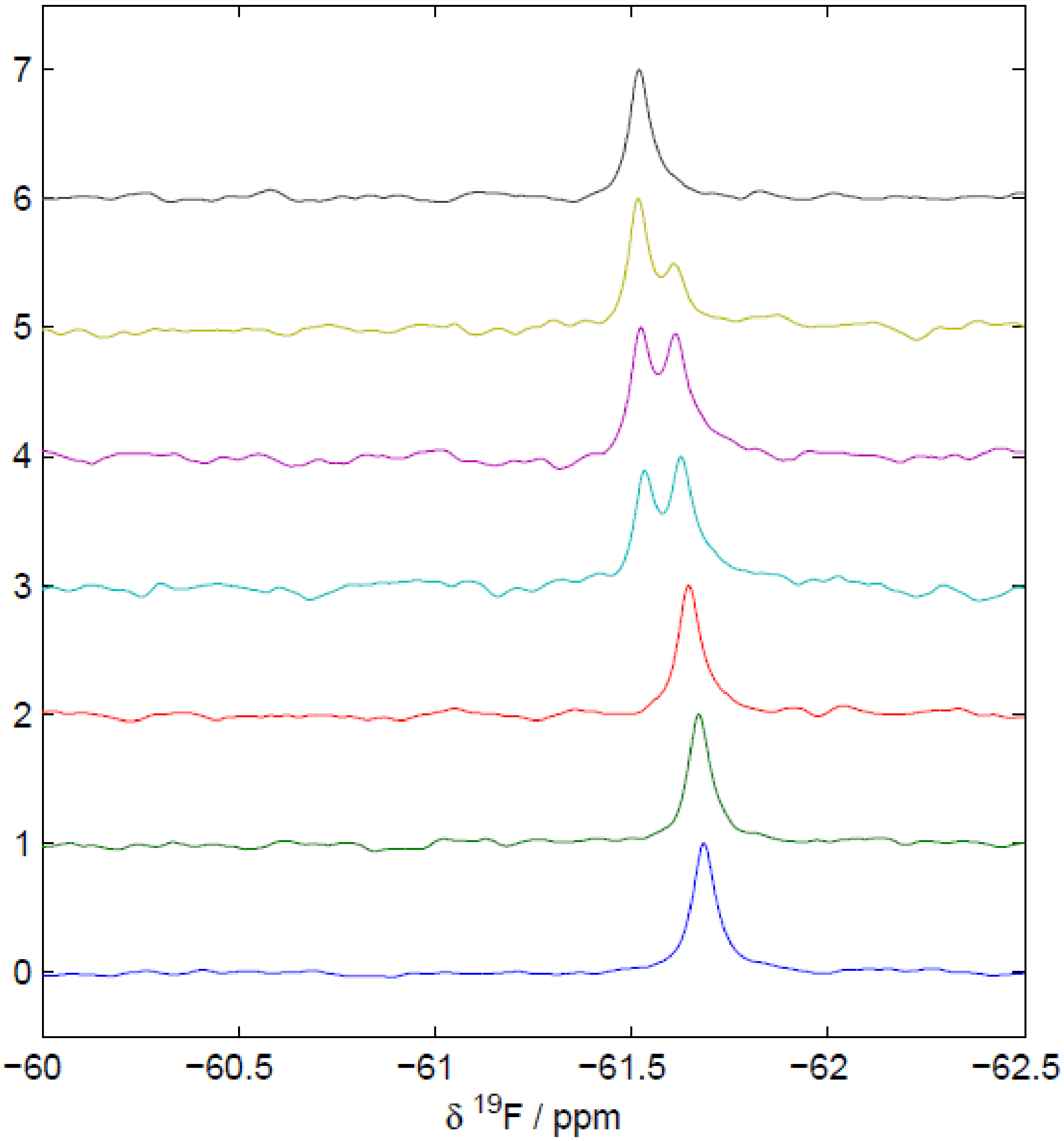


Figure VI-28. Unfolding of sfGFP130-**6**, by stepwise addition of GndCl, characterized by ^{19}F NMR. Y axis indicates the concentration of GndCl from 0 to 6 M by in increments of 1 M GndCl.

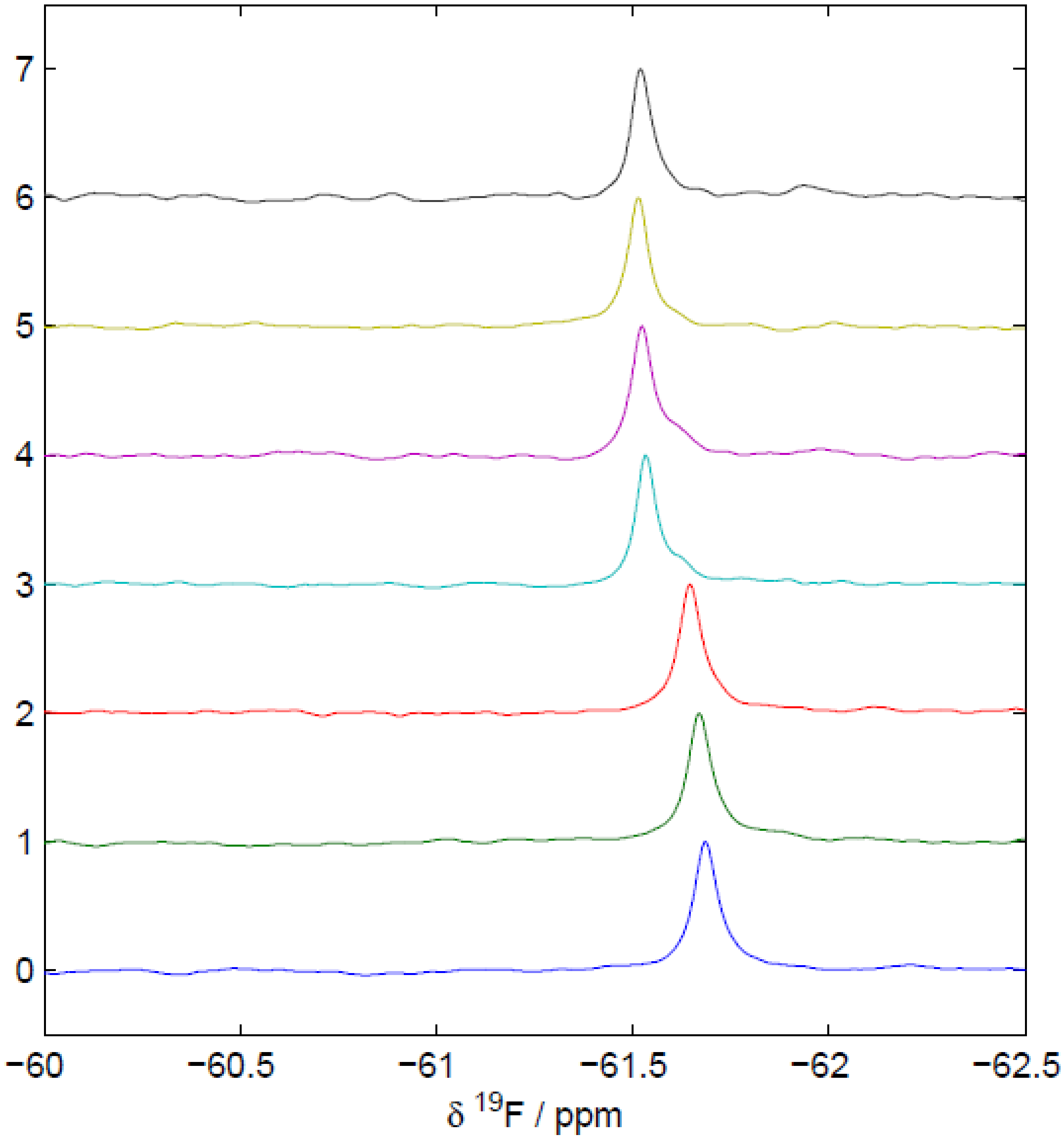


Figure VI-29. Unfolding of sfGFP135-**6**, by stepwise addition of GndCl, characterized by ^{19}F NMR. Y axis indicates the concentration of GndCl from 0 to 6 M by in increments of 1 M GndCl.

Conclusion

In summary, we have demonstrated genetic incorporation of seven *meta*-substituted phenylalanine derivatives into proteins at amber mutation sites in *E. coli* using a rationally designed PylRS mutant N346A/C348A and $\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$. These NAAs contain functional groups that can serve as biophysical and biochemical probes and therefore enable a variety of studies involving proteins. Given that all seven NAAs contain *meta*-substituents that are substantially smaller than the hydrophobic pocket in N346A/C48A that accommodates them, the current study also leads the way to search for phenylalanine derivative substrates of N346A/C348A that have larger *meta*-substituents and better binding affinities to the enzyme. Currently, we are synthesizing other *meta*-substituted phenylalanine derivatives, their genetic encoding using the N346A/C348A-$\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair will be reported later. In the present study, we also demonstrated the application of a site-specifically incorporated **6** in protein folding/unfolding analysis using ^{19}F NMR. Given its strong NMR signal, free background NMR detection, and high chemical shift variation due to local environment change, this ^{19}F probe can be potentially applied for investigating thermodynamics and kinetics of protein-ligand, protein-protein, and protein-DNA/RNA interactions, in addition to protein structure and dynamics.

CHAPTER VII

CONCLUDING REMARKS AND FUTURE OUTLOOK

We have recently demonstrated that the PylRS-$\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair can be evolved for genetic incorporation of phenylalanine derivatives[68], *p*-iodophenylalanine and *p*-bromophenylalanine and guide the direction to evolve this pair for the incorporation of sulfotyrosine and *p*CMF. Since the PylRS-$\text{tRNA}_{\text{CUA}}^{\text{Pyl}}$ pair is orthogonal in both yeast and mammalian cells, direct applications of the evolved pairs in these cells will allow synthesizing proteins with defined tyrosine sulfation and phosphorylation mimics for their cellular function analysis. This work also achieved the incorporation of N^{ε}-methyllysine using indirect methods. Despite the efforts of other groups and us, direct incorporation of N^{ε}-methyllysine is not successful. We may argue that it is difficult to evolve an aminoacyl-tRNA synthetase that thermodynamically favors the binding to N^{ε}-methyllysine than lysine. But another explanation is also possible. N^{ε}-methyllysine may have a difficulty to be transported into *E. coli* cells by ABC transporters. Since no NAA with a positively charged side chain has been evolved so far, we can not rule out this possibility. If this was real, it would rule out the possibility to directly incorporate N^{ε},N^{ε}-dimethyllysine and N^{ε},N^{ε},N^{ε}-trimethyllysine into proteins. We have tried to evolve PylRS mutants specific for N^{ε},N^{ε}-dimethyllysine and N^{ε},N^{ε},N^{ε}-trimethyllysine but no positive clones have beenidentified. We think a thorough investigation of membrane transportation of these NAAs is necessary before further efforts are carried out. Since arginine also undergoes methylation with several distinctive patterns, an interesting question is whether we could evolve an aaRS-tRNA pair for genetic incorporation of methylated arginines. Given the structure similarity between lysine and arginine, it is highly possible to evolve PylRS mutants that are specific for protected or unprotected methylated arginines. Another challenge in this research area is the installation of multiple different modifications to proteins. Concomitant modifications in one protein are prevalent in cells. It is an interesting question whether these modifications crosstalk to each others, reinforce each others' effects, or counteract each others' effects. To synthesize a protein with two different modifications, two orthogonal aaRS-tRNA pairs and two different blank codons have to be used. Recently, we and Chin *et al.* independently developed two systems for genetic incorporation of two different NAAs. The method developed by Chin *et al.* uses one amber codon and one quadruple AGGA codon to code two different NAAs. Our method uses one amber codon and one ochre codon. Two orthogonal pairs in both systems were derived from the *Mj*TyrRS-*Mj*tRNA$_{\mathrm{CUA}}^{\mathrm{Tyr}}$ pair and the PylRS-tRNA$_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair. These two methods make it possible to synthesize a protein containing one sulfotyrosine/phosphotyrosine mimic and one N^{ε}-acetyllysine/N^{ε}-methyllysine. However, it will require developing an additional aaRS-tRNA pair orthogonal to the PylRS-tRNA$_{\mathrm{CUA}}^{\mathrm{Pyl}}$ pair for synthesis of a protein containing both N^{ε}-acetyllysine and N^{ε}-methyllysine. Searching additional orthogonal aaRS-tRNA pairs may be necessary.

REFERENCES

1. Walsh, C. T. (2005) *Posttranslational Modification of Proteins: Expanding Nature's Inventory*, Roberts & Company Publishers, Englewood, CO.

2. Walsh, C. T., Garneau-Tsodikova, S., and Gatto, G. J., Jr. (2005) Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew Chem Int Ed Engl 44*, 7342-7372.

3. Luo, J., Li, M., Tang, Y., Laszkowska, M., Roeder, R. G., and Gu, W. (2004) Acetylation of p53 augments its site-specific DNA binding both in vitro and in vivo. *Proc Natl Acad Sci U S A 101*, 2259-2264.

4. Jansson, M., Durant, S. T., Cho, E. C., Sheahan, S., Edelmann, M., Kessler, B., and La Thangue, N. B. (2008) Arginine methylation regulates the p53 response. *Nat Cell Biol 10*, 1431-1439.

5. Latham, J. A., and Dent, S. Y. (2007) Cross-regulation of histone modifications. *Nat Struct Mol Biol 14*, 1017-1024.

6. Yang, X. J., and Seto, E. (2008) Lysine acetylation: codified crosstalk with other posttranslational modifications. *Mol Cell 31*, 449-461.

7. Gu, W., and Roeder, R. G. (1997) Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell 90*, 595-606.

8. Dastugue, B., Tichonicky, L., and Kruh, J. (1972) Effect of enzymatic phosphorylation of histone on its ability to bind to RNA. *Biochimie 54*, 1435-1441.

9. Hicke, L. (2001) Protein regulation by monoubiquitin. *Nat Rev Mol Cell Bio 2*, 195-201.

10. Thrower, J. S., Hoffman, L., Rechsteiner, M., and Pickart, C. M. (2000) Recognition of the polyubiquitin proteolytic signal. *Embo J 19*, 94-102.

11. Hoffhines, A. J., Damoc, E., Bridges, K. G., Leary, J. A., and Moore, K. L. (2006) Detection and purification of tyrosine-sulfated proteins using a novel anti-sulfotyrosine monoclonal antibody. *J Biol Chem 281*, 37877-37887.

12. Puente, L. G., and Megeney, L. A. (2008) Isolation of phosphoproteins. *Methods Mol Biol 424*, 365-372.

13. Barlev, N. A., Liu, L., Chehab, N. H., Mansfield, K., Harris, K. G., Halazonetis, T. D., and Berger, S. L. (2001) Acetylation of p53 activates transcription through recruitment of coactivators/histone acetyltransferases. *Mol Cell 8*, 1243-1254.

14. Goodman, R. H., and Smolik, S. (2000) CBP/p300 in cell growth, transformation, and development. *Genes Dev 14*, 1553-1577.

15. Tang, Y., Zhao, W., Chen, Y., Zhao, Y., and Gu, W. (2008) Acetylation is indispensable for p53 activation. *Cell 133*, 612-626.

16. Lu, X., Simon, M. D., Chodaparambil, J. V., Hansen, J. C., Shokat, K. M., and Luger, K. (2008) The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. *Nat Struct Mol Biol 15*, 1122-1124.

17. Simon, M. D., Chu, F. X., Racki, L. R., de la Cruz, C. C., Burlingame, A. L., Panning, B., Narlikar, G. J., and Shokat, K. M. (2007) The site-specific

installation of methyl-lysine analogs into recombinant histones. *Cell 128*, 1003-1012.

18. Huang, R., Holbert, M. A., Tarrant, M. K., Curtet, S., Colquhoun, D. R., Dancy, B. M., Dancy, B. C., Hwang, Y., Tang, Y., Meeth, K., Marmorstein, R., Cole, R. N., Khochbin, S., and Cole, P. A. (2010) Site-specific introduction of an acetyl-lysine mimic into peptides and proteins by cysteine alkylation. *J Am Chem Soc 132*, 9986-9987.

19. Dawson, P. E., Muir, T. W., Clark-Lewis, I., and Kent, S. B. (1994) Synthesis of proteins by native chemical ligation. *Science 266*, 776-779.

20. McGinty, R. K., Kim, J., Chatterjee, C., Roeder, R. G., and Muir, T. W. (2008) Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. *Nature 453*, 812-U812.

21. Chin, J. W., Cropp, T. A., Anderson, J. C., Mukherji, M., Zhang, Z., and Schultz, P. G. (2003) An expanded eukaryotic genetic code. *Science 301*, 964-967.

22. Furter, R. (1998) Expansion of the genetic code: site-directed p-fluoro-phenylalanine incorporation in Escherichia coli. *Protein science : a publication of the Protein Society 7*, 419-426.

23. Wang, L., Brock, A., Herberich, B., and Schultz, P. G. (2001) Expanding the genetic code of Escherichia coli. *Science 292*, 498-500.

24. Wang, L., and Schultz, P. G. (2004) Expanding the genetic code. *Angew Chem Int Ed Engl 44*, 34-66.

25. Herring, S., Ambrogelly, A., Polycarpo, C. R., and Soll, D. (2007) Recognition of pyrrolysine tRNA by the Desulfitobacterium hafniense pyrrolysyl-tRNA synthetase. *Nucleic Acids Res 35*, 1270-1278.

26. Srinivasan, G., James, C. M., and Krzycki, J. A. (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science 296*, 1459-1462.

27. Gaston, M. A., Jiang, R., and Krzycki, J. A. (2011) Functional context, biosynthesis, and genetic encoding of pyrrolysine. *Current opinion in microbiology 14*, 342-349.

28. Liu, C. C., Choe, H., Farzan, M., Smider, V. V., and Schultz, P. G. (2009) Mutagenesis and evolution of sulfated antibodies using an expanded genetic code. *Biochemistry 48*, 8891-8898.

29. Liu, C. C., and Schultz, P. G. (2006) Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nat Biotechnol 24*, 1436-1440.

30. Stone, M. J., Chuang, S., Hou, X., Shoham, M., and Zhu, J. Z. (2009) Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins. *New biotechnology 25*, 299-317.

31. Bettelheim, F. R. (1954) Tyrosine-o-sulfate in a peptide from fibrinogen. *J Am Chem Soc 76*, 2838-2839.

32. Martin, K. A., Wyatt, R., Farzan, M., Choe, H., Marcon, L., Desjardins, E., Robinson, J., Sodroski, J., Gerard, C., and Gerard, N. P. (1997) CD4-independent binding of SIV gp120 to rhesus CCR5. *Science 278*, 1470-1473.

33. Huang, C. C., Lam, S. N., Acharya, P., Tang, M., Xiang, S. H., Hussan, S. S., Stanfield, R. L., Robinson, J., Sodroski, J., Wilson, I. A., Wyatt, R., Bewley, C. A., and Kwong, P. D. (2007) Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science 317*, 1930-1934.

34. Choe, H., Li, W., Wright, P. L., Vasilieva, N., Venturi, M., Huang, C. C., Grundner, C., Dorfman, T., Zwick, M. B., Wang, L., Rosenberg, E. S., Kwong, P. D., Burton, D. R., Robinson, J. E., Sodroski, J. G., and Farzan, M. (2003) Tyrosine sulfation of human antibodies contributes to recognition of the CCR5 binding region of HIV-1 gp120. *Cell 114*, 161-170.

35. Huang, C. C., Venturi, M., Majeed, S., Moore, M. J., Phogat, S., Zhang, M. Y., Dimitrov, D. S., Hendrickson, W. A., Robinson, J., Sodroski, J., Wyatt, R., Choe, H., Farzan, M., and Kwong, P. D. (2004) Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proc Natl Acad Sci U S A 101*, 2706-2711.

36. Pouyani, T., and Seed, B. (1995) PSGL-1 recognition of P-selectin is controlled by a tyrosine sulfation consensus at the PSGL-1 amino terminus. *Cell 83*, 333-343.

37. Wilkins, P. P., Moore, K. L., McEver, R. P., and Cummings, R. D. (1995) Tyrosine sulfation of P-selectin glycoprotein ligand-1 is required for high affinity binding to P-selectin. *J Biol Chem 270*, 22677-22680.

38. Young, T., and Kiessling, L. L. (2002) A strategy for the synthesis of sulfated peptides. *Angew Chem Int Ed Engl 41*, 3449-3451.

39. Liu, C. C., Brustad, E., Liu, W., and Schultz, P. G. (2007) Crystal structure of a biosynthetic sulfo-hirudin complexed to thrombin. *J Am Chem Soc 129*, 10648-10649.

40. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science 298*, 1912-1934.

41. Tong, L., Warren, T. C., Lukas, S., Schembri-King, J., Betageri, R., Proudfoot, J. R., and Jakes, S. (1998) Carboxymethyl-phenylalanine as a replacement for phosphotyrosine in SH2 domain binding. *J Biol Chem 273*, 20238-20242.

42. Xie, J., Supekova, L., and Schultz, P. G. (2007) A genetically encoded metabolically stable analogue of phosphotyrosine in Escherichia coli. *ACS chemical biology 2*, 474-478.

43. Chin, J. W., Santoro, S. W., Martin, A. B., King, D. S., Wang, L., and Schultz, P. G. (2002) Addition of p-azido-L-phenylalanine to the genetic code of Escherichia coli. *J Am Chem Soc 124*, 9026-9027.

44. Deiters, A., Cropp, T. A., Mukherji, M., Chin, J. W., Anderson, J. C., and Schultz, P. G. (2003) Adding amino acids with novel reactivity to the genetic code of Saccharomyces cerevisiae. *J Am Chem Soc 125*, 11782-11783.

45. Serwa, R., Wilkening, I., Del Signore, G., Muhlberg, M., Claussnitzer, I., Weise, C., Gerrits, M., and Hackenberger, C. P. (2009) Chemoselective Staudinger-phosphite reaction of azides for the phosphorylation of proteins. *Angew Chem Int Ed Engl 48*, 8234-8239.

46. Radi, R. (2004) Nitric oxide, oxidants, and protein tyrosine nitration. *Proc Natl Acad Sci U S A 101*, 4003-4008.

47. Yokoyama, K., Uhlin, U., and Stubbe, J. (2010) Site-specific incorporation of 3-nitrotyrosine as a probe of pKa perturbation of redox-active tyrosines in ribonucleotide reductase. *J Am Chem Soc 132*, 8385-8397.

48. Schopfer, F. J., Baker, P. R., and Freeman, B. A. (2003) NO-dependent protein nitration: a cell signaling event or an oxidative inflammatory response? *Trends Biochem Sci 28*, 646-654.

49. Sokolovsky, M., Riordan, J. F., and Vallee, B. L. (1966) Tetranitromethane. A reagent for the nitration of tyrosyl residues in proteins. *Biochemistry 5*, 3582-3589.

50. Neumann, H., Hazen, J. L., Weinstein, J., Mehl, R. A., and Chin, J. W. (2008) Genetically encoding protein oxidative damage. *J Am Chem Soc 130*, 4028-4033.

51. Yakovlev, V. A., Bayden, A. S., Graves, P. R., Kellogg, G. E., and Mikkelsen, R. B. (2010) Nitration of the tumor suppressor protein p53 at tyrosine 327 promotes p53 oligomerization and activation. *Biochemistry 49*, 5331-5339.

52. Neumann, H., Peak-Chew, S. Y., and Chin, J. W. (2008) Genetically encoding N(epsilon)-acetyllysine in recombinant proteins. *Nat Chem Biol 4*, 232-234.

53. Neumann, H., Hancock, S. M., Buning, R., Routh, A., Chapman, L., Somers, J., Owen-Hughes, T., van Noort, J., Rhodes, D., and Chin, J. W. (2009) A method for genetically installing site-specific acetylation in recombinant histones defines the effects of H3 K56 acetylation. *Mol Cell 36*, 153-163.

54. Lammers, M., Neumann, H., Chin, J. W., and James, L. C. (2010) Acetylation regulates cyclophilin A catalysis, immunosuppression and HIV isomerization. *Nat Chem Biol 6*, 331-337.

55. Huang, Y., Russell, W. K., Wan, W., Pai, P. J., Russell, D. H., and Liu, W. (2010) A convenient method for genetic incorporation of multiple noncanonical amino acids into one protein in Escherichia coli. *Mol Biosyst 6*, 683-686.

56. Huang, Y., Wan, W., Russell, W. K., Pai, P. J., Wang, Z., Russell, D. H., and Liu, W. (2010) Genetic incorporation of an aliphatic keto-containing amino acid into proteins for their site-specific modifications. *Bioorg Med Chem Lett 20*, 878-880.

57. Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964) Acetylation + Methylation of Histones + Their Possible Role in Regulation of Rna Synthesis. *P Natl Acad Sci USA 51*, 786-&.

58. Cheung, P., and Lau, P. (2005) Epigenetic regulation by histone methylation and histone variants. *Mol Endocrinol 19*, 563-573.

59. Chuikov, S., Kurash, J. K., Wilson, J. R., Xiao, B., Justin, N., Ivanov, G. S., McKinney, K., Tempst, P., Prives, C., Gamblin, S. J., Barlev, N. A., and Reinberg, D. (2004) Regulation of p53 activity through lysine methylation. *Nature 432*, 353-360.

60. Huang, J., Perez-Burgos, L., Placek, B. J., Sengupta, R., Richter, M., Dorsey, J. A., Kubicek, S., Opravil, S., Jenuwein, T., and Berger, S. L. (2006) Repression of p53 activity by Smyd2-mediated methylation. *Nature 444*, 629-632.

61. Huang, J., Sengupta, R., Espejo, A. B., Lee, M. G., Dorsey, J. A., Richter, M., Opravil, S., Shiekhattar, R., Bedford, M. T., Jenuwein, T., and Berger, S. L. (2007) p53 is regulated by the lysine demethylase LSD1. *Nature 449*, 105-108.

62. Masatsugu, T., and Yamamoto, K. (2009) Multiple lysine methylation of PCAF by Set9 methyltransferase. *Biochemical and biophysical research communications 381*, 22-26.

63. Nolz, J. C., Gomez, T. S., and Billadeau, D. D. (2005) The Ezh2 methyltransferase complex: actin up in the cytosol. *Trends Cell Biol 15*, 514-517.

64. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D., and Patel, D. J. (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol 14*, 1025-1040.

65. He, S., Bauman, D., Davis, J. S., Loyola, A., Nishioka, K., Gronlund, J. L., Reinberg, D., Meng, F. Y., Kelleher, N., and McCafferty, D. G. (2003) Facile synthesis of site-specifically acetylated and methylated histone proteins: Reagents for evaluation of the histone code hypothesis. *P Natl Acad Sci USA 100*, 12033-12038.

66. Nguyen, D. P., Garcia Alai, M. M., Kapadnis, P. B., Neumann, H., and Chin, J. W. (2009) Genetically encoding N(epsilon)-methyl-L-lysine in recombinant histones. *J Am Chem Soc 131*, 14194-14195.

67. Groff, D., Chen, P. R., Peters, F. B., and Schultz, P. G. (2010) A genetically encoded epsilon-N-methyl lysine in mammalian cells. *Chembiochem 11*, 1066-1068.

68. Wang, Y. S., Wu, B., Wang, Z., Huang, Y., Wan, W., Russell, W. K., Pai, P. J., Moe, Y. N., Russell, D. H., and Liu, W. R. (2010) A genetically encoded photocaged Nepsilon-methyl-L-lysine. *Mol Biosyst 6*, 1557-1560.

69. Ai, H. W., Lee, J. W., and Schultz, P. G. (2010) A method to site-specifically introduce methyllysine into proteins in E. coli. *Chem Commun (Camb) 46*, 5506-5508.

70. Wang, J., Schiller, S. M., and Schultz, P. G. (2007) A biosynthetic route to dehydroalanine-containing proteins. *Angew Chem Int Ed Engl 46*, 6849-6851.

71. Guo, J., Wang, J., Lee, J. S., and Schultz, P. G. (2008) Site-specific incorporation of methyl- and acetyl-lysine analogues into recombinant proteins. *Angew Chem Int Ed Engl 47*, 6399-6401.

72. Chau, V., Tobias, J. W., Bachmair, A., Marriott, D., Ecker, D. J., Gonda, D. K., and Varshavsky, A. (1989) A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. *Science 243*, 1576-1583.

73. Nagy, V., and Dikic, I. (2010) Ubiquitin ligase complexes: from substrate selectivity to conjugational specificity. *Biological chemistry 391*, 163-169.

74. Li, W., and Ye, Y. (2008) Polyubiquitin chains: functions, structures, and mechanisms. *Cell Mol Life Sci 65*, 2397-2406.

75. Chatterjee, C., McGinty, R. K., Pellois, J. P., and Muir, T. W. (2007) Auxiliary-mediated site-specific peptide ubiquitylation. *Angew Chem Int Ed Engl 46*, 2814-2818.

76. Chatterjee, C., McGinty, R. K., Fierz, B., and Muir, T. W. (2010) Disulfide-directed histone ubiquitylation reveals plasticity in hDot1L activation. *Nat Chem Biol 6*, 267-269.

77. Chen, J., Ai, Y., Wang, J., Haracska, L., and Zhuang, Z. (2010) Chemically ubiquitylated PCNA as a probe for eukaryotic translesion DNA synthesis. *Nat Chem Biol 6*, 270-272.

78. Li, X., Fekner, T., Ottesen, J. J., and Chan, M. K. (2009) A pyrrolysine analogue for site-specific protein ubiquitination. *Angew Chem Int Ed Engl 48*, 9184-9187.

79. Virdee, S., Ye, Y., Nguyen, D. P., Komander, D., and Chin, J. W. (2010) Engineered diubiquitin synthesis reveals Lys29-isopeptide specificity of an OTU deubiquitinase. *Nat Chem Biol 6*, 750-757.

80. Couture, J. F., and Trievel, R. C. (2006) Histone-modifying enzymes: encrypting an enigmatic epigenetic code. *Curr Opin Struct Biol 16*, 753-760.

81. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell 128*, 693-705.

82. Lall, S. (2007) Primers on chromatin. *Nat Struct Mol Biol 14*, 1110-1115.

83. Blight, S. K., Larue, R. C., Mahapatra, A., Longstaff, D. G., Chang, E., Zhao, G., Kang, P. T., Green-Church, K. B., Chan, M. K., and Krzycki, J. A. (2004) Direct charging of tRNA(CUA) with pyrrolysine in vitro and in vivo. *Nature 431*, 333-335.

84. Fekner, T., Li, X., Lee, M. M., and Chan, M. K. (2009) A pyrrolysine analogue for protein click chemistry. *Angew Chem Int Ed Engl 48*, 1633-1635.

85. Mukai, T., Kobayashi, T., Hino, N., Yanagisawa, T., Sakamoto, K., and Yokoyama, S. (2008) Adding l-lysine derivatives to the genetic code of mammalian cells with engineered pyrrolysyl-tRNA synthetases. *Biochemical and biophysical research communications 371*, 818-822.

86. Nguyen, D. P., Lusic, H., Neumann, H., Kapadnis, P. B., Deiters, A., and Chin, J. W. (2009) Genetic encoding and labeling of aliphatic azides and alkynes in recombinant proteins via a pyrrolysyl-tRNA Synthetase/tRNA(CUA) pair and click chemistry. *J Am Chem Soc 131*, 8720-8721.

87. Tatsu, Y., Shigeri, Y., Ishida, A., Kameshita, I., Fujisawa, H., and Yumoto, N. (1999) Synthesis of caged peptides using caged lysine: application to the synthesis of caged AIP, a highly specific inhibitor of calmodulin-dependent protein kinase II. *Bioorg Med Chem Lett 9*, 1093-1096.

88. Wan, W., Huang, Y., Wang, Z., Russell, W. K., Pai, P. J., Russell, D. H., and Liu, W. R. (2010) A facile system for genetic incorporation of two different noncanonical amino acids into one protein in Escherichia coli. *Angew Chem Int Ed Engl 49*, 3211-3214.

89. Wiejak, S., Masiukiewicz, E., and Rzeszotarsk, B. (2001) Improved scalable syntheses of mono- and bis-urethane derivatives of ornithine. *Chem Pharm Bull 49*, 1189-1191.

90. Dyer, R. G., and Turnbull, K. D. (1999) Hydrolytic Stabilization of Protected p-Hydroxybenzyl Halides Designed as Latent Quinone Methide Precursors. *The Journal of organic chemistry 64*, 7988-7995.

91. Yamamoto, H., and Yang, J. T. (1974) The thermal induced helix--beta transition of poly(N epsilon-methyl-L-lysine) and poly(N delta-ethyl-L-ornithine) in aqueous solution. *Biopolymers 13*, 1109-1116.

92. Santoro, S. W., Wang, L., Herberich, B., King, D. S., and Schultz, P. G. (2002) An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nat Biotechnol 20*, 1044-1048.

93. Wang, L., Zhang, Z., Brock, A., and Schultz, P. G. (2003) Addition of the keto functional group to the genetic code of Escherichia coli. *Proc Natl Acad Sci U S A 100*, 56-61.

94. Xie, J., Wang, L., Wu, N., Brock, A., Spraggon, G., and Schultz, P. G. (2004) The site-specific incorporation of p-iodo-L-phenylalanine into proteins for structure determination. *Nat Biotechnol 22*, 1297-1301.

95. Tam, J. P., Yu, Q., and Miao, Z. (1999) Orthogonal ligation strategies for peptide and protein. *Biopolymers 51*, 311-332.

96. Chen, P. R., Groff, D., Guo, J., Ou, W., Cellitti, S., Geierstanger, B. H., and Schultz, P. G. (2009) A facile system for encoding unnatural amino acids in mammalian cells. *Angew Chem Int Ed Engl 48*, 4052-4055.

97. Xie, J., Liu, W., and Schultz, P. G. (2007) A genetically encoded bidentate, metal-binding amino acid. *Angew Chem Int Ed Engl 46*, 9239-9242.

98. Umehara, T., Kim, J., Lee, S., Guo, L. T., Soll, D., and Park, H. S. (2012) N-acetyl lysyl-tRNA synthetases evolved by a CcdB-based selection possess N-acetyl lysine specificity in vitro and in vivo. *FEBS Lett 586*, 729-733.

99. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Gene Dev 16*, 6-21.

100. Jones, P. A., and Takai, D. (2001) The role of DNA methylation in mammalian epigenetics. *Science 293*, 1068-1070.

101. Laird, P. W., and Jaenisch, R. (1996) The role of DNA methylation in cancer genetics and epigenetics. *Annu Rev Genet 30*, 441-464.

102. Nakao, M. (2001) Epigenetics: interaction of DNA methylation and chromatin. *Gene 278*, 25-31.

103. Bhaumik, S. R., Smith, E., and Shilatifard, A. (2007) Covalent modifications of histones during development and disease pathogenesis. *Nat Struct Mol Biol 14*, 1008-1016.

104. Turner, B. M. (2002) Cellular memory and the histone code. *Cell 111*, 285-291.

105. Fischle, W. (2008) Talk is cheap--cross-talk in establishment, maintenance, and readout of chromatin modifications. *Genes Dev 22*, 3375-3382.

106. Lee, J. S., Smith, E., and Shilatifard, A. (2010) The language of histone crosstalk. *Cell 142*, 682-685.

107. Suganuma, T., and Workman, J. L. (2008) Crosstalk among Histone Modifications. *Cell 135*, 604-607.

108. Manohar, M., Mooney, A. M., North, J. A., Nakkula, R. J., Picking, J. W., Edon, A., Fishel, R., Poirier, M. G., and Ottesen, J. J. (2009) Acetylation of histone H3 at the nucleosome dyad alters DNA-histone binding. *J Biol Chem 284*, 23312-23321.

109. Muir, T. W., Sondhi, D., and Cole, P. A. (1998) Expressed protein ligation: a general method for protein engineering. *Proc Natl Acad Sci U S A 95*, 6705-6710.

110. Shogren-Knaak, M., Ishii, H., Sun, J. M., Pazin, M. J., Davie, J. R., and Peterson, C. L. (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science 311*, 844-847.

111. Nozawa, K., O'Donoghue, P., Gundllapalli, S., Araiso, Y., Ishitani, R., Umehara, T., Soll, D., and Nureki, O. (2009) Pyrrolysyl-tRNA synthetase-tRNA(Pyl) structure reveals the molecular basis of orthogonality. *Nature 457*, 1163-1167.

112. Polycarpo, C. R., Herring, S., Berube, A., Wood, J. L., Soll, D., and Ambrogelly, A. (2006) Pyrrolysine analogues as substrates for pyrrolysyl-tRNA synthetase. *FEBS Lett 580*, 6695-6700.

113. Bernardes, G. J., Chalker, J. M., Errey, J. C., and Davis, B. G. (2008) Facile conversion of cysteine and alkyl cysteines to dehydroalanine on protein surfaces: versatile and switchable access to functionalized proteins. *J Am Chem Soc 130*, 5052-5053.

114. Arnaud, O., Koubeissi, A., Ettouati, L., Terreux, R., Alame, G., Grenot, C., Dumontet, C., Di Pietro, A., Paris, J., and Falson, P. (2010) Potent and fully noncompetitive peptidomimetic inhibitor of multidrug resistance P-glycoprotein. *Journal of medicinal chemistry 53*, 6720-6729.

115. Karanewsky, D. S., Badia, M. C., Cushman, D. W., DeForrest, J. M., Dejneka, T., Loots, M. J., Perri, M. G., Petrillo, E. W., Jr., and Powell, J. R. (1988)

(Phosphinyloxy)acyl amino acid inhibitors of angiotensin converting enzyme (ACE). 1. Discovery of (S)-1-[6-amino-2-[[hydroxy(4-phenylbutyl)phosphinyl]oxy]-1-oxohexyl]-L -proline a novel orally active inhibitor of ACE. *Journal of medicinal chemistry 31*, 204-212.

116. Foster, S. J., Kraus, R. J., and Ganther, H. E. (1986) The metabolism of selenomethionine, Se-methylselenocysteine, their selenonium derivatives, and trimethylselenonium in the rat. *Arch Biochem Biophys 251*, 77-86.

117. Brouwer, A. J., and Liskamp, R. M. J. (2005) Synthesis of Novel Dendrimeric Systems Containing NLO Ligands. *European Journal of Organic Chemistry 2005*, 487-495.

118. Michael, F. E., Sibbald, P. A., and Cochran, B. M. (2008) Palladium-catalyzed intramolecular chloroamination of alkenes. *Organic letters 10*, 793-796.

119. Young, T. S., Ahmad, I., Yin, J. A., and Schultz, P. G. (2010) An enhanced system for unnatural amino acid mutagenesis in E. coli. *Journal of molecular biology 395*, 361-374.

120. Miyake-Stoner, S. J., Refakis, C. A., Hammill, J. T., Lusic, H., Hazen, J. L., Deiters, A., and Mehl, R. A. (2010) Generating permissive site-specific unnatural aminoacyl-tRNA synthetases. *Biochemistry 49*, 1667-1677.

121. Pedelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol 24*, 79-88.

122. Yanagisawa, T., Ishii, R., Fukunaga, R., Kobayashi, T., Sakamoto, K., and Yokoyama, S. (2008) Multistep engineering of pyrrolysyl-tRNA synthetase to genetically encode N(epsilon)-(o-azidobenzyloxycarbonyl) lysine for site-specific protein modification. *Chemistry & biology 15*, 1187-1197.

123. Chalker, J. M., Gunnoo, S. B., Boutureira, O., Gerstberger, S. C., Fernandez-Gonzalez, M., Bernardes, G. J. L., Griffin, L., Hailu, H., Schofield, C. J., and Davis, B. G. (2011) Methods for converting cysteine to dehydroalanine on peptides and proteins. *Chemical Science 2*, 1666-1676.

124. Reid, G. E., and McLuckey, S. A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *Journal of mass spectrometry : JMS 37*, 663-675.

125. Zhang, J., Dong, X., Hacker, T. A., and Ge, Y. (2010) Deciphering modifications in swine cardiac troponin I by top-down high-resolution tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry 21*, 940-948.

126. Zinnel, N. F., Pai, P. J., Russell, W. K., and Russell, D. H. (2012) *in preparation*.

127. Liu, C. C., and Schultz, P. G. (2010) Adding new chemistries to the genetic code. *Annu Rev Biochem 79*, 413-444.

128. Liu, W., Brock, A., Chen, S., and Schultz, P. G. (2007) Genetic incorporation of unnatural amino acids into proteins in mammalian cells. *Nature methods 4*, 239-244.

129. Wu, N., Deiters, A., Cropp, T. A., King, D., and Schultz, P. G. (2004) A genetically encoded photocaged amino acid. *J Am Chem Soc 126*, 14306-14307.

130. Gautier, A., Nguyen, D. P., Lusic, H., An, W., Deiters, A., and Chin, J. W. (2010) Genetically encoded photocontrol of protein localization in mammalian cells. *J Am Chem Soc 132*, 4086-4088.

131. Yanagisawa, T., Ishii, R., Fukunaga, R., Kobayashi, T., Sakamoto, K., and Yokoyama, S. (2008) Crystallographic studies on multiple conformational states of active-site loops in pyrrolysyl-tRNA synthetase. *Journal of molecular biology 378*, 634-652.

132. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M., and Chin, J. W. (2010) Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature 464*, 441-444.

133. Miyake-Stoner, S. J., Miller, A. M., Hammill, J. T., Peeler, J. C., Hess, K. R., Mehl, R. A., and Brewer, S. H. (2009) Probing protein folding using site-specifically encoded unnatural amino acids as FRET donors with tryptophan. *Biochemistry 48*, 5953-5962.

134. Guo, J., Melancon, C. E., 3rd, Lee, H. S., Groff, D., and Schultz, P. G. (2009) Evolution of amber suppressor tRNAs for efficient bacterial production of proteins containing nonnatural amino acids. *Angew Chem Int Ed Engl 48*, 9148-9151.

135. Chalker, J. M., Wood, C. S., and Davis, B. G. (2009) A convenient catalyst for aqueous and protein Suzuki-Miyaura cross-coupling. *J Am Chem Soc 131*, 16346-16347.

136. Anderson, J. C., Wu, N., Santoro, S. W., Lakshman, V., King, D. S., and Schultz, P. G. (2004) An expanded genetic code with a functional quadruplet codon. *Proc Natl Acad Sci U S A 101*, 7566-7571.

137. Wang, L., and Schultz, P. G. (2002) Expanding the genetic code. *Chem Commun (Camb)*, 1-11.

138. Melancon, C. E., 3rd, and Schultz, P. G. (2009) One plasmid selection system for the rapid evolution of aminoacyl-tRNA synthetases. *Bioorg Med Chem Lett 19*, 3845-3847.

139. Zhang, D., Vaidehi, N., Goddard, W. A., 3rd, Danzer, J. F., and Debe, D. (2002) Structure-based design of mutant Methanococcus jannaschii tyrosyl-tRNA synthetase for incorporation of O-methyl-L-tyrosine. *Proc Natl Acad Sci U S A 99*, 6579-6584.

140. Tanrikulu, I. C., Schmitt, E., Mechulam, Y., Goddard, W. A., 3rd, and Tirrell, D. A. (2009) Discovery of Escherichia coli methionyl-tRNA synthetase mutants for efficient labeling of proteins with azidonorleucine in vivo. *Proc Natl Acad Sci U S A 106*, 15285-15290.

141. Wang, Y. S., Russell, W. K., Wang, Z., Wan, W., Dodd, L. E., Pai, P. J., Russell, D. H., and Liu, W. R. (2011) The de novo engineering of pyrrolysyl-tRNA synthetase for genetic incorporation of L-phenylalanine and its derivatives. *Mol Biosyst 7*, 714-717.

142. Kavran, J. M., Gundllapalli, S., O'Donoghue, P., Englert, M., Soll, D., and Steitz, T. A. (2007) Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation. *Proc Natl Acad Sci U S A 104*, 11268-11273.

143. Zhang, M., Lin, S., Song, X., Liu, J., Fu, Y., Ge, X., Fu, X., Chang, Z., and Chen, P. R. (2011) A genetically incorporated crosslinker reveals chaperone cooperation in acid resistance. *Nat Chem Biol 7*, 671-677.

144. Kolb, H. C., Finn, M. G., and Sharpless, K. B. (2001) Click chemistry: Diverse chemical function from a few good reactions. *Angew Chem Int Edit 40*, 2004-+.

145. Wu, B., Wang, Z., Huang, Y., and Liu, W. R. (2012) Manuscript inpreparation.

146. Takimoto, J. K., Dellas, N., Noel, J. P., and Wang, L. (2011) Stereochemical basis for engineered pyrrolysyl-tRNA synthetase and the efficient in vivo incorporation of structurally divergent non-native amino acids. *ACS chemical biology 6*, 733-743.

147. Hancock, S. M., Uprety, R., Deiters, A., and Chin, J. W. (2010) Expanding the genetic code of yeast for incorporation of diverse unnatural amino acids via a pyrrolysyl-tRNA synthetase/tRNA pair. *J Am Chem Soc 132*, 14819-14824.

148. Greiss, S., and Chin, J. W. (2011) Expanding the genetic code of an animal. *J Am Chem Soc 133*, 14196-14199.

149. Xie, J., and Schultz, P. G. (2006) A chemical toolkit for proteins--an expanded genetic code. *Nature reviews. Molecular cell biology 7*, 775-782.

150. Wang, L., Xie, J., and Schultz, P. G. (2006) Expanding the genetic code. *Annu Rev Biophys Biomol Struct 35*, 225-249.

151. Bose, M., Groff, D., Xie, J., Brustad, E., and Schultz, P. G. (2006) The incorporation of a photoisomerizable amino acid into proteins in E. coli. *J Am Chem Soc 128*, 388-389.

152. Brustad, E., Bushey, M. L., Lee, J. W., Groff, D., Liu, W., and Schultz, P. G. (2008) A genetically encoded boronate-containing amino acid. *Angew Chem Int Ed Engl 47*, 8220-8223.

153. Brustad, E. M., Lemke, E. A., Schultz, P. G., and Deniz, A. A. (2008) A general and efficient method for the site-specific dual-labeling of proteins for single molecule fluorescence resonance energy transfer. *J Am Chem Soc 130*, 17664-17665.

154. Chin, J. W., Cropp, T. A., Chu, S., Meggers, E., and Schultz, P. G. (2003) Progress toward an expanded eukaryotic genetic code. *Chemistry & biology 10*, 511-519.

155. Chin, J. W., Martin, A. B., King, D. S., Wang, L., and Schultz, P. G. (2002) Addition of a photocrosslinking amino acid to the genetic code of Escherichiacoli. *Proc Natl Acad Sci U S A 99*, 11020-11024.

156. Deiters, A., Groff, D., Ryu, Y., Xie, J., and Schultz, P. G. (2006) A genetically encoded photocaged tyrosine. *Angew Chem Int Ed Engl 45*, 2728-2731.

157. Deiters, A., and Schultz, P. G. (2005) In vivo incorporation of an alkyne into proteins in Escherichia coli. *Bioorg Med Chem Lett 15*, 1521-1524.

158. Fleissner, M. R., Brustad, E. M., Kalai, T., Altenbach, C., Cascio, D., Peters, F. B., Hideg, K., Peuker, S., Schultz, P. G., and Hubbell, W. L. (2009) Site-directed

spin labeling of a genetically encoded unnatural amino acid. *Proc Natl Acad Sci U S A 106*, 21637-21642.

159. Schultz, K. C., Supekova, L., Ryu, Y., Xie, J., Perera, R., and Schultz, P. G. (2006) A genetically encoded infrared probe. *J Am Chem Soc 128*, 13984-13985.

160. Taskent-Sezgin, H., Chung, J., Patsalo, V., Miyake-Stoner, S. J., Miller, A. M., Brewer, S. H., Mehl, R. A., Green, D. F., Raleigh, D. P., and Carrico, I. (2009) Interpretation of p-cyanophenylalanine fluorescence in proteins in terms of solvent exposure and contribution of side-chain quenchers: a combined fluorescence, IR and molecular dynamics study. *Biochemistry 48*, 9040-9046.

161. Tippmann, E. M., Liu, W., Summerer, D., Mack, A. V., and Schultz, P. G. (2007) A genetically encoded diazirine photocrosslinker in Escherichia coli. *Chembiochem 8*, 2210-2214.

162. Tsao, M. L., Summerer, D., Ryu, Y., and Schultz, P. G. (2006) The genetic incorporation of a distance probe into proteins in Escherichia coli. *J Am Chem Soc 128*, 4572-4573.

163. Turner, J. M., Graziano, J., Spraggon, G., and Schultz, P. G. (2006) Structural plasticity of an aminoacyl-tRNA synthetase active site. *Proc Natl Acad Sci U S A 103*, 6483-6488.

164. Zeng, H., Xie, J., and Schultz, P. G. (2006) Genetic introduction of a diketone-containing amino acid into proteins. *Bioorg Med Chem Lett 16*, 5356-5359.

165. Xie, J., and Schultz, P. G. (2005) An expanding genetic code. *Methods 36*, 227-238.

166. Zhang, Z., Smith, B. A., Wang, L., Brock, A., Cho, C., and Schultz, P. G. (2003) A new strategy for the site-specific modification of proteins in vivo. *Biochemistry 42*, 6735-6746.

167. Wang, Y.-S., Fang, X., Wallace, A., Wu, B., and Liu, W. R. (2012) A Rationally Designed Pyrrolysyl-tRNA Synthetase Has a Broad Substrate Spectrum. *in preparation*.

168. Wang, Y. S., Fang, X., Wallace, A. L., Wu, B., and Liu, W. R. (2012) A rationally designed pyrrolysyl-tRNA synthetase mutant with a broad substrate spectrum. *J Am Chem Soc 134*, 2950-2953.

169. Wang, Y.-S., Wang, Z., Wan, W., Lee, Y.-J., Dodd, L. E., Russell, W. K., Pai, P. J., Russell, D. H., and Liu, W. R. (2010) The de novo engineering of pyrrolyl-tRNA synthetase for genetic incorporation of L-phenylalanine and its derivatives. *Mol Biosyst submitted*.

170. Andrews, B. T., Gosavi, S., Finke, J. M., Onuchic, J. N., and Jennings, P. A. (2008) The dual-basin landscape in GFP folding. *Proc Natl Acad Sci U S A 105*, 12283-12288.

VITA

Yane-Shih Wang received his Bachelor of Science and Master of Science degrees from National Cheng-Kung University, Taiwan in 1999 and 2001. He entered the Ph.D. program at Texas A&M University in 2007, where he joined the group directed by Dr. Wenshe Liu from Department of Chemistry. His research interests include bioinorganic chemistry, as well as the expanding genetic code for protein modifications of epigenetic researchs. Yane-Shih Wang graduated with his Ph.D. degree in Chemistry in August, 2012.

Yane-Shih Wang may be reached at No. 1-7, Neighborhood 15, Tayuan Twonship, Taoyuan, 337 Taiwan. His e-mail address is ericyswang@gmail.com.