A molecular engineering toolbox for the structural biologist

Galia T. Debelouchina and Tom W. Muir*

Department of Chemistry, Princeton University, Princeton, NJ 08540, USA

Quarterly Reviews of Biophysics (2017), 50, e7, page 1 of 41 doi:10.1017/S0033583517000051

Abstract. Exciting new technological developments have pushed the boundaries of structural biology, and have enabled studies of biological macromolecules and assemblies that would have been unthinkable not long ago. Yet, the enhanced capabilities of structural biologists to pry into the complex molecular world have also placed new demands on the abilities of protein engineers to reproduce this complexity into the test tube. With this challenge in mind, we review the contents of the modern molecular engineering toolbox that allow the manipulation of proteins in a site-specific and chemically well-defined fashion. Thus, we cover concepts related to the modification of cysteines and other natural amino acids, native chemical ligation, intein and sortase-based approaches, amber suppression, as well as chemical and enzymatic bio-conjugation strategies. We also describe how these tools can be used to aid methodology development in X-ray crystallography, nuclear magnetic resonance, cryo-electron microscopy and in the studies of dynamic interactions. It is our hope that this monograph will inspire structural biologists and protein engineers alike to apply these tools to novel systems, and to enhance and broaden their scope to meet the outstanding challenges in understanding the molecular basis of cellular processes and disease.

1. Introduction 2
2. Advances and challenges in structural biology 3
   2.1. X-ray crystallography 3
   2.2. Cryo-EM 4
   2.3. NMR spectroscopy 4
3. Molecular engineering toolbox for complex biological samples 4
   3.1. Cysteine chemistry 5
   3.2. Chemical modification of other amino acids 6
   3.3. Native chemical ligation 7
   3.4. Inteins 8
      3.4.1. The intein splicing mechanism 9
      3.4.2. Applications in protein engineering 9
      3.4.3. Toward fast and promiscuous inteins 10
   3.5. Sortases 12
3.6. Genetic code expansion 14
   3.6.1. Amber codon suppression in living cells 15
   3.6.2. The amber suppression toolbox 16
   3.6.3. Limitations and future directions 17
   3.7. Chemical modification of unnatural amino acids 18
   3.8. Enzymatic bioconjugation approaches 20
4. Protein engineering approaches for tackling outstanding challenges in structural biology 20

* Author for correspondence: T. W. Muir, Department of Chemistry, Princeton University, Princeton, NJ 08540, USA. Email: muir@princeton.edu


https://doi.org/10.1017/S0033583517000051 Published online by Cambridge University Press
1. Introduction

Since the first crystal structure of myoglobin (Kendrew et al. 1958), the three-dimensional (3D) reconstruction image of T4 bacteriophage tails by electron microscopy (EM) (De Rosier & Klug, 1968), and the solution nuclear magnetic resonance (NMR) structure of proteinase inhibitor IIA (Williamson et al. 1985), structural biology has made tremendous strides toward revealing intimate atomic level details that guide the function of biological molecules. We live at a time when we know the structures of more than 120,000 (and counting – Source: http://www.rcsb.org/pdb/statistics/holdings.do) biological molecules, when we can visualize the inner workings of the ribosome (Ben-Shem et al. 2011), or the nucleosome interactions that preserve the integrity and identity of our genome (Luger et al. 1997). At the same time, advances in instrumentation engineering have pushed the frontiers of structural biology methodologies and have allowed experiments and accomplishments that would have been unthinkable 30 years ago. Thus, it is now possible to record high-resolution movies of fast protein motions using X rays (Tenboer et al. 2014), obtain cryo-EM electron density maps at sub-Å resolution (Campbell et al. 2015; Merk et al. 2016), or record multidimensional NMR spectra of protein crystals (Igumenova et al. 2004). Yet, the task in front of the structural biologist is getting harder and harder. The wealth of structural, biochemical and biological data has revealed that many mammalian cellular proteins are very large (>50 kD) (Brocchieri & Karlin, 2005), that they are often part of complex assemblies composed of many interchangeable molecular players, and that their function is often defined and regulated by an intricate layer of post-translational modifications (PTMs). In addition, many disease-related biological macromolecules do not have a defined secondary or tertiary structure at all, and function, instead, through intrinsic disorder and numerous weak, transient interactions (Hyman et al. 2014; Tompa, 2012). To make sense of this complicated, multilayered and often chaotic biological world, the structural biologist will become more and more dependent on the ability of protein engineers to faithfully and efficiently reproduce this complexity in the test tube.

Analogous to the advances in instrumentation design and engineering that have allowed structural biology to travel far, the tools of protein engineering have also become much more sophisticated, efficient and ultimately broader in scope over time. It is now possible to routinely synthesize polypeptide chains that are 50 amino acids long, to stitch them together into much longer chains without leaving any chemical scars (Dawson et al. 1994), and to decorate them with PTMs, biophysical probes and chemical moieties that perturb or enhance their function. It is also possible to ‘persuade’ the cellular protein synthesis machinery to produce polypeptide chains incorporating completely unnatural amino acids, thus expanding the genetic code of engineered living organisms (Wang et al. 2001). The current protein engineering toolbox contains many biocompatible chemical reactions, proteins with unique polypeptide ‘stitching’ abilities, and concepts and ideas that might ultimately prove essential in solving the interesting and relevant structural biology problems of today (Fig. 1). As the structural biologist might not be aware of all the current developments in protein chemistry, we intend this review as a resource that describes the state-of-the-art protein engineering tools, keeping an eye on the past and future to provide context for their limitations and the exciting new possibilities that undoubtedly lie ahead. We start with a very brief overview of recent advances in X-ray crystallography, cryo-EM and NMR, and outline challenges where the tools of protein engineering might be the most impactful. We then describe the contents of the molecular engineering toolbox that allow the construction of large modified proteins and complex macromolecular assemblies. We continue with a discussion of the concepts and ideas that directly concern structural biology methodology development. Our monograph ends with an outlook toward emerging trends in structural and chemical biology and exciting new developments that will guide the two fields in the future.
2. Advances and challenges in structural biology

2.1 X-ray crystallography

The workhorse of structural biology, X-ray crystallography, is more than 100 years old and has contributed nearly 90% of the macromolecular structures deposited in the PDB (http://www.rcsb.org/pdb/statistics/holdings.do). Protein engineering has long been part of the everyday life of the crystallographer as mutations, truncations and fusion proteins are often required to ‘trick’ proteins into adopting a crystal form. There are many crystal structures of proteins containing PTMs or their analogs, and chemical approaches are often used to trap interesting functional states, stabilize dynamic interactions or aid the formation of crystals (e.g. racemic crystallography (Yeates & Kent, 2012)). Yet, the voracious need to test hundreds if not thousands of single crystal growth conditions has certainly challenged the protein chemist to optimize her tools and deliver relevant samples with much greater yields. Recent instrumentation developments such as X-ray free-electron lasers may potentially alleviate this need as these sources allow the acquisition of room-temperature data from easier to obtain micro-, nano- and 2D crystals (Neutze et al. 2015). Currently, there is also a growing demand for the construction of...
homogeneous, chemically well defined and stable complex biological assemblies such as those relevant for chromatin biology, for example.

### 2.2 Cryo-EM

Exciting developments in the last few years have propelled cryo-EM into the spotlight and turned this method into a mainstream and vital structural biology technique that can achieve crystallographic resolution (Cheng, 2015; Nogales, 2016). The commercialization of direct electron-detection cameras has allowed the acquisition of images with higher contrast and fast readouts that can overcome beam-induced motion and radiation damage (Brilot et al., 2012; McMullan et al., 2009). On the other hand, improvements in data analysis approaches have made it possible to characterize heterogeneous samples and even rare structural states (Fernandez et al., 2013; Scheres, 2012). Coupled with other cryo-EM advantages (no need for crystallization and only small amounts of sample required), these advances have made it possible to obtain subnanometer (and in some cases <3 Å) resolution maps of integral membrane proteins (Matthies et al., 2016), biological polymers (von der Ecken et al., 2015), chromatin (Song et al., 2014), as well as biological assemblies such as the transcription and translation initiation complexes (Fernandez et al., 2013; He et al., 2013; Plaschka et al., 2016). In this context, protein chemistry and engineering can have a tremendous benefit for the cryo-EM structural biologist in the design and construction of relevant biological samples such as post-translationally modified proteins. Perhaps more importantly, however, chemical biology approaches such as cross-linking can allow the preparation of samples that are more robust and do not fall apart during sample vitrification. The addition of cross-linkers can also be extremely useful in integrated cryo-EM/mass-spec structural approaches for samples where the high-resolution identification of protein–protein interfaces might not be possible (Leitner et al., 2016).

### 2.3 NMR spectroscopy

NMR spectroscopy detects the magnetic properties of nuclei in molecules, which in turn provide a window into their surrounding chemical environment. Uniquely suited to probe molecular structure and dynamics in solution at physiologically relevant conditions (temperature, pH and salts) and ultimately non-destructive in its readout, NMR spectroscopy has long been battered by its intrinsically low sensitivity. The introduction of ‘NMR-visible’ isotopic labels into biological macromolecules has become a standard practice in the field, and efficient molecular engineering approaches that allow the installation of nuclear isotopes at specific positions within the polypeptide or polynucleotide chain are highly desirable. Recent advances such as (methyl)-TROSY and dark-state exchange saturation transfer experiments have pushed the molecular size limits of solution NMR into the MDa regime (Fawzi et al., 2011; Pervushin et al., 1997; Tugarinov et al., 2003), while the rapid instrumentation and pulse sequence developments in magic angle spinning NMR have made it possible to pursue the structures of large biological polymers such as amyloid fibrils (Fitzpatrick et al., 2013; Lu et al., 2013; Wasmert et al., 2008), bacterial secretion needles (Loquet et al., 2012), membrane proteins embedded in their native lipid environments (Cady et al., 2010; Wang et al., 2013), or even the molecular composition of bones (Chow et al., 2014). Thus, molecular engineering approaches can have a profound impact on the assembly of homogeneous, isotopically labeled and yet native substrates for structural investigation by in vitro NMR. Also, uniquely suited to probe structure and dynamics in the cellular milieu (Burz et al., 2006; Frederick et al., 2015; Inomata et al., 2009; Sakakibara et al., 2009), NMR spectroscopy can benefit tremendously from chemical and molecular biology techniques that allow the specific isotopic labeling of macromolecules in the cell.

### 3. Molecular engineering toolbox for complex biological samples

Before we delve into the chemistry, it is important to note that the methods described below complement the well-established molecular biology framework that allows the manipulation of protein sequences at the genetic level. Such manipulations can now be achieved in several different organisms ranging from bacteria (Escherichia coli, Lactococcus lactis), yeast (Saccharomyces cerevisiae, Pichia pastoris), insect cells, and stable mammalian expression cell lines such as HEK293 and CHO. Thus, we will start with review of the methodologies that can selectively modify natural amino acids introduced at specific protein positions with site-directed mutagenesis. We will then describe tools that can be used to ligate modified peptides and proteins into longer polypeptide chains, including native chemical ligation (NCL), inteins and transpeptidases. We will then discuss the molecular engineering toolbox afforded by incorporation of unnatural amino acids by amber suppression. These chemical and genetic tools give chemists the ability to position bio-orthogonal reactive handles into polypeptide chains with extraordinary precision and control, and we will end this section with discussion of bioconjugation approaches that take advantage of this rapidly evolving expertise.
3.1 Cysteine chemistry

The field of protein chemistry would have been very different in a world without cysteines. Indeed, many of the protein chemistry tools that exist today have been developed to exploit the reactivity of the cysteine sulfhydryl group that uniquely stands out among a sea of other protein side-chains. Cysteines are relatively rare in Nature (<2% abundance), and their high nucleophilicity makes them good candidates for the development of selective chemical reactions that work well in protein compatible conditions (aqueous solution, physiological pH and temperature). Generally, these reactions can be divided into three types: (1) cysteine alkylation, (2) oxidations and (3) desulfurization reactions, each providing a unique way to exploit the reactivity of this amino acid and a pathway to build proteins with distinct and desirable properties (reviewed in (Chalker et al. 2009; Spicer & Davis, 2014)) (Fig. 2).

Alkylation reactions have long been used to modify cysteine containing proteins and rely on familiar reagents such as iodoacetamide and maleimide. By careful control of the buffer pH, these electrophiles can selectively react with the cysteine sulfhydryl group and tether additional functionalities to the protein of interest. Iodoacetamides and other α-halocarboxyls, for example, have been used to attach carbohydrate moieties to proteins and create glycoprotein mimics (Macmillan et al. 2001; Tey et al. 2010). Maleimides, on the other hand, are commercially available and user-friendly means to add spectroscopic probes to proteins, including fluorescent, electron paramagnetic resonance (EPR) and paramagnetic relaxation enhancement (PRE) labels. Aminomethylation of cysteines also provides a cheap and convenient way to generate lysine methylation analogs, available in the mono-, di- and tri-methylation states (Simon et al. 2007). In recent years, the chemist’s attention has turned toward the development of more advanced cysteine bioconjugation protocols that circumvent, for example, the reversibility of maleimide additions in the presence of external bases and thiols (Lyon et al. 2014), or provide the opportunity to conjugate additional functionalities (such as aryl groups) under an expanded range of reaction conditions (Vinogradova et al. 2015).

Cysteine oxidation in the protein context is usually associated with the formation of disulfide bonds, a unique structural transformation with often dramatic consequences to protein function. In the test tube, disulfide bonds are very easy to build – all that is required is basic pH and exposure to air. Furthermore, the reaction is fast and does not require large excess of reagents. Therefore, it comes as no surprise that protein chemists often exploit disulfide bonds as means to attach useful functionalities to proteins of interest. The challenging aspect of such protocols is to ensure that only the desired disulfide bonds are formed and the final product is not a mixture of homo- and heterodimer species. A common strategy used to alleviate this problem relies on controlled cysteine activation and disulfide exchange based on the lower pKa of aromatic thiols. In this case, a selected cysteine side-chain can be activated (and protected) at low pH with the aromatic thiol, and upon addition of the other thiol-containing component and increase in pH, the aromatic disulfide bond is exchanged with the desired connectivity (Pollack & Schultz, 1989; Rabanal et al. 1996). This strategy has found diverse applications ranging from the design of cytochrome peptides (Rabanal et al. 1996) to the tethering of ubiquitin moieties to various proteins (Chatterjee et al. 2010; Chen et al. 2010a; Meier et al. 2012). Other activating molecules include methanethiosulfonates, glycosyl and allicy thiols with applications in protein glycosylation, prenylation and sulfenylation (Gamblin et al. 2008; Grayson et al. 2005; van Kasteren et al. 2007).

While the formation of disulfide bonds is fast and facile, they are just as easily destroyed in the presence of common biochemical reducing reagents or in the cellular environment where glutathione is present at high concentrations. If compatible linkages are desired, it is possible to convert the disulfide to a thioether bond, which is stable under reducing conditions. This is essentially a desulfurization reaction that proceeds with the formation of a dehydroalanine intermediate. Dehydroalanines, on the other hand, can be useful stepping stones to a vast number of protein PTMs, reactions that will be discussed in more detail in Section 3.7. Another key desulfurization reaction involves the conversion of cysteine to alanine. This transformation, discussed in more detail in Section 3.3, is particularly important for NCL, as it allows the construction of polypeptide chains without cysteine ‘scars’.

Cysteines are relatively rare in polypeptide chains and are usually essential for protein function. Thus, the outstanding challenge for the protein chemist is to find new approaches and reaction conditions that target only the desired residues in a polypeptide chain. Many of these efforts are focused on the search of suitable peptide sequences that can provide the necessary amino acid context for tuning side-chain reactivity. For example, Tsien and co-workers have developed tetra-cysteine motifs that can be selectively targeted with biarsenic reagents, including in living cells (Griffin et al. 1998). Pentetle and co-workers, on the other hand, have reported the so-called ‘α-clamp’ sequence (Phe-Pro-Cys-Ph) that reacts preferentially with perfluoroaromatic moieties in aqueous solvents (Zhang et al. 2016a). Excellent selectivity can be obtained by replacing cysteine with selenocysteine (reviewed in (Metanis et al. 2009; Yoshizawa & Bock, 2009), a rare natural amino acid that shares many of the desirable properties of the cysteine side-chain, yet is more acidic and substantially more reactive at lower pH (selenol pKa is ∼5.2 versus 8.3 for thiols).
3.2 Chemical modification of other amino acids

In addition to cysteine, several other natural amino acids present functional groups that can be targeted for protein modification (reviewed in (Basle et al. 2010; Spicer & Davis, 2014)). These include lysine, tyrosine, arginine, glutamate, aspartate, serine, threonine, methionine, histidine and tryptophan side-chains, as well as N-terminal amines or C-terminal carboxyls (Basle et al. 2010; Hu et al. 2016; Lin et al. 2017). The modification of these residues is less precise as they are significantly more abundant than cysteine, yet, some selectivity can be achieved in a context-dependent manner. The primary amines of lysine side-chains are often popular targets due to favorable reaction kinetics that can be achieved with activated esters such as N-hydroxysuccinimide (Kalkhof & Sinz, 2008) (Fig. 3a), isothiocyanates (Nakamura et al. 2009), or aldehydes in a reductive alkylation reaction with sodium cyanoborohydride (Jentoft & Dearborn, 1979; McFarland & Francis, 2005). Since these reactions usually modify all accessible lysine side-chains, they can be used in applications that require multiple modifications (e.g. therapeutic protein conjugates) or in protein cross-linking for mass spectrometry analysis (Holding, 2015). An example of a more discriminating lysine-based modification strategy involves the 6α-aza-electrocyclization reaction with unsaturated aldehyde esters that targets solvent-accessible lysine residues with excellent selectivity and reaction kinetics, and has been used for the attachment of fluorescent or positron emission tomography probes (Tanaka et al. 2011). Yet, further selectivity can be achieved by discriminating the lower pKa of N-terminal amino groups (∼8) from the pKa of the ε-amine of a lysine side-chain (∼10−5). The identity of the N-terminal amino acid may further change the reactivity of the α-amine, although more general modification strategies have been developed. For example, functionalized ketenes can preferentially react with the α-amine in the context of 13 different N-terminal amino acids (Chen et al. 2012), while 2-pyridinecarboxyaldehydes provide efficient and specific N-terminal labeling for all amino acids except proline (MacDonald et al. 2015) (Fig. 3b). The unique reactivity of the N-terminus is also central to the mechanism of native chemical ligation, discussed in Section 3.3.

A set of chemical reactions has also been developed for the specific modification of the aromatic electron-rich tyrosine side-chain. The selectivity of tyrosine-focused reactions usually exploits the low exposure of this residue on native protein surfaces and its susceptibility to oxidation (ElSohly & Francis, 2015; Seim et al. 2011). In particular, modification of the ortho-position
can be achieved with diazonium salts (Schlick et al. 2005) or the three-component Mannich-type reaction with aldehydes and anilines (Joshi et al. 2004; McFarland et al. 2008) (Fig. 3c). While some side-reactions have been known to occur, it has been possible to optimize the selectivity of the Mannich-type reaction to achieve efficient modification of a single tyrosine residue with an EPR spin label in the presence of disulfide and tryptophan functionalities (Mileo et al. 2013). Proteins containing few surface exposed tyrosine residues can also be modified at low concentrations (5 µM) in aqueous solvents with π-allylpalladium reagents (Tilley & Francis, 2006).

Carboxyl groups in glutamate and aspartate side-chains can be targeted with water-soluble carbodiimides such as N-ethyl-3-N',N'-dimethylaminopropylcarbodiimide (EDC) (Fig. 3d). In what is essentially a standard peptide coupling reaction, this reagent pairs carboxyls and amines in a covalent amide bond under aqueous conditions in a pH-dependent manner (Gilles et al. 1990). Commonly employed as a cross-linking reagent for the identification of protein–protein interactions in biochemical assays and mass spectrometry, EDC can also be used to modify protein assemblies such as viral capsids with a variety of functional and biophysical probes (Schlick et al. 2005). While EDC-based reactions do not discriminate between side-chain and C-terminal carboxyl groups, unique reactivity at the C-terminus can be generated by replacing the carboxyl functionality with a thioester, as discussed in the sections below.

3.3 Native chemical ligation

The unique reactivity of the cysteine side-chain and the positional control afforded by the protein N- and C-termini are at the heart of a simple, yet powerful, chemical reaction that allows the construction of native polypeptide chains from peptide-building blocks. This methodology, called NCL (Dawson et al. 1994), grants unprecedented control over the chemical
functionalities that can be introduced in a polypeptide chain and is a fundamental tool in the repertoire of protein chemists. NCL simply requires two components with the following properties: (1) a peptide with a C-terminal thioester, and (2) a peptide with an N-terminal cysteine residue (or functional equivalent). When these components are mixed, the cysteine side-chain attacks the thioester of the other peptide, resulting in the formation of an intermolecular thioester (Fig. 4). This intermediate rapidly and irreversibly rearranges to form a native peptide bond, thus ligating the two fragments. Since both peptides can be produced by solid-phase peptide synthesis, virtually any modified natural or unnatural amino acid, spectroscopic probe, cross-linker or isotopic label can be incorporated at a well-defined position into the ligated sequence (Dawson & Kent, 2000). Moreover, in contrast to biosynthetic approaches such as amber suppression (see Section 3.6), there is no practical restriction on the number, or type, of unnatural amino acids that can be introduced – this point is driven home by the total synthesis of enantiomeric proteins composed of all D-amino acids (Mandal et al. 2012). The remarkable chemoselectivity of the NCL reaction means that ligations can be performed under aqueous conditions in the presence of internal cysteine residues. NCL is also compatible with protein denaturants or detergents allowing the construction of aggregation prone polypeptide chains or even polytopic membrane proteins (Hejjaoui et al. 2012; Kwon et al. 2015; Valliyaveetil et al. 2006). The fragment bearing the N-terminal cysteine can also be produced recombinantly, usually preceded by a cleavable tag or fusion protein to avoid N-terminal cysteine processing complications in bacteria. Alternatively, the thiostep component can be generated recombinantly with the help of proteins called inteins (Muir et al. 1998) (see Section 3.4). It is also possible to perform sequential NCL reactions, allowing three or more building blocks to be assembled in a regioselective fashion (Mandal et al. 2012; Torbeev & Hilvert, 2013). These advances thus allow the construction of considerably larger polypeptide chains that would be accessible from two synthetic peptides alone.

While NCL results in a native peptide bond, the required cysteine may still produce an unwanted ‘scar’ at the ligation junction – i.e. in cases where the target does not contain a native cysteine at an appropriate ligation point. If the final ligation product does not contain other cysteine residues, then the undesirable cysteine can be converted to alanine in a subsequent desulfurization step. Options for desulfurization reactions include reduction with metals or radical desulfurization mediated by the familiar tris(2-carboxyethyl)phosphine (TCEP) reagent (Dawson, 2011; Wan & Danishefsky, 2007; Yan & Dawson, 2001). Selective desulfurization reactions are also possible in the presence of other cysteine residues, provided that compatible protecting groups are used on the non-targeted cysteine side-chains (Ficht et al. 2007; Pentelute & Kent, 2007). Alternatively, selective ligation and desulfurization of selenocysteine can provide additional sequence positional control (Hondal et al. 2001; Reddy et al. 2016). Today, desulfurization has become a routine part of NCL protocols and the abundant alanine residue is a commonplace choice for a ligation site. To expand the junction amino acid set, a more advanced strategy involves the incorporation of β- and γ-thio amino acids. These moieties replace the cysteine-like residue at the N-terminal position, and provide a reactive thiol for trans-thioesterification. Desulfurization protocols can then produce the desired native side-chain. While ligations at such sites proceed more slowly due to increased steric hindrance, NCL can now be performed at phenylalanine (Crich & Banerjee, 2007), valine (Chen et al. 2008; Haase et al. 2008), leucine (Harpaz et al. 2010; Tan et al. 2010), threonine (Chen et al. 2010b), lysine (El Oualid et al. 2010; Kumar et al. 2009; Yang et al. 2009), proline (Shang et al. 2011), glutamine (Siman et al. 2012), arginine (Malins et al. 2013), tryptophan (Malins et al. 2014), aspartate (Thompson et al. 2013), glutamate (Cergol et al. 2014) and asparagine (Sayers et al. 2015). Work also continues on more streamlined ligation/desulfurization approaches that remove purification steps and increase the yield of ligated products (Moyal et al. 2013; Thompson et al. 2014).

The preceding discussion highlights just a few of the many refinements and extensions the NCL strategy has undergone since its introduction over 20 years ago (Harrmand et al. 2014; Malins & Payne, 2015). As a consequence of this massive effort, the technique has become a central tool in protein science, having been applied to literally hundreds of protein targets. Of particular relevance here, it has provided the raw materials for numerous structural biology studies that have employed a broad range of spectroscopic or crystallographic methods (Fig. 5) (Grosse et al. 2011; Kent et al. 2012; Muralidharan & Muir, 2006). In general, the ability to modify any atom in the protein of interest with the precision afforded by synthetic organic chemistry is enormously powerful for dissecting protein function, especially when combined with high-resolution structural approaches such as NMR spectroscopy and X-ray crystallography. Thus, we imagine that NCL will continue to evolve as an approach and be integrated into structural biology campaigns.

### 3.4 Inteins

Inteins (intervening proteins) are a peculiar group of proteins that can excise themselves from a larger precursor polypeptide chain, a process that leads to the formation of a native peptide bond between the flanking extein (external protein) fragments. This auto-processing event, called protein splicing, is analogous to the self-splicing of RNA introns and is spontaneous, i.e. it does not require external factors or ATP. Since they were first discovered in the early 1990s (Hirata et al. 1990; Kane et al. 1990), thousands of putative intein domains have been identified in the genomes of many unicellular organisms and viruses,
with some containing multiple inteins in their genomes or even within the same gene (Perler, 2002; Shah & Muir, 2014). A small fraction of the known inteins has an even more curious property—the intein is split into two fragments, each fused to a separate extein fragment (the N- and C-exteins) (Wu et al. 1998). These intein fragments, called split inteins, are transcribed and translated separately, and upon a spontaneous and non-covalent association in the cellular milieu, they carry out protein splicing in trans to unite the extein fragments into a single polypeptide chain. While it is known that many inteins are embedded within essential protein genes (such as DNA or RNA polymerase, ribonucleotide reductase or metabolic enzymes), their evolutionary origins and biological significance remain mysterious, and only a small percentage of the identified intein domains have been carefully characterized (Pietrokovski, 2001; Shah & Muir, 2014). Despite these big gaps in our knowledge, the unique reactivity of inteins has turned them into a versatile and transformative tool in protein chemistry and chemical biology. For a detailed overview of intein applications, we refer the interested reader elsewhere (Shah & Muir, 2014; Topilina & Mills, 2014; Volkman & Mootz, 2013; Wood & Camarero, 2014). Here, we will focus on aspects of intein function that would be of use to the structural biologist looking to install site-specific PTMs, segmentally label proteins with NMR isotopes, or aid the purification of recombinant polypeptides. Inteins have come a long way since their first applications in structural biology (Xu et al. 1999; Yamazaki et al. 1998), so we will end this section with a discussion of the current members of the intein toolbox and research directions taken to circumvent their limitations.

3.4.1 The intein splicing mechanism

Despite the low-sequence homology of known intein domains, they share a common protein splicing mechanism that relies on several conserved residues in the intein/extein polypeptide (Fig. 6a) (Volkmann & Mootz, 2013). One of these key residues is a cysteine (or in some cases a serine) at position 1 of the intein sequence. This nucleophilic side-chain attacks the amide carbon of the N-extein at position −1 (Fig. 6b) resulting in an N to S(O) acyl shift and the formation of a linear thio(oxy)ester intermediate. This intermediate is subject to a nucleophilic attack by a side-chain (cysteine, serine or threonine) at position +1 on the C-extein leading to trans-(thio)esterification and the generation of a branched intermediate. The branched intermediate is resolved through the cyclization of the C-terminal asparagine of the intein and results in intein excision from the polypeptide chain. Next, the spliced exteins quickly undergo an S(O) to N acyl shift to form a native peptide bond (i.e. identical to the last step in NCL). The protein splicing mechanism is facilitated by several conserved threonine and histidine residues occupying strategic positions in the intein structural fold (Frutos et al. 2010). The efficiency and kinetics of the splicing mechanism may also depend on the identity of the residues immediately flanking the intein placing important constraints on the choice of ligation junction (Cheriyan et al. 2013; Iwai et al. 2006; Shah et al. 2012).

3.4.2 Applications in protein engineering

Expressed protein ligation (EPL) is an extension of NCL that employs a contiguous intein to recombinantly generate a protein bearing a C-terminal thioester (Muir et al. 1998). In this case, the N-extein is fused to a modified intein construct lacking the ability to perform trans-thioesterification. Instead, this step is performed by an exogenously added thiol, resulting in cleavage of the N-extein α-thioester intermediate (Fig. 7a). The resultant thioester can be used in NCL reactions as described in Section 3.3, while the recombinant origin of this fragment allows the construction of much larger semi-synthetic proteins as compared...
to total chemical synthesis by NCL. The efficiency of thioester generation rests on the propensity of the intein to avoid unwanted side reactions that result in premature N-extein cleavage and hydrolyzed products, problems that can be alleviated by the use of streamlined EPL protocols (Vila-Perello et al. 2013). Alternatively, hydrolysis of the N-extein under slightly basic conditions can be exploited in the so-called tagless protein purification protocols (Batjargal et al. 2015; Guan et al. 2013; Southworth et al. 1999) (Fig. 7b). In this case, the protein of interest is fused to an intein carrying the appropriate mutations and a suitable purification tag. The tag can be used for affinity column enrichment of the construct, followed by increase in the buffer pH. This results in the release of the tagless protein while the intein remains on the column.

Harnessing the protein trans-splicing (PTS) process mediated by split inteins offers an alternative approach to the ligation of polypeptide building blocks (Fig. 7c). Natural split inteins are especially attractive in this regard due to the extremely high affinity between the fragments (Shah et al. 2011, 2013) – this renders the ligation reaction less dependent on reagent concentration as compared to strictly chemical processes like NCL/EPL. Using orthogonal split intein pairs, it is also possible to perform one-pot three-piece ligations (Carvajal-Vallejos et al. 2012; Shah et al. 2011; Shi & Muir, 2005) resulting in the regiospecific assembly of the associated extein building blocks. While most natural split inteins have N- and C-fragments that are relatively large, it is possible to generate artificially split inteins that are as short as six or eleven residues (Appleby et al. 2009; Ludwig et al. 2006). There is also an efficient natural split intein pair (Acel–TerL) where the N-intein fragment is only 25 amino acids long (Thiel et al. 2014). Thus, it is now possible to use PTS with both synthetic or recombinant intein fragments and to install a wide range of N- or C-terminal chemical modification, including biophysical probes. One of the most important applications of PTS in structural biology is protein segmental isotopic labeling discussed in Section 4.2. Other applications include the cyclization of proteins and peptides (Lennard & Tavassoli, 2014; Scott et al. 1999), conditional protein splicing (Mootz et al. 2003; Schwartz et al. 2007), and protein semi-synthesis in cells (David et al. 2015).

### 3.4.3 Toward fast and promiscuous inteins

The first intein tools were introduced in the mid-to-late 1990s. One of the ‘early’ inteins, the 198-residue gyrase A intein from *Mycobacterium xenopi* (Mxe GyrA) (Southworth et al. 1999), still used today, exhibits many desirable properties for...
applications in protein engineering – it is relatively small, can be efficiently expressed in *E. coli*, works in moderate concentrations of denaturants and its activity can be controlled with temperature. The first ‘natural’ split intein was discovered in 1998 in the cyanobacterium *Synechocystis* sp. strain PCC6803 (Ssp) where it was found to ligate two fragments of the catalytic subunit of DNA polymerase III (DnaE) (Wu et al. 1998). This discovery paved the road to more efficient protein trans-splicing and opened the way to using split inteins in other applications such as the cyclization of proteins and peptides and the creation of large cyclized libraries for potential therapeutic applications (Scott et al. 1999).

While these first generation intein tools were certainly enabling (Vila-Perello & Muir, 2010), they were not without their limitations – in retrospect, one could even say they were rather fussy and slow. For example, depending on the fusion partners, splicing (or thiolysis) could take hours to days and was often inefficient (Muralidharan & Muir, 2006). These ‘idiosyncrasies’ constrained the application of inteins in structural biology and fueled the search for faster, more promiscuous and efficient inteins. Several important discoveries in the mid-2000s challenged the view that all natural inteins are inefficient and slow. A genomic study of cyanobacterial genes expanded the DnaE intein family (Caspi et al. 2003) and the characterization of one newly discovered member, the DnaE intein from *Nostoc punctiforme* (Npu) revealed a few surprises. This split intein could perform protein trans-splicing reactions in *vitro* on a minute timescale and was much more tolerant to sequence deviations on the attached exteins than Ssp ((Iwai et al. 2006; Zettler et al. 2009). Now we know that many members of the DnaE family are fast (Shah et al. 2012), thus greatly expanding the choice of natural intein tools for the efficient generation of protein *α*-thioesters for EPL or the ligation of protein fragments in PTS (Table 1). Furthermore, efficient natural split inteins that are not part of the DnaE family have also been discovered and these include the gp41-1 and gp41-8 inteins (with insertion sites in the gp41 DNA gyrase gene), the IMPDH-1 intein (splitting a gene coding for inosine-5′-monophosphate dehydrogenase), the NrdJ intein (splitting the gene coding for the ribonucleotide reductase subunit NrdJ) (Carvajal-Vallejos et al. 2012), and the AceL–TerL pair (discovered in metagenomics data from the antarctic permanently stratified saline Ace Lake) (Thiel et al. 2014). These proteins bring more diversity to the intein molecular engineering toolbox, including splicing rates that are

---

Fig. 6. Intein structure and mechanism. (a) Intein/extein residues important for splicing. (b) Protein splicing mechanism of contiguous inteins. In some cases, the hydroxyl groups of Ser/Thr act as nucleophiles in the first two steps.
an order of magnitude faster than those for NpuDnaE; N- and C-intein fragments that are relatively short and can be made by peptide synthesis rather than recombinantly; the option of utilizing serine instead of cysteine at the +1 position; and the possibility of exploiting orthogonality in one-pot multi-piece ligations.

Careful biochemical characterization of newly discovered inteins has provided insights into the principles governing fast splicing and extein tolerance. For example, a batch mutagenesis approach that compared the slow split intein DnaE family member Ssp and the fast Npu split intein revealed that speed is determined by a handful of ‘accelerator’ residues located in the second shell of the folded protein, adjacent to the intein active site (Stevens et al. 2016). These residues were used as a filter in aninformatics analysis of the DnaE sequence database, leading to the identification of several dozens of other split inteins predicted to support ultrafast splicing. A consensus split intein sequence, termed Cfa, was then derived from this putative fast set and was found to possess quite remarkable properties; in addition to splicing faster than Npu at ambient conditions, Cfa is extremely robust, maintaining efficient activity at 80 °C or in the presence of up to 4 M guanidinium chloride or 8 M urea. As a result of these attributes, Cfa was found to be a superior tool for several PTS applications (Stevens et al. 2016).

3.5 Sortases

Sortases are a class of cysteine transpeptidases responsible for the attachment of virulence proteins to the cell wall of Gram-positive bacteria (Mazmanian et al. 1999). They are also involved in the polymerization of pilin subunits to form the pilus structures responsible for bacterial attachment to the host and biofilm formation (Mandlik et al. 2008). As important players in bacterial virulence, they have evolved to recognize a specific sorting sequence (LPXTG in the case of *Staphylococcus aureus*) and to attach the virulence factor to the cell wall using a pentaglycine cross-bridge (Ton-That et al. 2000). Naturally,
sortases are of a considerable interest as drug targets, but they have also become an important and versatile protein engineering tool (Mao et al. 2004). Most sortase-based applications utilize the soluble fragment of wild-type or modified sortase A from S. aureus. These enzymes recognize the LPXTG motif and use their catalytic cysteine residue to cleave between the threonine and glycine backbone within the recognition sequence (Fig. 8). The cleavage reaction involves a thioacyl intermediate similar to the intermediates generated by cysteine proteases (Aulabaugh et al. 2007). Unlike the water molecules employed by proteases, however, sortases use a nucleophilic attack from the N-terminus of an oligoglycine motif to create a peptide bond between the acyl donor and acceptor. This results in the ligation of polypeptide chains that are subsequently connected with a LPXT(G)<sub>n</sub> linker. The sortase mechanism also requires binding of Ca<sup>2+</sup> to a dynamic loop of the enzyme, an event that slows down the loop motion and allows enough time for the substrate to 

![Table 1. Intein toolbox for protein semi-synthesis](https://doi.org/10.1017/S0033583517000051)

<table>
<thead>
<tr>
<th>Intein</th>
<th>Temperature (°C)</th>
<th>t&lt;sub&gt;1/2&lt;/sub&gt;</th>
<th>Comments</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>MxeGyrA</td>
<td>25</td>
<td>10 h</td>
<td>Commonly used contiguous intein for EPL applications</td>
<td>Frutos et al. (2010); Southworth et al. (1999)</td>
</tr>
<tr>
<td>SspDnaE</td>
<td>37</td>
<td>76 min</td>
<td>The first discovered natural split intein</td>
<td>Shah et al. (2012); Wu et al. (1998)</td>
</tr>
<tr>
<td>NpuDnaE</td>
<td>37</td>
<td>19 s</td>
<td>Fast split intein; synthetically accessible C-fragment</td>
<td>Shah et al. (2012); Vila-Perello et al. (2013)</td>
</tr>
<tr>
<td>AsoDnaE</td>
<td>37</td>
<td>23 s</td>
<td>Fast split intein; higher protein expression yields reported</td>
<td>David et al. (2015); Shah et al. (2012)</td>
</tr>
<tr>
<td>Cfa (consensus</td>
<td>30</td>
<td>20 s</td>
<td>Fast split intein; works at high temperatures and in denaturants</td>
<td>Stevens et al. (2016)</td>
</tr>
<tr>
<td>DnaE split intein)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp41-1</td>
<td>45</td>
<td>4 s</td>
<td>Fast and orthogonal to other split inteins</td>
<td>Carvajal-Vallejos et al. (2012)</td>
</tr>
<tr>
<td>gp41-8</td>
<td>37</td>
<td>15 s</td>
<td>Fast and orthogonal to other split inteins</td>
<td>Carvajal-Vallejos et al. (2012)</td>
</tr>
<tr>
<td>NrdF-1</td>
<td>37</td>
<td>7 s</td>
<td>Fast and orthogonal to other split inteins</td>
<td>Carvajal-Vallejos et al. (2012)</td>
</tr>
<tr>
<td>IMPDH-1</td>
<td>37</td>
<td>8 s</td>
<td>Fast and orthogonal to other split inteins</td>
<td>Carvajal-Vallejos et al. (2012)</td>
</tr>
<tr>
<td>AceL-TerL</td>
<td>8</td>
<td>7-2 min</td>
<td>Synthetically accessible N-intein</td>
<td>Thiel et al. (2014)</td>
</tr>
</tbody>
</table>

* Optimal splicing kinetics in the presence of native sequences at the immediate intein-extein junctions. Variation from this sequence context can lead to less efficient splicing. See also ref. (Shah & Muir, 2014).

Analogous to intein technology development, sortases have considerably improved as protein engineering tools since their introduction in 2004 (Antos et al. 2016). The sortase-based protein engineering toolbox now contains evolved variants that exhibit much faster kinetics or that eliminate Ca<sup>2+</sup>-dependence (albeit at the cost of slightly reduced enzyme activity) (Chen et al. 2011; Hirakawa et al. 2015; Wueethrich et al. 2014). There are also alternatives based on S. aureus sortase A or homologs from other organisms that can recognize variations of the LPXTG motif and/or allow non-glycine amino acids as the acyl acceptor (Antos et al. 2016; Dorr et al. 2014; Glasgow et al. 2016). To increase the yields of the ligation reaction, several clever strategies have been employed. In situations where the released aminoglycine peptide fragment is relatively small, it can be removed by dialysis or centrifugation while the reaction is proceeding (Freiburger et al. 2015). Affinity immobilization strategies or flow-based platforms have also been used for the selective removal of reaction components (Policarpo et al. 2014; Warden-Rothman et al. 2013). Alternatively, the equilibrium of the reaction can be controlled by ligation product or by-product deactivation. In the first case, a WTWTW motif was added to the donor and acceptor, and upon ligation this sequence promoted a stable hairpin at the ligation junction, rendering the site inaccessible for cleavage (Yamamura et al. 2014). In the latter case, the acyl donor glycine was chemically modified such that upon release, chemical rearrangements occurred on the by-product transforming it into a poor nucleophile (Liu et al. 2014a; Williamson et al. 2014).

One important advantage of sortase-based ligations is that the acyl donor and acceptor polypeptide chains can be very short (only the LPXTG tag is required on the donor, and the oligoglycine motif is necessary on the acceptor) and thus are easily accessible by solid-phase peptide synthesis. Therefore, N- and C-terminal labeling reactions of large proteins are relatively straightforward (assuming by-products are efficiently removed) (Guimaraes et al. 2013; Theile et al. 2013). Larger polypeptides, on the other hand, can be expressed recombinantly with the appropriate donor and acceptor tags and the ligation reaction unites them in a single polypeptide chain with an LPXT(G)<sub>n</sub> ‘scar’. In such cases it is recommended that the...
ligation junction is chosen on an unstructured region where it will not affect the function and/or fold of the protein and will be accessible to the sortase catalytic site (Guimarães et al. 2013). Such reactions can be applied to create polymers and cyclized polypeptides (van ‘t Hof et al. 2015), or to stitch together domains into bifunctional or segmentally labeled proteins (Matsumoto et al. 2016; Williams et al. 2016; Witte et al. 2012). Sortase-based labeling has also been used in the functionalization of solid supports, nanoparticles, antibodies or cell surfaces, as well as in the labeling of proteins in vivo. We refer the interested reader to several comprehensive reviews on the subject (Popp & Ploegh, 2011; Ritzefeld, 2014; Schmohl & Schwarzer, 2014).

The success of sortase-based ligations has stimulated efforts to discover other protein ligases with expanded capabilities. A promising candidate is butelase-1, which was isolated from the plant *Clitoria ternatea* (Nguyen et al. 2014). Butelase-1 is the fastest ligase known with catalytic efficiencies as high as 542 000 M$^{-1}$ s$^{-1}$. Furthermore, it only requires the recognition sequence NHV on the acyl donor and produces ligation junctions with a minimal ‘scar’ (NX). Currently, the major limitation of this technology is that the enzyme is not available in recombinant form, and therefore has to be extracted and purified from the native plant (Nguyen et al. 2015). There are, however, evolutionary related ligases that may be more amenable to protein engineering approaches (Yang et al. 2017).

### 3.6 Genetic code expansion

After billions of years of evolution, Nature has engineered extraordinary functional diversity into proteins with only 20 amino acid building blocks. Yet, many of these building blocks are often modified post-translationally, clearly indicating the need of living organisms to enhance and modulate their protein repertoire with additional chemical functionalities. There are also organisms from all domains of life that can produce and incorporate other building blocks into their proteins. This includes selenocysteine, often called the 21st amino acid that provides a unique reactive site for precise tuning of biological function in cells. Interestingly, this amino acid is incorporated into proteins by a natural reassignment of the UGA stop codon coupled with the recognition of a specific structural element on the mRNA transcript known as the selenocysteine-insertion sequence (reviewed in (Metanis et al. 2009; Yoshizawa & Bock, 2009)). Similarly, there are methane producing Archaea species that have evolved a specialized tRNA/aaRS synthetase (tRNA/aaRS) pair to exploit the UAG stop codon and insert pyrrolysine site-specifically into certain methyltransferase proteins (Srinivasan et al. 2002). Exploiting the natural translation machinery, protein engineers have worked hard to ‘persuade’ living organisms to incorporate additional building blocks into polypeptide chains (reviewed in (Liu & Schultz, 2010)). One approach involves the use of cell lines that are auxotrophic for...
one of the 20 amino acids, for example methionine, and that will only grow when the missing amino acid is included in the culture medium. Replacing this amino acid with a close structural analog that is utilized by the wild-type aminocatyl-tRNA synthetase, results in incorporation of the unnatural amino acid (UAA) into overexpressed proteins. Since this results in global incorporation of the UAA, this approach is often used for the replacement of rare amino acids with their structural analogs. For example, methionine can be substituted with selenomethionine to provide a heavy atom for phasing crystallographic data (Barton et al. 2006; Yang et al. 1990).

A second approach involves the semi-synthesis of tRNAs that are pre-loaded with the UAA of interest (Hecht et al. 1978; Noren et al. 1989). These tRNAs have been used in in vitro translation systems that bypass the need for a matching aaRS, and since the identity of the UAA is decoupled from the information content of the tRNA, any coding or blank codon can be used for reassignment (Cornish et al. 1994; Judice et al. 1993; Koh et al. 1997). While the semi-synthesis of acylated tRNAs can be technically challenging, more efficient production strategies have been developed. This includes flexizymes, flexible tRNA acylation ribozymes that accept a versatile range of aminocatyl substrates and tRNAs with different sequences (Goto et al. 2011). Pre-loaded tRNAs can also be injected or transfected directly into living cells (England et al. 1999; Kohrer et al. 2003), although the success of such strategies has been limited by their short lifetimes in the cellular environment and challenges associated with in-cell delivery.

Today, UAA incorporation in living cells is almost exclusively performed following the strategy introduced by Peter Schultz and co-workers in 2001 (Wang et al. 2001). This methodology, commonly referred to as amber suppression, relied on the development of an orthogonal tRNA/aaRS pair that could be expressed in E. coli and used to incorporate O-methyl-L-tyrosine into dihydrofolate reductase with 99% fidelity. In the 15 years since this landmark study, the unnatural building block palette for genetic incorporation has grown more than 100 amino acids strong (Lang & Chin, 2014; Liu & Schultz, 2010; Neumann-Staubitz & Neumann, 2016). This includes amino acids carrying natural modifications (e.g. phosphoserine or acetyltysine), biophysical and structural probes, cross-linkers, reactive handles for bio-orthogonal reactions, and site-specific protein engineering functionalities that can modify their attendant proteins upon a specific cellular or chemical cue. Here, we describe the basic principles of the technology, review UAA of particular interest to the structural biologist, and discuss current limitations and efforts to improve the efficiency of UAA incorporation.

### 3.6.1 Amber codon suppression in living cells

The successful incorporation of an UAA into a protein synthesized by a living cell requires several important considerations and components (Fig. 9a). First and foremost, the UAA of interest must be chemically and metabolically stable, cell permeable or otherwise biosynthetically accessible in the cellular environment. It also must be tolerated by the ribosome and the cellular elongation factors without being recognized as a substrate by any of the endogenous synthetases. The UAA then requires its own unique codon, with the amber stop codon (UAG) being a popular choice due to its low occurrence in both prokaryotic and eukaryotic systems. The successful site-specific incorporation of the UAA, however, rests on the presence of a dedicated tRNA/aaRS pair that is highly specific for the UAA of interest, yet orthogonal in the context of all endogenous tRNA/aaRS pairs. Developing such pairs for a chemically diverse set of UAA is currently one of the time consuming and difficult steps of this technology. Since tRNA recognition by aaRS is often species specific, it is sometimes possible to import a heterologous pair into the cell of interest and use it as a starting point to build orthogonality and specificity into the system. For example, many UAA incorporation systems in E. coli are based on the heterologous tRNA<sup>Tyr</sup>/TyrRS pair from Methanococcus jannaschii, while the tRNA<sup>Leu</sup>/LeuRS and tRNA<sup>Lys</sup>/LysRS pairs from E. coli have been used in eukaryotic cells (reviewed in (Chin, 2014)). The tRNA<sup>Phe</sup>/PyrRS pair from methanogenic bacteria that can incorporate pyrrollysine has also been a very useful tool, as it is orthogonal in E. coli, yeast and mammalian cell lines, and has allowed the incorporation of many lysine-based UAAs, including acetyltysine (Neumann et al. 2008). While these systems provide a useful starting point, it is usually necessary to use mutagenesis and rounds of negative and positive selection to improve on the selectivity and orthogonality of the pair. Directed evolution approaches can also be used for the generation of de novo tRNA/aaRS pairs, or to expand the function of other components of the translational machinery (reviewed in (Chin, 2014)). Once an appropriate tRNA/aaRS pair is developed, however, the practical implementation of amber suppression for the UAA is relatively straightforward. E. coli cells, for example, can be transformed with two plasmids: (1) a plasmid encoding the protein of interest and an appropriate point mutation with the amber TAG codon, and (2) a plasmid carrying the appropriate DNA sequence to produce the optimized tRNA/aaRS pair. After addition of UAA to the media, gene expression is induced for both plasmids and the UAA is incorporated into the protein of interest by the bacterial translational machinery. To separate the full-length protein from prematurely truncated species, often a purification tag is added to the protein C-terminus – notably, these can involve ‘silent’ intein- or sortase-based purification tags (Batjargal et al. 2015; Warden-Rothman et al. 2013). Well-established protocols for UAA incorporation are now available for yeast (Hancock et al. 2010), mammalian (Chen et al. 2009) and insect cells (Koehler et al. 2016; Mukai et al. 2016).
and amber suppression can even be performed in multicellular organisms including C. elegans (Greiss & Chin, 2011), D. melanogaster (Bianco et al. 2012), mice (Ernst et al. 2016; Kang et al. 2013) and plants (Li et al. 2013b).

### 3.6.2 The amber suppression toolbox

The amber suppression toolbox contains many UAAs designed with structural biology applications in mind (Fig. 9b–e). For example, heavy atoms can be incorporated site-specifically for solving the phase problem in X-ray crystallography – appropriate...
UAAs include p-iodo-L-phenylalanine (Xie et al. 2004) and 3-iodo-L-tyrosine (Sakamoto et al. 2009), as well as metal-ion chelating amino acids (Lee et al. 2009b). For NMR spectroscopists, amber suppression offers a relatively cheap and efficient way to install site-specific isotopic labels in otherwise unlabeled proteins. Many of the NMR ‘friendly’ UAAs are fluorinated derivatives that exploit the unique spectroscopic properties of 19F as a reporter of global protein folding and dynamics (Jones et al. 2010; Yang et al. 2015). Similarly, amber suppression has been used to install nitroxide spin labels for distance measurements by EPR, thus overcoming the problems often associated with cysteine-based approaches (Park et al. 2015; Schmidt et al. 2015). More importantly, however, amber suppression is a living cell protein engineering tool; thus, it is ideally suited for NMR or EPR studies designed to follow the structural fate of proteins in the cellular milieu.

Of particular interest to the structural biologist are UAAs carrying natural PTMs. Currently, amber suppression can directly incorporate the following modifications: phosphoserine (Rogerson et al. 2015), acetylated lysine (Neumann et al. 2008), several lysine acylations (Gattner et al. 2013; Kim et al. 2012), phosphotyrosine (Fan et al. 2016) and sulfonated tyrosine (Liu et al. 2007). To expand this toolbox, however, amber suppression can be used to install a reactive handle at the position of interest, and then the appropriate modification can be chemically generated after protein purification. Using this strategy, for example, the UAA δ-thiol-lysine can be incorporated, followed by traceless attachment of a ubiquitin moiety with NCL (Virdee et al. 2011). Similarly, methylated lysines can be generated by the incorporation of a suitable pre-cursor UAA (Nguyen et al. 2010; Wang et al. 2017). Phosphoserine, on the other hand, can serve as a starting point for the generation of dehydroalanine, which in turn can be converted into a number of modified side-chains (Wright et al. 2016).

Amber suppression also allows the facile installation of site-specific cross-linkers, a valuable tool for the identification of protein–protein and protein–ligand interactions both in vitro and in vivo. There are several options for UV-activatable cross-linkers that exploit different cross-linking mechanisms, and afford temporal and spatial control of the reaction. The oldest members of this toolbox are p-benzophenylalanine (Chin et al. 2002a) and p-azido-L-phenylalanine (Chin et al. 2002b), both available for bacterial and eukaryotic systems, and extensively used for cross-linking experiments of purified proteins or in the cellular environment. More recently, diazirine-modified lysine-based amber suppression systems have been developed (Ai et al. 2011; Chou et al. 2011; Zhang et al. 2011), exhibiting superior cross-linking efficiency, smaller structural perturbation effects and ideally suited for cross-linking experiments of lysine-rich proteins such as histones. UAAs that can cross-link proteins to nucleic acids are also available, and these include p-benzophenylalanine (Winkelmann et al. 2016), and a furan-based cross-linker activated upon red-light irradiation (Schmidt & Summerer, 2013). Interestingly, p-azido-L-phenylalanine can also serve as an infrared (IR) spectroscopy probe, and as such can be used to report on the structural transitions of a protein along its activation pathway (Ye et al. 2010). Similarly, para-cyanophenylalanine can be installed as a site-specific and sensitive vibrational probe of ligand binding (Schultz et al. 2006).

A significant fraction of the amber suppression UAA library has been designed for imaging and fluorescence-based biophysical applications. Some fluorescent UAAs can be incorporated directly into polypeptide chains and these include coumarin- and dansyl-based modifications (Kuhn et al. 2011; Summerer et al. 2006), as well as environmentally sensitive aminomethylphalene-based probes available for both yeast and mammalian applications (Chatterjee et al. 2013a; Lee et al. 2009a). More commonly, however, optical probes are attached site-specifically using bio-orthogonal reactions. In this case, amber suppression is used to install a reactive handle at a site in the polypeptide chain, while the fluorescent probe is modified with a compatible reactive functionality. The conjugation reaction can be carried out in vitro with purified components, or the optical label can be added to the media and/or delivered into cells for bio-orthogonal reaction chemistry within the cellular milieu. While this places important UAA-fluorescent label design constraints with respect to cell permeability, chemical stability and reaction kinetics, this approach allows the incorporation of optical labels that work at a variety of wavelengths, amid reduced background fluorescence. Currently, UAAs with a diverse set of functionalities for bio-orthogonal reactions are available (reviewed in (Lang & Chin, 2014)), and it is likely that this list will become much more expansive in the future as new fast and specific bio-compatible approaches are developed (Section 3.7).

3.6.3 Limitations and future directions

The remarkable plasticity of the natural and evolved cellular protein synthesis machinery has allowed chemical biologists to create a large and diverse set of UAAs that can be incorporated into biological polymers assembled in vivo. Yet, the incorporation of many of these UAAs is inefficient, context dependent and essentially limited to the inclusion of a single modification per protein. Since UAA incorporation relies on the reassignment of natural stop codons, the suppressor tRNAs compete for binding sites with the endogeneous release factors that terminate translation. Therefore, truncations of the desired protein are often produced, resulting in compromised yields, complicated purification protocols and potential toxicity for the recombinant cell. In E. coli, translation termination involves release factor protein 1 (RF1) responsible for recognition of ochre (UAA) and amber (UAG) stop codons, and release factor protein 2 (RF2) that identifies ochre (UAA) and opal

https://doi.org/10.1017/S0033583517000051 Published online by Cambridge University Press
pressed codon schemes and free up unique codons for reassignment (Ostrov
messages (Neumann
codons can in principle provide 256 blank codons, thus dramatically expanding the capabilities of the natural 64-codon based
phine oxide at the ligation junction, to form a stable amide bond (Saxon & Bertozzi, 2000). While the
classical Staudinger reduction reaction, azides react with triarylphosphines bearing an electrophilic trap (usually an ester)
to form a red probe, this small moiety is also at the heart of several important bio-orthogonal reactions. For example, in an extension of
suppression, total chemical synthesis and/or peptide ligation. In addition to serving as a light-activatable cross-linker or infra-
reader to several excellent reviews on the subject (Lang & Chin, 2014; Shieh & Bertozzi, 2014; Spicer & Davis, 2014;
neered to recognize quadruplet codons has been developed (Neumann et al. 2010; Wang et al. 2014a). The use of quadruplet
codons can in principle provide 256 blank codons, thus dramatically expanding the capabilities of the natural 64-codon based
genomes. Such systems, however, require the engineering of orthogonal ribosomes that can efficiently recognize quadruplet
messages (Neumann et al. 2010). Alternatively, the genomes of organisms such as E. coli can be reprogrammed to create compressed
codon schemes and free up unique codons for reassignment (Ostrov et al. 2016; Wang et al. 2016). In the future, it may also be possible to create new coding schemes by expanding the genetic alphabet of living organisms and utilizing synthetic DNA bases in new synthetic organisms (Malysh et al. 2014).

3.7 Chemical modification of unnatural amino acids

The ability to install unnatural amino acids in polypeptide chains with exquisite specificity has opened widely the doors to a
post-cysteine world of bio-orthogonal reactions. There is now a rapidly expanding collection of reactions that work efficiently and selectively with low concentrations of reactants in physiological buffers or the cellular milieu. Here, we will review a small sample of these tools focusing primarily on reactions that might be used to install site-specific PTMs or their mimics, as well as probes for structural investigations. For a comprehensive discussion of the bio-orthogonal literature, we refer the interested reader to several excellent reviews on the subject (Lang & Chin, 2014; Shieh & Bertozzi, 2014; Spicer & Davis, 2014; Stephanopoulos & Francis, 2011).

The azido functional group is one of the most versatile modifications that can be installed on peptides and proteins by amber suppression, total chemical synthesis and/or peptide ligation. In addition to serving as a light-activatable cross-linker or infrared probe, this small moiety is also at the heart of several important bio-orthogonal reactions. For example, in an extension of the classical Staudinger reduction reaction, azides react with triarylphosphines bearing an electrophilic trap (usually an ester) to form a stable amide bond (Saxon & Bertozzi, 2000). While the first version of this methodology generated a residual phosphine oxide at the ligation junction, ‘traceless’ Staudinger ligation variants are now available (Nilsson et al. 2000; Saxon et al. 2000) (Fig. 10a), as well as reactions based on azobenzene or caged phosphine reagents that can be activated with light (Shah et al. 2013b; Wang et al. 2016). This reaction can be used to ligate a peptide fragment containing a C-terminal thioester to another fragment carrying an N-terminal azide resulting in an extended native polypeptide chain and bypassing the need for a cysteine residue at the ligation junction (Nilsson et al. 2000). While displaying slower reaction kinetics compared with some other bio-orthogonal reactions, the Staudinger ligation can be impactful in protein engineering applications where versatility and orthogonality are desired. For example, it has been used to create native isopeptide linkages between ubiquitin and other proteins (Andersen & Raines, 2015), install multiple probes into proteins such as RNA polymerase or GPCRs (Chakraborty

https://doi.org/10.1017/S0033583517000051 Published online by Cambridge University Press
et al. 2015; Huber et al. 2013), prepare glycoprotein or phosphoprotein mimics (Bernardes et al. 2011; Serwa et al. 2009), and turn on the fluorescence of optical probes upon protein labeling (Lemieux et al. 2003).

Another popular and versatile bio-orthogonal reaction is the azide-alkyne 1,3-dipolar cycloaddition (Rostovtsev et al. 2002; Tornoe et al. 2002), commonly referred to as ‘click chemistry’ (Kolb et al. 2001). The cycloaddition can be promoted either by the presence of Cu(I) ligands or the use of highly strained cyclooctyne systems (Agard et al. 2004). Both versions are specific, exhibit relatively fast reaction kinetics and are easy to use, which has led to numerous applications in the protein conjugation area. Recent examples include the incorporation of dual EPR or FRET probes for distance measurements in proteins or RNA (Kucher et al. 2016; Lavergne et al. 2016), and the attachment of rigid lanthanide tags to large proteins for paramagnetic relaxation experiments (Mallagaray et al. 2016). In vivo applications of the Cu(I) version have been limited by copper toxicity, although extracellular labeling of cells has been reported (Link et al. 2004; Uttamapinant et al. 2012). The strained cyclooctyne systems, on the other hand, are compatible with living systems and improved versions that exhibit faster labeling kinetics are available for genetic incorporation through amber suppression (Dommerholt et al. 2010; Lang et al. 2012b; Plass et al. 2011) (Fig. 10b). These copper-free cycloadditions have thus been valuable in imaging and proteomic applications in living cells and organisms (Laughlin et al. 2008; Smits et al. 2016; Xie et al. 2014), and can be useful for the structural biologist interested in understanding the dynamics of macromolecules in the cellular environment either by FRET or EPR (Kucher et al. 2016).

The success of click chemistry bioconjugation has inspired the search for faster and more efficient reactions for labeling in the cellular milieu. An important development along this direction involves the use of inverse-electron demand Diels-Alder reactions between strained dienophiles and tetrazine dienes (Blackman et al. 2008; Devaraj et al. 2008). The dienophile system in

---

**Fig. 10.** Chemical modification of unnatural amino acids.

---

https://doi.org/10.1017/S0033583517000051 Published online by Cambridge University Press
this case can be a trans-cyclooctene or a norbornene, both exhibiting significantly faster reaction kinetics as compared to the 1,3-dipolar cycloadditions described above, and thus allowing labeling reactions of proteins at much lower concentrations (Fig. 10c). These bio-orthogonal reactive handles are also available for genetic incorporation by amber suppression (Lang et al. 2012a, 2012b; Seitchik et al. 2012) and provide an expanded and rapidly evolving toolbox for bio-orthogonal protein manipulations and modifications in the cellular environment.

While most of the reactions described above are incredibly useful for the conjugation of small molecules and biophysical probes to proteins in vitro and in vivo, they generally leave a relatively large chemical ‘scar’ at the conjugation site to be practical in the generation of PTMs or their mimics. A versatile approach for the site-specific installation of PTMs involves the UAA dehydroalanine. This UAA can be generated from a cysteine residue installed by site-directed mutagenesis (Chalker et al. 2012) or from an O-phosphoserine precursor incorporated by amber suppression (Yang et al. 2016). Under biocompatible conditions, dehydroalanine can then be reacted with a variety of alkyl halides, via a radical-mediated process, to produce an impressive list of more than 50 modified functionalities at the site of interest (Wright et al. 2016; Yang et al. 2016). These include methylated lysine and arginine residues, as well as fluorinated, glycosylated, phosphorylated, alkylated or isotopically labeled natural and UAA side chains. It should be noted that this methodology currently leads to the generation of both D and L configurations at the mutation site (i.e. the reaction lacks stereospecific control), a potential limitation for some applications.

3.8 Enzymatic bioconjugation approaches

The site-specific incorporation of bio-orthogonal handles on proteins can also be achieved enzymatically. These enzymes choose their targets based on the recognition of a specific amino acid sequence and can perform various chemical modifications of cysteine, lysine, serine, glutamine or glycine residues within their target sequences. This toolbox, for example, includes the formylglycine generating enzyme (FGE) that recognizes the CXPXR sequence motif and converts the cysteine residue to formylglycine, thus introducing an aldehyde functional group (Wu et al. 2009) (Fig. 11a). The aldehyde handle can be subsequently elaborated with various probes via bio-orthogonal transformations such as oximation and Hydrazino-Pictet-Spengler reactions (Agarwal et al. 2013; Dirksen & Dawson, 2008). Another popular tool is lipoic acid ligase, an enzyme that modifies a lysine side-chain within the 13-residue target sequence (Uttamapinant et al. 2010). Engineered versions of this enzyme can accommodate lipoic acid analogs and have been used to introduce bio-orthogonal handles, including azides (Plaks et al. 2015; Uttamapinant et al. 2012) (Fig. 11b), aryl aldehydes and hydrazines (Cohen et al. 2012), p-iodophenyl derivatives (Hauke et al. 2014), norbornenes (Best et al. 2015) an trans-cyclooctenes (Liu et al. 2012). It is also possible to engineer fluorescent lipoic acid analogs that can be installed directly on the protein of interest (Uttamapinant et al. 2010). Other enzymes of the enzymatic toolbox include biotin ligase, farnesyltransferase, transglutaminase and N-myristoyltransferase (reviewed in (Rashidian et al. 2013)).

Enzymatic bioconjugation is an increasingly useful tool for cell imaging applications, functionalization of therapeutic proteins, immobilization of proteins on solid supports or the preparation of protein—polymer or protein—nanoparticle conjugates (Hu et al. 2016; Slavoff et al. 2011; Walper et al. 2015). For the structural biologist and biophysicist, it offers an orthogonal strategy for protein modification that is based on genetically encodable peptide tags (Stephanopoulos & Francis, 2011). For example, it provides a conceptually straightforward way to introduce a second modification to a construct that already contains an UAA incorporated by amber suppression. Enzymatic bioconjugation can also be useful in the multiplex labeling of complex protein mixtures, and the modification of constructs that are large, difficult to purify or not easily amenable to the approaches outlined above. Since the recognition sequence is fused to the protein of interest, the modification site is usually limited in location to the N- and C-termini, or to a surface exposed and flexible loop of the protein. Other important considerations include the reaction kinetics, the stability and solubility of the enzymes, and the relatively large size of the modifications that are installed and that can perturb the function of the protein target.

4. Protein engineering approaches for tackling outstanding challenges in structural biology

4.1 X-ray crystallography

The molecular engineering toolbox presented here can aid crystallographers in all stages of the structure determination process. For example, when phase values cannot be determined by direct methods or molecular replacement, protein crystallographers can incorporate heavy atoms (e.g. Se, I or Br) into the protein crystal and use their anomalous diffraction patterns to solve the structure. Heavy atom incorporation can be accomplished in several ways (reviewed in (Pike et al. 2016). For
example, cysteine residues can be derivatized with mercury by pre-labeling, co-crystallization or soaking of the protein crystals in mercury salts (Martinez et al. 1993). If necessary, cysteine accessibility can be pre-evaluated with Ellman’s reagent (5,5′-dithiobis-2-nitrobenzoic acid) (Li et al. 2015). Another popular heavy atom labeling methodology relies on the incorporation of the modified amino acid selenomethionine (Hendrickson et al. 1990). In this case, selenomethionine is added to the culture of a methionine-auxotrophic bacterial strain resulting in substitution of all methionine residues in the protein. Selenomethionine incorporation is also possible in other non-bacterial expression systems, albeit with somewhat lower substitution efficiency (Cronin et al. 2007; Nettleship et al. 2010). Site-specific incorporation of appropriately modified unnatural amino acids (e.g. p-iodo-L-phenylalanine, p-bromo-L-phenylalanine and 3-iodo-L-tyrosine) can be performed with amber suppression or chemical synthesis (Kwon et al. 2006; Sakamoto et al. 2009; Xie et al. 2004; Yeung et al. 2016). Solving the phase problem can also be aided by the development of metal-binding peptide tags (e.g. Tb³⁺) that can be genetically fused to the protein of interest (Silvaggi et al. 2007).

Total chemical synthesis of proteins is an important tool for racemic and quasi-racemic crystallography (Yeates & Kent, 2012) (Fig. 12). These methods are based on the observation that mixing D- and L-forms of proteins can aid crystallization as macromolecules have access to a much larger set of crystallographic space groups, including those that contain mirror or center of inversion operations (Wukovitz & Yeates, 1995). D-enantiomers of proteins can be made using solid-phase peptide synthesis and NCL, with synthesis efficiencies that have allowed the structure determination of constructs in the 200-amino acid range (Mandal et al. 2012; Pan et al. 2016). The synthetic origin of such polypeptides also allows the facile co-incorporation of amino acids containing heavy atoms (Yeung et al. 2016), or chemically well-defined post-translational modifications (Okamoto et al. 2014).

For structure characterization of larger biological macromolecules and complexes, it is often more practical to use mimics rather than the native PTM. For example, methylation mimics are accessible through the alkylation of cysteine residues, and produce constructs with high yields and chemical purity in all possible methylation states (Simon et al. 2007). Such approaches have been invaluable in elucidating the impact of histone methylation on the nucleosome surface, a structural problem that requires the efficient crystallization of a biological assembly comprised of four different proteins and DNA (Lu et al. 2008). Sometimes, non-native linkages and modifications are essential in trapping an important functional state of a protein complex, for example between an ubiquitylated protein and the corresponding deubiquitinating enzyme (Morgan et al. 2016). The ongoing optimization of amber suppression systems for in-cell and cell-free protein synthesis has also made it easier to produce certain native modifications in high-enough yield for protein crystallization. There are, for example, several crystal structures of acetylated and phosphorylated proteins prepared by amber suppression in E. coli (Arbely et al. 2011; Huguenin-Dezot et al. 2016; Kuhlmann et al. 2016; Lammers et al. 2010). This technology has also enabled the structural analysis of ubiquitin chains (Virdee et al. 2010). Crystallographic studies of proteins containing multiple modifications have been successfully addressed with cell-free protein synthesis (Wakamori et al. 2015) and EPL (Wang et al. 2014b).

Beyond the incorporation of PTMs, the modern protein engineering toolbox provides the means to precisely alter the covalent structure of proteins, importantly allowing access to both the amino acid side-chains and the polypeptide backbone. Crystallographic analysis of such modified proteins can be extremely powerful, yielding insights into processes as diverse as...
as ion conduction (Grosse et al. 2011; Valiyaveetil et al. 2006), enzyme catalysis (Torbeev et al. 2011; Wang et al. 2014b) and protein–protein interactions (Lu et al. 1999; Morgan et al. 2016).

4.2 Nuclear magnetic resonance

The site-specific incorporation of NMR probes into biological macromolecules presents two challenging requirements to the protein chemist. First, similarly to X-ray crystallography, large amounts of sample are required, and the incorporation methods should be robust and efficient. Second, the isotopic labeling precursors are often prohibitively expensive. It therefore comes as no surprise that most NMR labeling strategies exploit the metabolic pathways in bacteria to produce uniform or sparse isotopic labeling schemes, or incorporate isotopic labels in an amino acid specific fashion (for a comprehensive series of reviews on isotopic labeling of biomolecules, the interested reader is referred to volume 565 of Methods in Enzymology, 2015). Here, we will focus on labeling methodologies that allow more control over the placement of spectroscopic probes in a polypeptide chain, and that have historically been demonstrated to produce sufficient protein amounts for NMR analysis.

4.2.1 Segmental isotopic labeling

Segmental isotopic labeling is a powerful protein engineering approach that allows the generation of an intact and typically natively folded and functional protein where only a certain segment of the polypeptide chain is ‘visible’ by NMR (Xu et al. 1999; Yamazaki et al. 1998). This strategy is particularly impactful if the protein of interest is large and/or has a degenerate amino acid sequence, thus simplifying assignment and interpretation of crowded and overlapped NMR spectra. Such constructs can be crucial in generating unambiguous structural constraints in large proteins and assemblies, and have been exploited in structural investigations both by solution and solid-state NMR (Mehler et al. 2015; Schubeis et al. 2015; Tremblay et al. 2015; Williams et al. 2016). Several approaches can be used to achieve segmental labeling of polypeptide chains. For example, a protein domain can be fused to a contiguous intein and produced recombinantly in labeled media (e.g. supplemented with 15N and 13C-enriched nutrient sources). Following thiolysis of the intein, the thioester derivative of the labeled domain can then be ligated, via EPL, to a protein fragment that does not contain NMR isotopes (Xu et al. 1999). In principle, the NMR silent segment can be a synthetic peptide carrying a PTM or a paramagnetic tag, thus allowing the facile incorporation of C-terminal protein modifications. For example, this approach has been used to introduce a phosphorylated tyrosine residue at position 125 of the amyloid-related protein α-synuclein, while residues 1–106 were prepared recombinantly and isotopically labeled for NMR analysis (Hejjaoui et al. 2012). In practice, however, the second fragment is usually produced recombinantly due to cost and yield considerations.

PTS offers an alternative intein-based approach for segmental isotopic labeling. The high affinity of naturally split intein fragments means that the splicing reactions can be conducted at very low concentrations of reactants (low μM), a capability that distinguishes the PTS strategy from EPL where reactant concentrations in the mM range are typically required. In practice, the target protein is divided into two segments, for example at domain boundaries, each fused to an appropriate intein fragment (Fig. 13a). These constructs are expressed separately, allowing the differential incorporation of NMR-active isotopes. After purification, the two extein–intein fragments are mixed, protein trans-splicing takes place and the leftover intein components can be purified using an appropriate affinity tag or chromatographic separation. Notably, the splicing reaction can also be performed directly in the cellular environment. In this case, both constructs are transformed into the same cell, but with

---

**Fig. 12.** Principle of racemic crystallography. The L- and D- forms of the polypeptide chains are prepared separately by solid-phase peptide synthesis and native chemical ligation. The proteins are subsequently mixed and co-crystallized, thus gaining access to a much larger set of crystallographic space groups.
different promoters so that the expression of each construct can be controlled separately, and therefore timed with the addition of isotopically enriched nutrients to the media (Muona et al. 2010) (Fig. 13b). This strategy can be particularly helpful for labeling proteins that are hard to refold in vitro, e.g. membrane proteins (Mehler et al. 2015). The first segmental labeling protocols relied on artificially split or natural inteins that were not very efficient, thus limiting the utility of the technology to few favorable cases (Yamazaki et al. 1998). Today, however, the toolbox of useful split inteins has greatly expanded, and it is now possible to choose from inteins with diverse properties, e.g. those giving better expression yields for the protein of interest, capable of carrying out ligations at higher temperatures or denaturant concentrations, or more forgiving to the choice of extein junction site (see Section 3.4.3).

Another option for segmental isotopic labeling is provided by sortase-mediated ligation (Fig. 13c). As discussed in Section 3.5, this method relies on the use of Sortase A, a transpeptidase that can ligate two constructs provided that one contains an LPXTG recognition sequence, and the other, a glycine repeat motif. Since it results in an LPXTG ‘scar’ at the ligation junction, sortase-mediated ligation is best suited for the segmental labeling of protein domains separated by a flexible, mutation-tolerant linker (Williams et al. 2016). The reversibility of the sortase ligation also necessitates the removal of cleaved peptide byproducts during the reaction course for optimal yields (Freiburger et al. 2015). On the other hand, segmental labeling using sortase works well with low μM concentrations of reagents, and is a good alternative for constructs that do not express well as intein fusions.

The segmental labeling approaches described here have also been used to generate protein constructs with NMR silent solubility enhancement tags (Kobashigawa et al. 2009; Zuger & Iwai, 2005). To keep proteins more stable in solution during a long multidimensional experiment, NMR spectroscopists can resort to fusing the protein of interest to a soluble domain such
as SUMO, GB1, MBP or thioredoxin. If prepared as a single construct, however, the solubility tag will be isotopically labeled and will contribute to the NMR spectrum, a complication that can be avoided with segmental labeling.

4.2.2 Site-specific incorporation of magnetic resonance probes

The introduction of a single isotopic label at a well-defined position in a polypeptide chain provides NMR spectroscopists with a benign and unambiguous reporter of protein structure, dynamics and ligand binding. Strategically positioned isotopes can also be crucial in disentangling higher order structural interactions in complex biological assemblies such as amyloid fibrils (Debelouchina et al. 2013; Petkova et al. 2006). The incorporation of such probes is relatively straightforward in short polypeptides prepared by solid-phase peptide synthesis, and many protected isotopically labeled amino acids are available commercially at a reasonable cost. Longer polypeptide chains containing specific labels can be prepared by NCL as exemplified by the synthesis of the membrane protein M2 containing five $^{13}$C,$^{15}$N-labeled amino acids dispersed throughout the sequence (Kwon et al. 2015). A serendipitously positioned native cysteine residue in this case provided a convenient ligation site without the need for desulfurization.

In recent years, the biomolecular NMR field has seen a surge in the popularity of $^{19}$F as a tool to investigate large and complex systems both by solution and solid-state NMR spectroscopy. This nucleus is bio-orthogonal, has magnetic properties that ensure high sensitivity (~80% of the sensitivity of $^1$H nuclei) and a wide chemical shift range (~100-fold larger than the range of $^1$H). It can therefore serve as a highly sensitive probe of protein and peptide structure, folding and aggregation, dynamics and ligand binding (reviewed in (Chen et al. 2013; Sharaf & Gronenborn, 2015). Specific $^{19}$F labels can be installed by peptide synthesis, conjugation or fluorination of cysteine and lysine residues, or by supplementing the culture media with fluorinated amino acids (which results in amino acid specific labeling). Increasingly, amber suppression is becoming the site-specific incorporation method of choice ((Marsh & Suzuki, 2014; Sharaf & Gronenborn, 2015). There are several orthogonal tRNA/synthetase systems and fluorinated UAs available, each imparting different spectroscopic signatures with respect to chemical shift anisotropy and relaxation. These include 3,5-difluorotyrosine (Li et al. 2013a; Yang et al. 2015), p-trifluoromethoxyphenylalanine (Cellitti et al. 2008) and p-trifluoromethylphenylalanine (Jackson et al. 2007; Loscha et al. 2012). While most protein $^{19}$F NMR applications rely on amber suppression in E. coli, magnetic resonance imaging studies are already exploiting the unique spectroscopic properties of $^{19}$F in mammalian systems or whole organisms (Yu et al. 2013). Site-specific $^{13}$C and $^{15}$N probes can also be installed by amber suppression, including $^{13}$C/$^{15}$N-labeled p-methoxyphenylalanine and $^{15}$N-labeled o-nitrobenzyl-tyrosine (Cellitti et al. 2008). Ligand-binding studies of large molecular systems may also benefit from incorporation of o-tert-butylyrosine, where rapid bond rotations of the tert-butyl group ensure a narrow linewidth of the characteristic peak in solution NMR spectra (Chen et al. 2015).

Another problem of particular relevance to the magnetic resonance spectroscopist concerns the site-directed installation of EPR and PRE probes. These probes carry stable nitroxide radicals or chelated metal ions such as Mn$^{2+}$, Cu$^{2+}$ or Gd$^{3+}$, and can be used to measure distances by EPR (up to 80 Å), or change the relaxation properties of nuclear spins in their vicinity (up to 35 Å). Usually, paramagnetic tags are incorporated at available cysteine residues using alkylation with maleimides, and many appropriately functionalized reagents are available commercially. As discussed in Section 3.1, however, cysteine-based approaches often have limited selectivity, the generated maleimide-based linker can be unstable, and is generally not applicable to cellular investigations. The development of amber suppression and bio-orthogonal reactions has certainly provided new options for EPR and PRE label incorporation both in vitro and in vivo. For example, orthogonal tRNA/tRNA synthetase systems are available for direct incorporation of nitroxide-containing (Jones et al. 2010; Schmidt et al. 2015) and metal chelating UAs (Park et al. 2015). Another amber suppression incorporation route involves the UAA p-acetyl-L-phenylalanine that can react with paramagnetic tags functionalized with a hydroxylamine moiety (Fleissner et al. 2009). An NMR-specific and often prohibitive requirement for the incorporation of PRE tags with amber suppression, however, is the concurrent need to introduce isotopic labels in the protein of interest. Thus, general protocols that provide efficient UAA incorporation under suboptimal bacterial growth conditions (minimal $^{13}$C/$^{15}$N-supplemented or perdeuterated media, for example) are desperately needed (Evans & Millhauser, 2015; Venditti et al. 2012).

Another demanding problem concerns the site-specific installation of isotopically labeled PTMs such as different methylation states, glycosylation, ubiquitination and various acylations (phosphorylation being somewhat of an exception as the naturally abundant $^{18}$O is the NMR active isotope). Coupled with the often problematic behavior of the modified proteins (e.g. aggregation prone intrinsically disordered and amyloidogenic peptides and proteins, insoluble transmembrane domains or chromatin effectors), PTM installation in itself is a challenge even without the extra complication and associated cost of isotopic labeling. Not surprisingly, therefore, most NMR studies involving PTMs are performed with constructs that have been modified enzymatically. Enzymatic modification, however, is often incomplete and can result in a heterogeneous set of modifications, thus complicating the interpretation of NMR data. Well-established non-enzymatic site-directed protocols are available

https://doi.org/10.1017/S0033583517000051 Published online by Cambridge University Press
for ubiquitylation, taking advantage of the possibility to prepare and isotopically label ubiquitin recombinantly. The labeled ubiquitin moiety can then be installed at a specific position of the modified polypeptide using a variety of strategies including an asymmetric disulfide (Debeilouchina et al. 2017), a condensation reaction of a C-terminal ubiquitin thioester and the ε-amino of a targeted lysine residue (Castaneda et al. 2011a), or a combined recombinant expression–chemical synthesis approach (Castaneda et al. 2011b). Site-specific phosphorylation can be incorporated by solid-phase peptide synthesis, followed by EPL to construct segmentally labeled and modified proteins (Hejaoui et al. 2012). Studies of glycosylated proteins also require the attachment of homogeneous glycans that are often isotopically labeled to facilitate resonance assignments and structure determination (Skrivoskova et al. 2010). Cysteine-based 13C-labeled methyllysine analogs are a convenient and affordable way to generate methylated constructs, and have been used in the NMR structural analysis of proteins such as HP1 and p53 (Cui et al. 2012; Munari et al. 2012).

4.3 Studies of dynamic interactions

4.3.1 Incorporation of optical probes

The advances in single-molecule optical spectroscopy and imaging have propelled the search for better fluorescent probes and the development of more efficient and targeted strategies for their incorporation in vitro and in cells. The information obtained from such experiments can complement high-resolution structural studies and provide data not easily accessible by other methods, including dynamic protein–protein interactions, conformational states and protein folding pathways. There are, of course, many excellent reviews on the subject (Dimura et al. 2016; Haney et al. 2015; Minoshima & Kikuchi, 2017; Nikic & Lemke, 2015). Therefore, we will limit our discussion to the problem of incorporating two small FRET probes for biophysical and structural investigations, still a surprisingly challenging task for the protein chemist.

A protein construct prepared for FRET experiments can be labeled with two different fluorophores (A and B) by cysteine mutagenesis at two appropriately selected sites. Incubation with the maleimide, iodoacetamide or methyl bromide derivatives of the two probes results in a statistical mixture of labeled constructs even if labeling is efficient (A–A, A–B, B–A, B–B). Since the mixture of products can complicate data analysis (Husada et al. 2015), more selective strategies for labeling are highly desirable and have been explored in the literature. A popular approach, for example, combines cysteine labeling with amber suppression that can introduce a second fluorescent moiety either directly or through bio-orthogonal chemistries (Brustad et al. 2008; Haney et al. 2016; Milles et al. 2012; Ratzke et al. 2014). Suitable UAAAs include p-acetylphenylalanine for labeling with commercially available hydroxylamine probe derivatives (Fig. 14a), as well as azido- and alkynyl-functionalized phenylalanine, tyrosine and lysine residues for click reactions. Cysteine labeling can also be combined with the site-specific incorporation of 1,2-aminothiols that react selectively with cyanobenzothiazole fluorescent-probe derivatives (Nguyen et al. 2011). A conceptually different cysteine-based labeling approach involves the FLAsH system that targets biresonant reagents to the genetically encoded tetracysteine motif CCPGCC (Griffin et al. 1998). This technology has the advantage that the small molecule ligands are cell permeable and virtually non-fluorescent until they bind their recognition motif, and are therefore well-suited for in-cell applications. In the context of dual labeling, FLAsH has been combined with amber suppression (Perdios et al. 2017) (Fig. 14b) or fluorescent proteins such as CFP (Hoffmann et al. 2005) for the investigation of protein conformational states in vitro or in the cellular environment.

When the protein construct of interest contains functionally or structurally important cysteines, dual labeling can be attempted with NCL, EPL, the incorporation of genetically encodable UAAAs and peptide tags for bioenzymatic conjugation, or a combination of these approaches. For example, a fluorescent probe and a thioamide quencher can be installed using a combination of amber suppression and protein ligation, where the quencher is a backbone modification that is only accessible through protein synthesis (Wissner et al. 2013) (Fig. 14c). As discussed in Section 3.6.3, incorporation of two UAAAs in a polypeptide chain by genetic means is still difficult; however, the growing numbers of successful applications are almost exclusively developed for FRET-labeling. For example, using cells containing the orthogonal ribosome ribo-Q1, two fluorescent labels could be incorporated efficiently into calmodulin, one in response to the amber TAG codon, the other at the newly assigned AGTA quadruplet codon (Sachdeva et al. 2014) (Fig. 14d). Rapid, quantitative, one-pot labeling was achieved by using mutually orthogonal cycloadditions at UAAAs bearing a terminal alkyne and a cyclopropane. Alternatively, amber–ochre or amber–opal dual incorporation schemes can be used (Chatterjee et al. 2013b; Wan et al. 2010; Xiao et al. 2013).

4.3.2 Incorporation of vibrational probes

Two-dimensional infrared spectroscopy (2D IR) has emerged as a powerful technique to characterize the dynamic states and interactions of biological macromolecules (Baiz et al. 2013; Le Sueur et al. 2015). Dynamic and structural information is encoded in the vibrational modes of functional groups such as backbone amides and carbonyls, as well as side-chains with aromatic, carbonyl and guanidinium groups. Since even polypeptides of moderate length contain many overlapping
signals, site-specific incorporation of vibrational reporters with distinct spectroscopic signatures is essential for data interpretation. The installation of such probes can take advantage of several of the protein engineering approaches described above (Zhang et al. 2016b). Segmental isotopic labeling, for example, can be used to distinguish the spectroscopic behavior of protein domains (Moran et al. 2012). In this case, one domain is labeled with $^{13}$C, while the other domain remains at natural abundance. The higher mass of the $^{13}$C isotope changes the vibrational modes of the backbone carbonyls, and results in a downward shift of the vibrational frequency, effectively decoupling the spectroscopic signatures of the two domains. Higher resolution in linear and multidimensional IR spectra can be achieved by the site-specific installation of isotopic labels such as $^{13}$C–$^{18}$O pairs by solid-phase peptide synthesis, EPL, NCL or the incorporation of appropriately labeled methionine during recombinant protein expression (Courter et al. 2014; Davis et al. 2015; Dhayalan et al. 2016; Marecek et al. 2007; Zhang et al. 2016b). Resolution and specificity can also be achieved by the introduction of orthogonal vibrational probes such as nitrile, cyano, azido and thiocyanate functional groups. To this end, several UAAAs suitable for amber suppression are available, including 4-cyano-, 4-azido- and 4-azidomethyl-phenylalanine (Bazewicz et al. 2013; Schultz et al. 2006; Ye et al. 2010). $^{13}$C, $^{15}$N-labeled thiocyanates, on the other hand, can be obtained by cyanylation of cysteine residues (van Wilderen et al. 2014).

5. Outlook

Our goal in this review has been to examine the contents of the modern protein engineering toolbox, with the particular needs of the structural biologist in mind. As is hopefully evident from the preceding sections, there are now many highly versatile
strategies to manipulate and decorate protein constructs, allowing the creation of molecules and assemblies that faithfully represent the complexity of biological systems. We hope that this resource will serve as an inspiration to the structural biologist looking to improve on sample preparation protocols, construct interesting and relevant samples, or devise new strategies that push the boundaries of modern structural biology methods. Undoubtedly, these tools will continue to improve and evolve, expanding the range of options for both in vitro and in vivo applications. Yet, many challenges still lie ahead. For example, better tools are desperately needed to reign in obstinate systems such as membrane proteins, intrinsically disordered domains or amyloidogenic polypeptides (Butterfield et al. 2012; Úversky, 2015; Zuo et al. 2015). These constructs, traditionally recalcitrant to protein engineering and structural biology analysis alike, beg for more efficient (semi)-synthetic, genetic or intein/sortase-based approaches that circumvent low expression yields and poor solubility issues. Next, the site-specific manipulation of large proteins and assemblies (especially those prepared in eukaryotic systems) is still an overwhelmingly difficult task, and new chemical tools and ideas are needed to supplement the existing genetic and enzymatic approaches. At the same time, the demand for constructs bearing multiple modifications will continue to increase. For example, dissecting the molecular basis of biological cross-talks requires access to homogeneous samples of proteins carrying multiple chemically diverse PTMs (Allis & Muir, 2011; Bah & Forman-Kay, 2016). Many biological molecules and assemblies are also too complex for structural analysis by one technique alone (Cramer, 2016; McGinty & Tan, 2015; Tynan et al. 2016), and the increasing need for method integration will necessitate the construction of samples suitable for multi-modal studies.

Concurrently, structural biologists are devising new strategies to improve the sensitivity of their respective methodologies and reduce the amounts of precious biological materials required for structural analysis. New technological developments such as X-ray free-electron lasers (Neutze et al. 2015), polarization enhancement strategies for NMR spectroscopy (Ardenkjaer-Larsen et al. 2015; Maly et al. 2008), and the enhanced capabilities of cryo-EM instrumentation (Nogales, 2016) will certainly be central to these efforts. The rise of mass spectrometry based structural approaches, including hydrogen–deuterium exchange, ion mobility-mass spectrometry or cross-linking based analysis, also holds the bright promise to reveal structural information from samples in the picomole regime (Lossl et al. 2016). At the same time, structural biologists have started to turn to the cellular milieu as the future arena of structural endeavors (Beck & Baumeister, 2016; Freedberg & Selenko, 2014). In this undoubtedly daunting task, structural biologists are not alone. Protein engineers, with their long track record of successes in the selective manipulation of complex systems both in vitro and in vivo, are ready to meet these exciting new challenges through innovations that will continue to push the boundaries of chemical biology.

Acknowledgements

We would like to thank Adam Stevens, Antony Burton, Felix Wojcik and Robert Thompson for many helpful discussions. We also acknowledge the contributions of many colleagues whose work could not be included in this review due to space limitations. This work was supported by US National Institutes of Health grants R37-GM086868, R01-GM107047 and P01-CA196539.

References


