Biotechnology and Genetic Engineering Reviews - Vol. 24, 129-146 (2007)

Protein Ligation: Applications in NMR Studies of Proteins

HIDEO IWAI[1,]* AND SARA ZÜGER[2]

[1]*Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki. P.O. Box 65, Helsinki, FIN-00014, Finland and* [2]*Biochemisches Institut, Universität Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland*

Introduction

One of the ultimate goals of structural biology is to understand how proteins exert their functions at atomic resolution in order to modulate their activities for applications such as drug discovery. Nuclear Magnetic Resonance (NMR) spectroscopy is increasingly becoming an important method for characterizing the three-dimensional structure of proteins under near physiological conditions. NMR is not only a powerful method for elucidating the three-dimensional structures of proteins in aqueous solution, but also a convenient method for studying protein-ligand interactions as well as protein dynamics. NMR studies, however, require the assignment of individual signals to individual NMR-active atoms in proteins. This so-called resonance assignment of proteins can be a time-consuming and laborious process. In addition, the number of NMR signals increases proportionally to the molecular size of the proteins. Therefore, NMR studies of larger proteins are increasingly time-consuming and challenging. The use of stable isotopes such as ^{15}N and ^{13}C and multidimensional NMR spectroscopy has overcome the problem of overlapping signals by extending the dimensions of NMR spectra. Currently, the structure of proteins of up to 20-30 kDa in size can be determined with multi-dimensional NMR spectroscopy, even though it still requires considerable effort (Clore & Gronenborn, 1994). In contrast, recent improvements

*To whom correspondence may be addressed (hideo.iwai@helsinki.fi)

Abbreviations: Abl, Abelson protein tyrosine kinase; CBD, chitin-binding domain; EPL, expressed protein ligation; GPCR, G-protein-coupled receptor; IPL, intein-mediated protein ligation; MBP, maltose-binding protein; NCL, native chemical ligation; NMR, Nuclear Magnetic Resonance; *Npu*, *Nostoc punctiforme*; PDB, Protein Data Bank; PTB, polypyrimidine tract binding protein; SH2 and SH3, src homology type 2 and 3 domains; *Ssp*, *Synechocystis sp*. PCC6803.

of software and automated high-throughput crystallization have speeded up the structure determination of proteins by X-ray crystallography and this does not require isotope-labeled proteins and has no size limit. Indeed, more than 10,000 X-ray structures of proteins (and assemblies) larger than 300 amino acids length – in stark contrast to only a few NMR structures of proteins of more than a 300 amino acid residues length – are now deposited in the Protein Data Bank (PDB). This suggests that NMR structure-determination of proteins over 35 kDa has currently no practical use, unless a special isotope-labeling technique is applied (Kainosho *et al.*, 2006). Even though it may not be practical to determine the three-dimensional structure of large proteins over 35 kDa, NMR has the great advantage over X-ray crystallography for the study of structural changes, ligand interactions, and protein dynamics in near physiological conditions. In addition, NMR spectroscopy does not require the time-consuming process of crystallization. In fact, NMR has been widely used as a useful tool in drug discovery (Pellecchia *et al.*, 2002) and recent technological advances have pushed the size limit by solution NMR spectroscopy close to 1 MDa (Riek *et al.* 2000; Fiaux *et al.*, 2002). The application of NMR would be even wider if one can obtain the resonance assignments of larger proteins more easily by overcoming signal overlaps.

A number of approaches have been traditionally applied to overcoming overlapping signals that arise with increasing size of the protein (*Figure 1*). The most successful approach is to use uniformly ^{15}N and ^{13}C labeled samples (*Figure 1B*). This uniform double-labeling has made it possible to perform sequential resonance assignments using heteronuclear coupling (Clore & Gronenborn, 1994). It is now routinely used for the resonance assignments of proteins and extends the molecular size of proteins by NMR to up to 20-30 kDa. However, for larger proteins over 30 kDa, the resonance overlaps become increasingly problematic even with uniform double-labeling and multidimensional NMR spectroscopy. Selective labeling approaches would be attractive as they can reduce the number of signals by incorporating specific amino acids with isotope-labeling (*Figure 1C*). One of the problems with selective amino acid labeling is that not all the amino acid residues can be used for selective labeling, because of the amino acid conversions that occur in the cells due to their biosynthesis. The recent development of cell-free protein expression circumvents this problem (Kigawa *et al.*, 1995). However, the more profound and intrinsic problem of selective amino acid labeling is the difficulty to assign individual signals of the selective amino acids in the primary sequence, when there is more than one residue for one amino acid type as it loses sequential connectivity that can be used for resonance assignment. Although cell-free protein synthesis can offer the possibility to incorporate isotope-labeled amino acids into a specific site (*Figure 1E*), the use of this site-directed isotope labeling in NMR might be unpractical, because of the complicated procedure involved in preparing the necessary samples for the NMR assignment (Yabuki *et al.*, 1998). The most successful application of selective amino acid labeling is the combinatorial approach to use two different amino acid types with two different isotopes (^{15}N for one amino acid and ^{13}C for the other amino acid), which is successfully used for the resonance assignment of large proteins of up to 150 kDa (Kainosho and Tsuji, 1982; Arata *et al.*, 1994). This double-labeling approach enables the assignment of the signals to be sequence-specific provided that the combination of the two amino acid types is unique in the primary sequence. However, the preparation of the samples for the assignment could be enormously laborious.

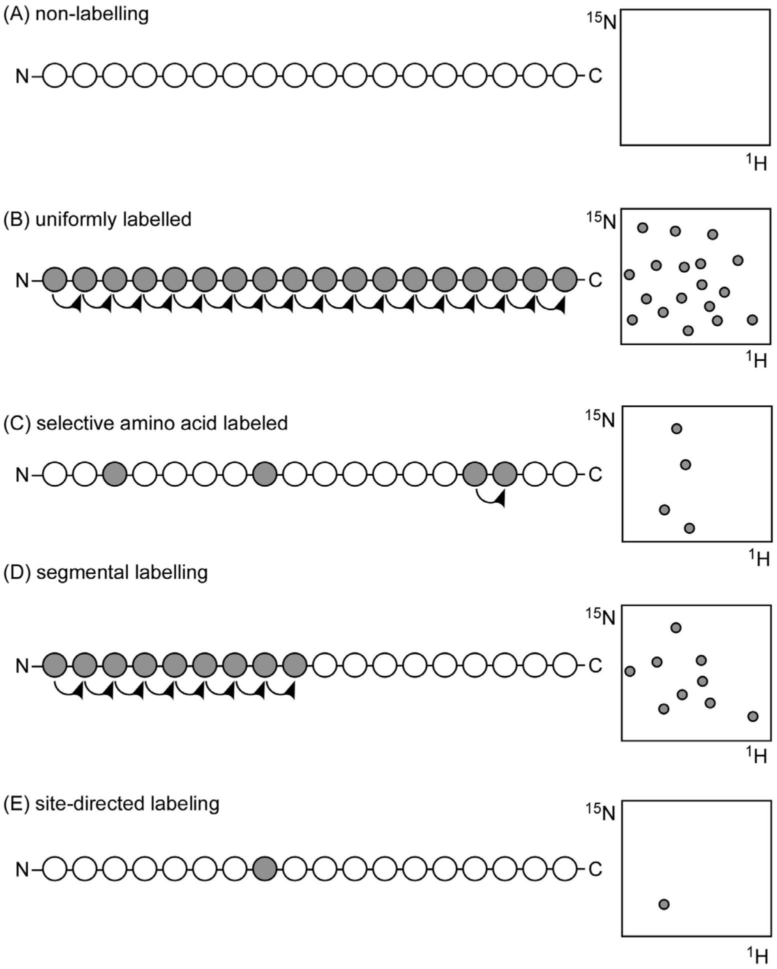

Figure 1A-E: Isotopic labeling patterns of a protein and their expected spectra. Filled circles indicate labeled amino acid residues, open circles indicate unlabeled amino acid residues. The arrows depict the sequential connectivity between the amino acid residues.

An alternative way is to use segmental isotopic labeling, in which a region of a protein is isotopically labeled (*Figure 1D*). Segmental isotope-labeling can reduce the number of signals and at the same time it is still possible to obtain sequential connectivity using conventional NMR techniques. This method is even more attractive for studying multi-domain proteins that contain several homologous domains, since a domain with a similar sequence and structure tends to produce a similar NMR spectrum, providing extra difficulty for the analysis. Segmental isotope-labeling could, therefore, offer an opportunity for the study of a domain of a protein in its full length context (Walters *et al.*, 2003; Vitali *et al.*, 2006).

How can one produce such segmental isotope-labeling in a protein? Advances in chemistry and protein engineering have opened a new avenue to create proteins with segmental-isotope labeling, i.e. native chemical ligation and protein splicing approaches

(*Figure 2*). Unlike other biophysical methods such as fluorescence spectroscopy, protein NMR spectroscopy requires large quantities of isotopically labeled samples (> 5 mg for a 20 kDa protein) despite the recent improvement in sensitivity of NMR instruments. Thus, the production of segmental isotope-labeled proteins for NMR studies is still not yet a routine procedure and still technically challenging, limiting the number of successful applications of segmental isotope-labeling. In this article, we summarize the current technologies that have been - and could be - used for segmental isotopic labeling of proteins and the applications of segmental isotopic labeling in NMR studies.

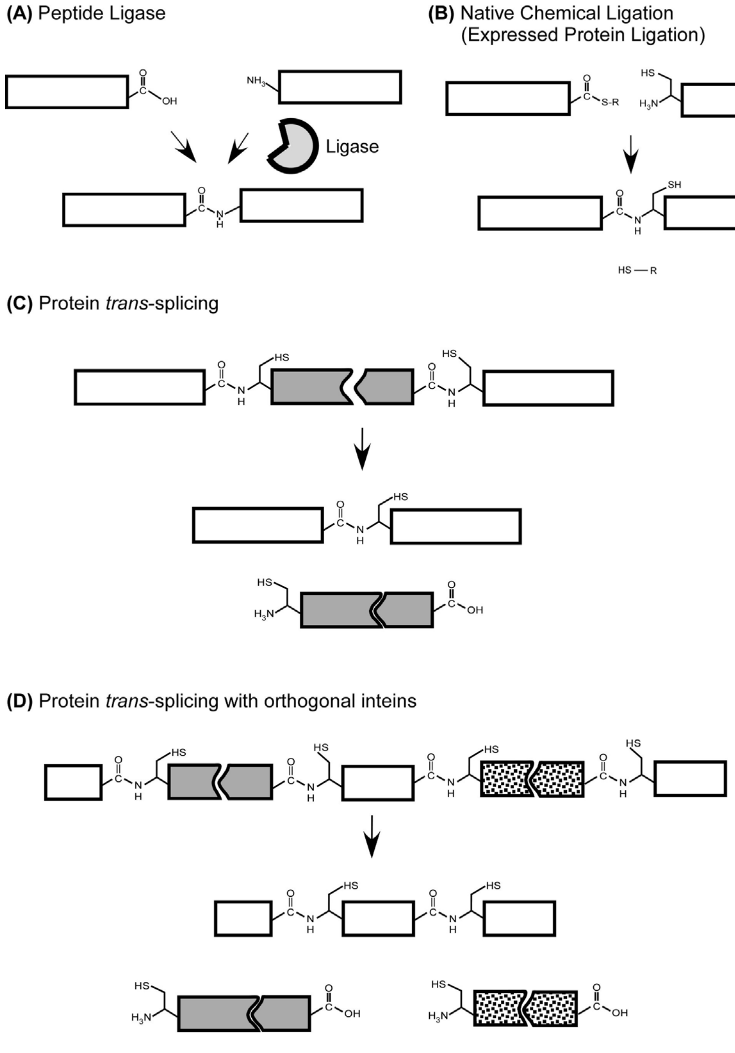
(A) Peptide Ligase
(B) Native Chemical Ligation (Expressed Protein Ligation)
Ligase
O
C
OH
NH3
N
H
HS
S-R
H3N
SH
HS — R
(C) Protein trans-splicing
(D) Protein trans-splicing with orthogonal inteins

Figure 2A-D: Methods of protein ligation. **A)** A ligase enzymatically ligates two polypeptide chains by a peptide bond. **B)** Native chemical ligation: a C-terminal thioester in a polypeptide chain reacts with a N-terminal cysteine of another polypeptide chain chemoselectively in aqueous solution, ligating the two fragments by a peptide bond. **C)** Protein *trans*-splicing is carried out by a split intein. Upon reconstitution of the two fragments, the reconstituted intein can exert its catalytic function and excises itself out from the adjacent polypeptides, concomitantly ligating them together by a peptide bond. **D)** A three-fragment ligation by using two orthogonal inteins.

Protein ligation

Chemical synthesis can incorporate isotopes site-specifically or segment-specifically into peptides, which is commonly used for solid-state NMR spectroscopy. It is, however, limited to peptides or small proteins that do not suffer from overlapping signals in solution NMR spectroscopy. Additionally, the isotope-labeled amino acids for automatic peptide synthesizers are considerably expensive. Ligation of two or more chemically synthesized peptide fragments can overcome the drawback of chemical synthesis approaches. Such peptide/protein ligation has been of special interest. One of these approaches is to use enzymes that can ligate polypeptide chains. One idea is to simply reverse the reaction of proteases that hydrolyze peptide bonds. Significant efforts in protein engineering have been made in this direction such as converting peptidases into ligases. The most successful enzyme is "subtiligase" that is derived from the serine protease subtilisin by protein engineering (Jackson *et al.*, 1994). Another possibility is to use a transpeptidase such as sortase, which has been demonstrated to be capable of ligating peptide chains (Mao *et al.*, 2004). These approaches have not been applied for segmental isotope-labeling of proteins for structural studies. If the ligation efficiency can be improved, these approaches would be of special interest in the future. The most successful and simplest ligation method is native chemical ligation (NCL) developed by Dawson and Kent (see Dawson et al., 1994). The chemoselective reaction between two unprotected peptides in aqueous solution, one containing an α-thioester and the other an N-terminal cysteine, produces a single covalently linked ligation product (*Figure 2B*). This approach has been widely used for the synthesis of more than 100 proteins (Dawson and Kent, 2000). However, chemical synthesis of isotope-labeled peptides is still too expensive to be used in the routine NMR studies. After the company *New England Biolabs* developed a novel purification method (IMPACT system), in which a fusion protein is formed with a protein splicing domain (Chong *et al.*, 1997), it was realized soon after that not only chemical synthesis but also a bacterial expression system can be used to create peptides and proteins with an α-thioester, opening a possibility for preparing isotopically enriched materials for native chemical ligation in an inexpensive manner. Because the protein splicing domain in the IMPACT system is modified in a way that it can undergo specific self-cleavage and release the target protein as an α-thioester from the fusion upon addition of thiols, one can easily prepare proteins that can be used for native chemical ligation. The combination of the IMPACT system and native chemical ligation is often called Expressed Protein Ligation (EPL) or Intein-mediated Protein Ligation (IPL). This approach has now become a powerful method to introduce segmental isotope-labeling into proteins.

Protein splicing

Protein splicing is a posttranslational modification that is auto-catalytically processed by an intervening sequence called intein (*Figure 3*). Up to date there are more than 300 intein genes found in eucarya, eubacteria and archaea (Perler, 2002). They are inserted into specific genes and after translation the intein excises itself out, concomitantly ligating the two flanking sequences of the host protein (exteins) into one mature protein. Soon after the discovery of inteins (Hirata *et al.*, 1990; Kane *et*

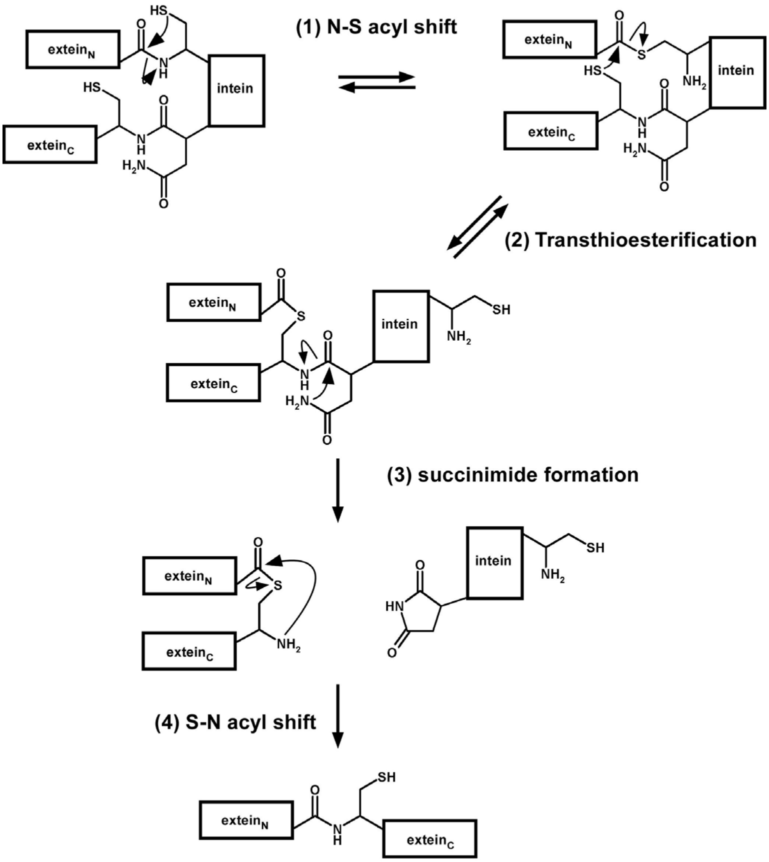

Figure 3: Reaction mechanism of protein splicing. The first step (**1**) in the splicing mechanism is an N-S acyl shift: the thiol group of the first residue of the intein attacks nucleophilically the carbon atom of the C-terminal carbonyl group of the N-terminal extein, creating a thioester group. The next step (**2**) involves a transesterification, where the thiol group of the first residue of the C-terminal extein attacks the carbonyl group of the N-terminal extein, leading to a rearrangement of the thioester, which is followed by a succinimide formation (**3**). The connected extein is released upon the succinimide formation. The ligated polypeptide chains undergo an immediate S-N acyl shift (**4**) and form a stable polypeptide bond.

al., 1990), their amazing nature was exploited for biotechnological applications (reviewed in Xu & Evans, 2005). Inteins are capable of ligating foreign extein sequences, even though a few amino acids of the native extein sequence seem to be required in many cases. The protein splicing reaction is comprised of four steps (*Figure 3*): (1) N-S acyl shift, (2) transesterification, (3) succinimide formation, and (4) S-N acyl shift. Inteins can also be found to be split into two fragments – the so-called split inteins – catalyzing protein splicing in *trans*, thereby ligating two separated extein sequences in the two polypeptide chains into one polypeptide chain (*Figure 2C*). Protein *trans*-splicing can also be performed by using artificially split inteins that are derived by splitting inteins into two fragments at a linker region. Artificially split inteins usually have to be purified under denaturing conditions and renatured in order to restore the splicing activity for the ligation of the two flanking polypeptide chains. The refolding process could be a problem for protein ligation. However, artificially split inteins have been often used for protein ligation, as there has been until recently only one natural split intein characterized (Wu *et al.*, 1998). In particular, for a three-fragment ligation as used for isotope-labeling of a central piece of a protein (Otomo *et al.*, 1999b), it is necessary to utilize two orthogonal inteins so as to selectively ligate the three fragments together (*Figure 2D*). The disadvantage of *trans*-splicing using split inteins is that splicing often requires an insertion of a few amino acid residues from the native extein sequence that might be included in the ligated product. Even though most of the inteins are promiscuous with respect to the extein sequences, the enzymatic activity seems to be dependent on the junction sequences. For example, in the case of *Synechocystis sp.* (*Ssp*) DnaE, an insertion of the natural extein sequence CFN is required for an efficient protein splicing reaction. This essential sequence seems to depend on the individual inteins. It is well-understood that the first C-terminal extein residue is the key residue in the transesterification reaction during splicing (*Figure 3*). Thus, the transesterification step requires a C-terminal first extein residue that contains a thiol or hydroxyl group, i.e. Cys, Ser or Thr, meaning that one of these residues must be present for protein *trans*-splicing. Proteins ligated by protein splicing are likely to contain an insertion of at least one non-native amino acid sequence at the splicing junction, requiring the assessment of the functionality of the ligated proteins prior to the structural analysis.

Segmental isotope-labeling using NCL

Even though NCL is quite a simple chemoselective reaction that can be achieved in aqueous solution, segmental isotope-labeling using NCL had not been applied until the IMPACT system had been developed. With the IMPACT system, the required α-thiol group can be produced from a recombinant fusion protein with an engineered intein, which is mutated such that only the first step in protein splicing (N-S acyl shift) can occur (Chong *et al.*, 1997). As the intein initiates the protein splicing reaction, the polypeptide of interest can be released with addition of an alkylthiol as an alkyl thioester. Importantly, this protein fragment can be easily prepared in an isotope-labeled form by growing bacteria in isotope-enriched culture medium. Proteins containing the α-thiol group prepared by the intein fusion can be ligated readily with the protein/peptide bearing an N-terminal cysteine using NCL. Thus, segmental isotope labeling using this NCL approach should be straightforward. Nevertheless, there have been only a few examples reported (Xu *et al.*, 1999; Camarero *et al.*, 2002; Walters *et al.*, 2003; Vitali *et al.*, 2006). This is most likely to be because of the low yield of ligation as well as the cumbersome preparation of the polypeptide bearing the N-terminal cysteine residue. Xu and coworkers prepared the polypeptide with the N-terminal cysteine from a fusion protein with a factor Xa cleavage site (IEGR/C). Thereby, the protein of interest can be released with an N-terminal cysteine upon treatment with the protease. The seemingly easy way of producing proteins bearing an N-terminal cysteine residue just by introducing cysteine as the second amino acid after the start-methionine has not been widely used (Iwai and Plückthun, 1999; Camarero *et al.*, 2001). It is presumably because the posttranslational modification by endogenous methionyl aminopeptidases is often incomplete, reducing the yield of α-cysteine proteins. A third strategy of producing polypeptides with an N-terminal cysteine again uses the ability of inteins (the so-called TWIN system from *New England Biolabs*): modified inteins can be fused to the N-terminus of the protein of interest and will release the protein bearing now an N-terminal cysteine by a temperature or pH shift (Evans *et al.*, 1999). In the following paragraphs the application of NCL for segmental isotope-labeling of proteins is described in more detail.

(1) ABL SH2-SH3 DOMAIN

Chemical ligation was applied to the src homology type 2 and 3 domains of Abelson protein tyrosine kinase (Abl-SH2-SH3) for joining the two protein fragments SH2 (12.0 kDa) and SH3 (6.3 kDa) with different isotope content (Xu *et al.*, 1999). The two domains were expressed and purified separately. The SH3 domain containing an α-thioester was prepared from the IMPACT system by cleaving it off from the intein fusion protein by adding ethanethiol (Chong *et al.*, 1997). The SH2 domain was expressed as a fusion protein with a factor Xa cleavage site and was released from the fusion with an N-terminal cysteine upon proteolysis. The two correctly folded domains were then ligated in an aqueous buffer (pH 7.2) containing thiophenol and benzyl mercaptan as thiol catalysts at a final concentration of 1.5 % (vol/vol). The ligation reaction was incubated for about 90 hours and yielded approximately 2.5 mg (137 nmol) of segmentally isotope-labeled Abl-SH2-SH3 out of 2 mg (317 nmol) Abl-SH3 α-thioester and 8 mg (667 nmol) uniformly ^{15}N-labeled Abl-SH2.

(2) BACTERIAL σ^A FACTOR

Segmental labeling by chemical ligation was applied to a bacterial σ factor of 399 amino acid residues, σ^A from *Thermotoga maritima* (Camarero *et al.*, 2002) to investigate the auto-inhibiting function of the N-terminal 1.1 domain that was believed to bind to a conserved region within the protein. By using NMR spectroscopy analysis of the segmentally labeled protein, it has been suggested that this was not the case. The same strategy as Abl-SH2-SH3 was applied for ligating the first 348 residues of σ (σ[1-348], 38.3 kDa) or residues 137-348 (Δ1.1-σ[137-348], 23.2 kDa), and the C-terminal 51 residues of σ (σ[349-399], 5.5 kDa). The α-thioester was prepared from the intein fusion protein (IMPACT system) by adding ethanethiol. The C-terminal fragment of σ[349-399] was expressed as a fusion protein with a factor Xa cleavage site. The protein carrying the N-terminal cysteine was released upon proteolysis by factor Xa. The two fragments were then ligated in a ligation buffer of 25 mM sodium phosphate (pH 7.2), 200 mM guanidinium hydrochloride, 250 mM NaCl, 1 mM EDTA containing 0.2 % octyl glucoside and ethanethiol as thiol catalysts at a final concentration of 3 % (vol/vol). The equimolar amounts of the two fragments were incubated overnight at room temperature at a concentration of 50 μM. The reaction was rapid with very efficient ligation (>90 %).

(3) HHR23A

The 40-kDa DNA repair protein hHR23a that belongs to the Rad23 family has been segmentally isotope labeled by protein ligation (Walters *et al.*, 2003). The two fragments of the protein (12.6 and 27.0 kDa) were prepared by using the IMPACT-TWIN protein fusion system from *New England Biolabs* (Evans *et al.*, 1999), however the details of the experiment were not reported. The results showed that the segmental labeling allowed the unambiguous investigation of the inter-domain interactions within the protein.

(4) C-TERMINAL RNA RECOGNITION MOTIFS OF PTB

Allain and co-workers used expressed protein ligation to segmentally label two C-terminal domains of the polypyrimidine tract binding protein (PTB) to investigate the interdomain interface between those two RNA recognition motifs (Vitali *et al.*, 2006). They also used the IMPACT-TWIN system to achieve the segmental labeling. The two protein domains (13.0 and 10.7 kDa) were fused either N- or C-terminally to two different intein constructs, respectively. One intein is fused to a chitin-binding domain (CBD) domain at the C-terminus and the other at the N-terminus, resulting in the following constructs: domain1-intein-CBD and CBD-intein-domain2. The expressed fusion proteins, one uniformly labeled with isotopes, were bound to the same affinity column. First, the protein thioester was produced by inducing the cleavage from the C-terminally located intein by addition of 2-mercaptoethanesulfonic acid (MESNA). Simultaneously, the temperature was elevated to induce N-terminal cleavage of the second protein fragment carrying a cysteine residue at the N-terminus from the second intein. The subsequent on-column chemical ligation of the protein α-thioester with the cysteine-containing second protein fragment produced the segmental isotopic-labeled polypeptide. 9 mg (380 nmol) segmentally isotope-labeled protein was obtained from the protein α-thioester precursor produced in 0.4 l LB medium and the α-Cys containing C-terminal precursor protein produced in 2 l minimal M9 medium containing the isotopes. Since the two PTB domains interact with each other, it is worth noting that this possibly forces the ligation sites to be closer during the splicing reaction. The remarkably good production yield could well be due to this effect. The use of on-column ligation could also improve the ligation efficiency because of the increased local concentration.

Segmental isotope-labeling using protein trans-splicing *in vitro*

Even though NCL is a simpler method for segmental isotopic labeling, only two fragments can be ligated at one time due to the chemoselective reaction between the α-thioester and the N-terminal cysteine residue. Therefore, for a three-fragment ligation, the individual fragments would have to be ligated in a stepwise manner, making the procedure more cumbersome. On the other hand, segmental isotopic labeling using protein *trans*-splicing has the advantage that it could easily extend to a three-fragment ligation simply by using two orthogonal inteins: the ligation is achieved by the protein *trans*-splicing upon the formation of an active conformation of the reconstituted split intein, which is fused with the two individual fragments. These two precursor fragments fused with the split intein can be produced independently in cell culture medium with different isotope contents. Thus, the ligated product can be segmentally labeled by reconstituting the split intein prepared independently with different isotope.

The production of NMR quantities of such a segmental isotope-labeled protein was first demonstrated using an artificially split intein (Yamazaki *et al.*, 1998). Several reports of segmental isotope-labeling using protein *trans*-splicing were published as summarized in the following (Otomo *et al.*, 1999a; Otomo *et al.*, 1999b; Yagi *et al.*, 2004). In contrast to NCL, with which domains of proteins were ligated, segmental isotope-labeling was also applied to a single-domain protein such as RNA polymerase α, MBP, or the ß-subunit of F1F0-ATPases.

(1) C-TERMINAL DOMAIN OF RNA POLYMERASE α

The C-terminal domain of RNA polymerase α subunit (81 amino acids) was the first protein to be segmentally isotope-labeled (Yamazaki *et al.*, 1998). The protein consisting of four α-helices was split into two pieces (5.5 and 3.8 kDa) in a short flexible loop between helix 3 and 4. The resulting fragments were genetically fused to either the N- or the C-terminal part of the artificially split intein PI-*Pfu*I intein (intervening the ribonucleoside-diphosphate reductase α-subunit from *Pyrococcus furiosus*). After purification of the inclusion bodies that were independently expressed in different isotope-containing medium and hence differently labeled, the two fragments were mixed under denaturing condition and refolded in order to obtain the functionally active intein and properly folded protein. The protein *trans*-splicing was then triggered by heating the samples at 70 °C. Several amino acids from the natural extein sequence were included into the linker between the intein and extein fragments in order to reconstitute protein splicing activity (GGG for N-extein and TG for C-extein), resulting in an 88 amino acid polypeptide after ligation. After the final purification only 0.14 mg (15 nmol or 0.05 mM in 0.3 ml) of segmentally isotope-labeled product was obtained (from 1 and 0.25 liters of culture for the N- and C-terminal parts, respectively). The low yield was due to the expression level and chemical instability of the fragments rather than to the ligation efficiency of the intein. The protocol was improved and in presence of 2 M urea a final yield of 2.0 mg (210 nmol) of segmentally isotope-labeled protein was achieved from 0.5 l M9 minimal medium (Otomo *et al.*, 1999a). The effectiveness of the splicing reaction was judged to have been between 70 and 90 % of the precursor proteins after 8 hours at 37 °C or 2 hours at 70 °C.

(2) MALTOSE-BINDING PROTEIN (MBP)

In vitro segmental isotopic-labeling was even extended to label a central segment of a polypeptide by a three-fragment ligation (Otomo *et al.*, 1999b). Two artificially split inteins, PI-*Pfu*I and PI-*Pfu*II (another intervening sequence from the ribonuleotide-diphosphate reductase gene from *Pyrococcus furiosus*), which are orthogonal to each other, were combined to perform this reaction. The N-terminal part of the protein MBP (10.8 kDa) was fused to the N-terminus of the N-terminal part of PI-*Pfu*II, the central fragment of MBP (15.5 kDa) was fused to the C-terminus of the C-terminal part of PI-*Pfu*II and to the N-terminus of the N-terminal part of PI-

*Pfu*I, and the C-terminal fragment of MBP (14.3 kDa) was fused to the C-terminus of the C-terminal part of PI-*Pfu*I. All of the separately expressed three precursors (the central fragment was expressed in the isotope-containing medium) were solubilized from inclusion bodies and refolded. *Trans*-splicing of the split inteins was induced by incubating at 70 °C for two hours. The ligated protein precipitated due to the harsh conditions, but was successfully refolded. 2.5 mg (60 nmol) central-segmentally labeled protein was obtained from 0.5 l culture in this experiment.

(3) ß SUBUNIT OF F1-ATPASE

Akutsu and co-workers (Yagi *et al.*, 2004) applied the segmental isotope-labeling to an even larger system. The 52 kDa β-subunit of F1-ATPase from the thermophilic bacterium PS-3 was segmentally-labeled in different ways generating four different fragments to be analyzed by NMR (13.5 kDa, 9.6 kDa, 29.3 kDa, and 22.6 kDa). They used the artificially split intein PI-*Pfu*I according to the protocol developed by Yamazaki and coworkers (Yamazaki *et al.*, 1998). NMR measurements were carried out with 0.2-0.4 mM segmentally labeled protein (this corresponds to about 3-6 mg): They were successfully able to monitor the conformational change of the ß-subunit monomer induced by nucleotide binding.

***In vivo* segmental isotope labeling**

NMR spectroscopy is a non-invasive type of spectroscopy that can analyze structures at atomic resolution in physiological conditions. Recent improvements of NMR instrumentation have enabled us to study proteins in living cells (Serber *et al.*, 2005). Thus, it should in principle provide a powerful tool for improving our understanding of the structure-function relationship of proteins in living organisms, provided one can incorporate stable isotopes selectively into the target of interest. In this line of research, an *in vivo* segmental isotope-labeling method has been developed by introducing a dual vector system into *Escherichia coli* that expresses the protein fragments fused to either the N-terminal or C-terminal part of the *Ssp* DnaE split intein (naturally occurring split intein intervening the DNA Polymerase III a subunit of *Synechocystis sp.* PCC6803) (Züger and Iwai, 2005). Because of the two different inducible promoters - T7/lac and *araBAD* - it is possible to selectively express the precursors in media containing different isotopes, providing that one of the fragments is stable in the cells (*Figure 4*). In more detail, the C-terminal protein precursor was fused to the C-terminal part of *Ssp* DnaE and a short linker of 6 amino acids (4 amino acids (CFNK) coming from the natural extein context to ensure effective protein splicing and 2 amino acids introduced from the restriction endonuclease site (GT)). The N-terminal protein precursor was fused to the N-terminal part of the intein and a linker of 2 amino acids (GS) that had been introduced for cloning purpose. After induction of the first promoter and incubation for 3 hours, the cells were harvested from the medium and re-suspended in a new medium that contained different isotopes. The second promoter was subsequently induced to produce the second precursor fragment. The physiological conditions in the *E. coli* cells allowed the split intein to reconstitute efficiently and subsequently ligate the two precursor proteins, producing a segmental isotopic-labeled protein. However, the *trans*-splicing reaction was often not completed, as unprocessed precursors were found after the purification. We obtained about 2 mg (140 nmol) of segmentally isotope-labeled pure protein per liter of culture medium. Compared with the existing segmental isotope-labeling methods, this procedure is much simpler and more robust in terms of the work required. Also, the yield of the ligated product is comparable or better, making this approach very attractive.

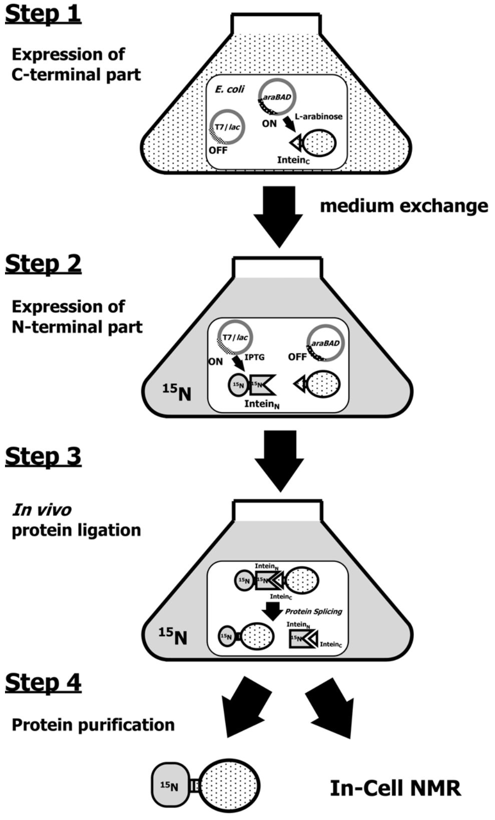

Figure 4: *In vivo* segmental isotope-labeling. *Escherichia coli* cells harboring two plasmids with different inducible promoters produce the two split intein fragments fused to the polypeptides of interest independently upon induction. **(1)** First, the induction of the C-terminal fragment fused to the C-terminal part of the split intein is carried out in unlabeled medium. **(2)** After exchange to isotope-enriched medium, the second promoter of the other plasmid is induced. **(3)** Upon expression of the second half of the intein, the split intein is reconstituted, ligating the two flanking polypeptides of interest in the cells. Either the cells can be used directly for in cell NMR or the protein can be purified and used for NMR studies.

We have also compared the efficiency of *in vivo* and *in vitro* ligation by reconstituting the split intein *in vitro* by mixing the two independently purified precursor fragments. Even with the same intein and extein context, the ligation efficiency of the *in vitro* ligation is much lower than the *in vivo* ligation (*Figure 5*). Although protein splicing itself does not require any cofactor, other cellular components such as chaperones seem to assist the reconstitution of the split intein to be in an active form. Thus, the *in vivo* segmental isotope labeling approach can provide an efficient and easy preparation of segmentally isotope-labeled protein for NMR studies despite some drawbacks such as isotope scrambling and an insertion of non-natural sequence.

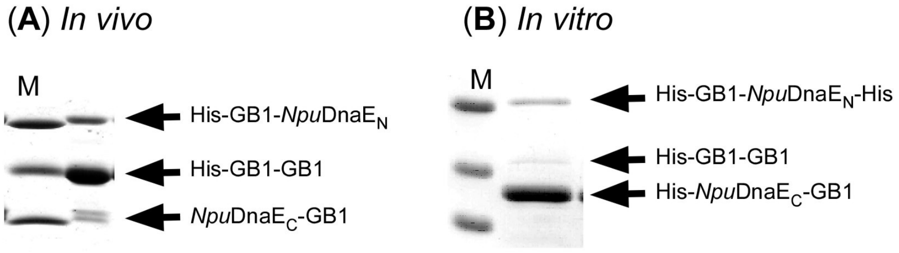

Figure 5: Comparison between *in vivo* protein ligation and *in vitro* ligation using protein *trans*-splicing. **A)** *In vivo* ligation using *Npu* DnaE intein. The precursor and ligated product were purified via the N-terminal His-tag from the cell lysate. **B)** *In vitro* ligation using *Npu* DnaE intein. The individual fragments were purified individually and mixed at room temperature for overnight incubation.

Improving protein ligation

The laborious procedure required for protein ligation and its yield are the principal factors which limit the practical application of this method in segmental isotope-labeling. In the case of NCL the stability and reactivity of the α-thioester as well as of the alkanethiols that are used as catalysts could play an important role in improving the reaction. The better catalyst and the choice of the C-terminal residue such as less bulky residues i.e. Gly, where the thioester will be created, could improve the ligation efficiency (Johnson and Kent, 2006). In addition, a higher local concentration of the two fragments, for example achieved by on-column ligation or ligation of interacting fragments, seems to improve their ligation yields.

Segmental isotope-labeling by protein splicing using the artificially split inteins PI-*Pfu*I and PI-*Pfu*II was performed under harsh conditions such as 70 °C in order to optimize the splicing reaction (Otomo *et al.*, 1999a). Not many proteins could survive such conditions. Even though the efficiency of the ligation was estimated to be rather high (70-90 %), the application of this system might be limited to only a few examples. Finding a suitable intein system that is more robust under milder conditions might thus be necessary. Another problem using split inteins is that the sequence at the splicing junction affects the protein splicing reaction. This sequence dependence at the splicing junction has not been well-studied. Currently, the ligation efficiency is not predictable for individual systems. Moreover, this effect seems to be strongly dependent on the inteins. Even though an intein can generally ligate two flanking foreign sequences, it is often the case that it requires at least a small native extein sequence in order to retain the splicing activity. Although more than 300 putative inteins have been reported, only a few inteins have been well-characterized in detail (Perler, 2002). Thus, it is still of importance to understand how the splicing junction would affect the splicing reaction as well as to identify inteins with robust properties such as high splicing efficiency in order to make segmental isotope-labeling more attractive. Recently, we have investigated the naturally occurring split intein DnaE from *Nostoc punctiforme* (*Npu*) that is more tolerant of amino acid substitutions at the splicing junction and more efficient compared with the widely used naturally split *Ssp* DnaE intein (Iwai *et al.*, 2006). Interestingly, *Npu* DnaE is not orthogonal with respect to *Ssp* DnaE, suggesting that the interaction between the N-terminal and C-terminal parts of the split DnaE inteins might not be as strong as has been previously reported (Shi and Muir, 2005). It is likely that there might be other inteins that are more suitable for protein ligation purposes. Therefore, further biochemical characterization of other protein splicing domains remains a prerequisite for improving segmental isotope-labeling and making it useful for a wider range of proteins.

In the case of *in vivo* protein ligation for segmental isotopic labeling, there are a few considerations required. Because the two fragments containing split intein parts must be able to interact in order to obtain splicing activity, each protein fragment should be soluble after the sequential expression in the bacterial cells. Moreover, the protein fragment expressed prior to the second induction of the other protein fragment has to be stable enough so that it will not degrade in the cells before or during the second expression. Even though we have reported previously a better selective isotope-labeling for the N-terminal protein fragment (Züger and Iwai, 2005), the order of the expression of the two fragments is important and strongly depends on the target proteins. With *in vivo* protein ligation, there are several parameters which one can optimize in order to obtain better expression and solubility, such as expression temperature, bacterial strains, or co-expression of chaperones. The optimization of those conditions can certainly improve the overall yield of the ligated product.

Application of protein ligation for structural studies: Improving solubility of proteins

Another application of *in vivo* ligation is to incorporate fusion tags. It is a common approach to add a fusion tag in order to increase protein solubility and stability *in vivo* as well as *in vitro* (reviewed by Waugh, 2005). Although biological assays are often done without removing the fusion tag, it is usually necessary to remove the fusion tag for structural studies, in particular for NMR studies, due to the increased molecular weight and signal overlaps. In many cases, the solubility and stability of proteins can be the limiting factor for NMR studies as extensive measurement time is often required and could last for weeks. Small solubility enhancement tags have been found to be useful for the elucidation of protein structures by NMR (Zhou *et al.*, 2001). *In vivo* protein ligation can provide a unique opportunity to ease the spectra analysis by being able to incorporate a non-labeled fusion tag into isotope-labeled target proteins (Züger and Iwai, 2005). The use of the dual vector system also provides a rapid screening method of an optimal fusion protein by applying protein ligation, as the tagging is done not on the DNA level but on the protein level. The most interesting challenge of this application is to increase solubility of membrane proteins such as G-protein-coupled receptors (GPCR). Structural studies of membrane proteins, particularly such as GPCRs, have been very slow despite their enormous pharmaceutical importance. Improving solubility and stability of such membrane proteins using segmental isotope-labeling for NMR studies could be of importance to understand the structure-function relationship of membrane proteins.

Conclusion

The applications of segmental isotopic labeling using protein splicing are currently not widely used but are becoming increasingly important. In particular for large systems of > 30 kDa, the application of segmental isotope-labeling will play a crucial role for studying interactions and dynamics of larger proteins by NMR spectroscopy. However, the techniques are still technically challenging and the success of protein ligation seems to be dependent on the targets. Despite the fact that the protein splicing mechanism is now believed to be largely understood, little is still known about the biophysical properties as well as why some inteins are more promiscuous with respect to their extein sequence at atomic level. In order to be able to use protein splicing for general applications in structural studies, it is of advantage to be able to understand and to be able to control protein splicing for wider applications. Further understanding of the mechanism of protein splicing in detail is likely to improve the protein ligation scheme in the future. Moreover, simple and robust methods such as *in vivo* segmental isotope-labeling might be of special importance for NMR studies of proteins in *in vivo* systems (Serber *et al*., 2001; Selenko *et al*., 2006). Thus, protein ligation using protein splicing is expected to play an important role in structural biology using NMR spectroscopy.

Acknowledgments

H.I. thanks the Sigrid Jusélius Foundation, Biocentrum Helsinki, and the Academy of Finland for funding. The authors thank Tanel Mahlakõiv for the production of *Figure 5*.

References

ARATA Y., KATO K., TAKAHASHI H., AND SHIMADA I. (1994) Nuclear magnetic resonance study of antibodies: a multinuclear approach. *Methods in Enzymology* **239**, 440-464.

CHONG, S., MERSHA, F.B., COMB, D.G., SCOTT, M.E., LANDRY, D., VENCE, L.M., PERLER, F.B., BENNER, J., KUCERA, R.B., HIRVONEN, C.A., PELLETIER, J.J., PAULUS, H. AND XU, M.Q. (1997) Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. *Gene* **192**, 271-281.

CAMARERO, J.A., FUSHMAN, D., COWBURN, D. AND MUIR, T.W. (2001) Peptide chemical ligation inside living cells: *in vivo* generation of a circular protein domain. *Bioorganic and Medicinal Chemistry* **9**, 2479-2484.

CAMARERO, J.A., SHEKHTMAN, A., CAMPBELL, E.A., CHLENOV, M., GRUBER, T.M., BRYANT, D.A., DARST, S.A., COWBURN, D. AND MUIR, T.W. (2002) Autoregulation of a bacterial sigma factor explored by using segmental isotopic labeling and NMR. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 8536-8541.

CLORE, G.M. AND GRONENBORN, A.M. (1994) Structures of larger proteins, protein–ligand and protein–DNA complexes by multidimensional heteronuclear NMR. *Progress in Biophysics and Molecular Biology* **62**, 153-184.

DAWSON, P.E., MUIR, T.W., CLARK-LEWIS, I. AND KENT, S.B. (1994) Synthesis of proteins by native chemical ligation. *Science* **266**, 776-779.

DAWSON, P.E. AND KENT, S.B. (2000) Synthesis of native proteins by chemical ligation. *Annual Review of Biochemistry* **69**, 923-960.

EVANS, T.C.JR., BENNER, J. AND XU, M.Q. (1999) The cyclization and polymerization of bacterially expressed protein using modified self-splicing inteins. *Journal of Biological Chemistry* **274**, 18359-18363.

FIAUX, J., BERTELSEN, E.B., HORWICH, A.L. AND WÜTHRICH, K. (2002) NMR analysis of a 900K GroEL-GroES complex. *Nature* **418**, 207-211.

HIRATA, R., OHSUMK, Y., NAKANO, A., KAWASAKI, H., SUZUKI, K. AND ANRAKU, Y. (1990) Molecular structure of a gene, *VMA1*, encoding the catalytic subunit of H^+-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *Journal of Biological Chemistry* **265**, 6726-6733.

KANE, P.M., YAMASHIRO, C.T., WOLCZYK, D.F., NEFF, N., GOEBL, M. AND STEVENS, T.H. (1990) Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H^+-adenosine triphosphatase *Science* **250**, 651-657.

IWAI, H. AND PLÜCKTHUN, A. (1999) Circular ß-lactamase: stability enhancement by cyclizing the backbone. *FEBS Letters* **459**, 166-172.

IWAI, H., ZÜGER, S., JIN J. AND TAM, P.-H. (2006) Highly efficient protein *trans*-splicing by a naturally occurring split DnaE intein from *Nostoc punctiforme*. *FEBS Letters* **580**, 1853-1858.

JACKSON, D.Y., BURNIER J., QUAN C., STANLEY M., TOM J. AND WELLS J.A. (1994) A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. *Science* **266**, 243-247.

JOHNSON, E.C. AND KENT, S.B. (2006) Insights into the mechanism and catalysis of the native chemical ligation reaction. *Journal of the American Chemical Society* **128**, 6640-6646.

KAINOSHO, M. AND TSUJI, T. (1982) Assignment of the three methionyl carbonyl carbon resonances in *Streptomyces* subtilisin inhibitor by a carbon-13 and nitrogen-15 double-labeling technique. A new strategy for structural studies of proteins in solution. *Biochemistry* **21**, 6273-6279.

KIGAWA, T., MUTO, Y. AND YOKOYAMA, S. (1995) Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. *Journal of Biomolecular NMR* **6**, 129-134.

MAO, H., HART, S.A., SCHINK, A. AND POLLOK, B.A. (2004) Sortase-mediated protein ligation: a new method for protein engineering. *Journal of American Chemical Society* **126**, 2670-2671.

OTOMO, T., TERUYA, K., UEGAKI, K., YAMAZAKI T. AND KYOGOKU, Y. (1999a) Improved segmental isotope labeling of proteins and application to a large protein. *Journal of Biomolecular NMR* **14**, 105-114.

OTOMO, T., ITO, N., KYOGOKU, Y. AND YAMAZAKI, T. (1999b) NMR observation of selected segments in a larger protein: central-segment isotope labeling through inein-mediated ligation. *Biochemistry* **38**, 16040-16044.

PELLECCHIA, M., SEM, D.S. AND WÜTHRICH, K. (2002) NMR in drug discovery. *Nature Reviews Drug Discovery* **1**, 211-219.

PERLER, F.B. (2002) InBase, the intein database. *Nucleic Acids Research* **30**, 383-384.

RIEK, R., PERVUSHIN, K. AND WÜTHRICH, K. (2000) TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution. *Trends in Biochemical Sciences* **25**, 462-468.

SELENKO, P., SERBER, Z., GADEA, B., RUDERMAN, J. AND WAGNER, G. (2006) Quantitative NMR analysis of the protein G B1 domain in *Xenopus laevis* egg extracts and intact oocytes. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 11904-11909.

SERBER, Z., LEDWIGE, R., MILLER, S.M. AND DÖTSCH, V. (2001) Evaluation of parameters critical to observing proteins inside living *Escherichia coli* by in-cell NMR spectroscopy. *Journal of the American Chemical Society* **123**, 8895-8901.

SERBER, Z., CORSINI, L., DURST, F. AND DÖTSCH, V. (2005) In-cell NMR spectroscopy. *Methods in Enzymology* **394**, 17-41.

SHI, J. AND MUIR, T.W. (2005) Development of a tandem protein trans-splicing system based on native and engineered split inteins. *Journal of the American Chemical Society* **127**, 6198-6206.

VITALI, F., HENNING, A., OBERSTRASS, F.C., HARGOUS, Y., AUWETER, S.D., ERAT, M. AND ALLAIN, F.H. (2006) Structure of the two most C-terminal RNA recognition motifs of PTB using segmental isotope labeling. *EMBO Journal* **25**, 150-162.

WALTERS, K.J., LECH, P.J., GOH, A.M., WANG, Q. AND HOWLEY, P.M. (2003) DNA-repair protein hHR23a alters its protein structure upon binding proteasomal subunit S5a. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12694-12699.

WAUGH, D.S. (2005) Making the most of affinity tags. *Trends in Biotechnology* **23**, 316-320.

WU, H., HU, Z. AND LIU, X.Q. (1998) Protein *trans*-splicing by a split intein encoded in a split DnaE gene of *Synechocystis sp*. PCC6803. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 9226-9231.

XU, R., AYERS, B., COWBURN, D. AND MUIR, T.W. (1999) Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 388-393.

XU, M.Q. AND EVANS, T.C.JR. (2005) Recent advances in protein splicing: manipulating proteins *in vitro* and *in vivo*. *Current Opinion in Biotechnology* **16**, 440-446.

YABUKI, T., KIGAWA, T., DOHMAE, N., TAKIO, K., TERADA, T., ITO, Y., LAUE, E.D., COOPER, J.A., KAINOSHO, M. AND YOKOYAMA, S. (1998) Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. *Journal of Biomolecular NMR* **11**, 295-306.

YAMAZAKI, T, OTOMO, T., ODA, N., KYOGOKU, Y., UEGAKI, K., ITO, N., ISHINO, Y. AND NAKAMURA, H. (1998) Segmental isotope labeling for protein NMR using peptide splicing. *Journal of American Chemical Society* **120**, 5591-5592

ZHOU, P., LUGOVSKOY, A.A. AND WAGNER, G. (2001) A solubility enhancement tag (SET) for NMR studies of poorly behaving proteins. *Journal of Biomolecular NMR* **20**, 11-14.

ZÜGER, S. AND IWAI, H. (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. *Nature Biotechnology* **23**, 736-740.